

Nutritional Systems Biology of Fat

**Integration and modeling of transcriptomics datasets
related to lipid homeostasis**

Mohammad Ohid Ullah

Thesis committee

Thesis supervisor

Prof. dr. M.R. Müller
Professor of Nutrition, Metabolism and Genomics
Wageningen University

Thesis co-supervisor

Dr. G.J.E.J. Hooiveld
Assistant Professor, Division of Human Nutrition
Wageningen University

Other members

Prof. dr. C.J.F. ter Braak, Wageningen University
Prof. dr. ir. V.A.P. Martins dos Santos, Wageningen University
Prof. dr. C. Evelo, Maastricht University
Dr. B. van Ommen, TNO, Zeist

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Nutritional Systems Biology of Fat

**Integration and modeling of transcriptomics datasets
related to lipid homeostasis**

Mohammad Ohid Ullah

Thesis

submitted in fulfillment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus

Prof. dr. M.J. Kropff,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Monday 08 October 2012

at 1.30 p.m. in the Aula.

Mohammad Ohid Ullah

Nutritional Systems Biology of Fat - Integration and modeling of transcriptomics datasets related to lipid homeostasis, 158 pages.

Thesis Wageningen University, Wageningen, NL (2012)
With references, with summaries in English and Dutch.

ISBN: 978-94-6173-381-8

Abstract

Fatty acids, in the form of triglycerides, are the main constituent of the class of dietary lipids. They not only serve as a source of energy but can also act as potent regulators of gene transcription. It is well accepted that an energy rich diet characterized by high intakes of dietary fat is linked to the dramatic increase in the prevalence of obesity in both developed and developing countries in the last several decades. Obese individuals are at increased risk of developing the metabolic syndrome, a cluster of metabolic abnormalities that ultimately increase the risk of developing vascular diseases and type 2 diabetes. Many studies have been performed to uncover the role of fatty acids on gene expression in different organs, but integrative studies in different organs over time driven by high throughput data are lacking. Therefore, we first aimed to develop integrative approaches on the level of individual genes but also pathways using genome-wide transcriptomics datasets of mouse liver and small intestine that are related to fatty acid sensing transcription factor peroxisome proliferator activated receptor alpha (PPAR α). We also aimed to uncover the behavior of PPAR α target genes and their corresponding biological functions in a short time series experiment, and integrated and modeled the influence of different levels of dietary fat and the time dependency on transcriptomics datasets obtained from several organs by developing system level approaches.

We developed an integrative statistical approach that properly adjusted for multiple testing while integrating data from two experiments, and was driven by biological inference. By quantifying pathway activities in different mouse tissues over time and subsequent integration by partial least squares path model, we found that the induced pathways at early time points are the main drivers for the induced pathways at late time points. In addition, using a time course microarray study of rat hepatocytes, we found that most of the PPAR α target genes at early stage are involved in lipid metabolism-related processes and their expression level could be modeled using a quadratic regression function. In this study, we also found that the transcription factors NR2F, CREB, ERF and RXR might work together with PPAR α in the regulation of genes involved in lipid metabolism. By integrating time and dose dependent gene expression data of mouse liver and white adipose tissue (WAT), we found a set of time-dose dependent genes in liver and WAT including potential signaling proteins secreted from WAT that may

induce metabolic changes in liver, thereby contributing to the pathogenesis of obesity.

Taken together, in this thesis integrative statistical approaches are presented that were applied to a variety of datasets related to metabolism of fatty acids. Results that were obtained provide a better understanding of the function of the fatty acid-sensor PPAR α , and identified a set of secreted proteins that may be important for organ cross talk during the development of diet induced obesity.

Table of contents

	<i>Abstract</i>	7
Chapter 1	<i>General introduction</i>	11
Chapter 2	<i>An integrated statistical approach to compare transcriptomics data across experiments: a case study on the identification of candidate target genes of the transcription factor PPARα</i>	31
Chapter 3	<i>Integrative analysis and modeling of PPARα function in murine liver and small intestine</i>	49
Chapter 4	<i>Characterization and modeling of acute effects of PPARα activation in rat liver cells</i>	67
Chapter 5	<i>Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice</i>	87
Chapter 6	<i>General discussion</i>	119
	<i>References</i>	127
	<i>Summary</i>	141
	<i>Samenvatting (Summary in Dutch)</i>	145
	<i>Acknowledgements</i>	149
	<i>Curriculum Vitae</i>	153
	<i>List of publications</i>	155
	<i>Overview of completed training activities</i>	157

Chapter 1

General introduction

Fatty acids and PPAR α

Fatty acids are the most important macronutrient components for mammals that function as a fuel in the cell to provide energy. After digestion and absorption of dietary fat in the small intestine, chylomicrons are formed that are packaged with triacylglycerol (TG), cholesterol esters, phospholipids, free cholesterol, and apoproteins [1]. Chylomicrons are then secreted by the intestinal epithelial cells and transported via the lymphatic system to the blood. The chylomicrons then circulate throughout the blood stream and reaches capillaries where LPL (lipoprotein lipase) captures these particles and hydrolyzes the TG. As a result tissues such as the adipose tissue and muscle take up the free fatty acids (FFA), which are then converted into cellular energy. Excess FFAs may cause obesity and its associated diseases like type 2 diabetes mellitus, dyslipidemia, atherosclerosis, hypertension and hepatic steatosis [2,3]. If adipose tissues exceed their capacity to store the FFA as TG, then it may travel to the liver where it may cause NAFLD (nonalcoholic fatty liver disease) [4,5]. The chylomicron remnants that remain after the hydrolysis of TG ultimately travel to the liver (Figure 1).

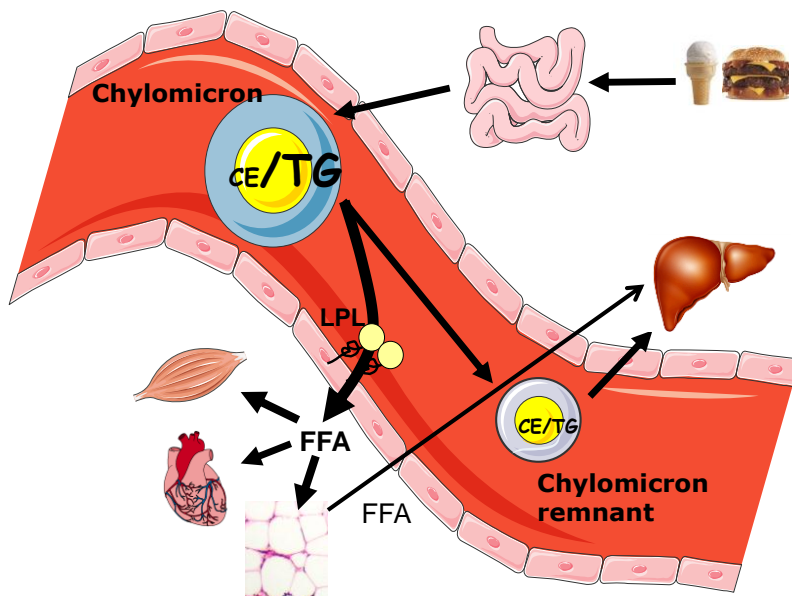


Figure 1: Digestion and metabolism of dietary fat

The FFAs derived from dietary fat, or any synthetic component, enter the cell and can bind specific transcription factors after which gene expression can be activated or suppressed [6]. The peroxisome proliferator-activated receptors (PPARs, NR1C) [7] is a family of transcription factors which are activated by dietary fatty acids. There are three PPAR isotypes: PPAR α (NR1C1), PPAR δ (also called β) (NR1C2) and PPAR γ (NR1C3). All are targets for treating type 2 diabetes, dyslipidemia and obesity [8]. Fatty acids and their derivatives bind to PPAR α with the greatest attraction [9]. Every PPAR heterodimerizes with the retinoid X receptor (RXR) and subsequently binds to specific regions on the DNA of target genes [10,11]. Since early 1990s, when the peroxisome PPAR α was discovered, the function of PPAR α has been studied broadly [12]. Currently PPAR α is well-known for its control of metabolism in response to diet. PPAR α is highly expressed in tissues with a high catabolic rate such as the liver, kidneys, heart, intestine and skeletal muscle [13,14]. The identification PPAR α target genes have concentrated mostly on cellular lipid metabolism in the context of the hepatocyte.

A comprehensive expression profiling analysis of PPAR α dependent regulation of hepatic lipid metabolism was done by [15], using the synthetic ligand WY14643. They found that the role of PPAR α in hepatic lipid metabolism was much more extensive than previously envisioned and uncovered novel PPAR α regulated genes and pathways, after 24 hours and 5 days exposure to WY14643. A genome wide analysis of PPAR α activation in murine small intestine was performed by [16]. They showed that PPAR α influences the immune and inflammatory response in the mouse intestine, which may be of particular importance for the development of fortified food and valuable for patients with inflammatory bowel diseases. A comparative analysis of gene regulation by the transcription factor PPAR α between mouse and human was conducted by [17], they showed that PPAR α regulates a mostly divergent set of genes in mouse and human hepatocytes. Taken together, PPAR α is considered a crucial fatty acid sensor that mediates effects of numerous fatty acids and its derivatives on gene expression and therefore is a master regulator of lipid metabolism in mouse and human [18]. Although much is already known about the PPARs, still gaps in our knowledge remain. The biological role of PPAR α , its target genes, pathways and biological processes have merely been investigated at only one or two time points after activation in isolated systems (tissues). However, no studies have been performed to integrate and model the different time points- and tissue-related gene

expression data. Therefore, system approaches are needed to integrate and model different datasets of PPAR α dependently regulated genes and pathways at different time points, as well as with different dietary fat doses to explore the more functional behavior of dietary fat among the different organs. By applying microarray tools one can collect whole genome information and then multivariate statistical analytical tools can be applied to get better insights in the biological function.

Multivariate data analysis

A microarray study provides expression data on thousands of genes simultaneously in several conditions. Basically this can be considered as multivariate data. To find out differentially expressed genes and then to evaluate the association between them and other (supplementary) information (internal or external factors), univariate and multivariate analysis approaches can be applied. This kind of association and research can be performed not only at the level of individual genes, but also at the level of groups of genes and at the level of tissues. In biological sciences, researchers collect lots of data to fully explore their study. A main purpose of microarray data analysis is the identification of differentially expressed genes and corresponding biological processes. To this end mainly univariate techniques are utilized for hypothesis testing; these include Student's t-test, F test, ANOVA or mixed models have been applied for each gene. Transcriptomics data involve many tested genes (variables) and the control of the false positives rate is not enough, besides, there are not enough replications to obtain good estimations.

Generally, in an experiment with small number of replications for each gene, variances can be poorly estimated and therefore the results of the classical t- or F statistics can lead to an increase of false positives. However, it is well noticed that genes do not act alone; therefore, there is mutual information within microarray data that could be used to improve variances estimates. Borrowing information from the collective of genes can assist in the inference about each individual gene. Tusher *et al* developed a modification of the t-statistic by adding a constant in its denominator which improves the estimation of the variance, the method is

known as significance analysis of microarrays (SAM) [19]. Another method is known as linear models for microarrays data (LIMMA), and is based on empirical Bayes approach taking moderated t-statistic [20]. Recently, the package *limma* is one of the most used programs for microarray data [21]. To analyze factorial time course microarrays data with capturing dynamic gene expression profiles, the time course analysis of variance (TANOVA) method can be applied [22]. For pathway level analysis, several methods have already developed and the most used programs are gene set enrichment analysis (GSEA) [23], DAVID [24,25] and Ingenuity Pathway Analysis (IPA) [26]. The multiple testing problem is one of the most challenging topic in microarray data as well. To handle the false discovery rate, many methods have already developed but the most useful method was developed by [27].

To adjust the relationships between genes in each pathway, still proper and suitable methods need to be discovered. The univariate techniques assume that genes are independent, but in reality this is not the case. Furthermore, the univariate methods are not able to handle the relationships between several genes (responses), therefore, multivariate statistical techniques are becoming popular to perceive the more explorative analysis in systems biology [28]. For instance, the multivariate statistical software package *FactoMinerR* [29] is a powerful tool to handle and analyze several groups of datasets which based on principal component analysis (PCA), factor analysis (FA) or partial least squares (PLS). PCA is a dimension reduced approach which produces a set of orthogonal principal components (linear combinations of original variables) to account for the maximum variation of the data. Observing then the loading plot of the top principal components one can find out the most influential genes/variables. The more absolute value of loading indicates that the corresponding gene is more important or influential. Of course, it is important to see in which treatment groups the influential genes are located. For this, one can compare the parallel comparison between the score plot (individual plot) and the loading plot (correlation circle). The genes and the samples/group(s) located in the same quadrant shows the importance of the genes in that samples/group(s).

Like PCA, FA also involves the description of a set of observed variables in terms of a reduced number of latent variables which is known as explanatory factor analysis (EFA). The main difference between PCA and FA is that PCA represents

the latent variables as functions of the original variables whereas FA represents the observed variables as function of the factors or latent variables. Usually the use of PCA or EFA appeals more to an explanatory data analysis perspective whereas FA is also considered as a model building approach and hypothesis testing which is known as confirmatory factor analysis (CFA). However, PCA or FA cannot handle the causal relationships as well as more noisy and numerous predictor variables between the two or more groups (blocks) of variables. To overcome this problem, partial least squares path model (PLSPM) [30,31] was developed using partial least squares (PLS) [32,33] to structural equation modeling (SEM) which is also known as SEM-PLS or soft modeling. The PLSPM or soft modeling does not depend on any distribution pattern and a few cases can suffice [34]. Furthermore, it is a components based approach and robust against missing values, misspecification and multicollinearity problems. The maximum likelihood method in SEM is known as SEM-ML or hard modeling. It is a covariance based approach and depends on a specific distribution pattern and need more cases [35]. PLS has widely been used in high-dimensional genomic data [36,37] to find out the influential genes that highly correlate with the response variable(s) and recently PLSPM has also been applied for genome wide association studies [38]. The PLSPM is able to handle several groups of data to identify inter- and intra- relationships based on inner and outer measurement model respectively, and it can be applied for microarray data in multivariate pathway levels. Applying several multivariate statistical tools in the different microarray experiments to elucidate the biological relation between organs may enable the generation of new hypotheses in biology.

Systems Biology

Systems biology is a holistic approach merging various experimental data, from the genome, proteome and metabolism in single cells and organs with the use of computational methods and predictive mathematical models [39,40]. Recently, systems biology has been referred to as a 'burgeoning field' [41] and 'executable biology' [42]. At first, a systems approach to biology was predicated on theoretical considerations of complex systems. Wiener introduced mathematical models of

complex systems control and communication in the 1940s [43]. In the 1960s and 1970s, Biochemical Systems Theory and Metabolic Control Theory were attempted to create simple mathematical models of biological systems [44]. However, such systems level approaches were not able to handle to connect the experimental molecules; molecules could be gene, modules/pathways, organs etc. In 1990s various omics platforms were developed to collect quantitative molecular data [45]. In 2002, Kitano mentioned that one should examine the structure and dynamics of cellular and organismal function instead of isolated parts of cell and organ to understand biological systems and this may have an impact on the future of medicine [46]. The combination of computational, experimental and observational enquiry in systems biology is highly relevant to drug discovery [47]. Basically, from 2002 the modern systems biology has started its modeling and network in different parts of the research. In order to understand biological systems, Aderem mentioned three basic concepts: emergence, robustness, and modularity. The details of these three concepts are mentioned in [48]. Systems biology is the combination of omics measurements, bioinformatics, statistics, metabolic engineering, computational sciences and mathematics. It is an attempt to detect a more integrated and hierarchical pattern that facilitates to build new biological pathways and networks at the cellular level [49].

Top-down and Bottom-up systems biology

In systems biology, two distinct approaches have evolved (i) bottom-up systems biology, namely computationally-based systems biology [50,51] and (ii) top-down systems biology, namely data-driven systems biology [51,52].

The bottom-up systems biology depends on computational modeling and simulation tools. The ultimate targets of bottom-up approach are to integrate and formulate the molecules in order to predict systems behavior and to combine pathway models into a global model for the entire systems under consideration.

The top-down approach mainly utilizes datasets that are mined in a discovery manner for new knowledge using a variety of bioinformatics and statistical tools. This inductive approach aims to determine new molecular mechanisms employing integrated data acquisition and analysis based on correlation [53]. Data-driven systems biology [54] has attempted to develop a more applied methodology for systems biological analysis. In this approach, researchers have been used a variety

of omics platforms with sophisticated statistical and bioinformatics tools to transform the discovery process in complex relationships among genetic, genomic, proteomic and metabolic pathway and networks. Recently, Martins dos Santos *et al* mentioned in their review of systems biology of the gut that systems biology is an integrated, modular modeling framework that cross-links top-down and bottom-up approaches for the various levels of biological organs [55].

Regulatory network and modeling

Generally systems biology handles three major topics (i) dynamic modeling of biological systems [56], (ii) reconstruction of regulatory networks [57] and (iii) integration as well as molecular interaction [58]. The keystone of systems biological research is mostly the focus on molecular interaction and this can easily be analyzed and visualized by tools such as Cytoscape [59] and R-spider [60]. Recently, an application of graph theory and network theory of biology has proven to be a powerful approach to gain insights into biological complexity and the advancement of systems biology [61]. Regulatory network analysis provides a powerful tool for describing complex systems, their components and their interactions in order to identify their topology, as well as the structures and functions of the components in broad way. This approach has been successfully applied to the representation of various systems in different kinds of data, such as in engineering and technology [62], life sciences [63,64], and social studies [65]. Xu *et al* identified and verified critical components of a transcriptional network directing lipogenesis, lipid trafficking and surfactant homeostasis in the mouse lung [66].

Biological types of data can be related to one another. In the Gaggle Genome Browser [67], heterogeneous data are joined by their location on the genome to create information-rich visualizations yielding transcription and its regulation. Systems biology is a rising consciousness of the composite dynamics of existing systems. Some computational methods have already been developed that can deal with the nonlinearity in signaling pathways in relationships between genotypes and phenotypes [68]. Although systems biology tends to focus on molecular networks, it utilizes analytic techniques designed to account for mounting properties arising from the background, flexibility and plasticity of the

function [69] in signaling pathways that contribute to disease phenotypes and treatment awareness.

Regulatory networks are the most emerging part of the systems biology. These are modeled as graphs, where nodes can be a gene/protein/module and directed edges represent transcriptional regulatory interactions. The reconstructions of these networks identify the spatial and temporal regulatory interactions between transcription factors (TFs) and their targets [70]. For exploiting the causal gene-gene temporal relationships, time series gene expression data are essential. These provide the dynamical properties of the molecular networks. Time-series gene expression data can also help to detect the dynamical properties of molecular networks, by exploiting the causal gene-gene temporal relationships. In the recent literature several dynamic models, such as TimeDelay-ARACNE [71]; ARACNE [72]; Dynamic Bayesian Networks [73]; Hidden Markov Model [74]; Ordinary Differential Equations [75,76]; and pattern signal processing approaches [77] have been proposed for reconstructing regulatory networks from time-course gene expression data.

Computational models of intracellular networks as well as a quantitative predictive model of gene expression are a foundation of systems biology. The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project is working with the current state of systems biology modeling and it organizes reverse-engineering challenges to infer the connectivity of the molecular networks underlying the measurements, or related reverse-engineering [78-80]. Researchers have developed various methods/algorithms to figure out the structure of different biological and artificial networks [81]. Recently, high-throughput experimental techniques have resulted in rapid accumulation of a wide range of omics data of various forms, providing in-depth understanding of biological processes. It is widely renowned that systems and network biology has the potential to increase our understanding of how nutrition influences metabolic pathways and homeostasis and how this regulation is disturbed in a diet-related diseases.

Nutritional systems biology (NSB)

Systems biological analysis with focusing on nutrition is known as nutritional systems biology. After a nutrient or other dietary component enters a cell, it may

influence gene expression by activation of specific transcription factors [6]. These TFs are therefore also called nutrient sensors. As a result metabolism may be modified. So it's important to understand this whole process- how it works and what the influences on the function of the living beings are. This process can be modulated by a number of internal (disease) or external (environmental) factors. Also important to consider is the association between omics data and these internal or external factors [82] as well as multilevel computational models [83] that integrates physiological mechanisms and different space-time scales related data. In order to see the nutrient control of eukaryote cell growth, Gutteridge *et al*/ conducted a comprehensive study of transcriptome, proteome and metabolism responses of chemostat cultures of the yeast and in four different nutrient-limiting conditions [84]. Every omics dataset represents the complexities of nutrition, physiology and cell biology. These datasets have been acquired and analyzed to get insight on the biological processes such as homeostasis, disease onset and optimal nutrition. For instance, even at the cellular level, simple pathways are highly interconnected [50,85]. The emergence of systems biology is also referred to as pathway, network, or integrative biology [46,49]. The aim of understanding the behavior of the system is to see as a whole rather than the behavior of the individual components [86-88]. Systems biology is the integrated approach for studying biological systems at the level of cells, organs or organisms by measuring and integrating genomics, proteomics and metabolomics data [89]. Furthermore, it's also useful to find out the promoter/transcription factor binding sites (TFBS) associations [90], which so far has been most successful in the yeast system [91].

The potential of systems biology is to provide a new dynamic for investigating personalized medicine and nutrition [92]. Systems biology has opened up a new outlook in our understanding of complex biological systems together with information technology, bioinformatics, statistical knowledge, and mathematical models. The expansion and use of omics platforms, particularly transcriptomics, proteomics and metabolomics, was discussed in detail by [93]. Transcriptome and proteome analyses were conducted by [94] to identify the fundamental molecular changes in hepatic lipid metabolism in zinc-deficient rats. They provided evidence for a rather complex regulatory network of zinc-dependent alterations in hepatic

metabolism. An integrated analysis to identifying molecular effects of diet was done by [95] using transcriptome data from three tissues (liver, muscle and adipose tissue) of mice with metabolic disease. They used low-density lipoprotein receptor-deficient (Ldlr $-/-$) mice which were fed a high fat diet to mimic a westernized diet. The diets were supplemented with herring. Transcriptome data was collected from the above three organs with some phenotype measurements (body composition, plasma lipids and aortic lesion). They found that the effect of diet on metabolic function in different tissues shows very clear effects that have implication for disease development.

A systems approach to identify early molecular signatures predicting genetic risk to metabolic diseases (Type 2 diabetes and obesity) using two strains of mice was done by [96]. They integrated different metabolic characterization, gene expression, protein-protein interaction networks, RT-PCR and flow cytometry data of adipose, skeletal muscle, and liver tissue of diabetes-prone C57BL/6NTac mice and diabetes-resistant 129S6/SvEvTac mice at 6 weeks and 6 months of age. They found that insulin resistance in mice with differential susceptibility to diabetes and metabolic syndrome is preceded by differences in the inflammatory response of adipose tissue.

A study to the pathogenesis of obesity-related nonalcoholic fatty liver disease (NAFLD) [2] using reverse phase protein microarrays (RPA) for multiplexed cell signaling analysis of adipose tissue from patients with NAFLD was done by [97]. They found that PKC (protein kinase C) delta, AKT (protein kinase B), and SHC phosphorylation changes occur in patients with simple steatosis. They also found that the amounts of cleaved caspase 9 and pp90RSK S380 were positively correlated in patients with nonalcoholic steatohepatitis (NASH) using Pearson correlation coefficient and specific insulin pathway signaling events are altered in the adipose tissue of patients with NASH compared with patients with non-progressive forms of NAFLD.

Metabolomics of the interaction between PPAR α and age in the PPAR α -null mouse was done by [98]. They used a combined (1)H nuclear magnetic resonance (NMR) spectroscopy and gas chromatography-mass spectrometry metabolomics approach to examine metabolism in the liver, heart, skeletal muscle and adipose tissue in PPAR α -null mice and wild-type controls during ageing between 3 and 13 months. Their metabolomics study, using multivariate statistical techniques:

partial least squares (PLS) and partial least squares discriminate analysis (PLS-DA), demonstrated that a loss of PPAR α results in a marked reduction in hepatic glucose/glycogen and subsequent hepatic steatosis with age.

It is important to know the association/integration among the nutrients/diets, transcriptomics, proteomics, metabolomics and phenotypes such as weight and disease status etc. The following Figure 2 shows an overview of such integration among the different kinds of data. This kind of study belongs to the top-down systems biology to detect the association among the multi datasets using multivariate data mining techniques.

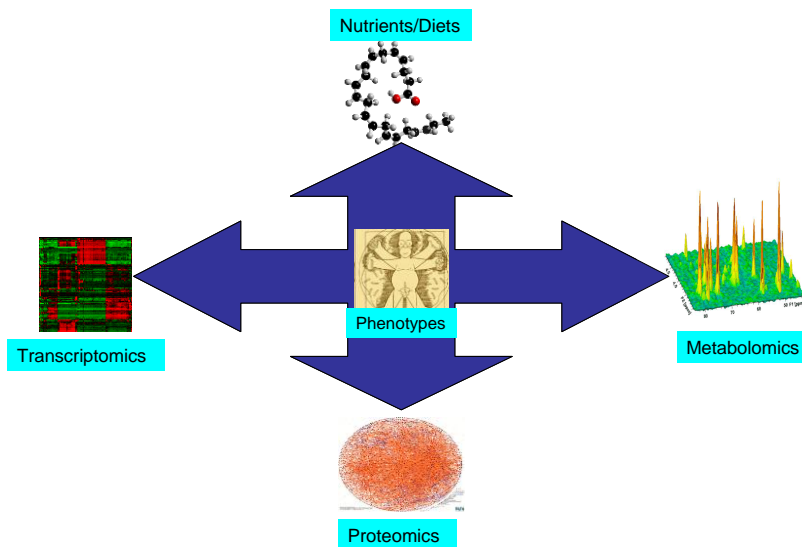


Figure 2: Integrate different omics datasets with phenotypes.

To identify the association between transcriptomics and proteomics data, the R package *mixOmics* can be applied by using regularized canonical correlation or sparse partial least squares [99]. Another way to integrate different datasets is by using the *FactoMineR* package [100], that is based on multiple factor analysis. In this package, one can handle different sets of high-throughput data with supplementary variables (e.g., plasma measurements, weight status etc.) To

uncover the causal relationships among the different blocks of variables partial least squares path model is very well known in chemometrics, econometrics and sociological data, and its relevance for the analysis of high throughput biological data such as microarray data is being reported [38,101]. This is a multivariate technique to collect the path coefficients and the loadings of the variables, and it is easy to analyze by *plspm* package [30] in R program.

Dynamic modeling of gene expression and gene regulatory network are the most useful terms in the field of systems biology. But it is also important to build meta gene modeling as well meta gene network using systems biological techniques to elucidate how biological process evolve over time instead of individual gene. This kind of meta gene or eigen gene modeling and network can be done using some data reduction technique like PCA (principal component analysis) to find out meta gene expression in different homogeneous genes cluster or modules. Afterwards, some models, for instance: statistical (linear or non-linear regression model), kinetic or mechanistic model can be built by repeated approach as well as meta gene network. These kinds of analyses can also be done for protein or metabolites. The types of models completely depend on data and the objectives of the research. Prifti *et al* developed an R package *FunNet* as well as web based tools to explore the transcriptional network on gene co-expression based on correlation [102]. But their approach does not cover the causal relationship between the gene co-expression networks; therefore, after creating meta gene expression by multivariate techniques the causal meta gene network can be done by *TDARACNE* [71].

Software tools

In order to implement the different approaches in the different fields, we need software and statistical tools. Of course, researchers like to analyze their data using valid tools that are freely available. Many software tools are available and almost every week some new tools are coming in this area. Among them some are easy to handle by biologists, some are free and some are not free. In this limited overview (Table 1), we list some useful software tools, mostly R packages (because these are free and easy to handle) and its function that are related to systems biological analyses. Some of these tools have already discussed above.

Table 1: Some useful and related software tools in the systems biological field

Software	Function	Reference
Top-down		
TDARACNE (R package)	Reverse engineering of gene network from time course data	[71]
GeneTS (R package)	Gene association network based on an empirical Bayes approach.	[103]
VAR network (R code)	Causal networks based on the vector autoregressive (VAR) process.	[77]
pcalg (R package)	Standard and robust estimation of the equivalence class of a Directed Acyclic Graph (DAG) via the PC-Algorithm. Predicting causal effects in large-scale systems from observational data.	[104]
Deal (R package)	Bayesian networks with continuous and/or discrete variables can be learned and compared from data.	[105]
BNArray (R package)	Constructing gene regulatory networks from microarray data by using Bayesian network	[106]
rHVDM (R package)	Hidden variable dynamic modeling to predict the activity and targets of a transcription factor	[107]
SysNet	For interactive analysis of molecular expression information in systems biology based on correlation	[108]
PathVisio	Presenting and exploring biological pathways	[109]
Cytoscape*	Integrate models of biomolecular interaction networks and visualization tool.	[59]
payao	It is a community-based, collaborative web service platform for gene-regulatory and biochemical pathway model creation	[110]
MetNet	Enable to visualize, statistically analyze and model a metabolic and regulatory network map of Arabidopsis, combined with gene expression profiling data.	[111]

VitisNet	"Omics" integration through grapevine molecular networks	[112]
GeneNet (R package)	Gene network based on partial correlation	[113]
MetNetGE	Visualization tool that organizes biological networks according to a hierarchical ontology structure	[114]
FactoMineR* (R package)	Able to handle multifactor data as well as to ingrate between them using multivariate statistical tools	[100]
integrOmics /mixOmics (R package)	Integrate two different datasets like transcriptomics and proteomics.	[99]
Simca-p	Able to handle multivariate normal and non-normal data by using PCA and PLS techniques	[115]
Unscrambler	Able to handle multivariate data as well as to integrate between them using different kinds of statistical tools	[116]
Plspm* (R package)	Handle several groups of multivariate data with causal relationships.	[117]
FunNet (R package as well as web based tool)	Transcriptional network (gene co-expression based on correlation)	[102]
Minet (R package)	Transcriptional network (gene to gene based on mutual information)	[118]
GeneAnswers (R package)	Provide an integrated tool for biological or medical interpretation of the given one or more groups of genes	[119]
CoGAPS (R package)	To identify patterns and biological process activity in transcriptomic data	[120]
iPath	Visualize the metabolic pathways	[121]
R spider*	Pathway network based on KEGG and Reactome	[60]
Ingenuity*	Pathway analysis and visualization of the interaction among the molecules	[26]
Genomatix*	Transcription factor binding sites and promoter analysis	[122]
Bottom-up		
Celldesigner	A modeling tool of biochemical networks	[123]
SBML (Matlab tool box)	Facilitates importing and exporting models	[124]

	represented in the Systems Biology Markup Language (SBML)	
Inferelator (R code)	Gene regulatory network based on sigmoidal, or Logistic model with the help of kinetic equation.	[125]
COPASI	For simulation and analysis of biochemical networks and their dynamics based ODE	[126]
DIPSBC	Data integration platform for systems biology collaborations by XML data format	[127]

**Used in this thesis*

The primary purpose of many biological research projects is to identify the gene(s), protein(s) or molecule(s) that are potentially related to a certain biological problem (disease). Once a list of potential target genes has been found using the proper statistical methods, the next task is to see the pathways or networks where these genes are significantly over-represented or not. Besides these, another most important thing is to visualize the output, for instance: gene-gene interactions, protein-protein interactions, pathway networks, metabolic pathways, visualize multivariate data and their interpretation to get better insight of biological function which leads to create another hypothesis. Recently, Ghelenborg *et al* have nicely been discussed how to visualize of omics data for systems biology [128].

The most of the packages and software in the above list are useful for both top-down and bottom-up approaches. CellDesigner, SBML, and COPASI are very useful for kinetic/mechanistic modeling. CellDesigner [123] is very handy to draw pathway/biological model and to produce SBML file, afterwards this SBML file can be used to simulate model and predicting by COPASI [126].

Perspectives

Until the end of last century, no significant developments occurred in the area of systems biology, although some modeling approach were conducted in 1960s and 70s. Initially, it was meant only in biological area. Recently, this concept is applied by other research fields as well, such as ecology, sociology, and medicine. In the eve of this century, some significant reviews and studies were published with software tools and approaches in systems biological field. However, it's not enough yet to uncover the function of whole organ in the livings beings. To integrate different datasets, it's very essential to use the proper statistical tools. Here we mentioned some well-known statistical tools for univariate and multivariate data analysis of microarray studies. Some PPAR α related articles also discussed in this overview to know the function of this transcription factor on the gene expression in different organs of mice. Especially, here we focused on the top-down systems biological literature on nutritional studies as well as some useful analytical software tools for systems biological analyses.

Based on literature, we may conclude that very few integrative analyses were performed by top-down systems biological approach focusing transcriptomics data of dietary fat in different organs or integrating omics data. Still more ideas and studies are necessary to reveal the function of nutrition in whole living beings. Here, we mentioned a schematic overview to integrate not only different omics data related with the nutrients but also with phenotype data. Besides the interaction among the transcriptomics, proteomics, metabolism and phenotypes, we need to know their behavior over time. Therefore, it would be more meaningful to produce such kind of omics data over time to reveal the evolution of nutritional components in the living beings. Still the research in nutritional systems biological is at its infancy, so many things need to be explored this aspect. However, this overview might be given us some clue to analyze nutritional systems biological analysis in future.

Aim and outline of this thesis

Although the function of the fatty acid sensor PPAR α has been extensively studied at the level of different organs, less is known about the systems-wide functional implications of PPAR α activation for metabolic health and plasticity of organs as well as its behavior over time under nutritional relevant conditions. Since the average Western diet, but also that increasingly more in developing countries, contains high amounts of fat, a comprehensive systems-wide understanding of the role of fatty acid-sensing mechanisms such as PPAR α is of great importance [46,49,92]. Applying or developing NSB approaches might help to interpret new experimental nutrigenomics data on transcriptional responses to dietary fat and may provide better insight into the biological implications of fat-specific responses in different metabolic organs as relevant for homeostasis, metabolic plasticity and prevention of metabolic diseases such as morbid obesity or diabetes type 2 [6] .

The aim of the research described in this thesis was to integrate and model different organ-specific transcriptomics datasets related to lipid-sensing by nutritional systems biological approaches, especially to characterize the function of PPAR α .

The different effect sizes (Fold Changes, i.e., mean differences between treatment and control groups) in experiments are a big challenge in the analysis of high throughput genomics studies and, therefore, in **chapter 2** an integrated statistical approach is presented to identify transcription factor target genes from transcriptomics data across different experiments. **Chapter 3** deals with the integration and modeling of multivariate data, an important challenge in top-down systems biology. An approach to characterize a pathway score and to integrate different time course and organ specific transcriptomics data by a path model are described here. In **chapter 4** characterization and modeling of acute effects of PPAR α activation in rat liver cells is investigated. In **chapter 5**, we focused to detect the time and dose dependently regulated genes in liver and white adipose tissue during the development of high-fat diet induced obesity in mice. Moreover, we studied the correlation of these genes with the different plasma factors (glucose, leptin, adiponectin, resistin, Il6 and tPAI-1) and weight

status indicators (BW at start of intervention, BW at section, BW gain, absolute liver weight, and relative liver weight). Finally, the general discussion and conclusions are presented in **chapter 6**.

Chapter 2

An integrated statistical approach to compare transcriptomics data across experiments: a case study on the identification of candidate target genes of the transcription factor PPAR α

Mohammad Ohid Ullah, Michael Müller, and Guido JEJ Hooiveld

Abstract

An effective strategy to elucidate the signal transduction cascades activated by a transcription factor is to compare the transcriptional profiles of wild type and transcription factor knockout models. Many statistical tests have been proposed for analyzing gene expression data, but most tests are based on pair-wise comparisons. Since the analysis of microarrays involves the testing of multiple hypotheses within one study, it is generally accepted that one should control for false positives by the false discovery rate (FDR). However, it has been reported that this may be an inappropriate metric for comparing data across different experiments. Here we propose an approach that addresses the above mentioned problem by the simultaneous testing and integration of three hypotheses (contrasts) using the cell means ANOVA model. These three contrasts test for the effect of a treatment in wild type, gene knockout, and globally over all experimental groups. We illustrate our approach on microarray experiments that focused on the identification of candidate target genes and biological processes governed by the fatty acid sensing transcription factor PPAR α in liver. Compared to the often applied FDR-based across experiment comparison, our approach identified a conservative but less noisy set of candidate genes with similar sensitivity and specificity. However, our method had the advantage of properly adjusting for multiple testing while integrating data from two experiments, and was driven by biological inference. Taken together, in this study we present a simple, yet efficient strategy to compare differential expression of genes across experiments while controlling for multiple hypotheses testing.

Introduction

Genome-wide transcriptional profiling, or transcriptomics, is extensively used to study how cells respond to certain stimuli or to diagnose and predict clinical outcomes [129-132]. Transcription factors (TFs) are the key effectors which control gene expression. From a variety of research fields, including nutrition sciences, there is a major interest in characterizing the genes and networks that are controlled by transcription factors. Advances in genome-wide expression profiling methodologies and the availability of model systems offered new, powerful tools to address this [6,133-138].

An effective strategy to elucidate the signal transduction cascades activated by transcription factors is through transcriptional profiling. Transcription profiling can be applied on gain- and loss-of-function TF mutants, and changes in global gene expression that are associated with the various phenotypes could then be used for a comprehensive understanding of TF function [133,134,138-140]. To this end, transcription factor target genes have to be efficiently and accurately identified from the transcriptomics dataset. It is important to realize that from a biological perspective, TF target genes are only those genes that do significantly respond upon treatment with a potent agonist or gain of function, in wild type but not mutant (knockout) models. However, from a statistical inference point of view the identification of biological relevant target genes from such 2x2 factorial experiments is less straight-forward.

It is generally accepted that statistical testing is required to reliably identify differentially expressed genes (reviewed in e.g. Allison *et al* [141]). Moreover, since the statistical analysis of microarrays involves the testing of multiple hypotheses (genes) within one study, it is necessary to control for false positives. A frequently used metric to quantify the level of confidence any particular gene is differentially expressed, that takes into account multiple testing, is the false discovery rate (FDR) [141]. Therefore in many studies a cutoff based on the FDR rather than p-value is used to select significantly regulated genes within experiments, which subsequently are compared across experiments to identify transcription factor target genes. However, Higdon *et al* [142] reported that the use of the FDR and its associated q-value may result in inconsistent and misleading interpretation of the comparisons across different experiments, especially when the effect sizes of the experiments vary dramatically, as for

example is the case when comparing effects of potent agonists in wild type and TF knockout models.

Therefore, the purpose of the work described in the current paper is to present a strategy that optimally integrates and controls for multiple hypotheses testing using data obtained from two biological systems that respond completely different to a treatment. We outline our approach using one of our datasets on the mouse peroxisome proliferator-activated receptor alpha (PPAR α) [15]. PPAR α is a TF belonging to the nuclear receptor superfamily, and is activated by a variety of compounds, including dietary fatty acids and their derivatives as well as synthetic agonists [7,9,143].

Material and Methods

Experimental data

We illustrate our approach (Figure 1) on one of our publicly available datasets (Gene Expression Omnibus (GEO) accession: GSE8295). This dataset was generated to identify PPAR α target genes in mouse liver [15], and was also used by Higdon *et al* [142] to illustrate the inappropriateness of using the FDR as cut-off metric when comparing two transcriptomics experiments with different effect sizes.

Briefly, pure bred wild type (129S1/SvImJ) and PPAR α -null (129S4/SvJae-Pparatm1Gonz/J) mice [144] were fed chow or chow supplemented with 0.1% WY14643 (Chemsyn, Lenexa, KS) for 5 days ($n = 4$ mice per group). WY14643, (4-Chloro-6-[(2,3-dimethylphenyl)amino]-2-pyrimidinyl)sulfanyl)acetic acid (CAS: 50892-23-4), is a chemical that was developed by the pharmaceutical industry to lower serum cholesterol. It is not used in clinical applications, but it is rather used as prototype chemical to induce peroxisome proliferation. WY14643 is a highly specific and potent agonist for PPAR α and is therefore often used in studies on this nuclear receptor [12,145]. On the sixth day, mice were anaesthetized and livers were excised. Total RNA was prepared using TRIzol reagent (Invitrogen, Carlsbad, CA) followed by purification using the RNeasy mini kit (Qiagen, Hilden, Germany). RNA integrity was checked by chip analysis (Agilent 2100 Bioanalyzer, Agilent Technologies, Amsterdam, the Netherlands) according to the manufacturer's instructions. RNA was judged as suitable for array hybridization

only if samples exhibited intact bands corresponding to the 18S and 28S ribosomal RNA subunits, and displayed no chromosomal peaks or RNA degradation products, and had a RNA integrity number (RIN) above 8.0). The Affymetrix GeneChip RNA One cycle Amplification Kit (Affymetrix, Santa Clara, CA) was used to prepare labeled cRNA from 5 μg of total RNA, which subsequently was hybridized on Affymetrix Mouse Genome 430 2.0 plus arrays. The animal study was approved by the Local Committee for Care and Use of Laboratory Animals.

Cell Means ANOVA Model

The dataset on the identification of PPAR α target genes in mouse liver has a 2x2 factorial design; that is factor ‘treatment’ has 2 levels (WY, Control), as has the factor ‘genotype’ (wild type, knockout). Analysis of variance (ANOVA) is commonly used for analyzing data from experiments with multiple categorical factors [146,147]. To appropriately identify candidate PPAR α target genes, we propose to perform and integrate three comparisons using the cell means ANOVA model [148]. For every probeset the model was defined as follows:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where Y_{ijk} is the expression of a probeset at i^{th} treatment (1 for WY, 2 for Control) in j^{th} strain of genotype (1 for WT, 2 for KO) and k^{th} replication ($n=4$), μ_{ij} is the mean value of i^{th} treatment and j^{th} strain of each gene, and ε_{ijk} is a random error term which follows normal distribution with mean = 0 and variance = σ^2 .

Formally, the definition of a contrast C is expressed below, using the notation μ_j for the j^{th} treatment mean:

$$C = c_1\mu_1 + c_2\mu_2 + \dots + c_j\mu_j + \dots + c_k\mu_k$$

$$\text{Where, } c_1 + \dots + c_j + \dots + c_k = \sum_{j=1}^k c_j = 0$$

As stated before, from a biological perspective, candidate PPAR α target genes are only those genes that do significantly respond upon treatment with the potent PPAR α agonist WY14643 in wild type but not in PPAR α knockout mice. Therefore three different contrasts (comparisons) from this 2x2 factorial experiment were

combined to infer the probesets that were significantly and PPAR α -dependently regulated. The different contrasts tested were (Table 1):

Contrast 1: $H_0: \mu_{11} - \mu_{21}=0$ versus $H_1: \mu_{11} - \mu_{21} \neq 0$, returning all probesets regulated in the wild type mice by the agonist WY;

Contrast 2: $H_0: \mu_{12} - \mu_{22}=0$ versus $H_1: \mu_{12} - \mu_{22} \neq 0$, returning all probesets regulated in the PPAR α knockout mice by the agonist WY; and

Global Contrast: $H_0: (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) = 0$ versus $H_1: (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) \neq 0$, returning the overall differential expressed probesets in wild type versus knockout mice groups after treatment with WY compared to control.

Table 1: The contrasts defining the different hypotheses.

μ_{ij}	Levels	Contrast 1 $H_0: \mu_{11} - \mu_{21}=0$	Contrast 2 $H_0: \mu_{12} - \mu_{22}=0$	Global Contrast $H_0: (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) = 0$
μ_{11}	WY, WT	1	0	1
μ_{12}	WY, KO	0	1	-1
μ_{21}	Con, WT	-1	0	-1
μ_{22}	Con, KO	0	-1	1

The PPAR α -dependently regulated probesets were then identified by extracting those probesets that were only significantly regulated in both Contrast 1 and Global Contrast, and subsequently corrected for multiple testing.

Implementation

All analyses were performed in R [149], using packages from the Bioconductor project [150]. Probesets were redefined according to Dai *et al* [151]. In this study, probes were reorganized based on Entrez Gene database, build 36, version 2 (remapped CDF version 12). Our workflow was as follows (note that since we used a remapped chip definition file based on the Entrez Gene database, the terms probeset and gene are used interchangeably):

1. Expression estimates were obtained by GC-robust multiarray (GCRMA) normalization, using the empirical Bayes approach to adjust background [152].

2. For each of the three above-mentioned contrasts, differentially expressed probesets (genes) were identified using linear models, as implemented in *limma* [153]. For each contrast probesets were selected based on $p < 0.05$.
3. Probesets that were common only in Contrast 1 and the combined Global Contrast were identified. This set of probesets represented only transcription factor regulated genes, and was designated X.
4. Multiple testing was corrected by using a false discovery rate method [27], based on the Global Contrast considering the output of all probesets. Probesets in X that satisfied the criterion $FDR < 5\%$ were considered to be transcription factor target genes.

A schematic overview of our implementation is also given in the Figure 1, and the R-code and other required files are available as supplemental material (<http://www.la-press.com/an-integrated-statistical-approach-to-compare-transcriptomics-data-acr-article-a3222>).

Validation

To validate our integrated approach, obtained results (Figure 2) were compared to results from the across experiment comparison (Figure 3) using two sets of well-established PPAR α target genes obtained from a recent review (Table 1 from Rakhshandehroo *et al* [18]).

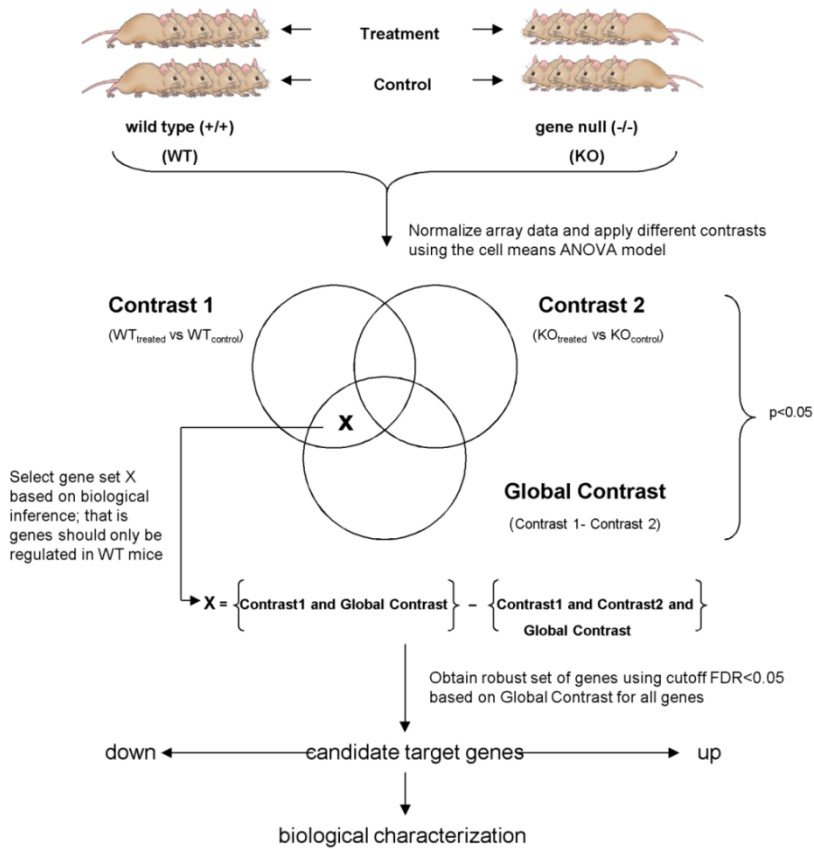


Figure 1: Overview of our integrated strategy. After normalization, transcriptome data are analyzed for differentially expressed probesets (genes) using three contrasts (comparisons): Contrast 1, representing probesets regulated by a specific treatment in wild type mice; Contrast 2, representing probesets regulated by the same treatment but in knockout mice, and Global Contrast, representing genes differentially regulated by the treatment between the WT and KO mice. Biologically irrelevant probesets, i.e., probesets that are also regulated by the treatment in the KO mice, are discarded, resulting in a set of probesets called X. To correct for multiple testing, FDR values of the probesets in X are calculated using the p-values obtained in Global Contrast for all probesets. A robust set of putative target genes regulated by the knocked-out gene is obtained by selecting those probesets from X that fulfill a Global Contrast-based FDR cutoff, e.g. FDR < 0.05. This set can subsequently be divided in up- and down-regulated genes.

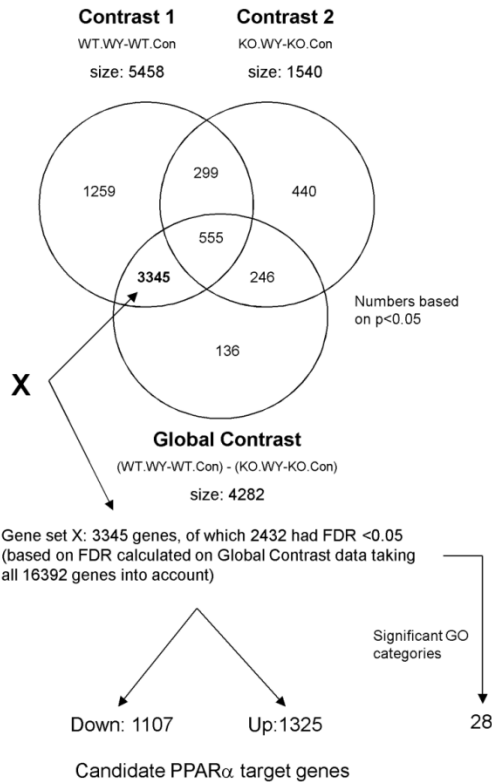


Figure 2: Application of the integrated approach on PPAR α dataset GSE8295.

Expression estimates were calculated by GCRMA normalization. Differentially expressed probesets were identified using three contrasts using p -value < 0.05 . Contrast 1, representing probesets regulated by the specific PPAR α agonist WY in wild type mice; Contrast 2, representing probesets regulated by WY14643 in PPAR α knockout mice, and Global Contrast, representing probesets differentially regulated by WY14643 between the WT and PPAR α KO mice. Biologically irrelevant probesets, i.e., those 854 probesets that were regulated by WY14643 in both WT and PPAR α KO mice, were discarded, resulting in a set of probesets called X of size 3345 that were only regulated in Contrast 1 and Global Contrast. To correct for multiple testing, FDR values (Benjamini Hochberg procedure) of the probesets in X were calculated based on the p -values for all probesets obtained in Global Contrast. A robust set of candidate PPAR α target genes was obtained by selecting those 2432 probesets from X that had Global Contrast-based FDR value < 0.05 . This set was divided in 1325 up- and 1107 down-regulated probesets.

These sets are available as supplemental material. The true positive rate (sensitivity) as function of the false positive rate (1-specificity) for different cutoff points was plotted for both the across experiment comparisons and our integrated approach using the R-library ROCR [154].

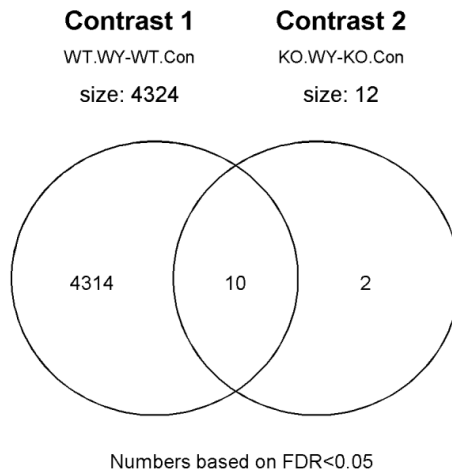


Figure 3: The FDR based across experiment comparison of PPAR α dataset GSE8295. Expression estimates were calculated by GCRMA normalization. Differentially expressed probesets were identified in each contrast. Using a FDR value < 0.05 criterion, 4324 probesets were regulated in the wild type experiment (Contrast 1), whereas 12 genes were changed in the knockout experiment (Contrast 2). Of these 12 genes, 10 were also regulated in the wild type experiment. Thus, when comparing across experiments with a FDR value cutoff level of 0.05, 4314 genes were considered PPAR α target genes.

The partial area under the ROC curve was calculated using $p=0.2$ (thus 1-specificity = 0.2) as cutoff. This cutoff value was chosen because for the identification of transcription factor target genes a high specificity is required (>80%) before considering its sensitivity [155]. In addition, the biological features that were overrepresented in the lists of candidate PPAR α target genes that were generated on the basis of both approaches were analyzed with the software tool Ontologizer [156], applying the 'parent-child-union' (PCU) algorithm and using the biological process ontology of Gene Ontology.

Results and Discussion

Identification of candidate PPAR α target genes

The application of transcriptomics to compare the effects of specific agonists, such as WY14643, in wild type and PPAR α knockout mice is a powerful approach to identify candidate PPAR α target genes [139,140]. However, when comparing across different experiments the use of FDR cutoff values may result in inconsistent and misleading interpretation of the data [142]. In this study we propose a simple yet effective strategy that avoids comparing probesets across experiments based on FDR values while still controlling for multiple testing. Testing three different hypotheses (contrasts) for each probeset allowed the robust identification of transcription factor target genes. Since only the interaction effects are of interest for identifying candidate target genes, the cell means ANOVA model was used to infer this 2x2 factorial design.

The number of probesets significantly regulated ($p < 0.05$) upon PPAR α activation by WY14643 in wild type mice (= Contrast 1) equaled to 5458, whereas in PPAR α -/- mice (Contrast 2) this number was only 1540 (Figure 2). Such a large difference was expected since the KO mice do not express any functional PPAR α . The Global Contrast, incorporating expression information for all probesets in all groups, identified 4282 significantly regulated probesets ($p < 0.05$) (Figure 2), representing genes that from an inferential perspective are differentially regulated by WY between the two mouse strains. However, these included genes that for example were only regulated in the KO mice, or were regulated in wild type and, although to a lesser extent, still in KO mice. To filter out these 'biological irrelevant' genes, only probesets that were common in Contrast 1 and Global Contrast were retained, resulting in a set of 3345 probesets, which was called set X. Thus, this set X contained only probesets that from a biological perspective fulfill the criterion of being candidate PPAR α target genes. To correct for multiple testing, FDR values of the 3345 genes in X were calculated based on all 16392 genes in Global Contrast, since in this comparison statistical inference was simultaneously adjusted for both experiments in wild type and knockout mice. Finally, a robust set of PPAR α target genes was obtained by selecting those 2432 probesets from set X that fulfilled the criterion $FDR < 0.05$ (Figure 2). Of these, 1325 probesets were induced and 1107 probesets were suppressed.

For comparison, we also generated a list of candidate PPAR α target genes that were generated on the basis of directly comparing the wild type and knockout experiment using a FDR cutoff (Figure 3). Note that this frequently used approach is criticized [142] and that it is in essence identical to the analysis strategy published and interpreted by Rakhshandehroo *et al* [15], except that these authors also employed a fold change cutoff. Using a FDR cutoff of 0.05, we identified 4324 probesets that were regulated in the wild type experiment (Contrast 1), whereas 12 probesets were changed in the knockout experiment (Contrast 2). Of these 12 probesets, 10 were also regulated in the wild type experiment. Thus, the FDR based comparison of these two experiments identified 4314 probesets that should be considered PPAR α target genes.

The number of FDR based selected probesets was about twice as large as the list of probesets obtained using our integrated approach (4314 versus 2432 probesets). Comparison of these two sets of candidate genes revealed that almost all (i.e., 99%) of the probesets obtained by our integrated approach were also identified when using a FDR cutoff (Figure 4). This indicates that while Global Contrast is more conservative it will identify similar if not identical biological features (see also section on validation).

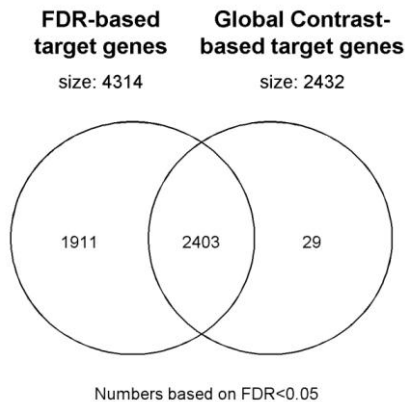


Figure 4: Venn diagram of the identified candidate PPAR α target genes obtained by our integrated approach or the FDR based across experiment comparison of PPAR α dataset GSE8295. Almost all (99%) of the candidate target genes identified by our proposed approach were also identified in the FDR based across experiment comparison.

It is important to realize that the results of statistical hypothesis testing are never free of error. Two types of error are distinguished: type I error, i.e., rejecting the null hypothesis when it is in fact true, and type II error, i.e., not rejecting the null hypothesis when in fact the alternative hypothesis is true. In other words, occurrence of the former leads to inclusion of false positives whereas the latter leads to inclusion of false negatives. Consequently, we cannot exclude that the set of 1911 probesets that were discarded by Global Contrast contained false negatives that otherwise would have been retained. However, especially within the context of genome-wide screening studies for candidate genes, we believe that limiting type I error is of primary concern, and that of type II error is of secondary importance. Thus, to err on the safe side we prefer to control for false positives rather than for false negatives. Moreover, the probesets that were discarded by Global Contrast were characterized by a relatively low effect size compared to the probesets that were still included. The mean of the absolute coefficients (\log_2 of the fold-change) of the excluded probesets was 0.36 (equaling to a mean fold change of 1.28), and was 0.87 (mean FC = 1.83) for the included probesets. Taken together, we showed that compared to the FDR based across experiment comparison our approach identified a conservative set of more robustly regulated candidate PPAR α target genes. We believe this is advantageous because a clear overview of candidate genes and corresponding biological processes normally is aimed for.

Validation

To compare the performance of our integrated approach with that of the FDR based across experiments comparison, we first performed sensitivity versus specificity analysis. To this end two benchmark sets of well-established PPAR α target genes were selected from a review that summarized the latest literature on this topic [18]. We created two benchmark sets; one set containing only 32 genes, and another set containing 189 genes. The smaller benchmark set contained only genes that were demonstrated to be bona fide PPAR α target genes in both human and mouse liver and that do contain a functional PPAR response element (PPRE) in the regulatory regions. The larger benchmark set contained all genes that were demonstrated to be PPAR α -dependently regulated in mouse liver but for which no functional PPRE has yet been identified. We next plotted the true positive rate (sensitivity) as function of the false positive rate (1-specificity) for different cutoff

points for both our integrated approach and the across experiment comparison (Figure 5).

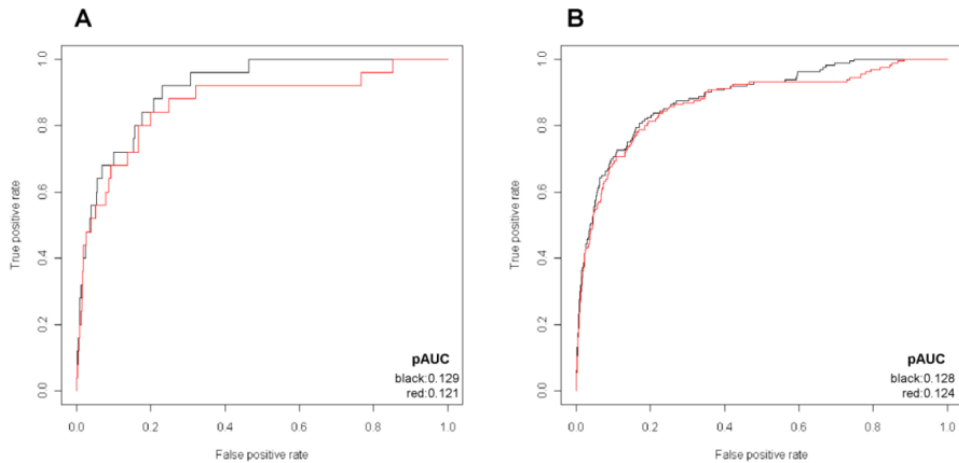


Figure 5: Sensitivity versus specificity of our proposed method and the across experiment comparison. The sensitivity versus specificity was analyzed using two benchmark lists of established PPAR α target genes derived from literature. Panel A: ROC curve for both methods using a set of 32 benchmark genes that were demonstrated to be PPAR α target genes in both human and mouse liver and that do contain a functional PPAR response element (PPRE) in the regulatory regions. Panel B: ROC curve for both methods using a set of 189 benchmark genes that were demonstrated to be PPAR α -dependently regulated in mouse liver but for which no functional PPRE has been identified yet. Red lines: ROC curves of our integrated approach; Black lines: ROC curves of the across experiment comparison.

Even though our approach identified a conservative list of candidate genes, we observed that it performed very similar to the across experiment comparison in identifying known PPAR α target genes, which was also reflected by almost identical partial area under the ROC curve (pAUC; $p=0.2$) for both methods. Values were 0.129 and 0.121, respectively for the across experiment comparison and our integrated approach when the smaller set of 32 PPAR α target genes was used, whereas these numbers were 0.128 and 0.124, respectively, for the larger set of 189 putative PPAR α target genes.

Next we detected and compared the biological features that were overrepresented in the lists of candidate PPAR α target genes that were either

generated by our approach or the across experiment comparison. Enriched biological processes were identified by overrepresentation analysis based on Gene Ontology (GO) categories, which is a generally accepted procedure to achieve this [141]. To this end the software tool Ontologizer was used [156], applying the ‘parent-child-union’ (PCU) algorithm. This algorithm takes the graph structure of GO into account, thereby reducing false-positive and biologically misleading results [157]. To this end the software tool Ontologizer was used [156], applying the ‘parent-child-union’ (PCU) algorithm. This algorithm takes the graph structure of GO into account, thereby reducing false-positive and biologically misleading results [157]. To this end the software tool Ontologizer was used [156], applying the ‘parent-child-union’ (PCU) algorithm. This algorithm takes the graph structure of GO into account, thereby reducing false-positive and biologically misleading results [157]. To this end the software tool Ontologizer was used [156], applying the ‘parent-child-union’ (PCU) algorithm. This algorithm takes the graph structure of GO into account, thereby reducing false-positive and biologically misleading results [157].

A							B								
GO ID	Name	NSP	P-value	Adj. P-value	Rank	Pop. Count	Study Count	GO ID	Name	NSP	P-value	Adj. P-value	Rank	Pop. Count	Study Count
GO:000152	metabolic process	B	4.16e-42	3.73e-43	1	6526	2105	GO:000152	metabolic process	B	7.91e-23	1.15e-23	1	6526	1199
GO:0042180	cellular ketone metabolic process	B	9.41e-22	8.39e-23	2	519	265	GO:0042180	cellular ketone metabolic process	B	3.74e-22	3.74e-22	2	519	170
GO:0006082	organic acid metabolic process	B	3.28e-20	3.36e-18	3	509	257	GO:0006082	organic acid metabolic process	B	1.56e-16	5.56e-15	3	509	164
GO:0051186	cofactor metabolic process	B	1.29e-18	6.39e-17	4	191	119	GO:0051186	cofactor metabolic process	B	7.07e-16	2.99e-14	4	191	81
GO:0006429	lipid metabolic process	B	1.03e-13	6.83e-12	5	681	295	GO:0006429	lipid metabolic process	B	4.41e-11	3.10e-09	5	681	182
GO:0009987	cellular process	B	1.43e-12	4.10e-11	6	9424	2694	GO:0009987	cellular process	B	6.84e-11	1.93e-09	6	308	91
GO:004237	cellular metabolic process	B	4.23e-12	1.04e-10	7	5651	1798	GO:0009987	cellular process	B	1.52e-09	3.47e-08	7	9424	1547
GO:0007038	amine metabolic process	B	8.10e-12	1.76e-10	8	308	141	GO:0006091	generation of precursor metabolites and energy	B	5.25e-08	1.02e-06	8	230	74
GO:0006091	generation of precursor metabolites and energy	B	1.27e-10	2.43e-09	9	230	119	GO:0006519	cellular amino acid and derivative metabolic pro...	B	5.75e-06	1.05e-06	9	281	87
GO:0006519	cellular amino acid and derivative metabolic pro...	B	2.29e-09	3.94e-08	10	281	136	GO:0051234	establishment of localization	B	8.75e-05	1.49e-06	10	2362	442
GO:0051234	establishment of localization	B	9.55e-09	1.49e-07	11	2362	742	GO:0044235	cellular lipid metabolic process	B	1.90e-07	2.93e-06	11	467	127
GO:0015021	protein transport	B	1.28e-07	1.83e-06	12	660	260	GO:0009976	carbohydrate metabolic process	B	4.94e-07	5.69e-06	12	417	112
GO:0050817	coagulation	B	3.18e-07	4.20e-06	13	85	39	GO:0051179	localization	B	2.59e-06	3.37e-05	13	2714	488
GO:0044235	cellular lipid metabolic process	B	1.86e-06	2.26e-05	14	467	196	GO:0050817	coagulation	B	0.000194	0.00185	14	85	23
GO:0051179	localization	B	3.46e-06	3.97e-05	15	2714	818	GO:0006518	peptide metabolic process	B	0.000237	0.00287	15	40	17
GO:0005975	carbohydrate metabolic process	B	9.23e-06	9.92e-05	16	417	169	GO:0050896	response to stimulus	B	0.000276	0.00282	16	2159	380
GO:0006396	RNA processing	B	5.08e-05	8.000514	17	367	116	GO:0044237	cellular metabolic process	B	0.000311	0.00309	17	5651	1009
GO:0050878	regulation of body fluid levels	B	6.58e-05	0.000629	18	123	45	GO:0006725	cellular aromatic compound metabolic process	B	0.000621	0.00564	18	132	39
GO:0006689	lipid transport	B	0.0001	0.00566	19	116	53	GO:0048771	tissue remodeling	B	0.00162	0.0136	19	70	18
GO:0007031	peroxisome organization	B	0.000273	0.00235	20	16	11	GO:0050878	regulation of body fluid levels	B	0.00162	0.0137	20	123	27
GO:0006725	cellular aromatic compound metabolic process	B	0.000362	0.00297	21	132	57	GO:0009225	nucleotide-sugar metabolic process	B	0.00264	0.0205	21	15	7
GO:0006996	organelle organization	B	0.000461	0.0361	22	1020	328	GO:0015021	protein transport	B	0.00267	0.0205	22	660	148
GO:0006096	response to stimulus	B	0.000482	0.0361	23	2159	625	GO:0006869	lipid transport	B	0.00302	0.0222	23	116	34
GO:0006805	xenobiotic metabolic process	B	0.000533	0.0382	24	17	11	GO:0007005	mitochondrion organization	B	0.00330	0.0232	24	92	28
GO:0005976	polysaccharide metabolic process	B	0.000607	0.0405	25	90	37	GO:0005976	polysaccharide metabolic process	B	0.00389	0.0263	25	90	24
GO:0007085	mitochondrion organization	B	0.000611	0.0405	26	92	41	GO:0006396	RNA processing	B	0.00410	0.0266	26	367	60
GO:0009225	nucleotide-sugar metabolic process	B	0.000668	0.0425	27	15	9	GO:0007031	peroxisome organization	B	0.00443	0.0274	27	16	8
GO:0016043	cellular component organization	B	0.000662	0.0530	28	1981	572	GO:0006681	cellular aldehyde metabolic process	B	0.00494	0.0274	28	20	9
GO:0016044	protein maturation	B	0.000860	0.0591	29	92	40	GO:0006996	organelle organization	B	0.0181	0.105	29	1020	182
GO:0007033	vasculature organization	B	0.0120	0.0660	30	30	16	GO:0016043	cellular component organization	B	0.0196	0.111	30	1981	330
GO:0006518	peptide metabolic process	B	0.0124	0.0660	31	40	20	GO:0006805	xenobiotic metabolic process	B	0.0210	0.114	31	17	7
GO:0015993	drug transport	B	0.0124	0.0660	32	21	12	GO:0005908	proteolysis	B	0.0249	0.132	32	510	106
GO:0005903	macromolecular complex assembly	B	0.0127	0.0660	33	391	136	GO:0051604	protein maturation	B	0.0269	0.138	33	92	24
GO:0048771	tissue remodeling	B	0.0171	0.0867	34	70	23	GO:0006818	hydrogen transport	B	0.0452	0.224	34	78	21
GO:0061301	cell division	B	0.0200	0.0984	35	312	106	GO:0006323	DNA packaging	B	0.0472	0.228	35	56	15
GO:0006811	cellular aldehyde metabolic process	B	0.0365	0.127	36	20	11	GO:0016044	cellular membrane organization	B	0.0532	0.280	36	305	61
GO:0001616	cytokine production	B	0.0309	0.143	37	162	45	GO:0004843	heterocycle metabolic process	B	0.0588	0.269	37	687	138
GO:0051640	organelle localization	B	0.0326	0.147	38	51	22	GO:0006800	oxygen and reactive oxygen species metabolic...	B	0.0631	0.281	38	30	11
GO:0007059	chromosome segregation	B	0.0446	0.197	39	52	21	GO:0051301	cell division	B	0.0768	0.333	39	312	61
GO:0042025	organelle fusion	B	0.0601	0.235	40	210	78	GO:0015993	drug transport	B	0.0806	0.341	40	21	7

Figure 6: Significantly enriched Gene Ontology categories found in the two lists of candidate PPAR α target genes. Enriched biological processes were identified in the two lists of candidate PPAR α targets genes generated by the across experiment comparison (panel A), or our integrated approach (panel B). All significant probesets identified by the respective methodologies were used as input. The ‘parent-child-union’ algorithm was applied followed by the Benjamini-Hochberg correction for multiple testing to identify enriched GO categories. In both lists the same underlying biology was identified. Abbreviations: NSP: name space (sub ontology), B: Biological process.

Similarly, using the same criteria 28 significantly enriched categories (out of 169 annotated categories) were scored in the list of 2432 genes generated by our integrated approach (Figure 6B). Twenty-five identified enriched biological processes were identical in both sets of genes. As expected, many processes that were enriched have been functionally demonstrated to be controlled by PPAR α , including cellular ketone metabolic process, lipid metabolic process, cellular amino acid and derivative metabolic process, peroxisome organization, and mitochondrion organization [9,13,158]. Thus, despite the drastically reduced number of candidate PPAR α target genes identified by our approach, GO enrichment analysis demonstrated a very similar functional characterization of these genes, again demonstrating the validity of our strategy.

Conclusions

Taken together, in this study we present a simple, yet efficient strategy to compare genes across experiments that controls for multiple testing and is able to properly detect differentially expressed genes. Compared to the conventional used FDR based across experiment comparison, our approach is more conservative and less noisy. Our approach is in particular suitable to identify candidate target genes of a transcription factor or signaling route from functional genomics experiments, but can be applied to any genomics experiment in which the effects of a treatment are compared between two genotypes.

Chapter 3

Integrative analysis and modeling of PPAR α function in murine liver and small intestine

Mohammad Ohid Ullah, Meike Bünger, Sander Kersten, Philip J de Groot, Michael Müller, and Guido JEJ Hooiveld

Submitted

Abstract

Systems biology approaches aim to discover biological systems in which the components work together and are connected to one another within and between organs. These components can be either genes or set of genes or organs. The peroxisome proliferator-activated receptor alpha (PPAR α) is a ligand activated nuclear receptor, which is activated by free fatty acids and their derivatives. Here, we propose a nutritional systems biology approach to identify and integrate PPAR α dependent pathways in mouse liver and small intestine from information obtained in different experiments using an array-wise pathway score. We also developed a partial least squares path model (PLSPM) to infer the effect of pathways' activities at early time points on late time points. We show that our approach enabled the identification of PPAR α dependent pathways as well as the type of regulation in mouse liver and small intestine, and that acutely induced pathways are the main drivers for regulation of pathways after long-term activation. Taken together, we show that our proposed methodology successfully identifies biological relevant PPAR α regulated processes and provides clues on the underlying mechanisms.

Introduction

The peroxisome proliferator-activated receptor alpha (PPAR α) is a ligand-activated transcription factor with diverse functions and is activated by a variety of synthetic compounds, including drugs used for the treatment of dyslipidemia and type 2 diabetes [13,159,160]. High affinity natural ligands include eicosanoids, unsaturated as well as long-chain fatty acids, and their activated derivatives (acyl-CoA esters) [161-166]. In analogy with other nuclear receptors, when activated, PPAR α forms obligate heterodimers with the retinoid X receptor and stimulates gene expression by binding to peroxisome proliferator response elements (PPREs) located in the promoter regions of target genes [13]. For efficient transcriptional regulation by PPAR α also co-regulators are required. These are molecules that assist PPAR α to positively or negatively influence the transcription of target genes, and thereby comprise an integral part of the transcriptional circuitry [167-169]. PPAR α is also able to repress transcription by directly interacting with other transcription factors and interfere with their signaling pathways, a mechanism commonly referred to as transrepression [13,170]. PPAR α is expressed in a variety of tissues, including liver and small intestine [14,16,171].

In liver PPAR α is critical for the coordinate transcriptional activation of genes involved in nutrient metabolism [13,159] and it is suggested that PPAR α is an important regulator of the hepatic acute phase response [172]. Even though the small intestine expresses PPAR α at high level and is frequently exposed to high levels of PPAR α agonists via the diet, the role of PPAR α in this organ was not investigated until recently. The intestinal PPAR α plays an important role, governing diverse processes ranging from numerous metabolic pathways and lipid handling to the control of apoptosis and cell cycle genes [16]. Thus, although PPAR α activation and target gene regulation has been studied in a range of organs, gaps in our knowledge remain.

In so far as the biological role of PPAR α is directly coupled to the function of its target genes, probing PPAR α -regulated genes via the application of genomics tools can greatly improve our understanding of PPAR function. By combining transgenic animal models with elaborate microarray analyses, a comprehensive understanding of the *in vivo* role of PPAR α can be obtained [139]. As a result many PPAR α target genes and PPAR α responsive pathways have been identified,

but it should be noted that these have been determined mainly after relatively long-term exposure (5 days and more) to potent PPAR α agonists. However, we have long term mixed effects by WY activation of PPAR leading to direct and indirect activation of numerous pathways. 6h treatment largely will lead to mainly PPAR α activation.

In the current study, we aimed to model the relation between PPAR α responsive genes and pathways in mouse liver and intestine using pathway scores. To this end, array data was used from acute (6h) and long-term (5d) exposure in combination with partial least squares path modeling.

Materials and methods

Experimental data

In this study we used datasets that were generated previously in our laboratory to identify PPAR α target genes in mouse liver and intestine [15,139,166,173]. Briefly, pure bred wild type (129S1/SvImJ) and PPAR α -null (129S4/SvJae-Pparatm1Gonz/J) mice [144] were dosed by oral gavage with 400 μ l of a 0.1% WY14643 suspension in 0.5% carboxymethyl cellulose (acute experiment), or fed chow or chow supplemented with 0.1% WY14643 (Chemsyn, Lenexa, KS) (long-term experiment). WY14643 is a highly specific and potent agonist for PPAR α and is therefore often used in studies on this nuclear receptor [12,145]. After 6 h (acute experiment) or 5 days (long-term experiment), mice were anaesthetized and livers and intestines were excised. Total RNA was prepared using TRIzol reagent (Invitrogen, Carlsbad, CA) followed by purification using the RNeasy mini kit (Qiagen, Hilden, Germany). RNA integrity was checked by chip analysis (Agilent 2100 Bioanalyzer, Agilent Technologies, Amsterdam, the Netherlands) according to the manufacturer's instructions. RNA was judged as suitable for array hybridization only if samples exhibited intact bands corresponding to the 18S and 28S ribosomal RNA subunits, and displayed no chromosomal peaks or RNA degradation products, and had a RNA integrity number (RIN) above 8.0). The Affymetrix GeneChip RNA One cycle Amplification Kit (Affymetrix, Santa Clara, CA) was used to prepare labeled cRNA from 5 μ g of total RNA, which subsequently was hybridized on Affymetrix Mouse Genome 430 2.0 plus arrays. For each

treatment, tissue and time point 4 replicate arrays were performed, so in total 64 arrays were included in this study. The animal study was approved by the Local Committee for Care and Use of Laboratory Animals.

NutriSysPath approach

It is well known that genes belonging to the same pathway (gene set) are related to each other, obviously from a biological but also statistical point of view. A large variety of pathway overrepresentation methodology has been published, and these include tools such as gene set enrichment analysis (GSEA) [23], MaxMean statistic [174], interaction-based gene set analysis (IB-GSA) [175], Hotelling's T2-statistic [176], Global Test [177], and so on [178]. A drawback of these methodologies is that from a statistical perspective they don't adjust for correlation between the genes [178]. Based on the first eigenvector from singular value decomposition that reflects pathway activity level, a method has been developed by [179] for pathway analysis, considering the correlation between the genes, and applying a t-test or analysis of variance to infer the significantly regulated pathways. As an alternative approach we propose to use principal component analysis (PCA) in combination with correlation analysis to identify 'relevant' pathways, an approach which we called NutriSysPath (**Nutritional Systems Biology of Pathway Analysis**).

Gene sets representing Gene Ontology categories, metabolic pathways or signaling transduction routes were extracted from well-recognized pathway databases (GO, KEGG, NCI, Biocarta, Pfam, Reactome and WikiPathways). A reference set of well-established PPAR α targets genes was derived from [18]. We limited our analyses to gene sets that contained at least 15, and maximally 500 genes, as very small or very large classes are unlikely to be as informative (either too specific or too general) [180]. In total 4588 pathways were included in the analysis. For each pathway (gene set), PCA was performed using the expression data of all samples. PCA involves a mathematical procedure that transforms a number of possibly correlated variables, in this case expression of genes, into a smaller number of uncorrelated variables considering the relationships among the variables, called principal components (meta genes) [128,181]. The first principal component accounts for as highest variation in the dataset. We considered the dominant principal component (PC1) as the pathway activity level, as was also done by [179]. It should be noted that a pathway activity level, i.e., PC1 score, is

calculated for each single sample. Next the correlation of all array-wise pathway activity scores with that of the reference set was calculated using the non-parametric Spearman correlation coefficient and the corresponding p-value. In this study, the cut-off point was considered P-value $\leq 10^{-6}$ and absolute correlation coefficient (r) ≥ 0.90 . The non-parametric Spearman correlation was used because it is more robust against outliers compared to parametric correlation measure [182]. Ultimately, this resulted in a list of pathways that had similar behavior as the reference set, and these were used as input for further analysis. An overview of this approach is given in Figure 1, and the R code used to calculate the pathway activity scores is available as supplemental data.

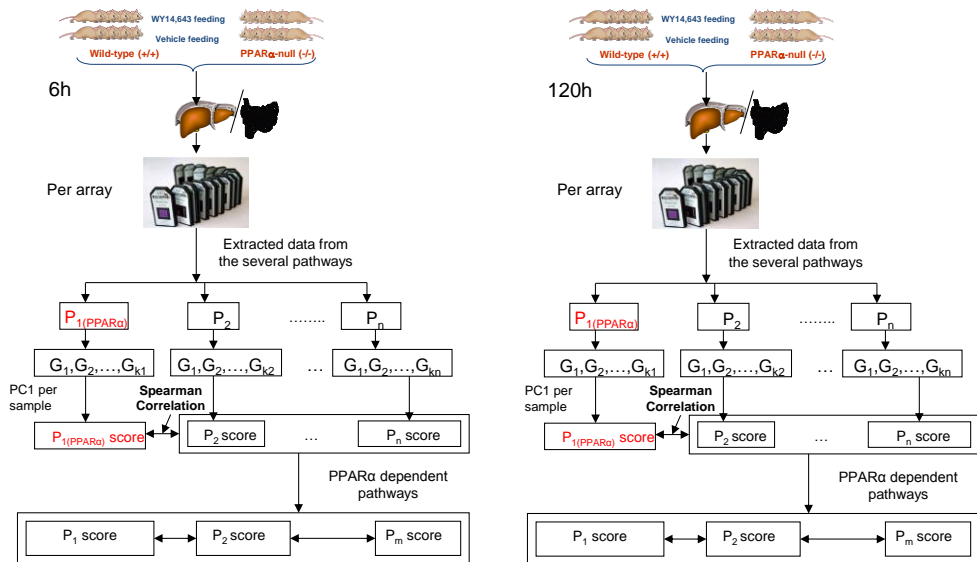


Figure 1: Overview of the NutriSysPath approach. After normalization the microarray data, genes were grouped based on the pathways or gene sets. Principal component analysis was applied at each pathway and collected principal component 1 (PC1) as array wise pathway activity level or pathway score. Afterwards, ran the Spearman correlation between the standard gene set or reference gene set (PPAR α target genes) and all other pathways scores and then identified the induced or suppressed pathways. The pathways that were positively correlated with the reference gene set were considered as PPAR α induced whereas the negatively correlated pathways were considered as suppressed pathways. The similar process was applied at each time point and each organ.

Implementation of the NutriSysPath approach

In the current study the NutriSysPath approach was implemented as follows: (i) expression estimates were obtained by GC-robust multiarray (GCRMA) normalization, using the empirical Bayes approach to adjust for background [152], (ii) for each pathway the expression data of contributing genes was extracted, (iii) PCA was used to calculate the PC1 score per sample for each pathway [29], and (iv) the correlation between PC1 scores for the reference set of well-established PPAR α target genes and all other pathways was calculated using Spearman correlation, and finally pathways that significantly correlated with the reference set ($p \leq 10^{-6}$, $r \geq |0.90|$) were retained. Pathways that had a positive correlation with the reference set were considered induced pathways, whereas anti-correlated pathways represented suppressed pathways. This procedure was applied in both organs for both time points. Results were visualized using heatmaps, and pathway interaction networks were created in Cytoscape using the Enrichment Map plugin [59,183].

Partial least squares-path model

Partial least squares-path modeling (PLSPM) proposed by [117] is a multivariate data analysis technique which provides a framework for analyzing multiple relationships between a set of blocks of variables. The PLS is robust against of missing values, model misspecification and violation of the statistical assumptions: normality and multicollinearity [184,185]. The PLSPM is an extension of the PLS. A detailed explanation about the PLSPM can be found in [31,34,117]. We used reflective way in the outer model and PLS regression in the inner model of the PLSPM with standardizing manifest variables (pathway scores). Analysis was performed in R using the library *pls* [30]. In this study, we evaluated how PPAR α -induced and -suppressed pathways that were regulated at 120h depended on the pathways regulated at 6h. After calculating the pathway scores, pathways were grouped in induced or suppressed pathways depending on their correlation with the reference set. Thus, in this study we had 8 groups of pathways scores, the groups details were as follows:

$Y_1 = \text{UP_Late_I}$; all induced pathways in small intestine 120h after intervention.

$Y_2 = \text{Down_Late_I}$; all suppressed pathways in small intestine 120h after intervention.

Y_3 =Up_Late_L ; all induced pathways in liver 120h after intervention.

Y_4 =Down_Late_L ; all suppressed pathways in liver 120h after intervention.

X_1 = Up_Early_L ; all induced pathways in liver 6h after intervention.

X_2 = Down_Early_L ; all suppressed pathways in liver 6h after intervention.

X_3 =Up_Early_I ; all induced pathways in small intestine 6h after intervention.

X_4 =Down_Early_I ; all suppressed pathways in small intestine 6h after intervention.

The PLSPM thus becomes:

$$Y_1 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 Y_3 + \beta_6 Y_4$$

$$Y_2 = \beta_7 X_1 + \beta_8 X_2 + \beta_9 X_3 + \beta_{10} X_4$$

$$Y_3 = \beta_{11} X_1 + \beta_{12} X_2 + \beta_{13} X_3 + \beta_{14} X_4$$

$$Y_4 = \beta_{15} X_1 + \beta_{16} X_2 + \beta_{17} X_3 + \beta_{18} X_4$$

Where, β s are the path coefficients.

Results

We applied the proposed NutriSysPath approach followed by PLSPM on datasets that aimed to identify PPAR α target genes in mouse liver and intestine, and which were generated previously in our laboratory [15,139,166,173]. After normalization, pathway activity scores were calculated for each using PCA as indicated in the Methods section. Part of the output is represented in Figure 2. The heatmap represents pathway activity scores of the reference set (PPAR α _targets) and 50 GO categories (rows) of the 16 liver samples that were 120h after start of the intervention (columns). It is clear that pathway activity scores varied within and between experimental groups (Figure 2A). As expected the activity score for the reference set was highest in the wild type mice treated with WY14643. Some GO categories displayed similar behavior as the reference set, whereas others behaved oppositely. In Figure 2B the loading plot for a GO category is displayed, which represented the contribution of each gene to the pathway score; the higher the loading, the more it contributed. This parameter can be used to identify genes that were most responsive to treatment with WY14643.

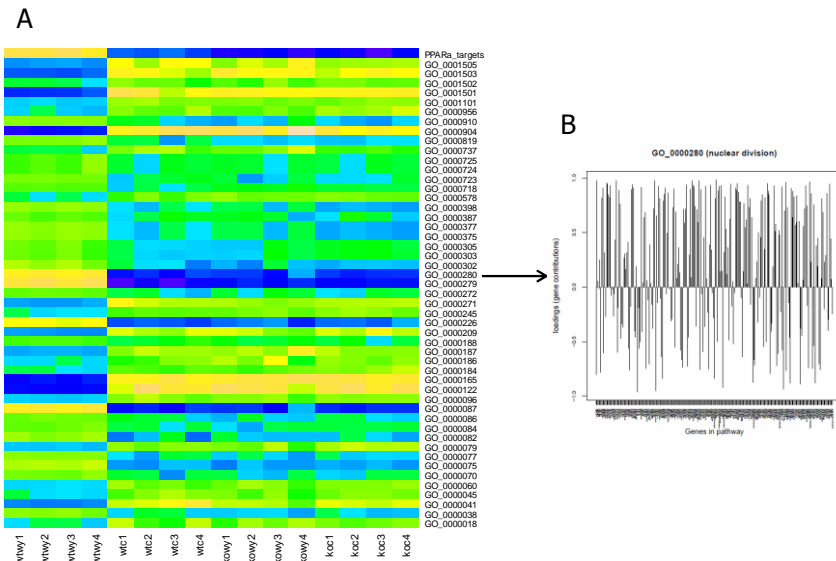


Figure 2: Array-wise pathway scores. (A) A sample pathway scores of liver 120h data. Yellow and blue colors indicated the high and low pathway scores respectively. Rows were the pathways and columns were the samples. (B) The loadings (correlation between the pathway scores and gene expressions) of the genes in a pathway. Higher bars indicated the more importance of the genes in the pathway.

Selected pathways in liver and small intestine at 6h and 120h

To select the pathways in liver and small intestine that were regulated both after acute (6h) and long-term (120h) treatment with WY14643, Spearman correlation analysis was performed. Pathways were selected if they highly and significantly (anti-)correlated with the reference set (absolute correlation coefficient ($r \geq 0.90$, $p \leq 10^{-6}$). Positively correlated pathways were considered to be induced by PPAR α , whereas negatively correlated pathways were considered to be suppressed. Using these cut-off criteria we found that in small intestine after acute activation PPAR α induced the activity of 80 pathways, and suppressed the activity of 27 pathways. After long-term activation the activity of 131 pathways was increased, and of 99 suppressed in small intestine. Similarly, acute activation of PPAR α resulted in the induction resp. suppression of 446 and 723 pathways in liver, and long-term activation induced resp. suppressed the activity of 115 and 229 pathways.

Pathway interaction networks for the induced pathways after acute a long-term activation for both organs are presented in Figure 3.

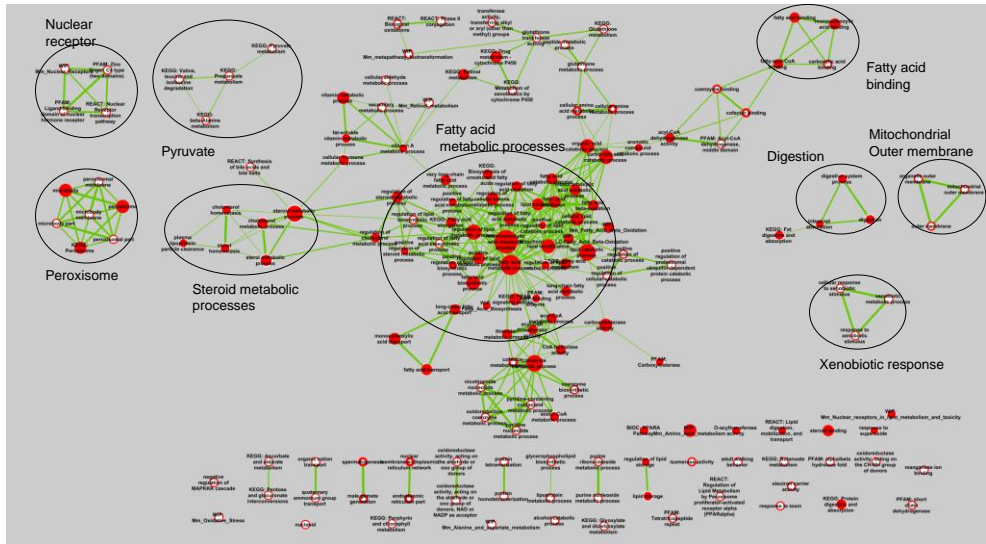


Figure 3A: Interaction map of PPAR α induced pathways in small intestine after acute and long-term activation. Red and white colors indicate pathways that were induced or not by PPAR α respectively. The inner circle and outer circle indicated the effects of acute (6h) and long-term (120h) activation, respectively. The sizes of the nodes is based on the number of genes belonging to the pathway; the bigger the nodes the more genes. Edges indicated the overlapped genes between the pathway; the thicker the edge, the more the pathways overlap. The network contains 162 nodes (pathways) and 353 edges.

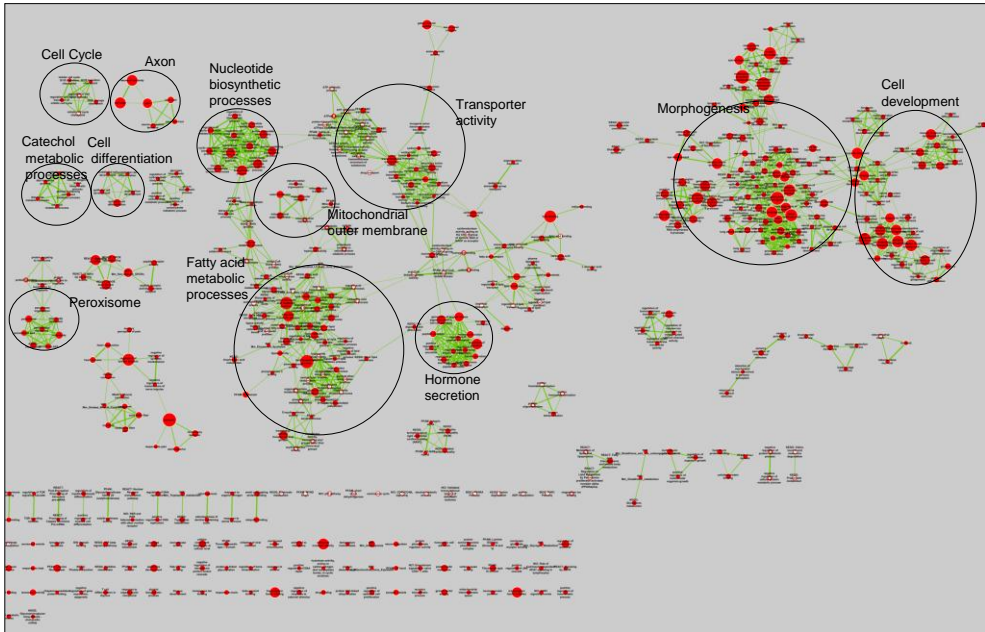


Figure 3B: Interaction map of PPAR α induced pathways in liver after acute and long-term activation. Red and white colors indicate pathways that were induced or not by PPAR α respectively. The inner circle and outer circle indicated the effects of acute (6h) and long-term (120h) activation, respectively. The sizes of the nodes is based on the number of genes belonging to the pathway; the bigger the nodes the more genes. Edges indicated the overlapped genes between the pathway; the thicker the edge, the more the pathways overlap. The network contains 521 nodes (pathways) and 1427 edges.

These networks revealed that in small intestine (Figure 3A) many pathways related to ‘fatty acid metabolic processes’ were induced at both time points, as was ‘peroxisome’ and ‘fatty acid binding’. Other processes such as ‘steroid metabolic processes’ and ‘digestion’ were induced only after acute treatment, whereas ‘nuclear receptor’, ‘pyruvate’, and ‘xenobiotic response’ were only induced after long-term activation.

Similar to the small intestine, many pathways related to ‘fatty acid metabolic processes’ and ‘peroxisome’ were induced in liver after both acute and long-term activation (Figure 3B). In addition to these, many pathways related to ‘morphogenesis’, ‘cell development’, ‘cell differentiation’ and ‘hormone secretion’

were induced after acute activation. We also observed that some pathways related to ‘cell cycle’ and ‘transporter activity processes’ were induced either after acute or long term activation. In general we observed for both time points that the number of significantly regulated pathways was higher for liver than small intestine, which suggested the activity of PPAR α in liver is higher than small intestine. In total 32 pathways were found to be commonly induced in liver and small intestine after acute activation, and this were 50 pathways after long term activation. The 20 pathways that were found to be induced in both time points and both organs are listed in Table 1, and functionally these reflect ‘fatty acid metabolic processes’, ‘fatty acid oxidation’ and ‘peroxisome’.

Table 1: Common 20 PPAR α induced pathways in liver and small intestine after both acute and long-term activation.

Name	Description
GO:0000038	Very long-chain fatty acid metabolic process
GO:0006631	Fatty acid metabolic process
GO:0006633	Fatty acid biosynthetic process
GO:0006637	Acyl-CoA metabolic process
GO:0009062	Fatty acid catabolic process
GO:0019217	Regulation of fatty acid metabolic process
GO:0019395	Fatty acid oxidation
GO:0030258	Lipid modification
GO:0032787	Monocarboxylic acid metabolic process
GO:0034440	Lipid oxidation
GO:0035383	Thioester metabolic process
GO:0044242	Cellular lipid catabolic process
GO:0046320	Regulation of fatty acid oxidation
GO:0072329	Monocarboxylic acid catabolic process
GO:0005777	Peroxisome
GO:0042579	Microbody
GO:0033293	Monocarboxylic acid binding
KEGG_Peroxisome	KEGG: Peroxisome
KEGG_PPARG signaling pathway	KEGG: PPAR signaling pathway
KEGG_Fatty acid metabolism	KEGG: Fatty acid metabolism

GO: Gene Ontology category identifier

Partial least squares-path modeling

To infer causal effects on pathway regulation, a PLSPM was designed that integrated the inter-organ regulation after acute and long-term PPAR α activation. To this end, pathways that included in this study were grouped per time point and organ based on pathway scores in blocks of induced or suppressed pathways. This resulted in 8 blocks groups of pathways, Up early liver (1807 pathways), Down early liver (2781 pathways), Up early intestine (1614 pathways), Down early intestine (2974 pathways), Up late liver (1079 pathways), Down late liver (3509 pathways), Up late intestine (1014 pathways) and Down late intestine (3574 pathways). We assumed that the pathway activity scores after long-term activation were (partially) driven by the pathway activity scores after acute activation, and that pathways in intestine and liver could influence each other. Next multivariate PLSPM was performed. The path coefficients that were obtained indicate how much of the regulation observed after long-term activation can be effected by the acute activation. For instance, the path coefficient of the model for Up_Early_I \rightarrow Up_Late_I was 0.94, and that of Up_Early_L \rightarrow Up_Late_L was 0.53 (Figure 4). This implies that induced pathways in intestine after long-term activation were more effected by the acutely induced pathways in intestine than in liver.

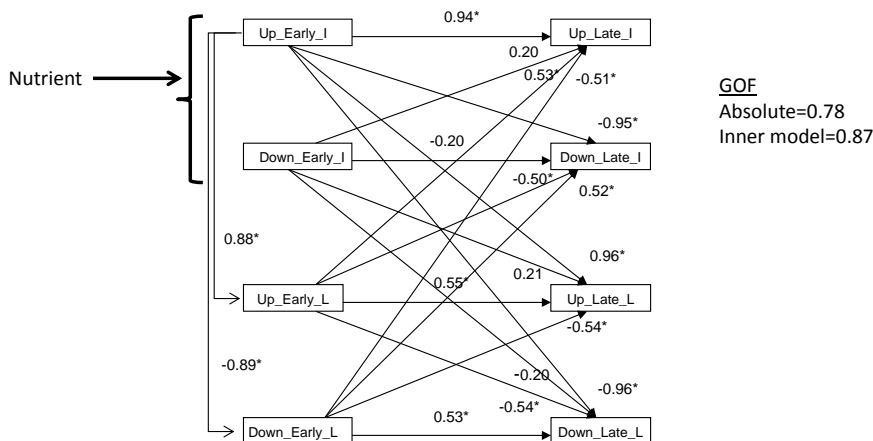


Figure 4: PLS-path coefficients (total effects) of PLS- path models for liver and small intestine by path-diagram. *path coefficients were significant at 5% level by bootstrap simulation.

We observed that the path coefficient of the model for Up_Early_I -> Up_Early_L was 0.88, indicates that acutely induced pathways in small intestine had a significant positive influence on acutely induced pathways in liver.

Moreover, we found no significant effects of the acutely suppressed pathways in intestine on any of the long-term regulated pathways in intestine or liver. We noticed that up regulated early pathways in liver had positive path coefficients on the up regulated late time point in small intestine (0.53), up regulated late time point in liver (0.55). On the other hand, acutely induced pathways in liver negatively influenced the long-term regulation of suppressed pathways in small intestine and in liver. We also noticed that acutely suppressed pathways in small intestine had no significant effect on long-term regulation in liver and small intestine.

Discussion

A key purpose of systems biology is to provide a systems level understanding by integrating, interconnecting and modeling of high-throughput datasets [186]. Here we applied this approach to a key nutrient sensing transcription factor and its target pathways. Multivariate data such as gene expression data is commonly generated in modern biology, and many tools have been developed to analyzed and visualize this kind of data [128]. To date many studies have examined the effect of PPAR α activation using gene expression profiling or metabolomics (see e.g. [17,187,98,179]). However, no systematic comparisons of the whole genome effects of PPAR α activation in mouse liver and small intestine have been reported. Here we presented a systematic comparison on PPAR α dependently regulated pathways utilizing array-wise pathway scores after acute and long-term activation in liver and intestine. Three main conclusions can be drawn from our work. First, our data support a more important role of PPAR α in mouse liver than in small intestine, as is evidenced by larger number of the list of significant pathways that were identified by our NutriSysPath approach at both time points. Secondly, acutely induced pathways in small intestine are suggested to major influence on acutely induced pathways in liver. Third, acutely induced pathways are the main drivers for regulation of pathways after long-term activation. To the best of our knowledge, we are one of the first to use PLSPM to infer the causal effect of early

measurement of the groups of up and down regulated pathways scores on that of the late measurements.

In conclusion, the approach used here allows to analyze different datasets either several time points or different omics datasets. When applied on PPAR α datasets, we obtained new insights on organ-specificity and time-dependency of PPAR α activation.

Supplementary

Table S1: The R code of our approach.

```

# Define some variables
Expression_File <- "int5_n.txt" # the input file where rows are the genes and columns are
samples/array after normalization.
GMT_File <- "GO_K_NCI_BIOC_PF_REACT_WIP_Mm_symbol.txt" # File saved via Excel to
make it square

PCA_Scores_File <- "int_5_pca_score_GMT-n.txt"
Gene_loadings_Output_File = "Gene_Loadings_int55-n"
Species_Annotation_Library <- "mouse4302mmentrezg.db"
Lower_Limit_Extracted_Genes_in_Pathway <- 15 # this number depends on researcher, it can be
changed
Upper_Limit_Extracted_Genes_in_Pathway <- 500 # this number depends on researcher, it can be
changed

# Remove extensions; these will be added automatically later on
# for multiple files
Gene_loadings_Output_File <- gsub("\\.", "_", Gene_loadings_Output_File)

# Load the input file first.

# Assumptions:
# 1. The first row contains the columns names that is samples/array!
# 2. The first column contains the gene identifiers!
## One can calculate the pathways score per array using our code just replacing there input file. If
one wants to calculate another species or another annotation or another choice of extraction limit,
s/he has to define it on the variables.
x <- as.matrix(read.delim(file=Expression_File, row.names=1))

# Load the GMT File
gmt <- read.delim(file=GMT_File, header=F, stringsAsFactors = FALSE)
pathways <- gmt[,3:dim(gmt)[2]]
rownames(pathways) <- gmt[,1]

# Load the utilized PCA library
library(FactoMineR)
library(Species_Annotation_Library, character.only = T)

# Replace the affy IDs with the symbols (if possible)
rownames(x) <- make.unique(toupper(mget(rownames(x), get(sprintf("%sSYMBOL", sub(".db", "",
Species_Annotation_Library))), ifnotfound=NA)))

```



```

# Open the PDF-file
pdf(file = paste(Gene_loadings_Output_File, ".pdf", sep=""), paper = "a4r", onefile=TRUE)

# Now extract the genes for each individual pathway
# We use a loop to do the calculations per pathway
Count <- 0
Rejected <- 0;
Final_loadings_Matrix <- vector("list", length(pathways))
for (i in 1:dim(pathways)[1])
{
  cat(sprintf("pathway %g: %s...", i, rownames(pathways[i,])))

  # Extract the matrix with the selected pathway intensities
  Current_Pathway_Indices <- which(nchar(pathways[i,]) > 1)
  # Some identifiers are not available. Filter them out here...
  Indices <- match(as.character(pathways[i, Current_Pathway_Indices]), rownames(x))
  if (sum(is.na(Indices)) > 0)
    Current_Pathway_Indices <- c(Current_Pathway_Indices[-which(is.na(Indices))])

  Number_Of_Valid_Genes_in_Current_Pathway <- length(Current_Pathway_Indices)
  if (!(Number_Of_Valid_Genes_in_Current_Pathway <
Lower_Limit_Extracted_Genes_in_Pathway)) && (!(Number_Of_Valid_Genes_in_Current_Pathway
> Upper_Limit_Extracted_Genes_in_Pathway))
  {
    cat("Included!\n")
    Count = Count + 1
    x.pathway <- x[as.character(pathways[i, Current_Pathway_Indices]),]
    transposed_matrix <- t(x.pathway)
    res <- PCA(transposed_matrix, scale.unit=TRUE, graph=FALSE)
    if (Count == 1)
    {
      Final_PCA_Matrix <- as.data.frame(res$ind$coord[,1])
      colnames(Final_PCA_Matrix)[Count] <- rownames(pathways)[i]
    } else {
      Final_PCA_Matrix <- cbind(Final_PCA_Matrix, as.data.frame(res$ind$coord[,1]))
      colnames(Final_PCA_Matrix)[Count] <- rownames(pathways)[i]
    }
  }
  Final_loadings_Matrix[[Count]] <- as.list(res$var$coord[,1])

  plot(res$var$coord[,1], main=sprintf("%s (%s)", rownames(pathways)[i], gmt[i,2]), xlab = "Genes
in pathway", ylab = "loadings (gene contributions)", type="h", xaxt="n", cex.axis=0.75)

  # Define the proper graph labels...
  labels.genenames <- sprintf("%s (%s)", names(res$var$coord[,1]), pathways[i,
Current_Pathway_Indices])

  axis(1, at=1:length(names(res$var$coord[,1])), labels=labels.genenames, las=3, cex.axis=0.2)
} else {
  Rejected <- Rejected + 1
  cat("Skipped!\n")
}

```

```
}  
}  
  
# Close the PDF-file  
dev.off()  
  
# Write the final file  
cat(sprintf("\nWriting PCA Scores file: %s.\n", PCA_Scores_File))  
write.table(Final_PCA_Matrix, file = PCA_Scores_File, sep="\t")  
cat("Script has succesfully ended!\n")  
### the end  
#####  
#####
```

Chapter 4

Characterization and modeling of acute effects of PPAR α activation in rat liver cells

Mohammad Ohid Ullah, Shohreh Keshtkar, Guido JEJ Hooiveld,
and Michael Müller

In preparation

Abstract

The peroxisome proliferator activated receptor alpha (PPAR α) is a transcription factor which is activated by natural and synthetic agonists. Studies in mouse, human and rat have shown that PPAR α plays an important role in liver and other organs. However, little is known on the genes and processes that are acutely regulated by PPAR α , and how these evolve over time. We therefore performed a time-course microarray study in rat hepatocytes to characterize the genome-wide effects of acute PPAR α activation. In this study, mRNA expressions in rat hepatocytes were measured at up to five time points (0, 1, 2, 3, and 4h) upon stimulation with WY14643. Including all time points, in total 386 genes were significantly induced by WY14643. Already 1h after stimulation, gene expression increased, and this stabilized after 3h. Several transcription factor binding sites were predicted to be involved with PPAR α activation, and these included recognition elements for NRF2 and RXR. Many genes were found that followed a quadratic model and were involved in lipid metabolic processes. Taken together, our systems approach identified a set of similar behaving genes with the evolution of gene network over time at early stage in rat hepatocytes and their potential common transcription factors with PPAR α . This information provides new details on the molecular mechanisms involved in PPAR α -dependent gene regulation.

Introduction

Dietary lipids, one of the main nutritional components, are able to excite their own catabolism through a set of nuclear receptors called the peroxisome proliferator activated receptors (PPARs) [188,189]. Among the three PPAR isoforms that exist (PPAR α , PPAR δ (also called β) and PPAR γ), fatty acids bind to PPAR α with highest affinity [9]. PPAR α is highly expressed in tissues with a high catabolic rate such as the liver, kidneys, heart, skeletal muscle and small intestine [14,139]. PPAR α is accountable for control lipid metabolism in many tissues, but its role has been best investigated in liver [190]. The liver plays an important role in the coordination of lipid metabolism and it actively metabolizes fatty acids as fuel. It is responsible for hepatic triglycerides export via synthesis of very low density lipoproteins (VLDL). An imbalance between lipid anabolic and catabolic processes may lead to triglycerides accumulation and as a consequence of hepatic steatosis [17]. Genes encoding peroxisomal and microsomal fatty acid oxidation in liver are transcriptionally regulated by PPAR α [13,191].

Hepatocyte performs most important function of the liver including lipid metabolism, regulation of urea and production of plasma proteins. To identify the temporal gene expression in toxicology of monolayer cultured rat hepatocytes cultures study has been done in 5 different time points (4, 12, 24, 48, and 72h) by [192]. In order to identify the effect of WY14643 at different markers of inflammation a study has been done by [193] at one time point (8 days) in a rat model of ligature-induced periodontitis. Several studies have been performed to detect the effect of WY14643 in gene expression level in time, such as: at two time points (6h and 120h) in mouse and human hepatocytes by [17], at 3 time points (1d, 7d, and 28d) by [194] in mouse liver and in one time point (5d) in mouse small intestine by [139]. To detect the effect on the expression of c-met, c-myc and PPAR-alpha in liver and liver tumors from rats has been done by [195]. Another study has been performed by [196] to see the differences between the promoting activities of the peroxisome proliferator agonist WY14,643 and phenobarbital in rat liver at 3 time points (11, 22, or 54 wk). Recently a study has been done by [197] for inferring statin-induced gene regulatory relationships in primary human hepatocytes over time (0, 6, 12,24, 48, and 72h). Usually, after 4h or 6h it's difficult to detect the direct effect of a treatment. Therefore, to avoid

toxicity and to detect the direct effect of WY14643, it's necessary to run the experiments at early stage. Until now such studies haven't been performed to detect the effect of WY14643 (a strong PPAR α agonist) of gene expression in rat hepatocytes at early time points using a nutritional systems biological approach. To fill this gape, we demonstrated an experiment using rat hepatocytes cell culture based on microarray experiments. In this study, we aimed to characterize the genome-wide effects of acute PPAR α activation by detecting the similar behavior genes which are activated by synthetic ligand WY14643, their biological functions and network at early stage (0-4h). Overall, the results reveal that PPAR α regulates a several profiles of genes over time in rat hepatocytes and most of the potential genes behave a quadratic model. Furthermore, several common transcription factors (TFs) also predict to bind with PPAR α , for instance: RXR, NR2F, EREF and CREB.

Materials and Methods

Cell culture

Rat hepatoma FAO cells were grown in DMEM containing 10% fetal bovine serum, 100U/ml penicillin and 100 μ g/ml streptomycin. FAO cells were seeded in 6-well culture plates at a 70% density. After 24 hrs cells were treated with the PPAR α agonist WY14643 (5 μ M) dissolved in DMSO (0.1% v/v). Incubations continued for 1, 2, 3, 4 hours. At each time point, including t=0 h, cells were harvested for RNA isolation from both WY14643 and DMSO treated cells; the latter served as control.

RNA isolation and quality control

Total RNA was isolated from FAO cells using Trizol reagent (Invitrogen, Brede, the Netherlands), followed by total RNA cleanup using RNEasy microkit (Qiagen, Venlo, the Netherlands). RNA quantity and quality was assessed spectrophotometrically (ND-1000, NanoDrop Technologies, Wilmington, USA) and with 6000 Nano chips (Bioanalyzer 2100; Agilent, Amstelveen, The Netherlands), respectively. RNA was judged as being suitable for array hybridization only if samples showed intact bands corresponding to the 18S and 28S ribosomal RNA subunits, displayed no chromosomal peaks or RNA degradation products, and had a RIN (RNA integrity number) above 8.0.

Microarray experiments and data processing

The Affymetrix GeneChip RNA One cycle Amplification Kit was used to prepare labeled cRNA from 5 µg of total RNA (Affymetrix, Santa Clara). Samples were hybridized on Affymetrix GeneChip Rat Genome 230 2.0 arrays. Hybridization, washing and scanning of the arrays was performed according to the manufacturer's recommendations. The raw intensity values applying the robust multiarray analysis (RMA) pre-processing algorithm [198,199] is not adjusted by background correction. Therefore normalized expression estimates were obtained from the raw intensity values using the GC-robust multi array (GCRMA) normalization, using the empirical Bayes approach to adjust background [152]. Probesets were redefined according to Dai *et al* [151]. In this study probes were reorganized based on the Entrez Gene database, build 37, version 1 (remapped CDF v14).

Identification of differentially expressed genes

Differentially expressed genes were determined by time course analysis of variance (TANOVA) [22]. TANOVA is a method to evaluate factor effects by pooling information across the time course while accounting for multiple testing and non-normality of microarray data.

After detecting differentially expressed genes by TANOVA, these were used for two complementary approaches, as depicted in Figure 1. Firstly, we focused on the induced genes on the basis of a fold change cut-off ($FC \geq 1.2$). This was done since PPAR α activation directly results in induced expression of target genes, whereas PPAR α -dependent suppression of gene expression is known to go through indirect mechanisms [170]. Per cumulative time point we performed pathway overrepresentation analysis and identification of *cis*-regulatory modules, i.e., combinations of transcription factor binding sites (TFBS).

Secondly, in a complementary approach we performed unsupervised clustering of the genes selected by TANOVA to identify genes that behaved similarly over time. This cluster analysis was performed to study the dynamics of the response (i.e., early and late response) to identify the corresponding genes. Afterwards, selected gene expression clusters were characterized with respect to biological function and polynomial model.

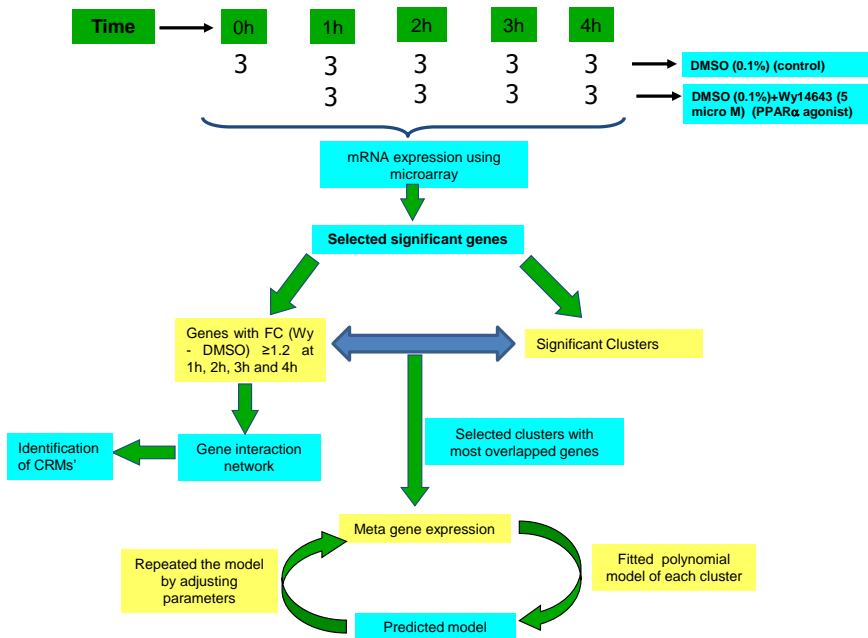


Figure 1: Overview of the experimental design and our analysis strategy.

After normalization the mRNA expression, the significant genes were selected by TANOVA. Afterwards, the analysis was done in two complementary ways. At first WY14643 induced were identified to detect the gene interaction network and to identify *cis*-regulatory modules (CRM). Secondly, a cluster analysis was performed by STEM using the all significant genes to find out the similar behavior of genes. The clusters that were mostly overlapped with the selected genes at the first step, were selected as final selected clusters. Finally, polynomial regression model was fitted for each selected cluster by adjusting parameters.

Clustering

Clustering or grouping the similar patterns of the gene expressions is a key issue to analyze the time series microarray data. Short Time-series Expression Miner (STEM) was used in this study to detect set of co-expressed genes [200]. STEM was specially designed for the analysis short time series gene expression data. This method implements to cluster, compare and visualize such data with its integration with the Gene Ontology [201]. The algorithm provides significant number of clusters whereas within profile genes are highly correlated according to Pearson correlation coefficient ($r \geq 0.80$) and correct the multiple tests by FDR.

In the STEM clustering method, we also assumed maximum number of model profiles is 50 and the maximum unit change in model profiles between time points is 2. To find out the effected biological processes, i.e., GO overrepresentation analysis for each significant profile, we assumed default option for minimum GO level and minimum number of genes and number of samples for randomized multiple hypothesis correction.

Modeling

Gene co-expression clusters were fitted using a polynomial regression model taking the average gene expression of the respective cluster. We ran the model several times considering different order and then selected the best model based on the root mean sum square of error (RMSE) and the adjusted R-square. A small RMSE with high adjusted R-square provides the best model for the respective cluster. Here, we only fitted the model for the most overlapped significant profiles. Fitting regression models was performed to enhance the biological interpretation of the gene expression cluster, since these models describe the shape of each gene expression cluster as a function of time. As a consequence, this provides insight into the underlying processes rather than simply identifying significant differences.

Gene interaction network analysis

Genes that were differentially expressed or co-expressed at several time points were used to infer gene interaction networks based on combining metabolic pathways from Reactome and KEGG databases using Rspider [60]. Interaction networks were visualized in Cytoscape [59].

Transcription factor binding sites

Identification of cis-regulatory modules (CRMs) in promoter regions of regulated genes was performed using the Genomatix software suite [122]. The Genomatix software suite is a collection of on-line tools for the retrieval and analyses of well annotated promoter sequences. In this study, at first we used Gene2Promoter to retrieve the promoter regions of regulated genes. Afterwards, FrameWorker [202] was used to identify common patterns of TF binding sites in the promoters.

Results

Differentially expressed genes

A nutritional systems approach, combining several statistical and bioinformatics tools, was used in this study to identify the temporal behavior of candidate PPAR α target genes and their corresponding biological function. After normalizing the microarray data, usually the first task is to find out the significantly regulated genes. In this study, we performed cross-sectional time-course study with two different conditions (Control [DMSO] and WY14643) at different time points. Sacrificing the time dependency, one can analyze this kind of data using conventional analysis of variance (ANOVA) to infer significantly regulated genes. Therefore, to capture the dynamic gene expression profiles, we used TANOVA to detect significant genes by pooling information across time points accounting for non-normality and multiple testing (FDR). Overall, we found 1177 significant genes with FDR-adjusted p-values < 0.05. Statistically significant does not always indicate biologically relevant. We therefore further refined our dataset by including only the genes that at one of each time point was more than or equal to 1.2-fold increased. This showed that the number of relevant significant genes increased over time (Figure 2). Total 79 genes were found common in all four time points, which represented highly sensitive genes that rapidly respond to WY14643 treatment. In total 386 genes were found to be increased in all four time points (Figure 2A). The temporal behavior of these 386 genes revealed that especially after 2h activation there was a strong response which leveled off at the later time points (Figure 2A). Moreover, most of the genes induced after 1h were also regulated after 2h, and this trend continued for the later time points (Figure 2B).

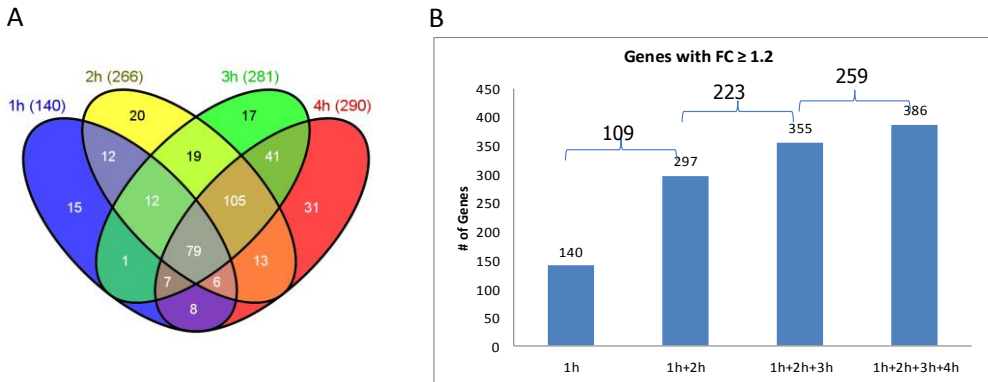


Figure 2: Differentially expressed genes. (A) Venn diagram of relevant regulated genes ($FC \geq 1.2$, $FDR < 0.05$) at four time points after stimulation with WY14643. (B) Bar diagram of cumulative gene regulation and at the four time points and their overlap.

Gene interaction network based on pathways

After having identified the genes that were regulated, we next analyzed per time point the functional implications of this regulation by analyzing which pathways were overrepresented in the sets of regulated genes (Table 1). These results were then combined with biochemical data to generate gene interaction networks.

Table 1: Pathways overrepresented in the sets of regulated genes

Pathways	1h (t=1h)	1h+2h (t=1 to 2h)	1h+2h+3h (t=1 to 3h)	1h+2h+3h+4h (t=1 to 4h)
Lipid Metabolic Process	1	1	1	1
Glycerophospholipid Metabolism		4	5	2
Fatty Acid Biosynthetic Process			6	3
Retinal Metabolism		2	2	4
Valine, Leucine and Isoleucine Degradation	2	5	3	5
Primary Bile acid Biosynthesis		3	4	6
Steroid Hormone Biosynthesis		7	8	7
Response to Glucose Stimulus			9	8
Histidine Metabolism			10	9
Lipid catabolic Process		8	11	10
Regulation of Fatty Acid Oxidation		6		
Transport		9	7	11

The 1st column indicated the name of the pathways which were found in the gene network. The columns 1h, 1h+2h, 1h+2h+3h and 1h+2h+3h+4h were indicated the existence of the pathways and their ranking based on the number of input genes in the network for the cumulative time points: 1h, 1h+2h, 1h+2h+3h and 1h+2h+3h+4h respectively. The number 1 meant the highest in the rank, 2 meant second highest and so on.

We identified only two pathways that were overrepresented in the genes induced after 1h, i.e., *lipid metabolic process* and *valine, leucine and isoleucine degradation*. This indicated that although a substantial number of genes were regulated, they likely were involved in a broad range of biological functions that therefore did not reach statistical significance. At later time points more pathways were identified, that almost all represent parts of lipid metabolism, and the number increased over time. The interaction network generated at 1h expanded

over time and always incorporated *lipid metabolic process*. The most evolved interaction network is presented in Figure 3 (the networks of the individual time points are available in the supplemental Figures).

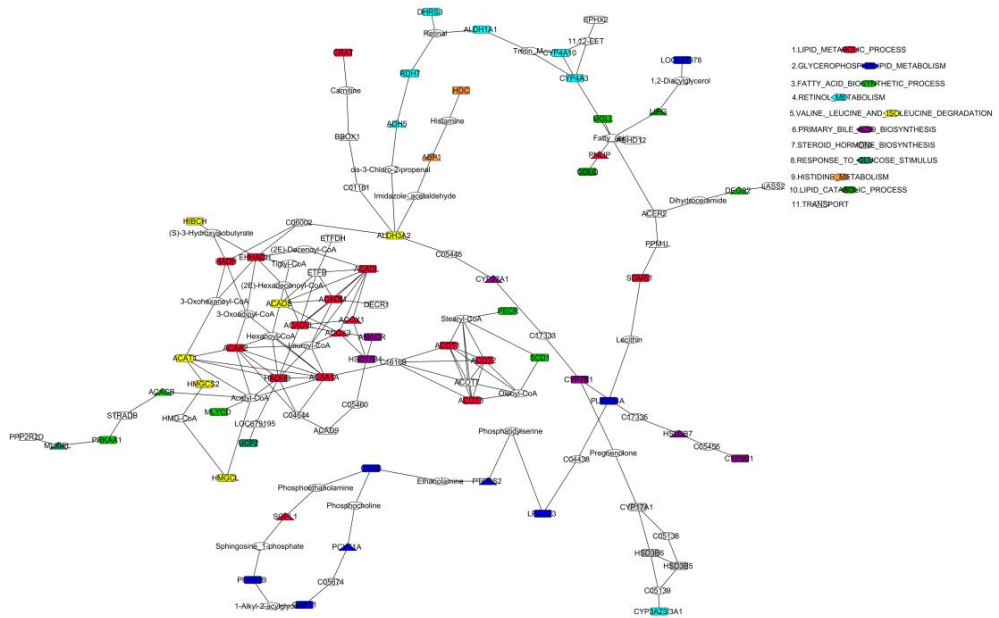


Figure 3: Gene interaction network at $t=1$ to 4h based on the metabolic pathways from Reactome and KEGG databases. The rectangular nodes indicated regulated genes; triangles represent intermediate genes used to link regulated genes; edges indicated biochemical reactions; and circles indicated chemical compounds that are reaction intermediates. Different colors indicated different pathways.

Transcription factor binding sites (TFBS)

To get more insight in the transcription factors that in addition to PPAR α are involved in the early induction and evolution of genes functionally related and represented by *lipid metabolic process*, we aimed to identify CRMs, i.e., the combinations of transcription factor binding sites, in promoter regions of regulated genes. We specifically aimed to identify CRMs since usually co-regulation of mammalian genes depends on sets of transcription factors rather

than individual factors alone [203,204]. Since WY14643 is a specific agonist for PPAR α , we searched for combinations putative TFBSs that included the TFBS V\$PERO in co-regulated genes (V\$PERO is the name of the PPAR response element in the Genomatix database). When promoter regions of five regulated genes at 1h that were annotated with *lipid metabolic process* (Ehhadh, Acot1, Acot2, Acot3, Acaa1a) were analyzed for CRMs, we found a framework that contained a recognition site termed CLOX in addition to the mandatory PERO element (Figure 4A). CLOX represented binding of CLOX and CLOX homology (CDP) factors, transcription factors known to be suppress transcription [205]. At 2h, 4 additional genes annotated with *lipid metabolic process* were found to be regulated (Crat, Acadvl, Hadh and Acaa2), and these were jointly analyzed with the 5 genes identified at 1h. By doing so we found two CRMs with 3 frameworks containing 3 elements each. The first consisted of binding sites for NR2F (nuclear receptor subfamily 2 factors) and HAND (twist subfamily of class B bHLH transcription factors) with the mandatory PERO element (Figure 4B). The NR2F motif corresponds to binding site for NRF2, which is a TF known to induce expression of antioxidant enzymes [206], and the HAND motif is recognized by a variety of TFs with basic function that induce transcription.

The second CRM contained RXR (RXR heterodimer binding sites) and CREB (cAMP-responsive element binding proteins) together with PERO (Figure 4C). RXR is the obligatory heterodimeric partner for PPAR α [10], and CREB proteins are important intracellular signaling factors [207]. An additional 4 genes (Acadl, Hadhb, Acadm and Sgms1) that were regulated at 3h and 4h were subsequently added to the analysis. We then found 2 CRMs with 5 major frameworks each containing 2 elements. The first one again contained RXR (Figure 4D) and second consisted of EREF (estrogen response elements) (Figure 4E) with PERO. Interestingly, in the framework identified at 1h the physical distance between two TFs was much larger than at the later time points.



Figure 4: Cis regulatory modules that were identified genes participating in lipid metabolic process at different time points. (A) CRM for 1h, TF CLOX bound with PERO in the promoters of Ehhadh, Acot3, and Acot4 genes. (B) CRM-1 for 1h+2h, the TFs NR2F and HAND were bound with PERO in the promoters of Sgms1, Acadvl and Acaa2 genes (C) CRM-2 for 1h+2h, the TFs RXR and CREB were bound with PERO in the promoters of Acot2, Acadvl and Acaa2 genes. (D) CRM-1 for 1h+2h+3h and 1h+2h+3h+4h, the TFs EREF was bound with PERO in the promoters of Acot2, Sgms1, Acadvl, Acadm and Hadhb genes (E) CRM-2 for 1h+2h+3h and 1h+2h+3h+4h, the RXR transcription factor was bound with PERO in the promoter of Acot2, Sgms1, Acadvl, Acaa2 and Acadm genes. In all CRMs the deep purple color indicated the mandatory PERO transcription factor.

Clustering and modeling

Time series expression data can be presented using a hierarchy of four systematic levels: experimental design, data analysis, pattern recognition and networks. Every level deals with a specific biological and computational issues, and also

provides as a pre-processing step for higher levels in the hierarchy [208]. Modeling is the key aspect of systems biology and it can comprise reaction models, mechanistic models, statistical models and stochastic models [209]. In the current study we aimed to model the evolution of gene expression over time, and we therefore used statistical non-linear regression models [210]. Since modeling of each individual gene is from a computational perspective challenging, we first searched for similar expression profiles using the STEM algorithm in the set of 1177 genes identified by TANOVA. We found 10 significant clusters (profiles), which are presented in Figures S4A and S4B. For example, cluster 1 contained the most genes (152), and its temporal pattern showed that expression of genes in this cluster increased up to 3h and then remained constant. In contrast, genes belonging to cluster 2, first decreased at 1h but the increased up to 3h and then remained constant. The other clusters likewise showed different behavior. Gene Ontology overrepresentation analysis revealed that genes in cluster 1 were functionally involved in *cellular lipid metabolic process*, *fatty acid metabolic process*, *lipid metabolic process*, *peroxisome*, and *fatty acid oxidation*. No distinct biological functions could be associated with the other clusters.

To further characterize the different clusters, we uncovered how many genes were overlapped between the selected 386 genes (Figure 2A) and the selected clusters (Figure S4B). We found that again most of the overlapped genes were found in cluster 1 (132 out of 152), followed by cluster 6 (55 out of 63). We observed that 4 clusters out of 10 showed comparatively higher overlap, and therefore we investigated their expression pattern by polynomial regression model (Figure 5C). To do this, we first calculated the average expression of all genes in each of the clusters and then fitted a regression model with different orders. Initially a simple regression model was fitted and then the adjusted R-square and root mean sum of square (RMSE) was calculated, after which this was repeated with a 2nd order model, and so on. The model with the highest adjusted R-square and the lowest RMSE was selected as a best predictive model. As expected, polynomial regression model gives better interpolation and better fitted pattern of the clusters than linear regression. We found that cluster 1 was best fitted using a quadratic model (R^2 (adj) = 0.987 and RMSE=0.047), indicating that 98.7% of the variation could be explained by the model. Likewise, we found that in clusters 2, 6, and 10 were the best fitted as a cubic, quadratic and cubic

model, respectively, with the highest adjusted R-square and the lowest RMSE (Figure 5C). The name of the genes with cytogenetic location of the overlapped genes in the four selected profiles were presented in the Figure S5.

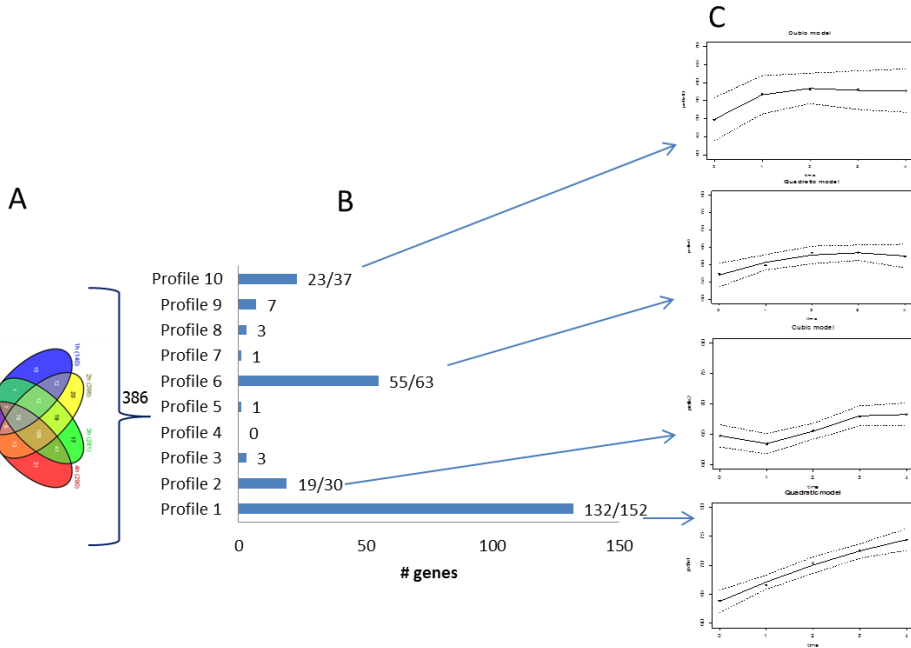


Figure 5: Overlapping genes between the selected 386 genes (with $FC > 1.2$ and $FDR < 0.05$) and the 10 selected clusters. (A) the selected 386 genes, (B) number of overlapping genes for each cluster (Profile), and (C) the fitted polynomial regression model of the four clusters. Profile 1: the fitted model is $\hat{Y} = 6.375 + 0.359 * t - 0.022 * t^2$, $R^2(adj) = 0.987$ and $RMSE = 0.047$, Profile 2: the fitted model is $\hat{Y} = 6.469 - 0.383 * t + 0.305 * t^2 - 0.047 * t^3$, $R^2(adj) = 0.995$ and $RMSE = 0.014$, Profile 6: $\hat{Y} = 5.704 + 0.438 * t - 0.075 * t^2$, $R^2(adj) = 0.899$ and $RMSE = 0.084$ and Profile 10: $\hat{Y} = 4.984 + 1.017 * t - 0.385 * t^2 + 0.045 * t^3$, $R^2(adj) = 0.982$ and $RMSE = 0.047$. The dashed line indicates the 95% confidence interval of the fitted model.

Discussion

Microarray technology has facilitated studies on the details of the gene expression data in a comprehensive way [129-132], and it has become a popular high-throughput screening platform in the area of systems biology [211]. Observing the change in expression patterns over time provides detailed information of different types of conditions instead of just observing at the terminal points of one or two time points [212]. Data from time series microarray experiments allows the unbiased comprehensive study of evolution, complex dynamics and interaction of regulated [208]. Although time-series microarrays experiments are highly relevant, still most temporal microarray data set contain only a limited number of time points, and these type of experiments are known as short-time-series data [213].

PPAR α governs the expression of a large set of genes and many of which are involved in fatty acid metabolism [15,18,166,214]. Although many studies have been performed on PPAR α regulation, no study has been performed using early time points in hepatocytes to identify the kinetics of PPAR α activation on target genes. Hence, our time-course study in rat hepatocytes represents an important advancement in our understanding of PPAR α function in hepatocytes.

A number of general conclusions can be drawn from our work. First, several sets of potential direct PPAR α target genes were identified in different profiles over time and most of them are significantly expressed already 2h after activation. Second, some novel candidate TFs were found that jointly with PPAR α regulate gene expression. Third, *lipid metabolic process* and *valine, leucine and isoleucine degradation* are the most important PPAR α target metabolic pathways. Fourth, most of the selected genes followed a quadratic model.

Genes coding for proteins which are involved in the same step of a metabolic pathway, are usually co-regulated and these genes mostly share common regulatory elements in their promoter sequences—so-called cis-regulatory modules (CRMs) [203,204]. A time course study [197] for inferring statin-induced gene regulatory relationships in primary human hepatocytes revealed a novel relationships of nuclear receptors NR2C2 and PPAR α on CYP3A4. In our study, the

results showed that *lipid metabolic process* is the most important pathways at all-time points in gene network analysis. Using the genes annotated with this functional process, we identified that the NR2F sequence was consistently closely located to the PERO recognition site in promoters of *Sgms1*, *Acadv1* and *Acaa2*. NR2F is bound by the TF NRF2 which plays an important role in controlling the response against oxidants [206]. Since activation of PPAR α induces fatty acid oxidation, hence increases oxidative stress, this would be expected. For three other genes (*Acot2*, *Acadv1* and *Acaa2*) we found RXR binding site closely located with PERO. This was envisioned since it is well known that PPAR α forms an obligatory heterodimer with RXR to function [215,216], and illustrates the biological validity of our approach.

The results from the clustering and polynomial regression modeling provide insight in the more subtle differences in temporal behavior of gene expression. Each gene cluster was reduced to a smaller set of parameters that are less noisy. The elimination of inherent variability in the data through the regression modeling approach allows a more precise comparison of the expression profiles of the various clusters. This enhances the generation of hypothesis on the molecular mechanisms that drive the observed gene expression responses [210].

Taken together, we conclude that our systems approach contributes to a better understanding of PPAR α function in rat hepatocytes. However, a series of future studies are required to investigate the different scientific issues in more detail.

Supplementary

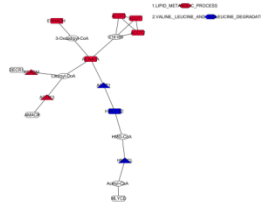


Figure S1: Gene network at 1h based on the metabolic pathways from Reactome and KEGG databases by Rspider. The rectangle nodes indicated the input genes from our list, circles were the compound, triangles were the intermediate genes, the edges indicated the biochemical reaction and different colors indicate the different pathways.

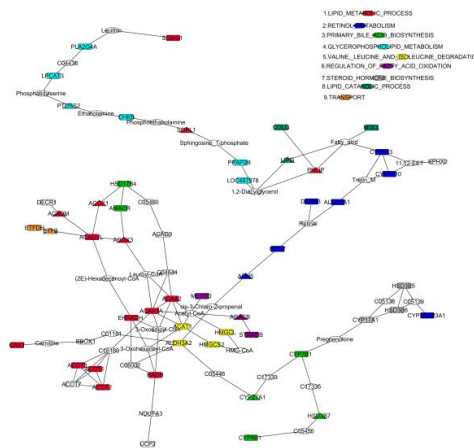


Figure S2: Gene network at 1h+2h based on the metabolic pathways from Reactome and KEGG databases by Rspider. The rectangle nodes indicated the input genes from our list, circles were the compound, triangles were the intermediate genes, the edges indicated the biochemical reaction and different colors indicate the different pathways.

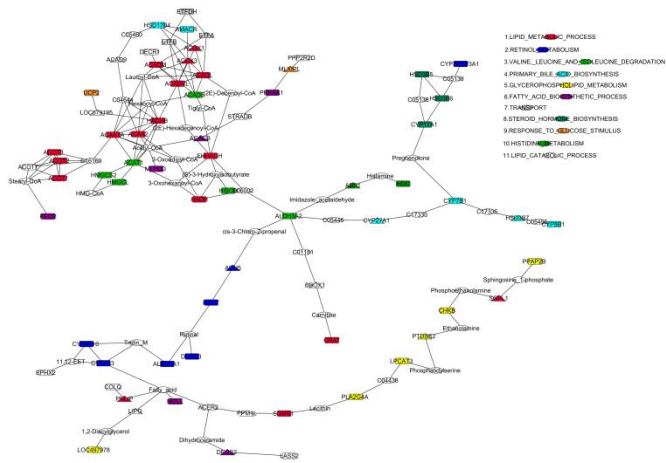


Figure S3: Gene network at 1h+2h+3h based on the metabolic pathways from Reactome and KEGG databases by Rspider. The rectangle nodes indicated the input genes from our list, circles were the compound, triangles were the intermediate genes, the edges indicated the biochemical reaction and different colors indicate the different pathways.

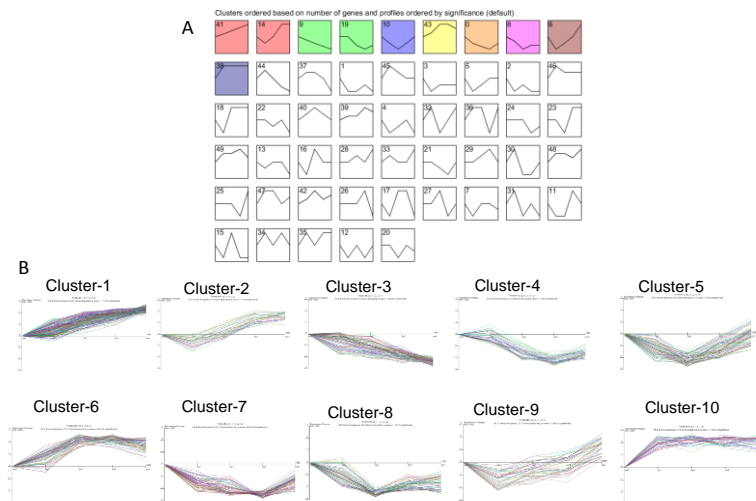


Figure S4: Clustering standardized WY14643 data of the selected genes by STEM. (A) Significant profiles (clusters) were shown by colors. (B) The details profiles of the significant profiles.

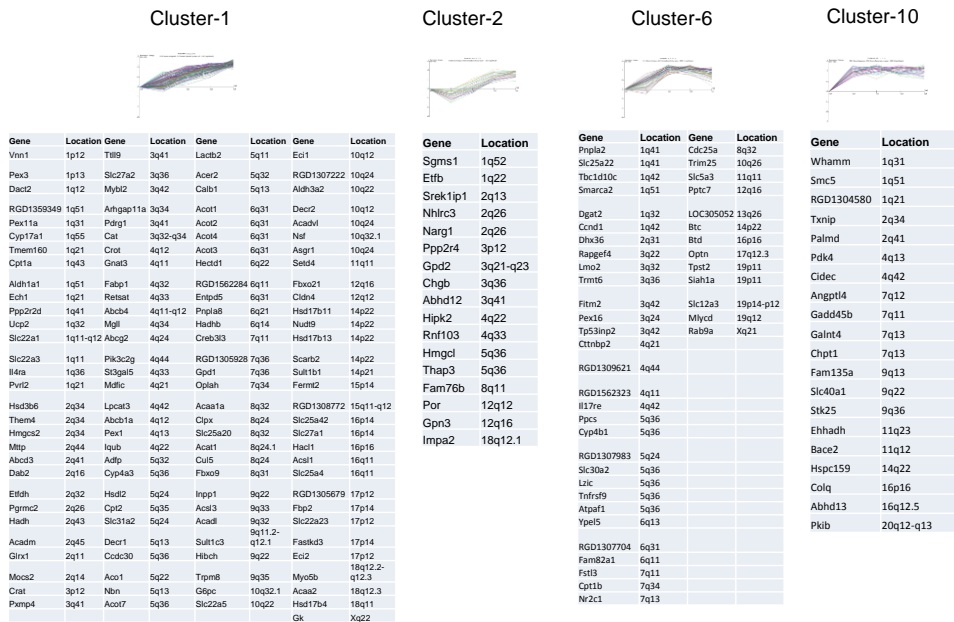


Figure S5: The genes with cytogenetic location of the overlapped genes in the four selected profiles. Cytogenetic locations were found by using 'rat2302rmentrezg.db' package in R program.

Chapter 5

Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice

Mohammad Ohid Ullah, Mark V Boekschoten, Annelies Bunschoten, Evert van Schothorst, Michael Müller, and Guido JEJ Hooiveld

Submitted

Abstract

Obesity is one of the main health problems world-wide. Excess dietary fat is stored in adipose tissue, but it has been suggested that this storage capacity is limited. Consequently, adipose tissue failure or dysfunction may drive progression of hepatic steatosis toward non-alcoholic steatohepatitis (NAFLD). However, knowledge on the functional link between adipose tissue dysfunction and NAFLD is currently limited. In this study we aimed to integrate and model the relationships between gene expression in white adipose tissue (WAT) and liver, weight status indicators as well as different plasma factors during the development of diet-induced obesity (DIO) in mice. Multiple factor analysis was used to determine the association between gene expressions in WAT and liver jointly with body weight gain and selected plasma proteins. Partial least squares-path modeling (PLSPM) for putative inter-organ signaling was used to reveal cause-effect relationships among the different blocks of multivariate observations. In both tissues a time and dose dependent effect on gene expression was observed that was most pronounced in WAT. A set of genes in both tissues and plasma leptin and insulin were found to be positively associated with body weight gain during the development of DIO. The PLSPM revealed that changes in WAT gene expression encoding for potentially secreted proteins were best explained by changes in the weight status indicators. In contrast, changes in liver gene expression were best explained by changed expression of potentially secreted proteins in WAT. Taken together, we showed that the development of DIO resulted in major changes in WAT and hepatic gene expression. The inter-organ PLSPM model identified a potential set of genes from WAT that may predict around 50% of induced metabolic changes in liver, thereby contributing to the pathogenesis of obesity.

Introduction

A vast body of literature has been published on the association between diet and chronic disease risk (see e.g. [217,218]). It is well known that an energy rich diet characterized by high intakes of dietary fat has been linked to the dramatic increase in the prevalence of obesity in both developed and developing countries in the last several decades [217-219]. Obese individuals are at increased risk of developing the metabolic syndrome, a cluster of metabolic abnormalities that ultimately increase the risk of developing vascular disease and type 2 diabetes [220-222].

In healthy situations excess energy is mainly stored as triglycerides in white adipose tissue (WAT). However, complications of obesity may in part be traced to aberrant storage of lipids in non-adipose tissues, such as liver, which can profoundly disturb organ function [223-225]. Moreover, it has been suggested that obesity starts to cause metabolic problems only when WAT cannot fully meet demands for additional storage of lipids [4,5]. As a consequence, the metabolic syndrome is often characterized by non-alcoholic fatty liver disease (NAFLD), which therefore is commonly considered as the hepatic manifestation of the metabolic syndrome [225,226].

Recently, it has been well accepted that WAT does not only serve as a storage organ, but also has an important endocrine function [225,227]. Adipokines secreted from WAT may have an important role in control of metabolism in organs other than WAT [228]. A well-known example of such an adipokine is leptin, which controls appetite in the central nervous system. Other adipokines include adiponectin, resistin, plasminogen activator inhibitor-1 (PAI-1), tumor necrosis factor alpha (TNF- α), interleukin 6 (IL-6), and estradiol. Their influence is not limited to modulation of metabolism, but includes regulation of inflammatory responses and hormone production [225,227]. However, the extent and means of inter-organ signaling between WAT and liver remains to be elucidated [228].

Genome-wide expression profiling allows an unbiased approach to the identification of genes regulated by a dietary intervention [6,139]. From the perspective of inter-organ communication, in addition to measuring levels of known adipokines, identification of potentially secreted factors whose expression

is regulated in WAT could identify novel molecules that may play a role during the development of diet-induced obesity (DIO) and its complications.

Therefore, in the present study we investigated the effects of development of DIO on gene expression in liver and WAT, and on plasma levels of selected adipokines. To this end, mice were fed 4 diets that differed in fat content for up to 12 weeks, body weight characteristics and selected adipokines were measured, and liver and WAT were subjected to microarray analysis. The various datasets were integrated using multivariate statistical tools, and specific focus was given on the interaction between WAT and liver by potentially secreted factors. We found that changes in weight status indicators mainly explained changes in plasma adipokines and gene expression in WAT, but not in liver, and we also identified a set of potentially secreted factors in WAT that explained most of the variation in hepatic gene expression.

Materials and Methods

Ethics statement

The institutional and national guidelines of the animal experiments were followed and the experiment was approved by the Local Committee for Care and Use of Laboratory Animals at Wageningen University.

Animals and diets

The animal study described here was conducted within the framework of the European Nutrigenomics Organisation NuGO, and has been described in detail by Baccini *et al* [229]. Briefly, male C57BL/6J mice were obtained from Charles River (Maastricht, The Netherlands) at three weeks of age and were housed in pairs. At twelve weeks of age, all mice received a low-fat control diet as a run-in for four weeks. This control diet contained 10 energy % (10 E%) of fat. After this run-in period, at t=0 week, mice were divided in four groups and fed diets containing 45, 30, 20, or the control diet of 10E % of dietary fat. Palm oil was the main fat source in the diets. The only other variable in the diets was the amount of corn starch. Mice were culled at the beginning of the study, after one week and four weeks.

After four weeks we continued with the 10 E% and 45 E% groups only until week 12 (n=7 to 10 per group). Body weight and food intake was recorded every week starting from t=0. Liver and WAT were harvested and subjected to transcriptomics analysis, and plasma samples were analyzed for glucose, insulin and adipokines.

Plasma adipokines and insulin

Plasma glucose concentrations were determined using a commercial device (Accu-Chek, Roche, Almere, the Netherlands). Plasma levels of insulin, leptin, resistin, monocyte chemoattractant protein-1 (MCP-1), interleukin 6 (IL-6), tumor necrosis factor alpha (TNF- α), total PAI-1 (tPAI-1), and adiponectin were measured using the mouse plasma multiplex Lincoplex Kit and Adiponectin singleplex (Linco Research, Nuclilab, Ede, the Netherlands), respectively, according to Van Schothorst *et al* [230] with slight modifications. Briefly, plasma samples were diluted 4x in HPE buffer (Sanquin, Amsterdam, the Netherlands) for the multiplex analysis and subsequently another 1,000x for the Adiponectin measurements. Assays were conducted according to the manufacturer's protocol and measured using the BioPlex X200 system and software (BioRad, Veenendaal, the Netherlands). All individual samples were analyzed in duplicate and averaged when the difference between the 2 measurements was $\leq 5\%$. Plasma levels of TNF- α and MCP-1 were below the detection levels of 3 pg/ml and 44 pg/ml, respectively, and were therefore not used in this study.

Transcriptome analysis

High quality total RNA was extracted from liver and white adipose tissue with TRIzol reagent (Invitrogen, Carlsbad, CA) and subsequently purified on columns with the RNeasy Mini Kit (Qiagen, Venlo, The Netherlands) including DNase treatment. RNA integrity was checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, Amsterdam, The Netherlands) with 6000 Nano Chips.

After isolation, RNA was labeled using the Affymetrix One-Cycle Target labeling Assay kit (Affymetrix, Santa Clara, CA). Due to the large number of samples, RNA labeling was performed in multiple rounds in a complete block design. Samples were hybridized on Affymetrix NuGO mouse arrays, washed, stained, and scanned on an Affymetrix GeneChip 3000 7G scanner. In total, 186 arrays from 93 mice were used in this study. Quality control of the datasets obtained from the scanned Affymetrix arrays was performed using Bioconductor packages [150], integrated in

an on-line pipeline [231]. Various advanced quality metrics, diagnostic plots, pseudo-images and classification methods were applied to ascertain only excellent quality arrays were used in the subsequent analyses [232]. An extensive description of the applied criteria is available upon request. Probesets were redefined according to Dai *et al* [151] utilizing current genome information. In this study probes were reorganized based on the Entrez Gene database, build 37, version 1 (remapped CDF v13). As a result, each array assays the expression of 15,501 unique genes. Normalized expression estimates were obtained from the raw intensity values using the GC-robust multi array (GCRMA) normalization, using the empirical Bayes approach to adjust background [152]. ComBat [233], an empirical Bayes method, was used to correct for the systematic error (batch effect) introduced during labeling. Differentially expressed probesets were subsequently identified using linear models, applying moderated t-statistics that implement intensity-dependent Bayes regularization of standard errors [20,234]. Only genes with a fold-change of at least 1.5 and a p-value < 0.01 were considered to be significantly regulated. Annotation information regarding biological function and cellular location of genes was analyzed through the use of IPA (Ingenuity Systems, Redwood City, CA)). The microarray dataset is deposited in the Gene Expression Omnibus (GEO) with accession number.

Multivariate data analysis

Multiple factor analysis (MFA) is an exploratory approach of multivariate data analysis to identify the association between two or more groups of sets of variables [235,236]. When more than one response variable are measured, multivariate data analysis is preferred over univariate data analysis to study how all variables are related to one another, and how they work in combination to distinguish between the cases on which the responses are made. In the current study we used three multivariate data sets; i.e., the transcriptome data from liver and WAT, and plasma levels of selected factors. The weight status indicators (body weight (BW) at start of intervention, BW at section, BW gain, absolute and relative liver weight) were used as supplementary variables. MFA was performed in R [149] using the library *FactoMineR* [29]. Before performing MFA, we filtered the transcriptome datasets by including per dataset (tissue) only those genes that were significantly different between the 11 dietary groups using ANOVA (*limma*,

moderated F-test, $p < 0.01$). Genes were considered to be associated with supplementary variable BW if their absolute correlation coefficient was significant ($p < 0.05$) and larger than 0.40.

Partial least squares path modeling

Partial least squares path modeling (PLSPM) [31,34], also known as structural equation modeling by partial least squares approach, is a methodology of multivariate data analysis that allows for modeling complex cause-effect relationships involving latent (unobserved) and observed variables. Generally speaking, these models seek to analyze the underlying causal process that is assumed to generate some phenomenon of interest. PLSPM is robust against missing values, model misspecification and violation of the statistical assumptions regarding normality and multicollinearity [184,185]. Initially, the PLSPM methodology was developed to analyze data from the chemometrics, econometrics and sociological fields, but more recently it has also been used to analyze high-throughput genomics data [36-38]. A detailed explanation on PLSPM can be found in [31,34,117]. In the current study PLSPM was used to investigate the cause-effect relationships between blocks of multiple regulated genes in adipose and liver tissues, plasma factors and weight parameters. Within PLSPM, three types of parameters were defined: (i) latent variable scores; representing 'Liver Activity', 'WAT Activity', 'plasma factors' and 'weight status' that were operationalized by reflective manifest variables, (ii) path coefficients between the endogenous and exogenous latent variables. These were the standardized regression coefficients by PLS regression of the output of inner model in PLSPM, and (iii) loadings of each block of manifest variables by reflective way; these were the output of the outer model and indicated the association between manifest and its latent variables. The significance of the path coefficients were analyzed by bootstrap sampling technique using 100 bootstrap samples. The contribution to coefficient of determination (R^2) [34] were calculated for each of the explanatory variables for predicting liver and WAT Activities. Analysis was performed in R using the library *pls* [30].

An overview of our analysis strategy is presented in Figure 1.

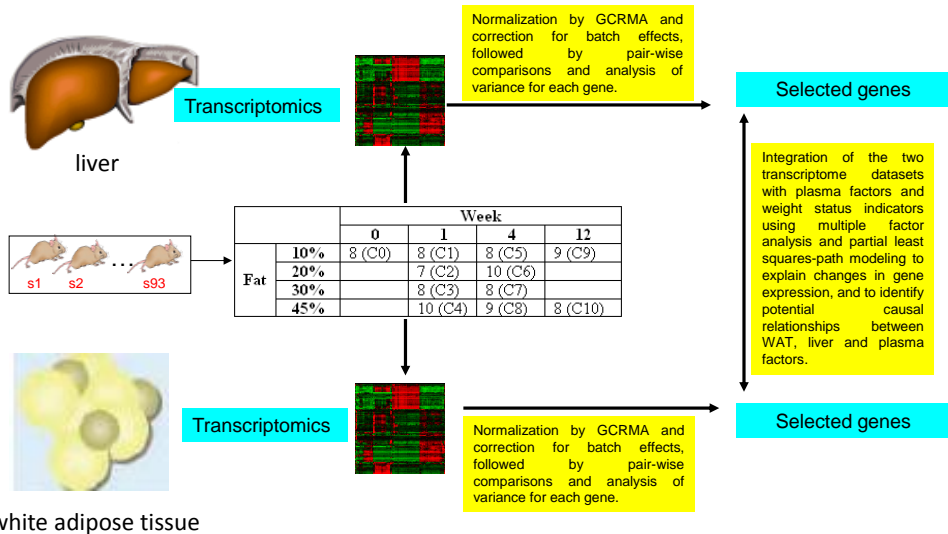


Figure 1: Overview of our analysis strategy. After normalization by GCRMA and correction for batch effects, pair-wise comparisons with the reference group (10 E% at week 0 [C0]) were made to identify genes in liver and WAT that were regulated during the development of diet-induced obesity. For each comparison, genes with moderated p -value < 0.01 were considered to be significantly regulated. The number of samples for each tissue in each diet group is listed in the table. For integrative multiple factor analysis, transcriptome data was first filtered based on p -value < 0.01 of the moderated F -test. These filtered transcriptome datasets were then combined with plasma factors (all as active variables) and weight status indicators (as supplementary variables). In addition, we also attempted to elucidate the cause-effect relationships between the blocks of variables using PLSPM.

Data representation

Weight and plasma factors are reported as mean (\pm standard deviation). Differences between the mean values of the groups (combination of diet and time) were tested for statistical significance by ANOVA with an additional Bonferroni post-hoc test (PASW Statistics 17.0 software, Chicago, Illinois). P -values < 0.05 were considered to be statistically significant.

Results

Physiological measurements

To determine the relationship between the hepatic and WAT transcriptome, physiological data (weight status indicators) and selected plasma factors during the development of DIO, mice were fed four diets containing increasing amounts of dietary fat for various time points up to 12 weeks. Mice were killed at the start of the diet intervention (10 E% at t=0) served as reference. Compared to the reference group, BW at section and as a result also BW gain were significantly different in 30 E% and 45 E% at 4 weeks and in 45 E% at 12 weeks (Table 1). These results show that body weight increased with energy percentage and time of intervention. A similar gain in body weight over time was observed between the 10 E% and 20 E% as well as 30 E% and 45 E% diet groups. Absolute liver mass was only significantly increased in 45 E% at 12 weeks, but this was not reflected in an altered liver to BW ratio. Recently, Duval *et al* [5] have showed that the hormone leptin plays an important role in the development of DIO in male C57BL/6J mice. Regarding the plasma factors, we only observed significant changes for leptin and glucose levels only in the 45 E% group at 12 weeks compared to reference group (Table 1).

Table 1: Weight status and plasma concentrations per experimental group.

Groups	Weight status					Plasma factor						
	BW at start of intervention (g)	BW at section (g)	BW gain (g)	Liver Weight at section (g)	Relative liver weight	Adiponectin (µg/ml)	Glucose (µmol/ml)	IL6 (pg/ml)	Insulin (pg/ml)	Leptin (pg/ml)	Resistin (pg/ml)	tPAI_1 (pg/ml)
W0_10 E%	26.43 (2.47)	26.43 (2.47)	.00 (.00)	.96 (.12)	.04 (.00)	11.76 (4.69)	8.95 (-1.90)	36.73 (68.51)	212.48 (83.25)	486.87 (271.62)	1920.29 (603.31)	945.10 (828.41)
W1_10 E%	26.84 (1.98)	26.66 (1.64)	-.18 (.75)	.95 (.13)	.04 (.00)	10.78 (2.21)	8.94 (2.78)	8.72 (7.70)	184.46 (169.13)	566.88 (407.48)	1364.10 (765.90)	739.66 (623.55)
W1_20 E%	26.27 (1.66)	27.33 (1.77)	1.06 (.44)	.97 (.11)	.04 (.01)	12.64 (2.28)	9.44 (1.85)	2.88 (1.41)	144.39 (140.02)	464.99 (487.19)	837.25 (782.97)	328.26 (206.50)
W1_30 E%	26.33 (1.39)	27.83 (1.69)	1.50 (.70)	1.07 (.18)	.04 (.01)	12.45 (3.22)	10.54 (2.34)	8.30 (3.48)	237.06 (136.85)	1058.33 (777.94)	1687.41 (706.54)	712.00 (304.62)
W1_45 E%	26.60 (2.18)	28.16 (1.53)	1.45 (.74)	.97 (.14)	.03 (.00)	13.03 (4.23)	9.64 (2.23)	13.83 (16.82)	270.98 (126.06)	1136.03 (789.03)	1864.62 (763.97)	1043.74 (955.01)
W4_10 E%	26.44 (1.85)	27.91 (1.86)	1.48 (.63)	.97 (.13)	.03 (.00)	10.68 (2.65)	9.49 (2.30)	14.40 (16.96)	203.60 (186.19)	733.80 (610.11)	1458.43 (647.62)	728.47 (799.44)
W4_20 E%	26.43 (1.57)	28.50 (1.83)	2.07 (1.32)	.91 (.15)	.03 (.00)	12.32 (3.21)	9.66 (2.70)	26.93 (46.47)	193.87 (97.19)	687.02 (491.72)	1164.47 (667.33)	721.98 (859.29)
W4_30 E%	26.49 (1.46)	30.33* (2.81)	3.84* (2.63)	1.07 (.12)	.04 (.00)	11.92 (2.28)	11.78 (2.33)	8.21 (5.19)	269.05 (173.17)	2520.02 (3820.22)	1464.78 (853.19)	540.66 (348.83)
W4_45 E%	26.26 (1.70)	30.58* (2.79)	4.32* (2.15)	1.05 (.19)	.03 (.01)	8.04 (1.40)	11.47 (2.64)	10.01 (9.56)	351.03 (154.86)	3478.59 (4193.36)	1729.37 (839.90)	674.40 (315.58)
W12_10E%	26.42 (1.27)	28.73 (1.53)	2.31 (1.32)	1.00 (.08)	.03 (.01)	14.57 (3.84)	10.19 (1.06)	8.65 (7.63)	114.82 (95.15)	1054.66 (1308.49)	1208.94 (770.43)	341.60 (261.12)
W12_45E%	26.25 (1.53)	38.89* (2.84)	12.64* (2.84)	1.28* (.08)	.03 (.00)	15.41 (3.58)	13.11* (0.71)	7.08 (5.16)	456.06 (288.92)	11423.67* (7729.92)	2215.81 (1191.53)	534.25 (319.49)

Values are represented as mean (SD), * indicated the group was significantly different from the reference group (W0_10E %) at $p < 0.05$ by Bonferroni post hoc test.

Differential gene expressions in liver and WAT

To identify genes in liver and WAT that were differential expressed during the development of DIO, pair-wise comparisons were made for each diet group with

the reference. Genes that satisfied the criteria of absolute $FC > 1.5$ and $p < 0.01$ were considered to be regulated. The numbers of differentially expressed genes in all pair-wise comparisons are presented in Table S1.

Generally speaking, we found that more genes were regulated in WAT than in liver. In adipose tissue most differentially expressed genes were identified in the 45 E% diet group at 12 weeks (Table S1). Remarkably, many genes were found to be significantly regulated in WAT in 10 E% group at 12 weeks, and a substantial number of these genes also overlapped with 45 E% diet group (Figure S1D). The number of differentially expressed genes in the liver was highest in the 45 E% group at week 4 and 12 compared with the 10 E% and 20 E% groups. However, the overlap of differentially expressed genes was limited. A previous study showed that the majority of WAT genes are down regulated by DIO in C57BL/6J male mice [237]. In line with these observations we also found that overall most of the genes were down regulated in both tissues.

For both tissues almost no overlap in genes regulated by dietary fat was observed at 1 and 4 weeks (Figures S1 and S2). In both tissues most of the differentially expressed genes overlapped between the 30 E% and 45 E% groups at week 4 (Figure S1F and Figure S2F) as well as between 45 E% group at week 4 and week 12 (Figure S1D and Figure S2D).

Associated genes of both tissues with weight status indicators

To identify genes and plasma factors that were associated with BW gain, we first reduced our liver and WAT transcriptome datasets by including only those genes that were significantly different in any of the 11 diet groups ($p < 0.01$, *limma* moderated F-test). Using this criterion we identified 1,421 genes in liver and 5,787 genes in WAT. Next we applied multiple factor analysis using all samples to reveal the associations of gene expression with weight status indicators and plasma factors. In this study, we considered liver and WAT gene expression levels and plasma factors as active variables, and the weight status indicators as supplementary variables.

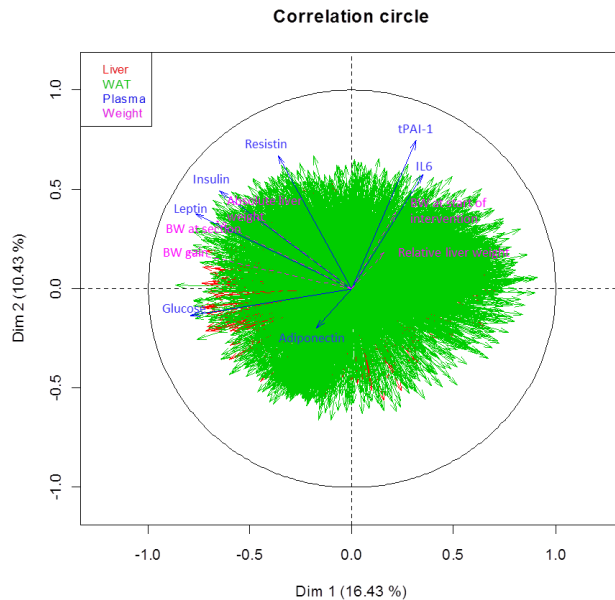


Figure 2: Loading plot of the multiple factor analysis. Visualization of the correlation coefficients between the 3 active and 1 supplementary variables and the first 2 principal components. Different colors represent the different groups of variables; red: hepatic gene expression, green: WAT gene expression, blue: plasma parameters, dashed purple: weight status indicators. The first 2 principal components explain ~27% of the variation in the dataset. The supplemental variable 'body weight gain' is highly positively correlated with a subset of genes expressed in liver and WAT, plasma leptin and insulin, and 'body weight at section'.

The loading plot obtained by MFA showed that plasma leptin and insulin levels, and to a lesser extent glucose, were positively associated with BW gain and BW at section (Figure 2). This was expected since these parameters are known to be related to BW and adipose tissue mass [5,238]. On the other hand, plasma levels of tPAI-1 and IL6 were highly associated amongst each other, but they did not correlate with BW gain, BW at section, and plasma leptin and insulin.

Since we were interested to identify genes in liver and WAT that play a role in DIO, we extracted the genes that were significantly correlated ($p < 0.05$, $r > 0.4$) with BW gain. This resulted in the selection of 2,643 genes in WAT, of which 1,037 and 1,606 genes were positively, resp. negatively associated with BW gain. Similarly, in

liver we identified 250 genes that were correlated with BW gain, of which 158 resp. 92 were positively resp. negatively associated with BW gain.

Time- and dose-dependently regulated genes in liver and WAT

Next we applied a simple linear regression model for evaluating the time- and dose-dependency of gene expression per organ. To this end, the 2,643 resp. 250 genes identified in the previous step were used as input. At first the average expression was calculated for each of these genes in each experimental group. To evaluate the time dependency of gene expression, a simple linear regression model was run using the average expression of a gene as dependent variable Y and time of intervention for each diet as independent variable X. Genes with the highest absolute regression coefficients represented genes that were most sensitive to time of intervention on each diet, where the positive and negative sign of the slope indicated increased resp. decreased expression over time. Since we did not have gene expression values for the 20 E% and 30 E% groups at 12 weeks of intervention, we excluded in the regression analyses all gene expression data for this specific time point. To infer the dose-dependency of the time-dependently regulated genes, a regression model was run that used the slopes of the initial regression models (i.e., time-dependency) as a dependent variable and the fat content of the diet as independent variable. Regression coefficients of this second model thus reflected the time- and dose-dependency of gene regulation.

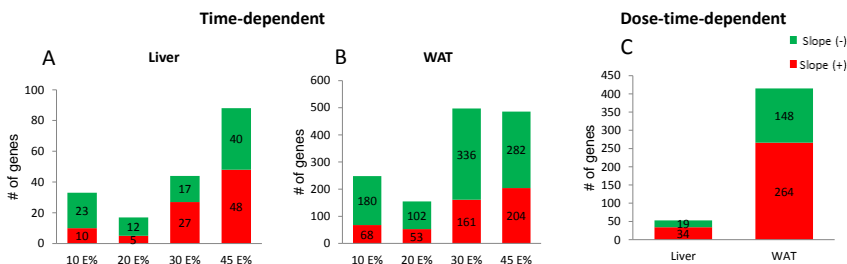


Figure 3: Time and dose dependently regulated genes. Number of time dependently regulated genes in A) WAT B) liver for each diet group based on the first 3 time points of diet intervention, and C) their dose dependency. Red and green indicated induced resp. suppressed expression. For time dependency the absolute threshold for coefficients was arbitrarily set ≥ 0.10 , and for dose dependency at ≥ 0.003 .

A number of genes were identified in liver and WAT that were dose and time dependently regulated (Figure 3). It should be kept in mind that these genes were also significantly associated with BW gain since these were initially identified in the MFA. Overall, we found that the number of genes regulated during the dietary intervention were much larger in WAT than in liver (Figure 3 A and B). Moreover, we noticed that more genes were time-dependently regulated in mice that received diets with highest amount of fat (30 E% and 45 E% versus 10 E% and 20 E% groups).

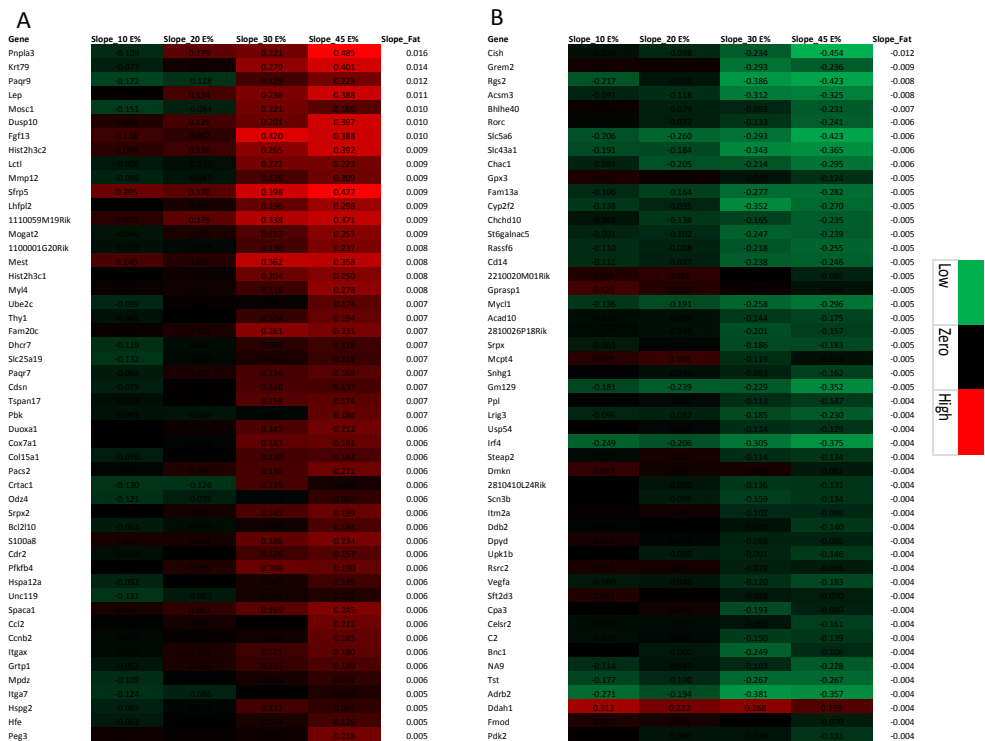


Figure 4: Heat map of regression coefficients for top time and dose dependently regulated genes in WAT. Top 50 time and dose dependently induced (panel A) and suppressed (panel B) genes in WAT identified by regression analysis. Green, black and red colors indicate the extent of time dependency, being negative, zero and positive, respectively. The column slope_Fat represents the time and dose dependency.

We observed that more genes were time and dose dependently regulated in WAT than in liver (Figure 3C). The top 50 regulated genes in WAT and top 15 regulated genes in liver are displayed in Figure 4 and Figure 5, respectively. The complete lists of regulated genes in WAT and liver are available in Table S2 and Table S3, respectively. Several genes were identified that were known to be regulated during the development of DIO, such as Pparg, Lep, Msc1 and Mest. At the functional level, the identified genes in WAT were associated with increased cell proliferation, inflammation, and fibrosis. Similarly, genes identified in liver were among others involved in lipid metabolism, development of connective tissue, steatohepatitis, and liver fibroses, all processes known to be associated with the development and progression of hepatic steatosis. Overall, in both tissues most of the genes were highly time dependently regulated in 30 E% and 45 E% groups, but not in the 10 E% and 20 E% groups.

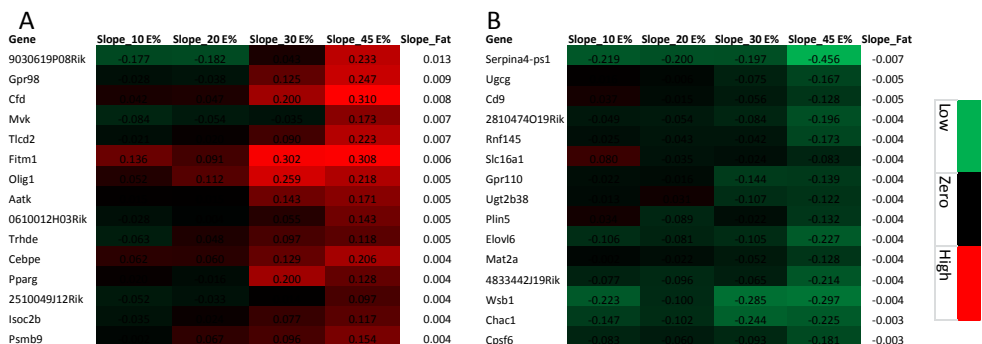


Figure 5: Heat map of regression coefficients for top time and dose dependently regulated genes in liver. Top 15 time and dose dependently induced (panel A) and suppressed (panel B) genes in liver identified by regression analysis. Green, black and red colors indicate the extent of time dependency, being negative, zero and positive, respectively. The column *slope_Fat* represents the time and dose dependency.

Results of the PLS-path model

Next we investigated the association and cause-effect relationships between the two transcriptome datasets, plasma factors and weight status indicators through PLSPM (Figure 6). We assumed that latent variables ‘WAT Activity’ and ‘Liver

Activity' in the PLS path model were reflected by the expression levels of selected genes in all conditions as determined by the microarrays. For WAT the set of dose and time dependently regulated genes was further filtered to include 69 potentially secreted genes only (Table S4), whereas for liver we used the expression data of all 53 dose and time dependent genes. The refinement for WAT on extracellular location was done since we hypothesized that cross talk between adipose and liver tissue could only occur through secreted factors. All plasma and body weight measurements were used as manifest variables for the latent variable 'plasma status' and 'weight status'. The fitted PLSPM inner model revealed that the latent variable 'WAT Activity' was more affected by 'weight status' (path coefficient = 0.63) than 'plasma status' (coefficient = 0.30). 'Liver Activity', in turn, was predicted to be more modulated by 'WAT Activity' (coefficient = 0.44) than by 'plasma status' (coefficient = 0.34) or 'weight status' (coefficient = 0.10). Thus, these outcomes indicated that especially an increase in 'weight status' would result in an increased 'WAT Activity'. Likewise, in particular an increase in 'weight status' was suggested to result in an increased 'plasma status'. The outcomes from the outer (measurement) model revealed that the manifest variables 'BW at section' (loading=0.97), 'BW gain' (0.96) and 'absolute liver weight' (0.79) correlated highly with their latent variable 'Weight status'. In contrast, 'relative liver weight' (-0.06) and 'BW at start of intervention' (0.15) only very weakly correlated with 'Weight status'. For latent variable 'plasma factors' we found a high correlation with plasma levels of leptin (0.93), glucose (0.77) and insulin (0.76). Along the same line, a set of potentially secreted genes could be identified that highly correlated with the latent variable 'WAT Activity'. Top positively correlated genes included *Serpinf1*, *Lep*, *Col6a2*, *Mest* and *Fgf13*, whereas *Ctf1*, *lqcb1* and *Grem2* were among the most negatively correlated genes. Similarly, a subset of genes was found that highly (anti-) correlated with the latent variable 'Liver Activity' (Figure 6 and Table S4).

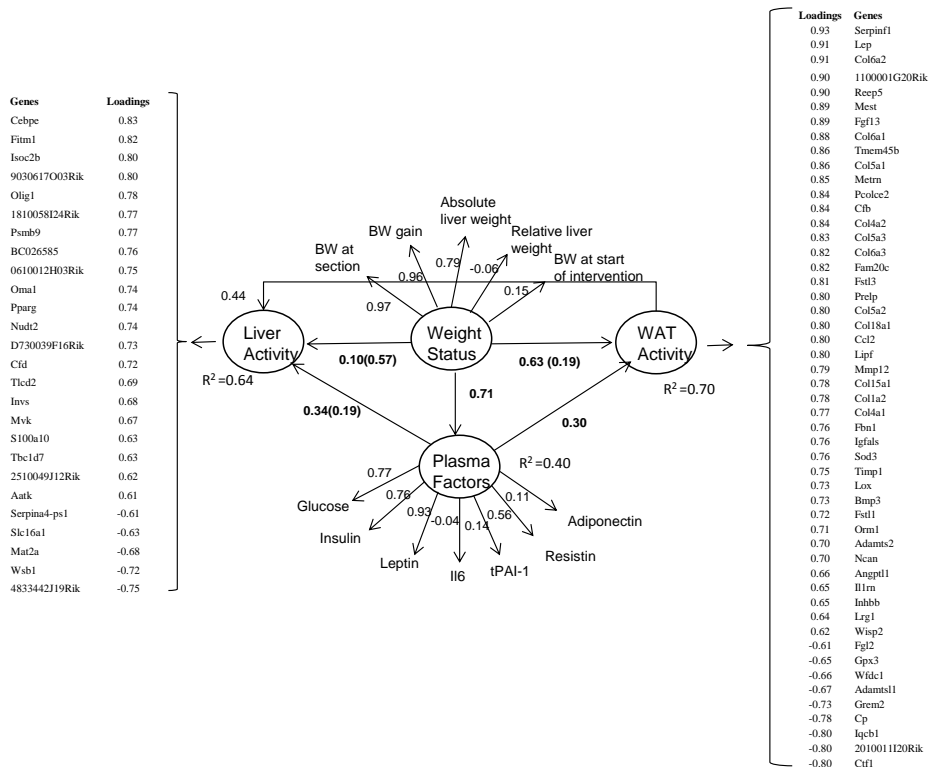


Figure 6: PLS-path model for time and dose dependently regulated genes in WAT and liver, plasma factors and weight indicators. The fitted PLS path model resulted in a good overall fit (absolute=0.52 and inner model=0.81). Indirect effects are shown in parentheses. The total effect of each of the path coefficient was significant at 5% level of significance as determined by bootstrap sampling. The plasma factors data was transformed by taking the natural logarithm, and all datasets were standardized when building the model. For WAT and liver genes, absolute loadings greater than 0.60 are presented in the figure. The complete list is available as Table S4. For weight status and plasma factors all the indicators are presented.

Next we checked the proportion of variability in the data set that could be accounted for by the model. The highest proportion of variation could be explained by the latent variable ‘WAT Activity’ ($R^2 = 0.70$), followed by ‘Liver Activity’ ($R^2 = 0.64$). Moreover, ‘WAT Activity’ was the most important variable in

the prediction of 'Liver Activity', contributing to 50.9% of the R^2 (Table 2), followed by 'plasma factors' (38.1%), whereas 'weight status' had a very low contribution on the prediction of 'Liver Activity'. On the other hand, 'weight status' (71.2%) was the most important variable in the prediction of 'WAT Activity', which suggested that weight status (especially BW gain) is an important determinant of gene expression in WAT during the development of DIO.

Table 2: Explanation of 'Liver Activity' and 'WAT Activity'

'Liver Activity'			
Explanatory variables for Liver	Path Coefficient	Correlation	Contribution to R^2 (%)
Weight Status	0.10	0.683	10.9
Plasma Factors	0.34	0.697	38.1
WAT Activity	0.44	0.722	50.9
'WAT Activity'			
Explanatory variables for WAT	Path Coefficient	Correlation	Contribution to R^2 (%)
Weight Status	0.63	0.813	71.2
Plasma Factors	0.30	0.691	28.9

From a biological perspective the results of the PLSPM analysis thus suggested that an increased gene expression profile in WAT can be mainly attributed to an increase in BW and to a lesser extent to changes in plasma factors. In turn, in this model the increased levels of potentially secreted gene products in WAT, such as leptin, are the main effectors of gene expression in the liver.

Discussion

C57BL/6J mice fed a high fat diet at different time points represent a popular animal model for human obesity and insulin resistance [239]. Nonalcoholic fatty liver disease (NAFLD) is strongly linked to obesity, and it has been suggested that proteins secreted from adipose tissue may be incriminated in the etiology of

NAFLD [228]. Moreover, it has been reported that a tight relationship exists between adipose tissue dysfunction and the pathogenesis of NAFLD [4,5,228], and previous work of our group pointed out several novel potential predictive biomarkers for NASH [5]. We extend this work in the current study, and found that changes in weight status indicators mainly explained changes in plasma adipokines and gene expression in WAT, but not in liver. Moreover, we also identified a set of potentially secreted factors in WAT that explained most of the variation in hepatic gene expression.

As expected, high-fat feeding induced body weight and adipose tissue mass that increased over time. This was also reflected in changes in gene expression that were more pronounced in adipose tissue than in liver. However, the most noticeable effects were observed in the 30 E% and 45 E% groups, and not in the 20 E% group. MFA analysis identified genes in WAT and liver that correlated with body weight gain. Several of these are already known to be regulated during DIO, such as Leptin and Pparg. At the functional level, the identified genes in liver were among others involved in lipid metabolism, development of connective tissue, steatohepatitis, and liver fibroses, all processes known to be associated with the development and progression of hepatic steatosis. Similarly, genes identified in WAT were associated with increased cell proliferation, inflammation, and fibrosis.

To identify potential causal relationships among the dose-time dependently regulated secreted WAT genes, liver genes, weight status indicators and plasma factors, we developed an inter-organ model that was analyzed by PLSPM. The PLSPM is a suitable multivariate statistical approach to handle multi-blocks of measurements and to the best of our knowledge we are one of the first to apply PLSPM to integrate and build a model for use with several transcriptomics and phenotypes related datasets. Related to our work is a study performed by Nock *et al* [240] that used structural equation modeling to define the genetic determinants of metabolic syndrome. Usually structural equation modeling with maximum likelihood (SEM-ML) approach [35] has been used to analyze multi block datasets, but it depends on a specific distribution pattern and needs more cases than variables. It is also known as hard modeling. On the other hand, a soft modeling technique such as PLSPM [31,34,117] does not depend on any specific distributional pattern and is superior for data sets that consist of fewer cases than

variables. In practice, omics experiments usually comprise of fewer cases than variables, are noisy, and suffer by multicollinearity. Therefore, the PLSPM can be used as a suitable approach for this kind of data. Recently, the concept of PLSPM has also been used by Xue *et al* [38] in their genetic association study. However, it should be realized that some biological mechanisms hidden in the gene expression data may not be revealed by PLSPM analysis. We have analyzed only linear relations among gene expression patterns, thus excluding non-linear relations which would need to be studied with extensions of the methods we used.

The partial least squares path model fitted in this study gives a good overall fit with significant path coefficients. We found that the activity of WAT played the most important contribution in the prediction of liver gene expression. On the other hand, the weight status played the largest role to predict the WAT gene expression, as was expected [42]. We also observed that liver genes were causally highly influenced by putatively secreted WAT genes that were dose and time dependently regulated, followed by plasma factors. In addition, the outcomes of PLSPM also showed that the weight status played a more important role on changes in WAT gene expression and plasma factors than hepatic gene expression. The plasma factors were found to be higher influential variables on changes in WAT gene expression than on hepatic gene expression.

In conclusion, we conclude that (i) dietary fat and time of intervention have a pronounced effect on WAT and liver as indicated by dose and time dependent changes in gene expression, (ii) the plasma factors leptin and glucose are associated with BW gain and are also associated with most of the positive changed time and dose dependent genes, and (iii) our data support the existence of a strong relationship between liver and WAT gene expression, followed by changes in plasma factors, including adipokines. All together, we conclude that the WAT gene expression profile predicts around 50% of liver gene expression profile. We also point out a set of liver genes that are strongly associated with a set of WAT secreted genes. The findings of this study give new insights on the exact role of WAT during the development of obesity and its effects on liver.

Supplementary

Figures

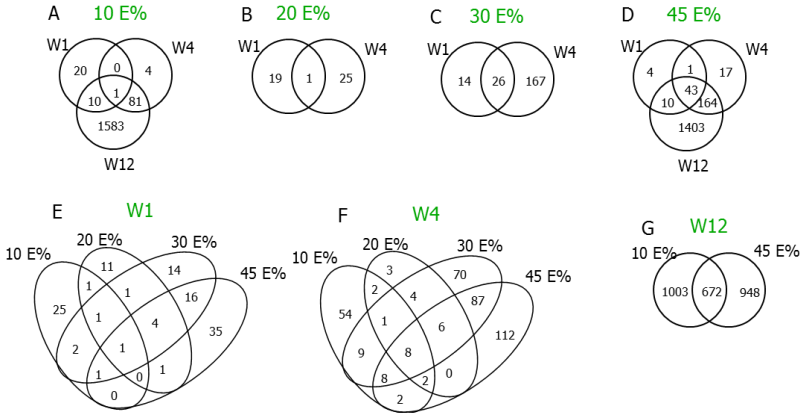


Figure S1: Venn plots representing the overlap among regulated genes per diet group in WAT. A-D: evolution of WAT gene regulation over time per diet group. E-G: dose-dependency of WAT gene regulation. Genes were considered to be regulated if the absolute FC was larger than 1.5 and $p < 0.01$.

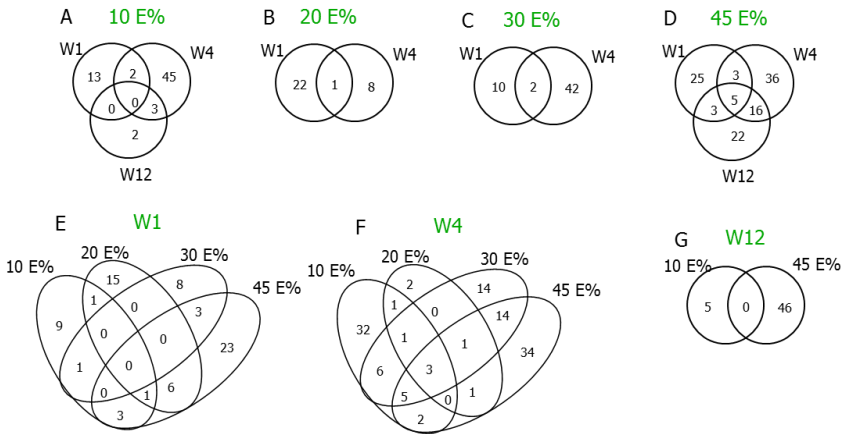


Figure S2: Venn plots representing the overlap among regulated genes per diet group in liver. A-D: evolution of hepatic gene regulation over time per diet group. E-G: dose-dependency of hepatic gene regulation. Genes were considered to be regulated if the absolute FC was larger than 1.5 and $p < 0.01$.

Tables

Table S1: Number of regulated gene per diet group compared to the reference group.

	Liver	1 Week			4 Weeks			12 Weeks			Overlap Gene
		Gene	Up	Down	Gene	Up	Down	Gene	Up	Down	
Fat E%	10%	15	11	4	50	25	25	5	1	4	0
	20%	23	13	10	9	3	6				1
	30%	12	6	6	44	23	21				2
	45%	36	15	21	60	25	35	46	33	13	5
Overlap Gene		0			3			0			
	WAT										
Fat E%	10%	31	19	12	86	55	31	1675	933	742	1
	20%	20	9	11	26	10	16				1
	30%	40	18	22	193	75	118				26
	45%	58	38	20	225	114	111	1620	733	887	43
Overlap Gene		1			8			672			

Table S2: Time and dose dependent genes in liver associated with BW gain.

Entrez	Gene (Liver)	Slope_10 E%	Slope_20 E%	Slope_30 E%	Slope_45 E%	Slope_Fat
110794	Cebpe	0.062	0.060	0.129	0.206	0.004
68680	Fitm1	0.136	0.091	0.302	0.308	0.006
67441	Isoc2b	-0.035	0.024	0.077	0.117	0.004
217830	9030617003Rik	-0.048	-0.001	0.029	0.085	0.004
50914	Olig1	0.052	0.112	0.259	0.218	0.005
67705	1810058124Rik	0.079	0.068	0.114	0.165	0.003
16912	Psmb9	-0.002	0.067	0.096	0.154	0.004
226527	BC026585	0.003	0.045	0.059	0.101	0.003
74088	0610012H03Rik	-0.028	0.004	0.055	0.143	0.005
67013	Oma1	-0.032	0.028	0.036	0.095	0.003
19016	Pparg	0.020	-0.016	0.200	0.128	0.004
66401	Nudt2	-0.056	-0.010	0.007	0.058	0.003
77996	D730039F16Rik	-0.011	0.038	0.026	0.127	0.004
11537	Cfd	0.042	0.047	0.200	0.310	0.008
380712	Tlcd2	-0.021	0.020	0.090	0.223	0.007
16348	Invs	-0.020	-0.007	0.029	0.073	0.003
17855	Mvk	-0.084	-0.054	-0.035	0.173	0.007
20194	S100a10	-0.035	0.012	0.056	0.099	0.004
67046	Tbc1d7	0.018	0.018	0.113	0.115	0.003
70291	251004912Rik	-0.052	-0.033	0.014	0.097	0.004
11302	Aatk	0.015	0.015	0.143	0.171	0.005
103140	Gstt3	-0.097	0.037	0.018	0.019	0.003
53901	Rcan2	0.050	0.008	0.158	0.157	0.004
110789	Gpr98	-0.028	-0.038	0.125	0.247	0.009
13009	Csrp3	-0.016	-0.022	-0.040	0.089	0.003
105892	9030619P08Rik	-0.177	-0.182	0.043	0.233	0.013
11761	Aox1	0.075	0.070	0.206	0.165	0.003
15957	Ifit1	0.081	0.070	0.086	0.198	0.003
237553	Trhde	-0.063	0.048	0.097	0.118	0.005
216551	1110067D22Rik	0.047	0.009	0.076	0.150	0.003
56219	Extl1	-0.018	0.035	0.072	0.103	0.003
68043	N6amt2	-0.014	0.032	0.023	0.110	0.003
80860	Ghdc	0.013	0.008	0.094	0.088	0.003
70028	Dopey2	-0.031	0.064	0.075	0.094	0.003
67246	2810474O19Rik	-0.049	-0.054	-0.084	-0.196	-0.004

74315	Rnf145	-0.025	-0.043	-0.042	-0.173	-0.004
170439	Elov16	-0.106	-0.081	-0.105	-0.227	-0.004
77596	Gpr110	-0.022	-0.016	-0.144	-0.139	-0.004
545487	Gm14439	0.009	0.000	-0.079	-0.072	-0.003
69065	Chac1	-0.147	-0.102	-0.244	-0.225	-0.003
66968	Plin5	0.034	-0.089	-0.022	-0.132	-0.004
268822	Adck5	0.084	0.018	0.000	-0.018	-0.003
75710	Rbm12	-0.049	-0.017	-0.077	-0.133	-0.003
100559	Ugt2b38	-0.013	0.031	-0.107	-0.122	-0.004
13211	Dhx9	0.012	-0.048	-0.092	-0.078	-0.003
22234	Ugcg	0.016	-0.006	-0.075	-0.167	-0.005
12527	Cd9	0.037	-0.015	-0.056	-0.128	-0.005
432508	Cpsf6	-0.083	-0.060	-0.093	-0.181	-0.003
321018	Serpina4-ps1	-0.219	-0.200	-0.197	-0.456	-0.007
20501	Slc16a1	0.080	-0.035	-0.024	-0.083	-0.004
232087	Mat2a	-0.002	-0.022	-0.052	-0.128	-0.004
78889	Wsb1	-0.223	-0.100	-0.285	-0.297	-0.004
320204	483344219Rik	-0.077	-0.096	-0.065	-0.214	-0.004

Table S3: Time and dose dependent genes in WAT associated with BW gain

Entrez	Gene (WAT)	Slope_10 E%	Slope_20 E%	Slope_30 E%	Slope_45 E%	Slope_Fat
116939	Pnpla3	-0.109	0.179	0.221	0.485	0.016
223917	Krt79	-0.077	0.035	0.279	0.401	0.014
75552	Paqr9	-0.172	-0.128	0.129	0.223	0.012
16846	Lep	0.009	0.134	0.236	0.388	0.011
66112	Mosc1	-0.151	-0.084	0.221	0.166	0.010
63953	Dusp10	0.063	0.129	0.201	0.397	0.010
14168	Fgf13	0.118	0.092	0.420	0.388	0.010
97114	Hist2h3c2	0.088	0.116	0.265	0.392	0.009
235435	Lctl	-0.056	-0.033	0.222	0.223	0.009
17381	Mmp12	-0.086	-0.047	0.135	0.209	0.009
54612	Sfrp5	0.205	0.170	0.398	0.477	0.009
218454	Lhfpl2	0.004	0.046	0.196	0.298	0.009
68800	1110059M19Rik	0.079	0.176	0.338	0.371	0.009
233549	Mogat2	-0.046	0.056	0.153	0.257	0.009
66107	1100001G20Rik	-0.037	-0.018	0.130	0.237	0.008
17294	Mest	0.140	0.071	0.362	0.358	0.008
15077	Hist2h3c1	-0.001	0.028	0.204	0.250	0.008
17896	Myl4	0.022	0.032	0.116	0.278	0.008
68612	Ube2c	-0.099	0.016	0.019	0.174	0.007
21838	Thy1	-0.042	0.005	0.104	0.194	0.007
80752	Fam20c	0.021	0.052	0.261	0.231	0.007
13360	Dhcr7	-0.119	-0.027	0.078	0.119	0.007
67283	Slc25a19	-0.132	-0.019	0.055	0.111	0.007
71904	Paqr7	-0.064	0.051	0.134	0.168	0.007
386463	Cdsn	-0.079	0.003	0.140	0.137	0.007
74257	Tspan17	-0.032	0.006	0.159	0.174	0.007
52033	Pbk	-0.041	-0.047	-0.014	0.184	0.007
213696	Duoxa1	0.002	0.028	0.147	0.212	0.006
12865	Cox7a1	0.001	-0.009	0.183	0.191	0.006
12819	Col15a1	-0.070	-0.001	0.130	0.142	0.006
21789	Pacs2	-0.011	0.065	0.136	0.211	0.006
72832	Crtac1	-0.130	-0.126	0.135	0.040	0.006
23966	Odz4	-0.121	-0.076	-0.013	0.093	0.006
68792	Srpx2	-0.003	0.041	0.145	0.199	0.006
12049	Bcl2l10	-0.064	-0.030	0.033	0.144	0.006
20201	S100a8	0.049	0.043	0.186	0.234	0.006
12585	Cdr2	-0.034	0.009	0.126	0.157	0.006
270198	Pfkfb4	-0.001	0.065	0.200	0.190	0.006
73442	Hspa12a	-0.092	-0.003	0.043	0.119	0.006
22248	Unc119	-0.131	-0.063	0.042	0.065	0.006
67652	Spaca1	0.049	0.093	0.169	0.245	0.006
20296	Ccl2	-0.006	0.041	0.008	0.212	0.006

Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice

12442	Ccnb2	-0.017	0.016	0.034	0.185	0.006
16411	Ilgax	-0.017	0.053	0.121	0.180	0.006
66790	Grtp1	-0.052	0.056	0.107	0.149	0.006
17475	Mpdz	-0.109	-0.005	0.045	0.092	0.006
16404	Ilg7	-0.124	-0.086	-0.003	0.060	0.005
15530	Hspg2	-0.083	-0.014	0.131	0.091	0.005
15216	Hfe	-0.063	0.000	0.054	0.129	0.005
18616	Peg3	0.028	0.016	0.024	0.218	0.005
16803	Lbp	0.032	0.085	0.192	0.210	0.005
12534	Cdk1	-0.024	-0.038	0.016	0.155	0.005
216343	Tph2	-0.086	-0.032	0.027	0.099	0.005
70083	Metrn	-0.055	0.012	0.058	0.133	0.005
67800	Dgat2	-0.101	-0.028	0.068	0.077	0.005
59126	Nek6	-0.064	0.006	0.084	0.117	0.005
77032	2610029I01Rik	0.002	0.029	0.098	0.178	0.005
20210	Saa3	-0.041	0.029	0.077	0.144	0.005
105892	9030619P08Rik	0.003	0.033	0.111	0.176	0.005
20167	Rtn2	-0.031	0.032	0.091	0.149	0.005
79221	Hdac9	-0.070	-0.029	0.038	0.105	0.005
18430	Oxtr	0.027	0.081	0.209	0.189	0.005
16835	Ldir	0.005	0.058	0.129	0.178	0.005
14086	Fscn1	-0.056	-0.001	0.088	0.113	0.005
20379	Sfrp4	-0.008	0.010	0.076	0.160	0.005
53313	Atp2a3	-0.013	0.053	0.090	0.167	0.005
53867	Col5a3	-0.061	-0.029	0.059	0.103	0.005
12575	Cdkn1a	0.000	0.037	0.068	0.176	0.005
14118	Fbn1	0.008	0.020	0.066	0.176	0.005
23796	Aplnr	-0.026	-0.010	0.023	0.143	0.005
53601	Pcdh12	0.014	0.062	0.078	0.188	0.005
116847	Prelp	-0.018	0.041	0.135	0.140	0.005
17314	Mgmt	-0.086	0.000	0.049	0.085	0.005
14733	Gpc1	0.111	0.142	0.291	0.254	0.005
257635	Sdsl	-0.031	-0.005	0.114	0.117	0.005
12832	Col5a2	-0.040	0.000	-0.016	0.137	0.005
67717	Lipf	-0.028	0.018	0.100	0.129	0.005
12035	Bcat1	-0.019	0.019	0.071	0.142	0.005
110075	Bmp3	0.168	0.084	0.220	0.293	0.005
72899	MacroD2	0.016	0.172	0.050	0.224	0.005
74186	Ccdc3	-0.065	0.011	0.047	0.101	0.005
29818	Hspb7	0.099	0.170	0.251	0.255	0.005
67956	Setd8	-0.069	-0.044	0.052	0.078	0.005
110208	Pgd	-0.092	-0.031	0.043	0.062	0.005
29815	Bcar3	-0.042	-0.006	0.091	0.103	0.004
22403	Wisps2	-0.014	0.001	0.052	0.138	0.004
76561	Snx7	-0.099	-0.013	-0.024	0.074	0.004
17200	Mc2r	-0.084	-0.048	-0.004	0.070	0.004
231070	Insig1	-0.072	-0.081	0.001	0.068	0.004
17345	Mki67	-0.039	-0.022	0.023	0.111	0.004
74107	Cep55	-0.036	0.012	-0.010	0.131	0.004
235135	Tmem45b	0.002	0.066	0.151	0.149	0.004
14211	Smc2	-0.028	0.031	0.012	0.139	0.004
18391	Sigmar1	-0.062	-0.016	0.044	0.087	0.004
18162	Npr3	0.025	0.067	0.140	0.170	0.004
75572	Acyp2	-0.060	-0.036	0.033	0.083	0.004
19713	Ret	0.101	0.138	0.159	0.254	0.004
11501	Adam8	-0.051	0.044	0.077	0.108	0.004
217431	Pqlc3	-0.068	-0.019	0.005	0.085	0.004
109042	Prkcdbp	-0.007	0.045	0.081	0.143	0.004
72433	Rab38	-0.045	0.012	0.062	0.103	0.004
12428	Ccna2	-0.026	-0.044	0.046	0.104	0.004
228966	Ppp1r3d	-0.103	0.051	-0.007	0.078	0.004
72713	Angptl1	-0.071	-0.023	0.036	0.070	0.004
20250	Scd2	-0.010	0.038	0.176	0.115	0.004
72033	Tsc22d2	0.000	0.018	0.069	0.139	0.004
54219	Cd320	-0.052	-0.010	0.009	0.095	0.004
239463	Fam83a	-0.071	-0.046	0.047	0.059	0.004

102294	Cyp4v3	0.048	0.084	0.157	0.180	0.004
18405	Orm1	-0.047	-0.066	0.063	0.066	0.004
69071	Tmem97	0.004	-0.043	0.046	0.122	0.004
85308	Fam158a	-0.094	0.009	-0.015	0.067	0.004
72169	Trim29	0.020	-0.011	0.108	0.134	0.004
11799	Birc5	-0.003	0.013	0.020	0.140	0.004
76905	Lrg1	-0.056	0.026	-0.015	0.104	0.004
20148	Dhrs3	-0.075	-0.025	-0.030	0.074	0.004
21857	Timp1	0.016	0.011	0.112	0.135	0.004
12831	Col5a1	-0.061	-0.010	0.003	0.082	0.004
22642	Rab71l1	-0.018	0.047	0.073	0.125	0.004
20317	Serpinf1	-0.025	0.018	0.078	0.108	0.004
66469	2810405K02Rik	-0.025	0.033	0.038	0.121	0.004
16948	Lox	0.005	0.040	0.045	0.147	0.004
13197	Gadd45a	0.105	0.111	0.141	0.237	0.004
13476	Reep5	-0.047	-0.002	0.054	0.085	0.004
67399	Pdlim7	0.208	0.202	0.194	0.345	0.004
58996	Arhgap23	-0.031	0.010	0.070	0.099	0.004
56496	Tspan6	0.061	0.057	0.196	0.168	0.004
78372	Snmp25	0.045	0.064	0.151	0.165	0.004
12835	Col6a3	-0.067	-0.017	-0.003	0.071	0.004
13038	Ctsk	-0.090	0.038	-0.040	0.077	0.004
16795	Large	-0.101	-0.019	-0.007	0.041	0.004
101437	Dhx32	-0.051	0.006	0.032	0.085	0.004
12505	Cd44	-0.032	-0.028	0.022	0.091	0.004
231123	Haus3	-0.114	-0.070	0.052	0.001	0.004
12834	Col6a2	-0.013	0.028	0.049	0.120	0.004
93726	Ear11	-0.005	0.015	0.091	0.115	0.004
66240	Kcne1l	-0.014	0.013	0.044	0.116	0.004
237847	Rtn4r1l	-0.006	0.068	0.111	0.126	0.004
16005	Igfals	-0.146	-0.114	-0.072	-0.020	0.004
67103	Ptgr1	0.161	0.185	0.108	0.310	0.004
71452	Ankrd40	-0.055	-0.031	0.050	0.062	0.004
17534	Mrc2	-0.055	-0.009	0.017	0.076	0.004
107173	Gpr137	-0.019	0.020	0.081	0.102	0.004
73379	Dcblid2	-0.063	-0.018	0.005	0.066	0.004
17909	Myo10	-0.006	0.028	0.117	0.108	0.004
12579	Cdkn2b	-0.091	-0.037	-0.068	0.048	0.004
12827	Col4a2	-0.053	-0.021	0.021	0.068	0.004
20419	Shcbp1	-0.051	-0.043	-0.011	0.068	0.003
66531	2310061C15Rik	-0.018	-0.013	0.051	0.094	0.003
114601	Ehbp1l1	0.036	-0.002	0.110	0.132	0.003
19362	Rad51ap1	-0.030	-0.018	-0.033	0.097	0.003
19348	Kif20a	-0.047	-0.021	0.022	0.070	0.003
14265	Fmr1	0.034	0.026	0.105	0.139	0.003
11749	Anxa6	0.000	0.018	0.066	0.115	0.003
75939	4930579G24Rik	-0.003	0.043	0.002	0.133	0.003
235497	Leo1	0.041	0.065	0.110	0.157	0.003
94187	Zfp423	-0.057	0.024	0.071	0.066	0.003
239436	Aard	0.041	0.045	0.075	0.157	0.003
16324	Inhbb	0.003	0.055	0.127	0.117	0.003
107373	Fam111a	-0.058	0.007	-0.026	0.078	0.003
53886	Cdkl2	-0.013	0.026	0.054	0.107	0.003
15460	Hr	-0.017	0.029	0.153	0.085	0.003
77772	Dcst1	-0.036	-0.003	0.004	0.087	0.003
107995	Cdc20	0.044	-0.007	0.023	0.150	0.003
210808	9030625A04Rik	-0.048	-0.021	0.013	0.067	0.003
20657	Sod3	-0.019	0.009	0.076	0.089	0.003
17161	Maoa	-0.063	-0.020	-0.096	0.075	0.003
26876	Adh4	-0.041	-0.010	0.047	0.067	0.003
70546	Zdhhc2	0.031	-0.001	0.106	0.121	0.003
83554	Fstl3	0.030	0.027	0.146	0.122	0.003
235587	Parp3	-0.061	-0.015	-0.002	0.057	0.003
226143	Cyp2c44	-0.039	0.037	0.083	0.077	0.003
381903	Alg8	-0.018	0.010	0.080	0.087	0.003
67468	Mmd	-0.042	-0.028	0.029	0.062	0.003
56401	Lepre1	-0.087	-0.021	-0.019	0.036	0.003

Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice

67041	Oxct1	-0.057	-0.016	0.028	0.053	0.003
75590	Dusp9	0.058	0.048	0.176	0.144	0.003
234729	Vac14	-0.047	-0.045	-0.003	0.058	0.003
67260	Lass4	-0.013	-0.003	0.017	0.097	0.003
67087	Ctnnbip1	-0.031	0.011	0.100	0.069	0.003
16403	Irga6	-0.027	-0.012	0.031	0.079	0.003
212111	Inpp5a	-0.044	-0.001	0.024	0.069	0.003
14314	Fstl1	-0.015	-0.018	0.016	0.089	0.003
67486	Polr3g	-0.051	-0.024	0.020	0.055	0.003
66427	Cyb5b	-0.018	0.009	0.072	0.084	0.003
56437	Rrad	-0.025	0.020	0.040	0.088	0.003
72759	Tmem135	-0.066	-0.029	0.018	0.040	0.003
103733	Tubg1	-0.116	-0.058	-0.042	-0.002	0.003
212898	Dse	0.009	0.043	-0.006	0.132	0.003
76820	Fam49a	-0.032	0.016	-0.027	0.091	0.003
434077	Gm5578	-0.027	0.017	0.068	0.075	0.003
217830	9030617O03Rik	-0.068	-0.014	-0.002	0.043	0.003
110542	Amhr2	-0.014	0.001	0.035	0.088	0.003
80909	Gatsl2	0.016	0.032	0.080	0.114	0.003
12615	Cenpa	-0.027	0.004	0.016	0.080	0.003
13004	Ncan	-0.032	-0.014	0.069	0.059	0.003
73569	Vgll3	0.149	0.156	0.175	0.251	0.003
69573	2310016C08Rik	0.081	0.046	0.107	0.169	0.003
70472	Atad2	-0.050	-0.038	0.032	0.043	0.003
208624	Alg3	0.023	0.012	0.119	0.105	0.003
72349	Dusp3	-0.022	-0.010	0.002	0.081	0.003
63993	Slc5a7	0.011	0.010	0.089	0.099	0.003
12826	Col4a1	-0.045	-0.010	-0.013	0.064	0.003
72345	Fam123b	-0.046	0.038	-0.015	0.079	0.003
76477	Pcolce2	-0.029	0.004	0.026	0.075	0.003
12822	Col18a1	0.011	0.078	0.050	0.129	0.003
16181	Ii1rn	-0.020	0.002	0.154	0.055	0.003
99730	Taf13	-0.003	0.026	0.030	0.102	0.003
12523	Cd84	-0.103	-0.035	-0.078	0.017	0.003
15945	Cxcl10	0.039	0.039	0.043	0.140	0.003
21991	Tpi1	-0.030	0.008	0.043	0.070	0.003
237253	Lrp11	0.012	0.022	0.026	0.114	0.003
108116	Slco3a1	-0.001	0.019	0.034	0.100	0.003
22031	Traf3	0.039	0.041	0.130	0.123	0.003
19038	Ppic	0.000	0.027	0.060	0.097	0.003
67196	Ube2t	-0.014	0.004	-0.040	0.098	0.003
68549	Sgol2	-0.020	0.008	0.023	0.081	0.003
68177	Ebpl	-0.057	0.016	-0.007	0.058	0.003
14251	Flot1	-0.048	0.010	0.023	0.056	0.003
72333	Palld	0.062	0.079	0.091	0.161	0.003
232599	Gm4876	-0.015	-0.020	0.011	0.077	0.003
14962	Cfb	-0.007	-0.006	0.028	0.085	0.003
104718	Ttc7b	-0.006	-0.003	0.086	0.075	0.003
108000	Cenpf	-0.019	-0.018	0.031	0.070	0.003
69094	Tmem160	-0.022	0.005	0.068	0.067	0.003
66508	2400001E08Rik	-0.035	0.023413	0.059	0.063	0.003
72119	Tpx2	-0.027	-0.043	-0.014	0.063	0.003
74241	Chpf	-0.024	0.000	0.065	0.064	0.003
17916	Myo1f	-0.061	-0.021	-0.046	0.047	0.003
211548	Nomo1	-0.043	-0.037	0.058	0.035	0.003
67046	Tbc1d7	-0.002	0.055	0.087	0.095	0.003
106795	Tcf19	-0.052	-0.034	-0.017	0.042	0.003
12843	Col1a2	-0.025	0.012	0.018	0.075	0.003
192193	Edem1	-0.004	0.000	0.048	0.082	0.003
20716	Bptf	-0.047	-0.017	0.002	0.048	0.003
235043	Tmem205	-0.028	-0.001	0.041	0.062	0.003
13605	Ect2	0.035	0.022	0.036	0.125	0.003
380711	Rap1gap2	0.058	0.090	0.099	0.155	0.003
232201	Arhgap25	-0.088	-0.038	-0.031	0.011	0.003
67739	Slc48a1	-0.007	0.010	0.074	0.077	0.003
13178	Dck	0.003	0.027	-0.041	0.114	0.003

17855	Mvk	-0.074	-0.042	-0.002	0.016	0.003
12038	Bche	-0.113	-0.060	-0.073	-0.010	0.003
20198	S100a4	-0.020	0.037	0.003	0.087	0.003
27214	Dbf4	-0.038	-0.005	-0.081	0.075	0.003
72709	C1qtnf6	-0.004	0.005	0.060	0.079	0.003
23934	Ly6h	-0.002	0.080	0.123	0.093	0.003
216974	Proca1	-0.124	-0.100	-0.099	-0.029	0.003
19360	Rad50	0.005	0.017	0.059	0.089	0.003
27029	Sgsh	-0.041	-0.059	-0.010	0.039	0.003
17769	Mthfr	0.038	0.131	0.090	0.150	0.003
74198	Dtx2	0.003	0.028	0.020	0.099	0.003
217946	Cdca7l	0.000	-0.016	0.091	0.067	0.003
109594	Lmo1	0.012	0.005	0.066	0.090	0.003
330260	Pon2	-0.021	0.009	0.021	0.071	0.003
211480	Kcnj14	0.005	-0.007	0.008	0.091	0.003
72607	Usp13	-0.005	0.006	0.041	0.079	0.003
16874	Lhx6	0.042	-0.020	0.047	0.110	0.003
12772	Ccr2	-0.010	0.009	-0.056	0.095	0.003
21672	Adams2	-0.039	-0.001	-0.034	0.063	0.003
68043	N6amt2	-0.001	0.032	0.055	0.088	0.003
20878	Aurka	-0.003	-0.016	0.031	0.075	0.003
12833	Col6a1	0.015	0.039	0.058	0.103	0.003
72017	Cyb5r1	0.015	0.030	0.068	0.098	0.003
50909	C1ra	-0.031	-0.040	-0.141	-0.102	-0.003
16497	Kcnab1	-0.002	-0.034	-0.105	-0.081	-0.003
23808	Ash2l	-0.002	-0.020	-0.048	-0.088	-0.003
56072	Lgals12	-0.098	-0.109	-0.157	-0.180	-0.003
66599	Rdm1	0.020	0.022	-0.040	-0.058	-0.003
19152	Prtn3	-0.115	-0.017	-0.229	-0.150	-0.003
269473	Lrig2	0.037	0.021	-0.035	-0.046	-0.003
14782	Gsr	-0.005	-0.015	-0.048	-0.093	-0.003
212073	4831426119Rik	-0.181	-0.258	-0.255	-0.284	-0.003
232236	C130022K22Rik	-0.060	-0.086	-0.140	-0.145	-0.003
13982	Esr1	-0.026	-0.033	-0.068	-0.114	-0.003
212943	Fam46a	-0.141	-0.101	-0.157	-0.219	-0.003
329470	Accs	-0.028	-0.009	-0.088	-0.104	-0.003
270110	Irf2bp2	0.021	-0.074	-0.056	-0.090	-0.003
59014	Rrs1	-0.008	-0.026	-0.068	-0.098	-0.003
320299	Iqcb1	0.011	0.000	-0.085	-0.069	-0.003
22361	Vnn1	0.120	0.028	0.023	0.013	-0.003
235441	Usp3	0.024	0.035	-0.069	-0.050	-0.003
234564	AU018778	-0.069	-0.088	-0.194	-0.147	-0.003
66869	Zfp869	0.055	0.045	-0.030	-0.028	-0.003
75750	Slc10a6	-0.134	-0.094	-0.204	-0.203	-0.003
235050	Zfp810	-0.024	-0.028	-0.148	-0.098	-0.003
12552	Cdh11	0.047	0.054	-0.088	-0.022	-0.003
218820	Zfp503	0.015	0.022	-0.131	-0.052	-0.003
217082	Hlf	-0.052	-0.036	-0.092	-0.137	-0.003
13019	Ctf1	-0.026	-0.032	-0.058	-0.120	-0.003
93834	Peli2	-0.013	-0.049	-0.106	-0.105	-0.003
53320	Folh1	0.255	0.198	0.234	0.141	-0.003
67866	Wfdc1	0.020	0.017	-0.028	-0.071	-0.003
20568	Slpi	-0.049	-0.079	-0.149	-0.139	-0.003
16601	Klf9	-0.161	-0.102	-0.250	-0.224	-0.003
21422	Tcfcp2	0.066	0.033	-0.015	-0.030	-0.003
106042	Prickle1	0.042	0.006	-0.056	-0.052	-0.003
99887	Tmem56	0.055	0.119	-0.015	-0.011	-0.003
73451	Zfp763	0.029	0.030	-0.044	-0.058	-0.003
213393	8430408G22Rik	-0.284	-0.192	-0.279	-0.357	-0.003
242608	Podn	-0.066	-0.067	-0.152	-0.152	-0.003
12298	Cacnb4	0.051	0.032	0.004	-0.048	-0.003
225372	Apbb3	-0.056	-0.045	-0.107	-0.145	-0.003
270035	Letm2	-0.014	-0.023	-0.098	-0.104	-0.003
106821	Al314976	0.054	0.012	-0.033	-0.047	-0.003
78329	2310010117Rik	0.010	-0.036	-0.097	-0.088	-0.003
20563	Slit2	0.038	0.050	-0.033	-0.048	-0.003
13078	Cyp1b1	0.094	0.108	-0.009	0.015	-0.003

Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice

75705	Eif4b	0.018	0.011	-0.001	-0.086	-0.003
19260	Ptpn22	0.215	0.078	0.174	0.073	-0.003
65099	Irak1bp1	-0.036	-0.033	-0.108	-0.126	-0.003
56448	Cyp2d22	-0.106	-0.050	-0.147	-0.184	-0.003
66966	Trit1	-0.123	-0.132	-0.180	-0.220	-0.003
107227	Macrod1	-0.033	-0.013	-0.067	-0.125	-0.003
75909	Tmem49	-0.046	-0.040	-0.086	-0.142	-0.003
320405	Cadps2	-0.055	-0.029	-0.076	-0.149	-0.003
66277	Klf15	-0.147	-0.145	-0.235	-0.235	-0.003
434234	2610020H08Rik	-0.118	-0.149	-0.243	-0.210	-0.003
16169	Ili5ra	-0.067	-0.102	-0.151	-0.169	-0.003
11622	Ahr	0.011	0.014	-0.069	-0.081	-0.003
18626	Perr1	-0.179	-0.175	-0.172	-0.288	-0.003
100647	Upk3b	-0.034	-0.092	-0.214	-0.126	-0.003
230903	Fbxo44	0.015	-0.001	-0.047	-0.087	-0.003
76454	Fbxo31	-0.129	-0.061	-0.223	-0.197	-0.003
16392	Isl1	0.002	-0.091	-0.229	-0.092	-0.003
20377	Sfrp1	0.092	0.108	0.027	0.001	-0.003
66300	Prr24	0.052	0.038	0.007	-0.053	-0.003
230751	Oscp1	-0.108	-0.095	-0.158	-0.203	-0.003
665033	Gm7455	0.013	0.045	-0.124	-0.061	-0.003
104582	Rprml	-0.045	-0.079	-0.058	-0.165	-0.003
78892	Crispld2	0.015	0.042	-0.066	-0.071	-0.003
268417	Zkscan17	-0.070	-0.051	-0.120	-0.167	-0.003
12808	Cobl	0.067	0.043	0.015	-0.045	-0.003
20442	St3gal1	-0.082	-0.135	-0.228	-0.185	-0.003
18595	Pdgfra	0.041	0.034	-0.040	-0.061	-0.003
20512	Slc1a3	-0.087	-0.115	-0.213	-0.186	-0.003
12870	Cp	-0.014	-0.050	-0.057	-0.132	-0.003
18619	Penk	0.033	0.048	-0.029	-0.065	-0.003
68695	Hddc3	-0.009	-0.058	-0.108	-0.121	-0.003
93737	Pard6g	-0.157	-0.129	-0.236	-0.248	-0.003
272428	Acsm5	-0.175	-0.162	-0.284	-0.265	-0.003
14313	Fst	0.014	-0.004	-0.095	-0.087	-0.003
13488	Drd1a	-0.076	-0.038	-0.214	-0.153	-0.003
74080	Nmnat3	-0.070	-0.064	-0.160	-0.169	-0.003
101488	Slco2b1	-0.089	-0.076	-0.170	-0.189	-0.003
244421	Lonrf1	-0.021	-0.038	-0.050	-0.141	-0.003
74155	Errfi1	-0.072	-0.043	-0.104	-0.177	-0.003
16548	Khk	0.057	0.080	0.025	-0.051	-0.003
13170	Dbp	-0.147	-0.155	-0.160	-0.270	-0.003
77739	Adamts1	0.031	0.086	-0.099	-0.048	-0.003
216505	Pik3ip1	-0.013	-0.046	-0.142	-0.124	-0.004
109828	C7	0.038	0.099	-0.044	-0.052	-0.004
14190	Fgl2	0.062	0.076	-0.109	-0.027	-0.004
70503	Ddo	-0.027	-0.094	-0.037	-0.176	-0.004
15483	Hsd11b1	-0.064	-0.054	-0.137	-0.176	-0.004
67378	Bbs2	0.051	0.016	-0.011	-0.078	-0.004
14872	Gstt2	-0.031	-0.041	-0.093	-0.152	-0.004
192199	Rspo1	0.036	0.012	-0.069	-0.081	-0.004
214804	Syde2	-0.095	-0.089	-0.198	-0.203	-0.004
67225	Rnpc3	0.006	0.020	-0.094	-0.101	-0.004
328330	D130037M23Rik	0.057	0.033	-0.062	-0.059	-0.004
68939	Rasl11b	0.014	-0.044	-0.103	-0.114	-0.004
67017	2010011I20Rik	-0.022	-0.068	-0.116	-0.151	-0.004
18604	Pdk2	0.010	-0.049	-0.046	-0.131	-0.004
14264	Fmod	0.057	0.044	0.004	-0.070	-0.004
69219	Ddah1	0.313	0.222	0.268	0.159	-0.004
11555	Adrb2	-0.271	-0.194	-0.381	-0.357	-0.004
22117	Tst	-0.177	-0.100	-0.267	-0.267	-0.004
12173	Bnc1	-0.002	-0.060	-0.249	-0.106	-0.004
12263	C2	-0.032	-0.017	-0.150	-0.139	-0.004
53883	Celsr2	-0.029	-0.027	-0.050	-0.161	-0.004
12873	Cpa3	0.008	0.039	-0.193	-0.080	-0.004
67158	Sft2d3	0.063	0.008	-0.058	-0.070	-0.004
22339	Vegfa	-0.060	-0.048	-0.120	-0.183	-0.004

20860	Rsrc2	0.053	0.043	-0.078	-0.065	-0.004
22268	Upk1b	-0.002	-0.065	-0.091	-0.146	-0.004
99586	Dpyd	0.049	-0.014	-0.088	-0.086	-0.004
107986	Ddb2	-0.007	0.000	-0.042	-0.140	-0.004
16431	Itm2a	0.013	0.035	-0.102	-0.099	-0.004
235281	Scn3b	0.000	-0.066	-0.159	-0.134	-0.004
100042332	2810410L24Rik	0.003	-0.050	-0.136	-0.131	-0.004
73712	Dmkn	0.097	0.032	0.045	-0.062	-0.004
74051	Steap2	-0.027	0.047	-0.114	-0.134	-0.004
16364	Irf4	-0.249	-0.206	-0.305	-0.375	-0.004
78787	Usp54	0.009	-0.016	-0.114	-0.129	-0.004
320398	Lrig3	-0.096	-0.082	-0.185	-0.230	-0.004
19041	Ppl	-0.004	-0.022	-0.113	-0.147	-0.004
229599	Gm129	-0.181	-0.239	-0.229	-0.352	-0.005
83673	Snhg1	-0.004	-0.039	-0.083	-0.162	-0.005
17227	Mcpt4	0.075	0.108	-0.119	-0.038	-0.005
51795	Srpx	-0.061	-0.013	-0.186	-0.183	-0.005
72655	2810026P18Rik	-0.024	-0.036	-0.201	-0.157	-0.005
71985	Acad10	-0.031	-0.038	-0.144	-0.175	-0.005
16918	Mycl1	-0.136	-0.191	-0.258	-0.296	-0.005
67298	Gprasp1	0.125	0.067	0.019	-0.043	-0.005
66528	2210020M01Rik	0.068	0.086	0.004	-0.085	-0.005
12475	Cd14	-0.111	-0.077	-0.238	-0.246	-0.005
73246	Rassf6	-0.110	-0.088	-0.218	-0.255	-0.005
26938	St6galnac5	-0.091	-0.102	-0.247	-0.239	-0.005
103172	Chchd10	-0.048	-0.138	-0.165	-0.235	-0.005
13107	Cyp2f2	-0.138	-0.095	-0.352	-0.270	-0.005
58909	Fam13a	-0.106	-0.164	-0.277	-0.282	-0.005
14778	Gpx3	0.055	0.030	-0.049	-0.124	-0.005
69065	Chac1	-0.084	-0.205	-0.214	-0.295	-0.006
72401	Slc43a1	-0.191	-0.184	-0.343	-0.365	-0.006
330064	Slc5a6	-0.206	-0.260	-0.293	-0.423	-0.006
19885	Rorc	-0.015	-0.072	-0.133	-0.241	-0.006
20893	Bhlhe40	0.021	-0.079	-0.093	-0.231	-0.007
20216	Acsm3	-0.091	-0.118	-0.312	-0.325	-0.008
19735	Rgs2	-0.217	-0.051	-0.386	-0.423	-0.008
23893	Grem2	0.037	0.027	-0.293	-0.236	-0.009
12700	Cish	-0.038	-0.098	-0.234	-0.454	-0.012

Table S4: Loadings of the PLS path model for time and dose dependent genes in liver and WAT.

Liver						WAT*					
Entrez	ID	Loadings	Entrez	ID	Loadings	Entrez	ID	Loadings	Entrez	ID	Loadings
110794	Cebpe	0.83	67246	2810474019Rik	-0.45	20317	Serpinf1	0.93	19152	Prtn3	-0.26
68680	Fitm1	0.82	74315	Rnf145	-0.47	16846	Lep	0.91	17227	Mcpt4	-0.26
67441	Isoc2b	0.80	170439	Elovl6	-0.53	12834	Col6a2	0.91	12873	Cpa3	-0.26
217830	9030617003Rik	0.80	77596	Gpr110	-0.53	66107	1100001G20Rik	0.90	73712	Dmkn	-0.26
50914	Olig1	0.78	545487	Gm14439	-0.54	13476	Reep5	0.90	50909	C1ra	-0.32
67705	1810058124Rik	0.77	69065	Chac1	-0.56	17294	Mest	0.89	14313	Fst	-0.37
16912	Psmb9	0.77	66968	Plin5	-0.57	14168	Fgf13	0.89	22339	Vegfa	-0.43
226527	BC026585	0.76	268822	Adck5	-0.58	12833	Col6a1	0.88	12263	C2	-0.48
74088	0610012H03Rik	0.75	75710	Rbm12	-0.58	235135	Tmem45b	0.86	56072	Lgals12	-0.50
67013	Oma1	0.74	100559	Ugt2b38	-0.58	12831	Col5a1	0.86	109828	C7	-0.54
19016	Pparg	0.74	13211	Dhx9	-0.58	70083	Metrn	0.85	18619	Penk	-0.55
66401	Nudt2	0.74	22234	Ugcg	-0.59	76477	Pcolce2	0.84	20563	Slit2	-0.56
77996	D730039F16Rik	0.73	12527	Cd9	-0.60	14962	Cfb	0.84	14264	Fmod	-0.58
11537	Cfd	0.72	432508	Cpsf6	-0.60	12827	Col4a2	0.84	230751	Oscp1	-0.58
380712	Tlcd2	0.69	321018	Serpina4-ps1	-0.61	53867	Col5a3	0.83	14190	Fgl2	-0.61
16348	lnvs	0.68	20501	Slc16a1	-0.63	12835	Col6a3	0.82	14778	Gpx3	-0.65
17855	Mvk	0.67	232087	Mat2a	-0.68	80752	Fam20c	0.82	67866	Wfdc1	-0.66
20194	S100a10	0.63	78889	Wsb1	-0.72	83554	Fstl3	0.81	77739	Adamts1	-0.67
67046	Tbc1d7	0.63	320204	4833442J19Rik	-0.75	116847	Prelp	0.80	23893	Grem2	-0.73
70291	2510049J12Rik	0.62				12832	Col5a2	0.80	12870	Cp	-0.78
11302	Aatk	0.61				12822	Col18a1	0.80	320299	lqcb1	-0.80
103140	Gstt3	0.60				20296	Ccl2	0.80	67017	201001120Rik	-0.80
53901	Rcan2	0.59				67717	Lipf	0.80	13019	Ctf1	-0.80
110789	Gpr98	0.58				17381	Mmp12	0.79			
13009	Csrp3	0.57				12819	Col15a1	0.78			
105892	9030619P08Rik	0.56				12843	Col1a2	0.78			
11761	Aox1	0.54				12826	Col4a1	0.77			
15957	lffit1	0.54				14118	Fbn1	0.76			
237553	Trhde	0.51				16005	Igfals	0.76			
216551	1110067D22Rik	0.51				20657	Sod3	0.76			
56219	Extl1	0.51				21857	Timp1	0.75			
68043	N6amt2	0.48				16948	Lox	0.73			
80860	Ghdc	0.45				110075	Bmp3	0.73			
70028	Dopey2	0.43				14314	Fstl1	0.72			
						18405	Orm1	0.71			
						21672	Adamts2	0.70			
						13004	Ncan	0.70			
						72713	Angptl1	0.66			
						16181	Il1rn	0.65			
						16324	Inhbb	0.65			
						76905	Lrg1	0.64			
						22403	Wisp2	0.62			
						20210	Saa3	0.59			
						72832	Crtac1	0.56			
						15945	Cxcl10	0.51			
						69071	Tmem97	0.35			

*For WAT dose-time dependent genes were categorized and only genes encoding secreted proteins were analyzed.

Chapter 6

General Discussion

Carbohydrate, protein and fat are the major nutritional components of living beings and these are the major metabolic fuel sources for the body. If the intake of metabolic fuels is greater than energy expenditure, the surplus is stored, largely as triacylglycerol in adipose tissue followed by in liver, leading to the development of obesity and its associated diseases. The products of the digestion and absorption of these carbohydrate, protein and fat are mainly glucose, amino acids, and fatty acids and mono-acyl glycerol respectively. All the products are metabolized to a common product, acetyl-CoA, which is oxidized by the citric acid cycle. Fatty acids are the major substrate. In this thesis we only focused on fat and its function in liver, small intestine and white adipose tissue over time as well as on their integration. To do these, we performed intervention studies with diets differing in the amount of fat, and also used a synthetic ligand (WY14643) to specifically activate the peroxisome proliferator-activated receptor alpha (PPAR α) [17,139,140,241-243]. PPAR α is a ligand activated transcription factor with diverse function and is activated by several synthetic compounds [7,9,13,18,143]. High affinity natural ligands include eicosanoids, unsaturated as well as long-chain fatty acids, and their activated derivatives (acyl-CoA esters) [161,162,164,166]. Moreover, it has been demonstrated that PPAR α is the major regulator of the effects of dietary fatty acids on gene expression in liver [166]. Since early 1990s, when PPAR α was discovered, its function has been studied broadly [12]. Several studies have been performed as well in last two decades; however such studies haven't been performed to integrate the function of PPAR α in different organs over time by a nutritional systems biological (NSB) approach. Therefore in this study we aimed to integrate different transcriptomics data by NSB approach, especially to characterize the function of PPAR α in different organs. NSB is the integrated approach for studying phenotypic variation and constructs prevalent models of cellular organization and function [6,46]. It also seeks to uncover how nutrition influences metabolic pathways and homeostasis [89].

Systems biology is a holistic approach that combines the knowledge of the different disciplines, such as biology, computer science, mathematics, statistics, physics and bioinformatics. Several methods and tools have already been developed to analyze and integrate high throughput omics data, the so-called top-down systems biology and model driven analysis, the so-called bottom-up systems biology. In **chapter 1** we reported an overview of systems biology and

discussed several statistical analytical approaches and software tools. Statistical tools are the most important to analyze all kinds of data. Depending on the data, design and research questions of the study, different statistical tools can be applied. Therefore, it's very essential to apply proper statistical tools and approaches in the proper design of the study. Whole genome microarray experiments are an essential part in genomics studies [129] and it produces thousands of gene expression in different experimental conditions. Many statistical tests and methods have been proposed for analyzing such data. Most tests are based on pairwise comparisons, however, the analysis of microarrays involves the testing of multiple hypotheses within one study, and it is usually known that one should control for false positives. Generally, a frequently used technique named is false discovery rate (FDR). However, the use of the FDR may be inconsistent and misleading interpretation of the comparisons across different experiments, especially when the effect sizes of the experiments vary dramatically, for instance, the case when comparing effects of potent agonists in wild type and transcription factor knockout models [142]. Therefore, we proposed an integrated statistical approach to identify transcription factor target genes from transcriptomics experiments by testing and integrating three hypotheses (contrasts) in cell means model of ANOVA (**chapter 2**). The three contrasts are based on the effect of a treatment in wild type, gene knockout, and globally over all experimental groups. We illustrated our approach using one of our datasets on the mouse [15] that focused on the identification of target genes and biological processes governed by the fatty acid sensing transcription factor PPAR α in liver, however our approach is also applicable to experiments with similar kind of design. The advantage of our method is that it properly adjusts for multiple testing while integrating data from two experiments, and it is driven by biological inference.

Integration is the key term in nutritional systems biology. Usually, fatty acids resulting from the dietary fat or synthetic ligand in small intestine are absorbed via the lymphatic system or directly through the hepatic portal vein. Fatty acids may be oxidized to acetyl-CoA (β -oxidation) or esterified with glycerol, forming triacylglycerol in the liver which is stored in adipose tissue as the body's main fuel reserve. This shows that there is a clear link between small intestine and liver. To uncover nutritional systems biology of fat in mouse liver and small intestine, we

integrated transcriptomics data of PPAR α activation in mouse liver and small intestine at the pathway level (**chapter 3**). To do this, we used WY14643 treated wild type and knock out microarray experimental data at 6h and 120h in mouse liver and small intestine [15,139,166,173]. At first, we developed an approach to collect array-wise pathway activity level by principal component analysis (PCA). PCA is able to reduce the dimension to create orthogonal components from the correlated genes. As we know in nature genes are correlated of each other, therefore to adjust their relationship is important to analyze the data in the pathway level. Since first principal component (PC) contains the most of the information of the data, therefore, we considered PC1 score as the pathway activity level or pathway score. We also developed the R code to automate the pathway score. If one has the list of pathway with gene names, then adjusting the corrected input files, it's very convenient to automate the calculation of pathway scores. We assumed that if any pathway is positively associated with a reference gene set, in our case known PPAR α target genes, then it is considered an activated pathway of PPAR α and if negatively associated with the reference gene set then it's a suppressed pathway. To find out the association we used Spearman correlation coefficients. We found that more pathways were regulated in liver than in small intestine. Afterwards, we visualized the overlapping pathways from the 6h (early) and 120h (late) time points experiments in mouse liver and small intestine to observe the temporal effects of PPAR α activation. Finally, a partial least squares path model (PLSPM) was analyzed to identify how regulation at late time points was influenced by the early regulated pathways, and what the importance of organ cross talk might be. We show that our approach enabled the identification of PPAR α dependent pathways as well as the type of regulation in mouse liver and small intestine, and that acutely induced pathways are the main drivers for regulation of pathways after long-term activation.

The partial least squares (PLS) method was originally developed by [32,33] and was used to analyze multivariate data in chemometrics, econometrics and sociological fields. Recently, it has also been widely used in high-throughput genomics data as a versatile tool [36,37]. However, the PLS approach can't handle multi block datasets. Therefore, PLS-path model (PLSPM) was developed by [31,34]. The PLSPM is an extension of PLS to handle multi block datasets to elucidate the causal relation among the different groups of data that includes

existing/prior knowledge. It is an alternate approach of structural equation modeling with maximum likelihood (SEM-ML) [35]. The PLSPM is also known as soft modeling because it doesn't depend on any distributional pattern and doesn't need more cases than variables. It is also robust against misspecification and multicollinearity problems. On the other hand, SEM-ML is known as hard modeling because it depends on distributional pattern and needs more cases than variables [34]. Recently, PLSPM has been used by [38] in their genome wide association study. Since in general the omics data are noisy and less number of cases than variables, therefore, PLSPM (soft modeling) can be used as a suitable approach for integrating and modeling multi-blocks datasets in top-down systems biology. To the best of our knowledge we are one of the first to apply PLSPM to integrate and deduce causal relationships from transcriptomics datasets based on existing knowledge.

Time-series microarrays experiments are essential to biologists for interpreting the nature of biological systems over time to several research groups [208,213]. The change in expression patterns over time provides profound information instead of just observing at the terminal points of one or two time points [212]. Although many studies have been performed on PPAR α regulation using transcriptome analysis, most of them incorporate only a single measurement in time, which often is in the order of days [139]. No study has been performed using early time points in hepatocytes to identify the kinetics of PPAR α activation on target genes. This is of particular relevance for nutrition, since the natural activators of PPAR α are rapidly metabolized. As a result, it can be envisioned that only for a limited time the concentrations of these agonists are sufficiently high levels. In other words, nutritional ligands are only briefly able to activate PPAR α mediated gene expression. It is therefore of relevance to investigate the short-term effects of PPAR α activation in a time series experiment. We therefore aimed in this study to characterize the genome-wide effects of acute PPAR α activation by detecting similar behaving genes, and analyze their biological functions, gene interaction network and transcription factor binding sites at early stage (**chapter 4**). Overall, the results reveal that PPAR α regulates a several profiles of genes over time in rat hepatocytes and most of the potential genes behave a quadratic model. Furthermore, several common transcription factors (TFs) were also predicted to bind together with PPAR α , for instance: RXR, NR2F, ERF and CREB.

Finally we showed the expansion of the gene interaction networks over time. Taken together, our study contributed important advancement in our understanding of PPAR α function for nutrition in hepatocytes.

Besides in liver and small intestine, lipids also play an important role in white adipose tissue (WAT). It has been suggested that obesity starts to cause metabolic problems only when WAT cannot fully meet demands for additional storage of lipids, which may contribute to the etiology of nonalcoholic fatty liver disease (NAFLD) [2,4,5]. This indicates that there is a clear link between WAT and liver or other organ where extra fat can store resulting obesity and its associated diseases, also called lipotoxicity [225]. Using the concept of integration in top-down systems biology, we used gene expression data from liver and WAT of mice that were subjected to diet-induced obesity. This data was integrated with data on plasma factors and weight status indicators (**chapter 5**). We identified sets of time- and dose-dependently induced genes in liver and WAT, and more genes were found to be regulated in WAT than in liver. We observed that most of the identified genes in liver involved in lipid metabolism, development of connective tissue, steatohepatitis, and liver fibroses, all processes known to be associated with the development and progression of hepatic steatosis. Likewise, genes identified in WAT were associated with increased cell proliferation, inflammation and fibrosis. Analysis by PLSPM showed that plasma factors (Leptin, Insulin, Glucose and Resistin) and the potential secreted proteins by WAT, such as leptin, Serpinf1, Mest, and Fgf13 etc. may regulate the gene expression in liver. The model also revealed that the potential set of genes from WAT that may predict around 50% of liver gene expression profile. Overall, the findings of this study give new insights on the role of WAT during the development of obesity and its effects on liver.

Taken together, we conclude that our developed approaches reported in this thesis are useful alternative ways to analyze multivariate transcriptomics datasets. When implemented in easy accessible analysis platform, such as MADMAX [231], this will promote the use of the developed approaches.

References

1. Lo CM, Nordskog BK, Nauli AM, Zheng S, Vonlehmden SB, et al. (2008) Why does the gut choose apolipoprotein B48 but not B100 for chylomicron formation? *Am J Physiol Gastrointest Liver Physiol* 294: G344-352.
2. Reddy JK, Rao MS (2006) Lipid metabolism and liver inflammation. II. Fatty liver disease and fatty acid oxidation. *Am J Physiol Gastrointest Liver Physiol* 290: G852-858.
3. Adiels M, Matikainen N, Westerbacka J, Soderlund S, Larsson T, et al. (2012) Postprandial accumulation of chylomicrons and chylomicron remnants is determined by the clearance capacity. *Atherosclerosis* 222: 222-228.
4. Strissel KJ, Stancheva Z, Miyoshi H, Perfield JW, 2nd, DeFuria J, et al. (2007) Adipocyte death, adipose tissue remodeling, and obesity complications. *Diabetes* 56: 2910-2918.
5. Duval C, Thissen U, Keshtkar S, Accart B, Stienstra R, et al. (2010) Adipose tissue dysfunction signals progression of hepatic steatosis towards nonalcoholic steatohepatitis in C57BL/6 mice. *Diabetes* 59: 3181-3191.
6. Muller M, Kersten S (2003) Nutrigenomics: goals and strategies. *Nat Rev Genet* 4: 315-322.
7. Sampath H, Ntambi JM (2004) Polyunsaturated fatty acid regulation of gene expression. *Nutr Rev* 62: 333-339.
8. Ament Z, Masoodi M, Griffin JL (2012) Applications of metabolomics for understanding the action of peroxisome proliferator-activated receptors (PPARs) in diabetes, obesity and cancer. *Genome Med* 4: 32.
9. Desvergne B, Wahli W (1999) Peroxisome proliferator-activated receptors: nuclear control of metabolism. *Endocr Rev* 20: 649-688.
10. Michalik L, Auwerx J, Berger JP, Chatterjee VK, Glass CK, et al. (2006) International Union of Pharmacology. LXI. Peroxisome proliferator-activated receptors. *Pharmacol Rev* 58: 726-741.
11. Feige JN, Gelman L, Tudor C, Engelborghs Y, Wahli W, et al. (2005) Fluorescence imaging reveals the nuclear behavior of peroxisome proliferator-activated receptor/retinoid X receptor heterodimers in the absence and presence of ligand. *J Biol Chem* 280: 17880-17890.
12. Issemann I, Green S (1990) Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators. *Nature* 347: 645-650.
13. Mandard S, Muller M, Kersten S (2004) Peroxisome proliferator-activated receptor alpha target genes. *Cell Mol Life Sci* 61: 393-416.
14. Braissant O, Fougere F, Scotto C, Dauca M, Wahli W (1996) Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha, -beta, and -gamma in the adult rat. *Endocrinology* 137: 354-366.
15. Rakhshandehroo M, Sanderson LM, Matilainen M, Stienstra R, Carlberg C, et al. (2007) Comprehensive analysis of PPARalpha-dependent regulation of hepatic lipid metabolism by expression profiling. *PPAR Res* 2007: 26839.
16. Bungler M, van den Bosch HM, van der Meijde J, Kersten S, Hooiveld GJ, et al. (2007) Genome-wide analysis of PPARalpha activation in murine small intestine. *Physiol Genomics* 30: 192-204.
17. Rakhshandehroo M, Hooiveld G, Muller M, Kersten S (2009) Comparative analysis of gene regulation by the transcription factor PPARalpha between mouse and human. *PLoS One* 4: e6796.

18. Rakhshandehroo M, Knoch B, Muller M, Kersten S (2010) Peroxisome proliferator-activated receptor alpha target genes. *PPAR Res*.
19. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98: 5116-5121.
20. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
21. Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, et al. (2010) Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS One* 5: e12336.
22. Zhou B, Xu W, Herndon D, Tompkins R, Davis R, et al. (2010) Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proceedings of the National Academy of Sciences of the United States of America* 107: 9923-9928.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545-15550.
24. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37: 1-13.
25. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4: 44-57.
26. Ingenuity Pathways Analysis. URL: [<http://www.ingenuity.com/>].
27. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B* 57: 289-300.
28. van Ommen B (2004) Nutrigenomics: exploiting systems biology in the nutrition and health arenas. *Nutrition* 20: 4-8.
29. Lê S, Josse J, Husson F (2008) FactoMineR: an R package for multivariate analysis. *J Stat Soft* 25: 1-18.
30. Sanchez G (2012) plsppm: Partial least squares data analysis methods, <http://cran.r-project.org/web/packages/plsppm>.
31. Lohmöller JB (1989) Latent variable path modeling with partial least squares. Heidelberg: Physica-Verlag.
32. Wold H (1985) Partial least squares. NewYork: Wiley.
33. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58: 109-130.
34. Tenenhaus M, Esposito Vinzi V, Y.M. C, Lauro C (2005) PLS path modeling. *Comput Stat Data An* 48: 159-205.
35. Joreskog KG (1970) A general method for analysis of covariance structure. *Biometrika* 57: 239-251.
36. Tan Y, Shi L, Hussain SM, Xu J, Tong W, et al. (2006) Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics* 22: 77-87.
37. Boulesteix AL, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* 8: 32-44.
38. Xue F, Li S, Luan J, Yuan Z, Luben RN, et al. (2012) A latent variable partial least squares path modeling approach to regional association and polygenic effect with applications to a human obesity study. *PLoS One* 7: e31927.
39. Krawetz S (2009) *Bioinformatics for Systems Biology* 2nd ed: Springer.

40. Chen L, Wang R-S, Zhang X-S (2009) *Biomolecular Networks: Methods and Applications in Systems Biology*. Wiley.
41. Aderem A (2007) Systems biology-editorial. *Curr Opin Biotech* 18: 331.
42. Fisher J, Henzinger TA (2007) Executable cell biology. *Nat Biotechnol* 25: 1239-1249.
43. Wiener N (1965) *Cybernetics or Control Communication in the Animal and the Machine*. MIT Press
44. Heinrich R, Schuster S (1996) *The Regulation of Cellular Systems*. Chapman & Hall
45. Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. *Nat Biotechnol* 22: 1249-1252.
46. Kitano H (2002) Systems biology: a brief overview. *Science* 295: 1662-1664.
47. Kitano H (2002) Computational systems biology. *Nature* 420: 206-210.
48. Aderem A (2005) Systems biology: its practice and challenges. *Cell* 121: 511-513.
49. Naylor S (2005) Systems biology, information, disease and drug discovery. *Drug Discov World* 6: 23-33.
50. Kitano H (2001) *Foundations of Systems Biology*. MIT Press.
51. Bruggeman FJ, Hornberg JJ, Boogerd FC, Westerhoff HV (2007) Introduction to systems biology. *EXS* 97 1-19.
52. van der Greef J, Martin S, Juhasz P, Adourian A, Plasterer T, et al. (2007) The art and practice of systems biology in medicine: mapping patterns of relationships. *J Proteome Res* 6: 1540-1559.
53. Clish CB, Davidov E, Oresic M, Plasterer TN, Lavine G, et al. (2004) Integrative biological analysis of the APOE*3-leiden transgenic mouse. *OMICS* 8: 3-13.
54. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.
55. Martins dos Santos V, Muller M, de Vos WM (2010) Systems biology of the gut: the interplay of food, microbiota and host at the mucosal interface. *Curr Opin Biotechnol* 21: 539-550.
56. Klipp E, Liebermeister W, Wierling C, Kowald Ae (2009) *Systems Biology: a Textbook*. Weinheim: Wiley-VCH.
57. Alon U (2007) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. London: Chapman & Hall/CRC.
58. Laukens K, Hollunder J, Dang TH, De Jaeger G, Kuiper M, et al. (2010) Flexible network reconstruction from relational databases with Cytoscape and CytoSQL. *BMC Bioinformatics* 11: 360.
59. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-2504.
60. Antonov AV, Schmidt EE, Dietmann S, Krestyaninova M, Hermjakob H (2010) R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic acids research* 38: W78-83.
61. Tieri P, Grignolio A, Zaikin A, Mishto M, Remondini D, et al. (2010) Network, degeneracy and bow tie integrating paradigms and architectures to grasp the complexity of the immune system. *Theor Biol Med Model* 7: 32.
62. Alderson D, L. L, Willinger W, Doyle J (2005) Understanding Internet topology: Principles, models, and validation. *IEEE-ACM Transactions on Networking* 13: 1205-1218.

63. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. *Proc Natl Acad Sci U S A* 104: 8685-8690.
64. Kiss HJ, Mihalik A, Nanasi T, Ory B, Spiro Z, et al. (2009) Ageing as a price of cooperation and complexity: self-organization of complex systems causes the gradual deterioration of constituent networks. *Bioessays* 31: 651-664.
65. Song C, Qu Z, Blumm N, Barabasi AL (2010) Limits of predictability in human mobility. *Science* 327: 1018-1021.
66. Xu Y, Zhang M, Wang Y, Kadambi P, Dave V, et al. (2010) A systems approach to mapping transcriptional networks controlling surfactant homeostasis. *BMC Genomics* 11: 451.
67. Bare JC, Koide T, Reiss DJ, Tenenbaum D, Baliga NS (2010) Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics* 11: 382.
68. Knox SS (2010) From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int* 10: 11.
69. Ahn AC, Tewari M, Poon CS, Phillips RS (2006) The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Med* 3: e208.
70. Baumbach J (2010) On the power and limits of evolutionary conservation--unraveling bacterial gene regulatory networks. *Nucleic Acids Res.*
71. Zoppoli P, Morganella S, Ceccarelli M (2010) TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics* 11: 154.
72. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1: S7.
73. Zou M, Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21: 71-79.
74. Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* 19 Suppl 1: i255-263.
75. Kim S, Kim J, Cho KH (2007) Inferring gene regulatory networks from temporal expression profiles under time-delay and noise. *Comput Biol Chem* 31: 239-245.
76. Bansal M, Della Gatta G, di Bernardo D (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* 22: 815-822.
77. Opgen-Rhein R, Strimmer K (2007) Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* 8 Suppl 2: S3.
78. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, et al. (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* 5: e9202.
79. Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann N Y Acad Sci* 1115: 1-22.
80. Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 Challenges. *Ann N Y Acad Sci* 1158: 159-195.
81. Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* 325: 429-432.
82. Hyman MA (2006) Systems biology: the gut-brain-fat cell connection and obesity. *Altern Ther Health Med* 12: 10-16.

83. de Graaf AA, Freidig AP, De Roos B, Jamshidi N, Heinemann M, et al. (2009) Nutritional systems biology modeling: from molecular mechanisms to physiology. *PLoS computational biology* 5: e1000554.
84. Gutteridge A, Pir P, Castrillo JI, Charles PD, Lilley KS, et al. (2010) Nutrient control of eukaryote cell growth: a systems biology study in yeast. *BMC Biol* 8: 68.
85. Palsson BO (2006) *Systems Biology-Properties of Reconstructed Networks*. Cambridge University Press.
86. Kirschner MW (2005) The meaning of systems biology. *Cell* 121: 503-504.
87. Grigorov MG (2006) Global dynamics of biological systems from time-resolved omics experiments. *Bioinformatics* 22: 1424-1430.
88. Vera J, Wolkenhauer O (2008) A system biology approach to understand functional activity of cell communication systems. *Methods Cell Biol* 90: 399-415.
89. Panagiotou G, Nielsen J (2009) Nutritional systems biology: definitions and approaches. *Annu Rev Nutr* 29: 329-339.
90. Werner T (2007) Regulatory networks: linking microarray data to systems biology. *Mechanisms of ageing and development* 128: 168-172.
91. Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics* 29: 153-159.
92. Naylor S, Culbertson AW, Valentine SJ (2008) Towards a systems level analysis of health and nutrition. *Curr Opin Biotechnol* 19: 100-109.
93. Kussmann M, Rezzi S, Daniel H (2008) Profiling techniques in nutrition and health research. *Curr Opin Biotechnol* 19: 83-99.
94. tom Dieck H, Doring F, Fuchs D, Roth HP, Daniel H (2005) Transcriptome and proteome analysis identifies the pathways that increase hepatic lipid accumulation in zinc-deficient rats. *J Nutr* 135: 199-205.
95. Nookaew I, Gabrielsson BG, Holmang A, Sandberg AS, Nielsen J (2010) Identifying Molecular Effects of Diet through Systems Biology: Influence of Herring Diet on Sterol Metabolism and Protein Turnover in Mice. *PLoS One* 5.
96. Mori MA, Liu M, Bezy O, Almind K, Shapiro H, et al. (2010) A Systems Biology Approach Identifies Inflammatory Abnormalities between Mouse Strains Prior to Development of Metabolic Disease. *Diabetes*.
97. Calvert VS, Collantes R, Elariny H, Afendy A, Baranova A, et al. (2007) A systems biology approach to the pathogenesis of obesity-related nonalcoholic fatty liver disease using reverse phase protein microarrays for multiplexed cell signaling analysis. *Hepatology* 46: 166-172.
98. Atherton HJ, Gulston MK, Bailey NJ, Cheng KK, Zhang W, et al. (2009) Metabolomics of the interaction between PPAR-alpha and age in the PPAR-alpha-null mouse. *Mol Syst Biol* 5: 259.
99. Le Cao KA, Gonzalez I, Dejean S (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25: 2855-2856.
100. de Tairac M, Le S, Aubry M, Mosser J, Husson F (2009) Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics* 10: 32.
101. Xie J, Bentler PM (2003) Covariance structure models for gene expression microarray data. *Struct Equ Modeling* 10: 566-582.
102. Prifti E, Zucker JD, Clement K, Henegar C (2008) FunNet: an integrative tool for exploring transcriptional interactions. *Bioinformatics* 24: 2636-2638.

103. Schafer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21: 754-764.
104. Maathuis MH, Colombo D, Kalisch M, Buhlmann P (2010) Predicting causal effects in large-scale systems from observational data. *Nat Methods* 7: 247-248.
105. Bøttcher SG (2003) Learning Bayesian networks with mixed variables. Aalborg University.
106. Chen X, Chen M, Ning K (2006) BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics* 22: 2952-2954.
107. Barengo M, Papouli E, Shah S, Brewer D, Miller CJ, et al. (2009) rHVDM: an R package to predict the activity and targets of a transcription factor. *Bioinformatics* 25: 419-420.
108. Zhang M, Ouyang Q, Stephenson A, Kane MD, Salt DE, et al. (2008) Interactive analysis of systems biology molecular expression data. *BMC Syst Biol* 2: 23.
109. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, et al. (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* 9: 399.
110. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H (2010) Payao: a community platform for SBML pathway model curation. *Bioinformatics* 26: 1381-1383.
111. Wurtele ES, Li J, Diao L, Zhang H, Foster CM, et al. (2003) MetNet: Software to Build and Model the Biogenetic Lattice of Arabidopsis. *Comp Funct Genomics* 4: 239-245.
112. Grimplet J, Cramer GR, Dickerson JA, Mathiason K, Van Hemert J, et al. (2009) VitisNet: "Omics" Integration through Grapevine Molecular Networks. *PLoS One* 4: e8365.
113. Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol* 1: 37.
114. Jia M, Choi SY, Reiners D, Wurtele ES, Dickerson JA (2010) MetNetGE: interactive views of biological networks and ontologies. *BMC Bioinformatics* 11: 469.
115. Wold H (1966) Estimation of principal components and related models by iterative least squares. New York: Academic Press.
116. Martens H, Naes T (1989) *Multivariate Calibration*. London: Wiley.
117. Wold H (1982) Soft modeling: the basic design and some extensions *Systems Under Indirect Observation: Causality, Structure, Prediction Part II*. K.G. Joreskog, H. Wold (Eds) ed. Amsterdam: North Holland pp. 1-54.
118. Meyer PE, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 9: 461.
119. Du P, Feng G, Flatow J, Song J, Holko M, et al. (2009) From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics* 25: i63-68.
120. Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF (2010) CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* 26: 2792-2793.
121. Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* 33: 101-103.
122. (2012) Genomatix Software Suite. <http://www.genomatix.de>.
123. Funahashi A, Tanimura N, Morohashi M, Kitano H (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*: 159-162.
124. Keating SM, Bornstein BJ, Finney A, Hucka M (2006) SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics* 22: 1275-1277.

125. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, et al. (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7: R36.
126. Mendes P, Hoops S, Sahle S, Gauges R, Dada J, et al. (2009) Computational modeling of biochemical networks using COPASI. *Methods in molecular biology* 500: 17-59.
127. Dreher F, Kreitler T, Hardt C, Kamburov A, Yildirimman R, et al. (2012) DIPSBC - Data Integration Platform for Systems Biology Collaborations. *BMC Bioinformatics* 13: 85.
128. Gehlenborg N, Seán IO, Baliga NS, Goesmann A, Matthew AH, et al. (2010) Visualization of omics data for systems biology. *Nature Methods* 7: S56 - S68.
129. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21: 33-37.
130. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249-255.
131. Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. *Nature genetics* 37 Suppl: S38-45.
132. Quackenbush J (2006) Microarray analysis and tumor classification. *The New England journal of medicine* 354: 2463-2472.
133. Kirmizis A, Farnham PJ (2004) Genomic approaches that aid in the identification of transcription factor target genes. *Experimental biology and medicine* 229: 705-721.
134. Gregory BD, Belostotsky DA (2009) Whole-genome microarrays: applications and technical issues. *Methods Mol Biol* 553: 39-56.
135. Babu MM, Lang B, Aravind L (2009) Methods to reconstruct and compare transcriptional regulatory networks. *Methods in molecular biology* 541: 163-180.
136. Brien GL, Bracken AP (2009) Transcriptomics: unravelling the biology of transcription factors and chromatin remodelers during development and differentiation. *Seminars in Cell & Developmental Biology* 20: 835-841.
137. Petricka JJ, Benfey PN (2011) Reconstructing regulatory network transitions. *Trends in Cell Biology* 21: 442-451.
138. Gorte M, Horstman A, Page RB, Heidstra R, Stromberg A, et al. (2011) Microarray-based identification of transcription factor target genes. *Methods Mol Biol* 754: 119-141.
139. Bungler M, Hooiveld GJEJ, Kersten S, Muller M (2007) Exploration of PPAR functions by microarray technology--a paradigm for nutrigenomics. *Biochim Biophys Acta* 1771: 1046-1064.
140. Woods CG, Heuvel JP, Rusyn I (2007) Genomic profiling in nuclear receptor-mediated toxicity. *Toxicologic pathology* 35: 474-494.
141. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* 7: 55-65.
142. Higdon R, van Belle G, Kolker E (2008) A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics* 24: 1225-1228.
143. Germain P, Staels B, Dacquet C, Spedding M, Laudet V (2006) Overview of nomenclature of nuclear receptors. *Pharmacological reviews* 58: 685-704.
144. Lee SS, Pineau T, Drago J, Lee EJ, Owens JW, et al. (1995) Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators. *Mol Cell Biol* 15: 3012-3022.
145. Willson TM, Brown PJ, Sternbach DD, Henke BR (2000) The PPARs: from orphan receptors to drug discovery. *J Med Chem* 43: 527-550.

146. Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. *Journal of computational biology* 7: 819-837.
147. Ayroles JF, Gibson G (2006) Analysis of variance of microarray data. *Methods in enzymology* 411: 214-233.
148. Kutner MH, Nachtsheim, C.J., Neter, J., & Li, W. (2005) *Applied Linear Statistical Models*. New York: McGraw-Hill.
149. Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* 5: 299-314.
150. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
151. Dai M, Wang P, Boyd AD, Kostov G, Athey B, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33: e175.
152. Wu Z, Irizarry RA, Gentleman R, Martinez- Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 99: 909-917.
153. Smyth G (2005) *limma: Linear Models for Microarray Data*. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer New York. pp. 397-420.
154. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21: 3940-3941.
155. Huber W, Scholtens D, Hahne F, Heydebreck A (2008) *Differential Expression*. In: Hahne F, Huber W, Gentleman R, Falcon S, editors. *Bioconductor Case Studies*. New York: Springer New York. pp. 89-102.
156. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* 24: 1650-1651.
157. Grossmann S, Bauer S, Robinson PN, Vingron M (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* 23: 3024-3031.
158. Kersten S, Mandard S, Escher P, Gonzalez FJ, Tafuri S, et al. (2001) The peroxisome proliferator-activated receptor alpha regulates amino acid metabolism. *The FASEB Journal* 15: 1971-1978.
159. Lefebvre P, Chinetti G, Fruchart JC, Staels B (2006) Sorting out the roles of PPAR alpha in energy metabolism and vascular homeostasis. *J Clin Invest* 116: 571-580.
160. Claudel T, Staels B, Kuipers F (2005) The Farnesoid X receptor: a molecular link between bile acid and lipid and glucose metabolism. *Arterioscler Thromb Vasc Biol* 25: 2020-2030.
161. Forman BM, Chen J, Evans RM (1997) Hypolipidemic drugs, polyunsaturated fatty acids, and eicosanoids are ligands for peroxisome proliferator-activated receptors alpha and delta. *Proceedings of the National Academy of Sciences of the United States of America* 94: 4312-4317.
162. Kliewer SA, Sundseth SS, Jones SA, Brown PJ, Wisely GB, et al. (1997) Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors alpha and gamma. *Proceedings of the National Academy of Sciences of the United States of America* 94: 4318-4323.
163. Krey G, Braissant O, L'Horsset F, Kalkhoven E, Perroud M, et al. (1997) Fatty acids, eicosanoids, and hypolipidemic agents identified as ligands of peroxisome proliferator-

- activated receptors by coactivator-dependent receptor ligand assay. *Molecular endocrinology* 11: 779-791.
164. Xu HE, Lambert MH, Montana VG, Parks DJ, Blanchard SG, et al. (1999) Molecular recognition of fatty acids by peroxisome proliferator-activated receptors. *Molecular cell* 3: 397-403.
 165. Hostetler HA, Kier AB, Schroeder F (2006) Very-long-chain and branched-chain fatty acyl-CoAs are high affinity ligands for the peroxisome proliferator-activated receptor alpha (PPARalpha). *Biochemistry* 45: 7669-7681.
 166. Sanderson LM, de Groot PJ, Hooiveld GJ, Koppen A, Kalkhoven E, et al. (2008) Effect of synthetic dietary triglycerides: a novel research paradigm for nutrigenomics. *PLoS One* 3: e1681.
 167. McKenna NJ, Lanz RB, O'Malley BW (1999) Nuclear receptor coregulators: cellular and molecular biology. *Endocrine reviews* 20: 321-344.
 168. Lonard DM, Lanz RB, O'Malley BW (2007) Nuclear receptor coregulators and human disease. *Endocrine reviews* 28: 575-587.
 169. Feige JN, Gelman L, Rossi D, Zoete V, Metivier R, et al. (2007) The endocrine disruptor monoethyl-hexyl-phthalate is a selective peroxisome proliferator-activated receptor gamma modulator that promotes adipogenesis. *J Biol Chem* 282: 19152-19166.
 170. Ricote M, Glass CK (2007) PPARs and molecular mechanisms of transrepression. *Biochimica et biophysica acta* 1771: 926-935.
 171. Escher P, Braissant O, Basu-Modak S, Michalik L, Wahli W, et al. (2001) Rat PPARs: quantitative analysis in adult rat tissues and regulation in fasting and refeeding. *Endocrinology* 142: 4195-4202.
 172. Stienstra R, Mandard S, Patsouris D, Maass C, Kersten S, et al. (2007) Peroxisome proliferator-activated receptor alpha protects against obesity-induced hepatic inflammation. *Endocrinology* 148: 2753-2763.
 173. de Vogel-van den Bosch HM, Bunger M, de Groot PJ, Bosch-Vermeulen H, Hooiveld GJ, et al. (2008) PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression. *BMC Genomics* 9: 231.
 174. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. *Ann Appl Stat* 1: 107-129.
 175. Zhang J, Li J, Deng HW (2009) Identifying gene interaction enrichment for gene expression data. *PLoS One* 4: e8064.
 176. Lu Y, Liu PY, Xiao P, Deng HW (2005) Hotelling's T² multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* 21: 3105-3113.
 177. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93-99.
 178. Nam D, Kim SY (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform* 9: 189-197.
 179. Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6: 225.
 180. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 29: 1213-1222.
 181. Hotelling H (1933) Analysis of complex statistical variables into principal components. *J Educ Psychol* 24: 417-441.

182. Dobrin R, Zhu J, Molony C, Argman C, Parrish ML, et al. (2009) Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol* 10: R55.
183. Merico D, Isserlin R, Bader GD (2011) Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. *Methods Mol Biol* 781: 257-277.
184. Cassel CM, Hackl P, Westlund AH (1999) Robustness of partial least-squares method for estimating latent variable quality structures. *J Appl Stat* 26: 435-446.
185. Cassel CM, Hackl P, Westlund AH (2000) On measurement of intangible assets: a study of robustness of partial least squares. *Total Qual Manag* 11: 897-907.
186. Hood L (2003) Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev* 124: 9-16.
187. Bunger M (2008) Probing the role of PPAR α in the small intestine: A functional nutrigenomics approach Wageningen University and Research Centre, the Netherlands.
188. Keller H, Dreyer C, Medin J, Mahfoudi A, Ozato K, et al. (1993) Fatty acids and retinoids control lipid metabolism through activation of peroxisome proliferator-activated receptor-retinoid X receptor heterodimers. *Proc Natl Acad Sci U S A* 90: 2160-2164.
189. Berger J, Moller DE (2002) The mechanisms of action of PPARs. *Annu Rev Med* 53: 409-435.
190. Peters JM, Cheung C, Gonzalez FJ (2005) Peroxisome proliferator-activated receptor-alpha and liver cancer: where do we stand? *Journal of molecular medicine* 83: 774-785.
191. Rao MS, Reddy JK (2001) Peroxisomal beta-oxidation and steatohepatitis. *Seminars in liver disease* 21: 43-55.
192. Baker TK, Carfagna MA, Gao H, Dow ER, Li Q, et al. (2001) Temporal gene expression analysis of monolayer cultured rat hepatocytes. *Chemical research in toxicology* 14: 1218-1231.
193. Briguglio E, Di Paola R, Paterniti I, Mazzon E, Oteri G, et al. (2010) WY-14643, a Potent Peroxisome Proliferator Activator Receptor-alpha PPAR-alpha Agonist Ameliorates the Inflammatory Process Associated to Experimental Periodontitis. *PPAR research* 2010: 193019.
194. Ross PK, Woods CG, Bradford BU, Kosyk O, Gatti DM, et al. (2009) Time-course comparison of xenobiotic activators of CAR and PPARalpha in mouse liver. *Toxicology and applied pharmacology* 235: 199-207.
195. Miller RT, Glover SE, Stewart WS, Corton JC, Popp JA, et al. (1996) Effect on the expression of c-met, c-myc and PPAR-alpha in liver and liver tumors from rats chronically exposed to the hepatocarcinogenic peroxisome proliferator WY-14,643. *Carcinogenesis* 17: 1337-1341.
196. Cattley RC, Popp JA (1989) Differences between the promoting activities of the peroxisome proliferator WY-14,643 and phenobarbital in rat liver. *Cancer research* 49: 3246-3251.
197. Schroder A, Wollnik J, Wrzodek C, Drager A, Bonin M, et al. (2011) Inferring statin-induced gene regulatory relationships in primary human hepatocytes. *Bioinformatics* 27: 2473-2477.
198. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.
199. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.

200. Ernst J, Bar-Joseph Z (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7: 191.
201. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29.
202. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21: 2933-2942.
203. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics* 34: 166-176.
204. Van Loo P, Marynen P (2009) Computational methods for the detection of cis-regulatory modules. *Brief Bioinform* 10: 509-524.
205. Pattison S, Skalnik DG, Roman A (1997) CCAAT displacement protein, a regulator of differentiation-specific gene expression, binds a negative regulatory element within the 5' end of the human papillomavirus type 6 long control region. *J Virol* 71: 2013-2022.
206. Baird L, Dinkova-Kostova AT (2011) The cytoprotective role of the Keap1-Nrf2 pathway. *Arch Toxicol* 85: 241-272.
207. Carlezon WA, Jr., Duman RS, Nestler EJ (2005) The many faces of CREB. *Trends Neurosci* 28: 436-445.
208. Bar-Joseph Z (2004) Analyzing time series gene expression data. *Bioinformatics* 20: 2493-2503.
209. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. *Nature cell biology* 8: 1195-1203.
210. Eastwood DC, Mead A, Sergeant MJ, Burton KS (2008) Statistical modelling of transcript profiles of differentially regulated genes. *BMC Mol Biol* 9: 66.
211. Panda S, Sato TK, Hampton GM, Hogenesch JB (2003) An array of insights: application of DNA chip technology in the study of cell biology. *Trends in cell biology* 13: 151-156.
212. Androulakis IP, Yang E, Almon RR (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual review of biomedical engineering* 9: 205-228.
213. Wang X, Wu M, Li Z, Chan C (2008) Short time-series microarray analysis: methods and challenges. *BMC systems biology* 2: 58.
214. Guo L, Fang H, Collins J, Fan XH, Dial S, et al. (2006) Differential gene expression in mouse primary hepatocytes exposed to the peroxisome proliferator-activated receptor alpha agonists. *BMC Bioinformatics* 7 Suppl 2: S18.
215. Kliewer SA, Umesono K, Noonan DJ, Heyman RA, Evans RM (1992) Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. *Nature* 358: 771-774.
216. Dreyer C, Krey G, Keller H, Givel F, Helftenbein G, et al. (1992) Control of the peroxisomal beta-oxidation pathway by a novel family of nuclear hormone receptors. *Cell* 68: 879-887.
217. Amine E, Baba N, Belhadj M, Deurenberg-Yap M, Djazayeri A, et al. (2003) Diet, nutrition and the prevention of chronic diseases. Geneva, Switzerland.: World Health Organization. 0512-3054 0512-3054. 1-149 p.
218. Roberts CK, Barnard RJ (2005) Effects of exercise and diet on chronic disease. *J Appl Physiol* 98: 3-30.
219. Kelly T, Yang W, Chen CS, Reynolds K, He J (2008) Global burden of obesity in 2005 and projections to 2030. *Int J Obes (Lond)* 32: 1431-1437.

220. Reaven GM (1995) Pathophysiology of insulin resistance in human disease. *Physiol Rev* 75: 473-486.
221. Moller DE, Kaufman KD (2005) Metabolic syndrome: a clinical and molecular perspective. *Annu Rev Med* 56: 45-62.
222. Malnick SD, Knobler H (2006) The medical complications of obesity. *QJM* 99: 565-579.
223. Lionetti L, Mollica MP, Lombardi A, Cavaliere G, Gifuni G, et al. (2009) From chronic overnutrition to insulin resistance: the role of fat-storing capacity and inflammation. *Nutr Metab Cardiovasc Dis* 19: 146-152.
224. Szendroedi J, Roden M (2009) Ectopic lipids and organ function. *Curr Opin Lipidol* 20: 50-56.
225. Cusi K (2012) Role of obesity and lipotoxicity in the development of nonalcoholic steatohepatitis: pathophysiology and clinical implications. *Gastroenterology* 142: 711-725 e716.
226. Perlemuter G, Bigorgne A, Cassard-Doulcier AM, Naveau S (2007) Nonalcoholic fatty liver disease: from pathogenesis to patient care. *Nat Clin Pract Endocrinol Metab* 3: 458-469.
227. Conde J, Scotecce M, Gomez R, Lopez V, Gomez-Reino JJ, et al. (2011) Adipokines: biofactors from white adipose tissue. A complex hub among inflammation, metabolism, and immunity. *Biofactors* 37: 413-420.
228. Wree A, Kahraman A, Gerken G, Canbay A (2011) Obesity affects the liver - the link between adipocytes and hepatocytes. *Digestion* 83: 124-133.
229. Baccini M, Bachmaier EM, Biggeri A, Boekschoten MV, Bouwman FG, et al. (2008) The NuGO proof of principle study package: a collaborative research effort of the European Nutrigenomics Organisation. *Genes Nutr* 3: 147-151.
230. van Schothorst EM, Bunschoten A, Schrauwen P, Mensink RP, Keijer J (2009) Effects of a high-fat, low- versus high-glycemic index diet: retardation of insulin resistance involves adipose tissue modulation. *FASEB J* 23: 1092-1101.
231. Lin K, Kools H, de Groot PJ, Gavai AK, Basnet RK, et al. (2011) MADMAX - Management and analysis database for multiple omics experiments. *J Integr Bioinform* 8: 160.
232. Heber S, Sick B (2006) Quality assessment of Affymetrix GeneChip data. *OMICS* 10: 358-368.
233. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118-127.
234. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, et al. (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* 7: 538.
235. Husson F, Lê S, Pagès J (2010) Exploratory multivariate analysis by example using R. Boca Raton, FL: Chapman & Hall/CRC Press. 240 p.
236. Escofier B, Pagès J (1994) Multiple factor analysis. *Comput Stat Data An* 18: 121-140.
237. Moraes RC, Blondet A, Birkenkamp-Demtroeder K, Tirard J, Orntoft TF, et al. (2003) Study of the alteration of gene expression in adipose tissue of diet-induced obese mice by microarray and reverse transcription-polymerase chain reaction analyses. *Endocrinology* 144: 4773-4782.
238. Panchal SK, Brown L (2011) Rodent models for metabolic syndrome research. *J Biomed Biotechnol* 2011: 351982.
239. Hariri N, Thibault L (2010) High-fat diet-induced obesity in animal models. *Nutr Res Rev* 23: 270-299.

240. Nock NL, Wang X, Thompson CL, Song Y, Baechle D, et al. (2009) Defining genetic determinants of the Metabolic Syndrome in the Framingham Heart Study using association and structural equation modeling methods. *BMC Proc* 3 Suppl 7: S50.
241. Wierzbicki M, Chabowski A, Zendzian-Piotrowska M, Gorski J (2009) Differential effects of in vivo PPAR alpha and gamma activation on fatty acid transport proteins expression and lipid content in rat liver. *Journal of physiology and pharmacology* 60: 99-106.
242. Motoki Y, Tamura H, Watanabe T, Suga T (1999) Wy-14,643, a peroxisome proliferator, inhibits compensative cell proliferation and hepatocyte growth factor mRNA expression in the rat liver. *Cancer letters* 135: 145-150.
243. Arnauld S, Fidaleo M, Clemencet MC, Chevillard G, Athias A, et al. (2009) Modulation of the hepatic fatty acid pool in peroxisomal 3-ketoacyl-CoA thiolase B-null mice exposed to the selective PPARalpha agonist Wy14,643. *Biochimie* 91: 1376-1386.

Summary

Several metabolic disorders including visceral obesity, insulin resistance, hypertension and dyslipidaemia, which increase the risk of cardiovascular diseases and diabetes are the main problems in the developed countries and are rising ones in the developing countries. These metabolic diseases are often associated with excess fat in the body. Nutritional systems biology of fat and fatty acids can enable the investigation of the relationship between genes and nutrients by integrating the organs and time specific data. One of the nuclear receptor super families, peroxisome proliferator activated receptors (PPARs), plays an important role in sensing nutrients and facilitating their effects on gene expression. PPAR α is one of them and it is an important transcription factor which is activated by free fatty acids and their derivatives. It is mainly involved in the regulation of lipid metabolism and storage as well as regulation of inflammation and immunity. Therefore it is highly interesting to identify the effect of fat via PPAR α by developing proper statistical tools and nutritional systems biological approaches.

Nutritional systems biology is a new biological research field where several biological levels are monitored by several ways. The aim of nutritional systems biology is to discover biological systems where the components work together and they are connected to one another within an organ and between organs. The components can be genes or set of genes or organs. It is essential to detect the proper transcription factor target genes by combining activation experiments performed in wild type and knockout mice. It is reported that most tests are based on pairwise comparisons in separate experiments and therefore adjusting the false discovery rate may interpret incorrectly because of a huge different in the effect sizes across experiments. Therefore, at first we aimed to develop an integrated statistical approach in **chapter 2**. We conclude that our integrated statistical approach successfully detect the transcription factor target genes with correcting for the multiple testing problem.

Analysis of gene expression data at the level of pathways is an important approach to unravel the biological function that is hidden in high throughput transcriptomics studies. We therefore developed a strategy to calculate pathway

activity level per arrays (**chapter 3**). Moreover, this data was used to study relationships between acute and long-term effects of PPAR α activation in intestine and liver. We found that PPAR α played a more important role in liver than in intestine, and that acutely induced pathways are the main drivers for regulation of pathways after long-term activation.

It is also relevant to uncover the evolution of gene expression and their function after acute PPAR α activation. Several studies have been performed to see the effect of treatment with the highly-specific PPAR α agonist WY14643 after relatively long time, but no study has been performed at earlier stages to detect the direct effects of PPAR α activation. Therefore, we conducted an experiment using rat hepatocytes cell cultured at 5 early time points (0, 1, 2, 3, and 4h) to identify the direct effect of WY14643 in **chapter 4**. We found that most of the acutely regulated genes were involved in lipid metabolism and they followed a quadratic pattern over time. We also found that transcription factors NR2F, CREB, ERF and RXR were closely bound with PERO in the genes involved in lipid metabolism process and these TFs may be bound with PPAR α . The results also revealed that the gene interaction networks were expanded over time. Taken together the time course study provides different sets of similar behaving genes with their potential common transcription factors with PPAR α .

It is well known that excess dietary fat is stored in adipose tissue, but it has been suggested that this storage capacity is limited. Subsequently, adipose tissue failure or dysfunction may drive progression of hepatic steatosis toward non-alcoholic steatohepatitis (NAFLD). However, knowledge on the functional link between adipose tissue dysfunction and NAFLD is currently limited. Therefore, in **chapter 5** we aimed to find out the relationships between gene expression in liver and white adipose tissue (WAT), weight status as well as different plasma factors in terms of the time and dose dependent effects of dietary fat during the development of obesity in C57BL/6J mice by developing a partial least squares-path model (PLSPM). We found that the exchange of carbohydrate for fat in the diet induces major changes in gene expression in both liver and WAT. Our analysis identified a set of potential signaling proteins secreted from WAT that may induce metabolic changes in liver, thereby contributing to the pathogenesis of obesity.

Taken together, our studies have further detailed the role of dietary fat on the transcriptome in small intestine, liver and white adipose tissue. To identify the detailed effects of dietary fat at the level of a whole organism, additional studies are required that integrate transcriptomics, proteomics, metabolomics datasets and phenotypes over time. The works of this thesis provide new approaches to integrate multiple datasets related to lipid homeostasis.

Samenvatting

Verschiedende metabole afwijkingen, zoals overgewicht, insulineon gevoeligheid, hoge bloeddruk en dyslipidemie, verhogen het risico op hart- en vaatziekten en diabetes; komen steeds vaker voor in zowel de westerse samenleving alsook in ontwikkelingslanden. Genoemde metabole ziekten zijn vaak geassocieerd met de aanwezigheid van overtollig vet in het lichaam. Systeembioogie is de wetenschap die biologische systemen als geheel bestudeerd en heeft als primair doel het kwantitatief achterhalen hoe moleculen, cellen en organen samenwerken om biologische processen te laten verlopen. Door de vooruitgang in de 'omics' disciplines komen steeds meer gegevens beschikbaar die geïntegreerd zullen worden in voorspellende modellen. In dit proefschrift worden systeembioologische benaderingen beschreven die als uiteindelijk doel hebben om de tijdsafhankelijke relatie tussen voeding, activiteit van genen in weefsels en fysiologische parameters te integreren. Dit heet nutritionele systeembioogie. In dit proefschrift hebben we ons gericht op de effecten van vet en vetzuren uit de voeding op de dunne darm en lever met speciale aandacht voor de rol van de transcriptiefactor PPAR α hierin. Hiertoe zijn voedingsstudies uitgevoerd met gewone muizen (wild type muizen) en muizen die geen functioneel PPAR α hebben (PPAR α knockout muizen), waarna de activiteit van alle 20.000 genen in de darm en de lever bepaald is met behulp van microarrays. Deze gegevens zijn vervolgens geïntegreerd met resultaten van andere kwantitatieve metingen door gebruik te maken van voorspellende multivariate statistische modellen.

Het is essentieel om op de juiste manier relevante genen te identificeren in grote datasets. Daartoe is onder andere de waarde van de zgn. false discovery rate (FDR) van belang. Deze FDR geeft aan wat de kans is dat er een vals-positief resultaat is opgepikt tijdens het analyseren van genexpressie data. Een veelgebruikte benadering om doelgenen van transcriptiefactoren te vinden is het vergelijken van genexpressieprofielen tussen wild type en knockout muizen voor en na activatie. Echter, de meeste statistische benaderingen die hiervoor gebruikt worden zijn gebaseerd op het combineren van resultaten uit paarsgewijze vergelijkingen van separate experimenten. Omdat de grootte van het effect van een behandeling per definitie verschilt tussen wild type en knockout muizen, is

het gebruik van de FDR als selectie criterium voor paarsgewijze vergelijkingen niet geschikt. In **hoofdstuk 2** stelden we een geïntegreerde benadering voor om relevante genen op biologisch alsook statistisch juiste manier te identificeren. Met behulp van gegevens van wild type en PPAR α knockout muizen lieten we vervolgens zien dat onze geïntegreerde statistische benadering met succes juiste PPAR α doelgenen detecteert waarbij ook wordt gecorrigeerd voor de kans op het includeren van vals positieven.

Analyse van genexpressie data op het niveau van metabole en signaaltransductie routes is een veelgebruikte en gevoelige manier om functionele informatie die in genexpressieprofielen verborgen is te ontrafelen. We hebben daarom een strategie ontwikkeld om per array de activiteit van deze metabole en signaaltransductie routes te kunnen berekenen (**hoofdstuk 3**). Deze resultaten zijn verder gebruikt om relaties tussen acute (6 uur) en lange termijn (5 dagen) effecten van PPAR α -stimulatie in de darmen en de lever te bestuderen. Wij vonden dat PPAR α een belangrijkere rol in de lever speelde dan in de darmen. Na toepassing van 'partial least squares- path modeling' (PLSPM) vonden we dat acuut geactiveerde metabole en signaaltransductie routes de belangrijkste aanstuurders waren voor de activiteit van deze routes na langdurige stimulatie.

Om een zo compleet mogelijk inzicht te krijgen in de rol van PPAR α is het ook van belang om de ontwikkeling van genexpressie en hun corresponderende functie te ontdekken na een acute stimulatie van PPAR α . Hoewel er diverse studies zijn gepubliceerd die het effect van behandeling met de specifieke PPAR α agonist WY14643 na relatief lange tijd bestuderen, zijn er tot nu toe geen studies uitgevoerd die de directe, acute effecten van PPAR α stimulatie analyseerden. Daarom hebben we in een experiment ratten hepatocyten gekweekt en hebben we na 5 vroege tijdstippen (0, 1, 2, 3 en 4 uur) na stimulatie met WY14643 genexpressieprofielen verzameld om de directe PPAR α stimulatie te bestuderen. Deze studie staat beschreven in **hoofdstuk 4**. Wij vonden dat het merendeel van de acuut geïnduceerde genen betrokken waren bij het metabolisme van vetten, en dat de expressieprofielen van deze genen een kwadratische patroon volgden in de tijd. Ook identificeerden we diverse transcriptiefactoren, zoals NRF2, CREB, ERF en RXR, die waarschijnlijk een rol speelden bij de acute inductie van genen

betrokken bij het vetmetabolisme. De resultaten lieten ook zien dat de gen-gen interactienetwerken uitgebreider werden in verloop van de tijd.

Het is bekend dat overtollig vet uit de voeding wordt opgeslagen in het vetweefsel, maar het is ook gesuggereerd dat deze opslagcapaciteit beperkt is. Vervolgens kan het niet goed functioneren of zelfs falen van het vetweefsel leiden tot progressie van vervetting van de lever en tot niet-alcoholische steatohepatitis (NAFLD). Er is echter weinig bekend over de functionele verbinding tussen vetweefsel dysfunctie en NAFLD. In **hoofdstuk 5** hebben we de relatie gemodelleerd tussen genexpressie in wit vetweefsel en de lever, gewicht status alsmede verschillende plasma factoren op het vlak van de tijd- en de dosisafhankelijke effecten van vet in de voeding tijdens de ontwikkeling van obesitas in muizen. Hiertoe hebben weer PLSPM toegepast. We vonden dat de het wisselen van koolhydraten voor vet in het dieet resulteert in grote veranderingen in genexpressie in zowel de lever als vetweefsel. Onze analyse identificeerde een lijst van mogelijke signaaleiwitten die worden afgescheiden door het vetweefsel en die kunnen leiden tot veranderingen in de stofwisseling in de lever, en op die manier bijdragen tot de negatieve effecten van overgewicht op de lever.

Samengevat hebben de studies beschreven in dit proefschrift de rol van vet in de voeding op de genexpressieprofielen in de dunne darm, lever en wit vetweefsel verder ontrafeld. Om de gedetailleerde effecten van vet in de voeding op het niveau van een volledig organisme te identificeren, zijn aanvullende studies nodig waarbij transcriptomics-, proteomics-, en metabolomics datasets en fenotypen geïntegreerd worden in de tijd. De inhoud van dit proefschrift leidt tot nieuwe benaderingen om meerdere datasets die verband houden met lipide homeostase te kunnen integreren.

Acknowledgements

It is really hard to explain how much I appreciate the educational and official systems and supports in the Netherlands that made my life easier to stay in the small city of Wageningen, “The City of Life Sciences”, and to gather knowledge. Thanks to WUR and all the people working here to keep its rank higher in the world. Of course, I would like to mention some specific persons with whom I was involved directly.

At first I would like to thank my supervisor Prof. Michael Müller. Prof. Michael, I am really pleased to be one of your PhD students, I find you a great man with brilliant thoughts and ideas. Especially your inspiration helped me to look at various issues from different angles to increase the quality of my work. I always got appreciation from you that helped me to finish the thesis on time. Your socialization in Twitter and Facebook about science is really fantastic; sometimes I easily found my research related articles from your tweets. Thank you very much for all the support and help.


My supervisor Dr. Guido Hooiveld, you are a fantastic and dynamic man, many, many thanks to you. Without your efforts and guidance it should not have been possible to finish my thesis on time. You always kept the things up to date in research, that really is appreciable, I admire your vast knowledge in many areas that inspired me really. You never minded disturbing you whenever I was at across the corridor regarding the way forward in the process of analyzing the dataset or other issues. Your appreciation and inspiration brought me in this position today. I'll never forget your efforts during my PhD period.

Prof. Sander Kersten, your vast knowledge about PPARs led to your name as Mr. PPAR. You are a great man and teacher. I was lucky that I had an opportunity to follow your class, from which I learned a lot. Thank you very much for your help and support during my PhD period. I also enjoyed to play soccer with you; you are a very good soccer player as well :). I will miss it!!

I would like to thank my paranimfs Mieke Poland and Mark Boekschoten for accepting my request on the big day in my educational life. Mieke, you are a very nice colleague and roommate. We have discussed several issues during the last

four years and I enjoyed a lot with talking and laughing, sometimes high thought philosophical things as well. You helped me a lot with translating Dutch letters and filling the tax related forms. Thank you very much for your help and support. Mark, you are a fantastic man and very helpful. I never forget your help and support during my PhD period, especially in the last chapter of my thesis (PPS3 study). Your contribution was very significant and you are a 2nd author of that chapter. Thanks a lot for your help and congratulation for being father of your second child.

Philip, you are a very good bioinformatician, nice roommate and you helped me many times to solve problems I had with R. Thank you very much for your support, especially for the R code you have written that automated our developed approach to calculate pathway scores in one of my chapters, of which you are a co-author as well. I am grateful for that, it has made my life easy. David, my former colleagues and roommate, you are a nice person and friend, we had pleasant talks on several issues. I got some biological knowledge from you as well, I miss you a lot. Thanks to Robert, former colleagues and roommates. Ya, Zheng, Fitim and Pierluigi, you are nice people, we discussed many things many times in the room and in the coffee corner. I liked to talk with you, thank you very much. Ya, hope you will finish your task with success. Michiel, you are a very nice man and helpful, thank you very much for your help and support at the end of my PhD period, especially translating the summary in Dutch as well as regarding the formatting and printing of my thesis and various other official processes.

Diederik, you are a fantastic man, good friend, good soccer player and organizer, Thank you, man! Often we discussed the so called “flower” :))) as fun, I will remember this. I hope you will finish your thesis soon, good luck!! Noortje, Danielle and Milene thank you very much, we had lovely discussions on many occasions. I wish you all the best with your projects. Rinke and Frits, both of you are nice guys, thank you very much for your support during my PhD. Frits, advance congratulation for being father of your first child!!

Lydia, Wilma, Mechteld and Carolien thank you very much for your support and help. At the end of 2008, when Mechteld was on leave then I spent the first week of my project at her desk- so you were my first roommates here. All of you are cordial and helpful. Jenny and Shohreh, thanks for the pleasant talks we had at

the coffee corner, you are nice ladies. Thank you very much for your support during my PhD, especially for all the work you did to generate the wet lab and microarray data. Without this I basically would have been jobless. Jessica and Sheril, thank you very much, I still enjoy the nice talk we had at the time of Prof. Sander's dinner party. My former colleagues Meike, Maryam, Anand, Susan, Els, Anastasia, Laeticia, Linda, thank you all! Special thanks goes to Meike and Maryam because I was able to use your data most of the time to develop my methods/approaches. Anand you are a very nice man and helpful, I will never forget the support I got the first time I came to Wageningen. You introduced me to many official and practical things I then had to go through. Thank you very much again. I would also like to thank Katja, Nicole, Karin, Jocelij, Nikkie, Jvalini, Fenni, and Inge. Katja, I liked it when we discussed biological topics, and I learned a lot from that. Thank you very much. Good luck with your project!! Nicole, thank you very much for your help. Dina, thank you very much for your help and documents about the course Advanced Metabolic Aspects of Nutrition from Prof. Sander, your materials helped me a lot to better understand the metabolism of macronutrients. Thanks also goes to Evert and Annelies from Human and Animal Physiology, WU, who helped me to finalize the manuscript of PPS3 study.

Prof. Edith, Olga, Monica, and Anouk - thank you very much for your wishes to finish my thesis. I appreciate the nice time we had at each stat advice meeting. I would like to thank Prof. Frans Kok and Prof. Renger Witkamp for their wishes during my PhD. Prof. Kok, you are a great man and cheerful, I enjoyed your accompany in Copenhagen during the PhD tour. Rina, Por, and Kevin, thank you as well. Rina and Por, I will remember your accompany in Denmark, Sweden and Finland at the PhD tour.

I would like to thank Gea, Didi, Maria, Lidwien, Cornelia, Lous, Eric, Adrienne, Yvonne, Vesna, and Jan for their support in many official tasks; without your support the completion of my PhD thesis would not have been possible. You are the key to smooth organizational and financial things. Especially Lidwien, you were very helpful and expert in human resource side, I really miss your absence. It's very mournful that you left us forever so early.

I would like to thank Prof. Jaap Molenaar and Apri from Biometris group, and Maria from Systems and Synthetic biology group for the useful discussions we had.

I would also like to thank the thesis committee members: Prof. dr. C.J.F. ter Braak, Prof. dr. ir. V.A.P. Martins dos Santos, Prof. dr. C. Evelo, and Dr. B. van Ommen for their time to evaluate my thesis and joining in my PhD defense day.

I have to commend the persistent patience, perseverance and support of my parents, family members, relatives, beloved daughters (Nabila and Sabiha) and especially my wife Mira during my PhD period. Thank you very much. At first I came here alone to do my PhD, half way my wife and my daughter (Nabila) joined me, and now I return to Bangladesh with a second daughter and PhD certificate :)). আলহামদুলিল্লাহ ।

I may have missed some people to mention their name here, thank you all.

“সবাইকে ধন্যবাদ”



08 October 2012, Wageningen

Curriculum Vitae

Mohammad Ohid Ullah was born on June 01, 1979 in Noakhali, Bangladesh. He completed his Bachelor of Science in Statistics in 2002 at Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh. He also completed his MSc in Statistics at SUST in 2003 and then in 2004 he joined as a faculty in the Statistics department. In 2006 he went to Belgium for higher study with a VLIR scholarship. At Hasselt University, Belgium he obtained an MSc in Applied Statistics and an MSc in Statistics Biostatistics in 2007 and 2008, respectively. In October 2008 he started his PhD study on the project “Nutritional Systems Biology of Fat” which was one of the IP/OP Systems Biology projects under the supervision of Dr. Guido Hooiveld and Prof. Michael Müller in the Nutrition, Metabolism and Genomics group, Division of Human Nutrition, Wageningen University. From 14 October 2012 he will be rejoining at SUST, Bangladesh as a faculty to do teaching and research. His personal Email ID is ohidullah@gmail.com .

List of Publications

Mohammad Ohid Ullah, Michael Müller, and Guido JEJ Hooiveld (2012) An integrated statistical approach to compare transcriptomics data across experiments: a case study on the identification of candidate target genes of the transcription factor PPAR α . *Bioinformatics and Biology Insights* 6:145–154.

Mohammad Ohid Ullah, Mark V Boekschoten, Evert van Schothorst, Michael Müller, and Guido JEJ Hooiveld (2012) Integrative multivariate modeling of the relationships between gene expression in white adipose tissue and liver during the development of obesity in mice (Submitted).

Mohammad Ohid Ullah, Meike Bünger, Sander Kersten, Philip J de Groot, Michael Müller, and Guido JEJ Hooiveld (2012) Integrative analysis and modeling of PPAR α function in murine liver and small intestine (Submitted).

Mohammad Ohid Ullah, Shohreh Keshtkar, Guido JEJ Hooiveld, and Michael Müller. Characterization and modeling of acute effects of PPAR α activation in rat liver cells (In preparation).

David L.M. van der Meer, Maryam Rakhshandehroo, **Mohammad O. Ullah**, Philip J. de Groot, Sacco C. de Vries, Michael Muller, Sander kersten (2010) Comparative microarray analysis of PPAR α induced gene expression in the human hepatoma cell line HepG2 and primary human hepatocytes, *In: Maryam Rakhshandehroo (2010) PPAR α : Master regulator of lipid metabolism in Mouse and human*, chapter 6, pp 159-180, *Thesis Wageningen University*, Netherlands. ISBN: 978-90-8585-771-6.

Mohammad Ohid Ullah (2010) Some statistical advices for the non-statisticians: Common examples of useful statistical tools by SPSS for the non-statisticians. *VDM-Verlag*, Saarbrücken, Germany. ISBN: 978-3-639-28013-5.

K.K. Islam, M. Hoogstra,, **M. Ohid Ullah**, N. Sato (2012) Economic contribution of participatory agroforestry program to poverty alleviation: a case from Sal forests, Bangladesh. *Journal of Forestry Research* 23(2):323-332.

M. Ohid Ullah, M. Jamal Uddin, Dr. M. Rahman, M. Nazrul Islam, and M. Taj. Uddin, (2009) A Study to detect the seasonal effect of chickenpox in Bangladesh. *Romanian Statistical Review*, 12:61-66.

M. Jamal Uddin, M. Zakir Hossain, **M. Ohid Ullah** (2009) Child Mortality in a Developing Country: A Statistical Analysis. *Journal of Applied Quantitative Methods*, 4(3):270-283.

M. Ohid Ullah, and M. Jamal Uddin (2008) A Study on Evolution of the pH Level over Time in Patients Suffering from Reflux. *Shiraz E Medical Journal*, 9(3):141-148.

M. Nazrul Islam, **M. Ohid Ullah**, and M. Taj Uddin (2006) A Study on Health Status of Urban Pregnant women of Bangladesh with Respect to Body Mass Index and Weight Gain. *J. Med. Sci.* 6(2):249-252.

Mohammad Taj Uddin, M.T. Islam and **M. Ohid Ullah** (2006) A Study of the Quality of Nurses of Government Hospitals in Bangladesh. *Proc. Pakistan Acad. Sci.* 43(2):121-129.

M. Nazrul Islam, and **M. Ohid Ullah** (2005) Knowledge and Attitude of Urban Pregnant Women of Bangladesh towards Nutrition, Health Care Practice and Delivery Place. *J. Med. Sci.* 5(2):116-119.

Overview of completed training activities

Discipline specific activities

Multivariate-Data Modeling for Systems Biology, Norwegian University of Life Sciences, January, 2009
Algorithms for Biological Networks, Netherlands Bioinformatics Centre, January, 2010
Modelling of Biological Processes, Netherlands Consortium for Systems Biology, 2010/2011
Bimolecular Principles of the Cells, Netherlands Consortium for Systems Biology, 2010/2011.
Advanced visualization, integration and biological interpretation of ~omics data, VLAG-WIAS, WUR, 2011
From Data to Models in Biological Systems. SystemsX.ch/SIB Summer School 2011, August 14-19, Kandersteg, Switzerland
SMBES2011 (Statistical modeling for biological and environmental systems) summer school, September 12-16, Warwick @ Venice, Italy (oral)
CMSB Symposium, Rotterdam, 1 day, October, 2009
NISB Symposium, Noordwijkerhout, Netherlands, 2 days, October, 2009
NCSB Kick-Off Symposium, Noordwijkerhout, Netherlands, 1day, October 2009
NCSB Symposium, Soesterberg, Netherlands, 2days, October, 2010 (poster)
NCSB Symposium, Soesterberg, Netherlands, 2 days, October, 2011
Symposium: Systems biology of gene transcription, Wageningen, 2009

General courses

VLAG PhD week 2009 (Bilthoven, the Netherlands)
Mini-symposium "How to write down a world-class article", Wageningen University library, October, 2010
Symposium: Statistical issues in nutritional research, Wageningen, 31 May 2012
Lectures : "How to manage your hard earned information efficiently", VLAG PhD council, WUR, 23-11-10
Lectures: "Statistics strategies" VLAG PhD council, WUR, 14-3-11
NuGo week 2011, Wageningen University, the Netherlands
Wageningen Nutritional science forum, 2009 (Arnhem , Netherlands)
WGS/VLAG course: "Philosophy and Ethics of Food Science & Technology", 2012, Wageningen University, The Netherlands

Optional activities

Preparing PhD research proposal
NMG scientific meeting, WUR (every week)
Stat-advice meeting as an advisor, HNE (every month)
PhD Tour 2009, Denmark, Sweden and Finland, Division of Human Nutrition. 18 Oct-31 Oct, 2009, (oral and poster)
Research retreat 2011, HNE, WUR

The research described in this thesis was financially supported by the Wageningen University program for Systems Biology (IP/OP SysBio).

Printing of this thesis was financially supported by the Division of Human Nutrition, Wageningen University.

Cover: Designed by Mohammad Ohid Ullah

Printing: Ipskamp Drukkers, Enschede, the Netherlands

Copyright © Mohammad Ohid Ullah, 2012.