LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz-Zentrum München

Bachelorarbeit
in Bioinformatik

Analysis Of The Mouse 200 Dataset By Integrating Transcriptomics And Metabolomics Data For Multilevel Ontology Analysis

Benedikt Rauscher

Aufgabensteller:	Prof. Dr. Fabian Theis
Betreuer:	Steffen Sass
Abgabedatum:	15. August 2013

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15. August 2013 _____

Benedikt Rauscher

Acknowledgment

Contents

List of Figures

List of Tables

1 Introduction

Metabolomics is a new field of biology which tries to comprehensively study organisms' metabolomes and their corresponding metabolic reactions. The field has been developing rapidly over the past couple of years. Recent discoveries in the field of metabolomics have led to a lot of new insights on epidemiology and molecular cell biology [22]. which is not surprising as the metabolome is thought of as the best indicator of an organism's phenotype [6]. This is a consequence of the fact, that metabolites function as immediate signatures of biochemical activity and thus are easier to link to a phenotype [27].

There is a number of advantages of analyzing metabolomics data over other "omics" data, such as the relatively small amount of metabolites in an organism compared to, for example, its number of gene products. Additionally, metabolites are biochemical phenotypes which result from genetic, transcriptomic and proteomic variability and thus represent biological status in an integrated way. Finally metabolomics can be used to identify possible toxicological effects of drug treatments [3].

The aim of metabolomics is to identify and quantify all metabolites in a tissue, cell or biofluid under different conditions [24]. This information can then be used to create a metabolic profile for a cell, providing a quick overview of its physiology. This information can then be used to contribute to the understanding of biological functions *in vivo* [16]. All in all the possibility of determining, how mechanistic biochemistry relates to cellular phenotype, is created [27].

Because of the rapid development of biological analysis techniques such as nuclear magnetic resonance (NMR) and mass spectrometry (MS) it has become possible to measure and quantify a large number of metabolites in blood and urine samples on a high throughput scale [21].

With the huge amount of metabolomics data provided through the rapid development of high-throughput experimental methods, it is a very important task for computational biology to make some sense of this data. Biochemistry shows that metabolite concentrations have a direct relationship with the activity of their respective enzymes. A popular way to determine enzyme activity is to measure the amount of mRNA in a cell, that translates to the enzyme. MRNAs, along with other types of RNAs are part of the transcriptome, which consists of the entirety of RNAs in a cell. The aim of transcriptomics approaches is to measure the transcriptome of a cell at a given point in time and under specific environmental circumstances [38].

All in all this leads to the conclusion, that integrating metabolomics data and transcriptomics data appears to be very promising in order to achieve biological insights and knowledge.

In this work, we present a novel approach to integrate our metabolomics and transcriptomics data called MONA metabolite, which is a gene set enrichment tool that can deal with transcriptomics and metabolomics data simultaneously in order to predict changes in metabolic pathways. We apply our methods to a data-set, which was very recently generated for a new study, called Mouse 200 and which consists of mRNA and metabolite measurements. Using our program, we try to gain insights on how diabetes and different medical treatments for diabetes affect the metabolism of the mouse.

2 Background

Metabolites represent small molecules. They are chemically transformed throughout the process of metabolism [27]. Thus,

> "metabolites are the end products of cellular regulatory processes and their levels can be regarded as the ultimate response of biological systems to genetic or environmental changes" [14].

Up to now, a number of up to 200 000 such metabolites are estimated to occur in the plant kingdom, and there are probability a lot more across multiple kingdoms [14]. The entirety of all metabolites synthesized by an organism represents its metabolome [26].

2.1 High-throughput metabolite measurment techniques

There are two kinds of approaches to experimentally measure metabolite concentrations. Which of these approaches is used, is dependent on the scientists personal motivation.

2.1.1 Targeted metabolomics

In this procedure the concentrations of a list of metabolites of distinct interest in a sample is measured. This is usually the case when experiment aims to shed light on certain metabolic pathways of interest [13]. Targeted metabolomics studies are best performed using nuclear magnetic resonance (NMR) [27].

2.1.2 Untargeted metabolomics

In contrast to targeted metabolomics, the focus in untargeted metabolomics is to create a metabolic profile of all metabolites in a sample. The methods which can fulfill this task the best are liquid chromatography and mass spectrometry (LC/MS) methods. The data files produced by untargeted metabolomics approaches are very complex and have file sizes of several gigabytes per sample [27].

As most metabolomics studies investigate on metabolomics levels depending on a given phenotypic state, e.g. a certain disease [24]. Measuring metabolomics levels under different conditions quickly results in a big, confusing bulk of data. In order to make sense out of such a large amount of available metabolomics data, it becomes more and more important to develop tools in order to automatically analyze this data using statistical methods, as it becomes impossible to gain any information out of the data by just simply looking at it.

2.2 Transcriptomics

The transcriptome is the term that describes the entirety of transcripts contained in one cell and how much of these transcripts are available at a given point in time or during a specific environmental condition. This makes the understanding of the transcriptome a key task in order to understand the genome and its functional elements as well as the development of diseases [38]. To do this, it is necessary to catalog all kinds of transcripts, including mRNAs, non-coding RNAs and small nuclear RNAs. We present a popular method which allows the measurements of these RNAs and thus provide the possibility to generate transcriptomics data.

Microarrays A microarray analysis always starts with a set of oligonucleotide probes. Each probe is complementary to a certain genomic DNA. The probes are designed according to known sequences. Once these probes have been constructed they are immobilized on a solid substrate and thus represent the microarray. After that, transcripts of interest, marked with fluorescent dyes, can hybridize to their targets and their expression can be measured using light intensity [25].

2.3 Integrating metabolomics and transcriptomics data

Changes in the environment of an organism results in alterations of this organism's metabolite concentration levels [28]. As transformation reactions between metabolites are catalyzed by specific metabolic enzymes, there is a strong correlation between the activity of those enzymes and the exchange rates of their corresponding metabolites [23]. From this follows, that altered metabolite abundances result in differential expression of their corresponding enzymatic genes in order to reestablish biological stability meaning the organism's adaption to the new environmental conditions [28]. Certain patterns in metabolite profiles along with expression rates of their corresponding enzymes can allow inspection off the organism's strategies of dealing with altered environmental conditions [28].

2.4 Recent attempts to integrate metabolomic and genomic data

Reconciling metabolomics and transcriptomics data is an important biological task for many reasons. As a consequence, a number of studies have been performed where metabolomics and transcriptomics have been examined in parallel. In the following we present three such approaches as examples.

Prediction of pathway co-memberships between metabolites and genes One method which tries to link metabolomics to transcriptomics data was presented by Henning Redestig and Ivan G. Costa in 2011 [28]. This approach tries to predict metabolites and genes which belong to the same metabolic pathway based on correlated response to applied stress. This method uses Pearson correlations and Hidden Markov Models (HMM) to capture the temporal abundance of each metabolite in a pathway and after that evaluates for all gene expression time courses the likelihood with each HMM where a high likelihood alludes to a correlation between gene and metabolite.

The O2PLS multivariate regression method Another approach in order to combine "omics" types of data is using the O2-PLS method [35] [36]. which is a further developed version of the orthogonal projections to latent structures (O-PLS) method. This method can be used to detect systemic variation overlaps across multiple platforms, and dissociate them from the systemic variation belonging to only one of the examined platforms. This theory can be applied to find overlaps between metabolomics and transcriptomics data. The O2PLS modeling process requires the scientists to estimate the complexity of the different platforms [9].

Network analysis of data from population studies The last attempt to integrate metabolomics and transcriptomics data we are going to briefly introduce, is the attempt to analyze networks built from population studies data. Here, first a genetic

co-expression network was built showing co-expressed gene modules. This information was then used to determine correlations between genetic expression profiles and metabolite distributions [18].

2.5 Gene set enrichment

Gene set enrichment is a method to interpret genome-wide expression data provided by high throughput experiments. In most gene expression experiments a big number of genes is measured in samples that belong to one of two different classes. To perform gene set enrichment, first a subset of genes is defined, that is differentially expressed between the two classes. After this a gene set enrichment tool determines for different sets of genes, for example genes that belong to a certain metabolic pathway or share the same GO-Term, whether these genes are statistically over-represented within the list of genes differentially expressed in the experiment and thus whether the gene set is correlated with the phenotypic class distinction [33]. A GO-Term is a term created by the Gene Ontology project. It is part of a fixed set of terms and it can represent a molecular function, a biological process or a cellular component [11].

Until now, gene set enrichment could only be performed for one kind of data on a single level, for example for gene expression rates or for metabolic concentrations. Abusing the correlations between different levels of data in order to improve the quality of gene set enrichment predictions, however, is a task which up to now, cannot be fulfilled to a satisfying degree.

2.6 MONA

That is the reason a new tool called MONA (Multilevel ONtology Analysis) has recently been developed. MONA is the first gene set enrichment tool that can take data from multiple "omics" levels into account, meaning that several input sets of different data types can be dealt with simultaneously, and at the same time is able to deal with term redundancies and multiple testing problems, resulting from the hierarchical arrangement of these terms. MONA models the gene set enrichment problem as a Bayesian network with multiple layers. The whole model consists of a base model, where a estimated probability is linked to every ontology term which are linked to a gene response layer. This base model extended with an additional model which can be chosen individually for a given type of input data. The model we are going to focus on here is the cooperative model where the gene response layer is is linked to an additional observation layer, which indicates for every measured gene whether it is differentially expressed within which of the input data sets. These observations of differential expression are linked to a final error rate layer which links to every observation the estimated probability whether it is a false positive or a false negative observation. After this Bayesian network model is built from the input, the probabilities for each ontology term to be active is estimated through inference of the network by probabilistic programming using the Infer.NET framework [29] [5].

2.6.1 Comparison with other gene set enrichment algorithms

In order to evaluate the performance of MONA for the integration of metabolomics and transcriptomics data, it was compared to two already existing popular methods for gene

set enrichment analysis. In the following the Methods MGSA and Fisher's exact test will be described briefly.

MGSA The Model-based Gene Set Analysis (MGSA) tool is a Bayesian modeling approach for gene set enrichment. It investigates all pathways simultaneously by modeling gene response as a function of the combination of active pathways. After that the Metropolis-Hasting algorithm is used to perform probabilistic inference and thus achieving probabilities for each pathway. Like MONA, MGSA takes category overlap into account and thus avoids the need for multiple testing correction [4].

Fisher's exact test Fisher's exact test is a statistical significance test. Contingency tables are used in order to make statements about the input data. It provides a p-value for each pathway, signaling whether the pathway is predicted to be altered or not. The pathways are analyzed one after another.

2.7 Mouse 200 - A systematic analysis of anti-diabetic drugs

In this thesis we investigate the biological effects of different anti-diabetic drugs. We do this by performing gene set enrichment for the data provided by the Mouse 200 project. This allows as to get insight on changes in metabolic pathways which can the be biologically interpreted. The Mouse 200 dataset can be divided into metabolite concentration and gene expression level measurements.

The first part of the data-set represents two tables. These tables include measurements for a total of 25,649 genes from different experimental mice. These measurements were generated from liver-cell samples using Illumina gene expression microarray chips. The observed mice can be divided into five categories. The first category represents healthy wild type mice. The second category consists of mice diseased with type II diabetes mellitus without any medical treatment. The mice in the third group also have type II diabetes but have been medically treated with a drug called metformin, which is a widely known and used drug for the treatment of type II diabetes. The fourth group is similar to the third group, but here, instead of metformin, a newly developed drug, which we will just call Drug x, was used to treat the sick mice. The fifth an last group includes diabetes mice who were treated with a combination of the previously mentioned medicaments. As mentioned before, this part of the data set contains two table, where one table includes measurements from samples, which were taken after four hours of treatment and the other table contains measurements from samples taken after two weeks of treatment.

The second part of the data set consists of metabolite concentration measurements. The numbers of metabolites measured ranges from about 20,000 to 30,000. The reason for the difference in metabolite numbers measured, is that a measurements were sometimes excluded due to quality control. These to tables just described exist in two versions. The data set comes as a number of tables containing the aforementioned metabolite concentration methods from gut-cell samples taken from the same kinds of mice as presented in the description of the first part of the data set. Again the measurements were made after four hours and after two weeks of the beginning of the treatment respectively.

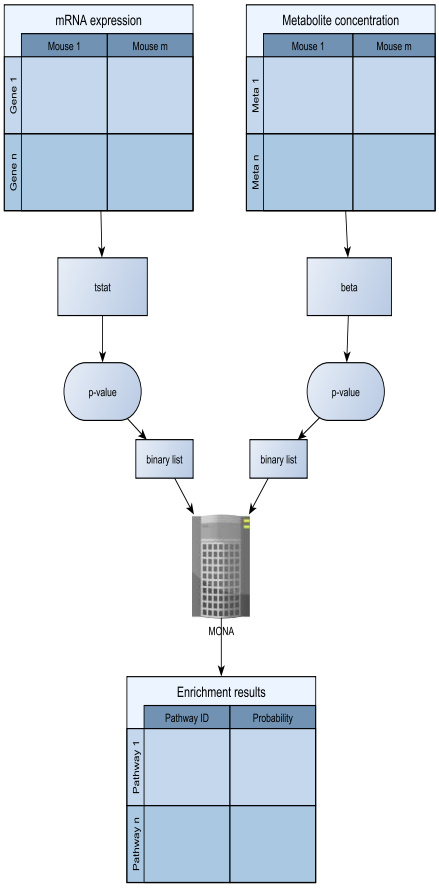

Figure 1: Workflow for the data MONA data pre-processing, starting with microarray data and ending with a valid and biologically meaningful input for the MONA core application.

3 Methods

For all the methods previously introduced in section 2.4, which try to integrate metabolomics and transcriptomics data, correlations between genetic expression and metabolite concentration patterns played an important role. This is also the case for our integration attempt we will present in this thesis. Our method will use these correlation between metabolite levels and gene expression in combination with gene set enrichment analysis, to gain information about which metabolic pathways might be perturbated as the consequence of the change in an organism's environment.

3.1 Integration of metabolomics data for MONA

As of now MONA allowed simultaneous input of data across multiple "omics" levels, however these data had to be of the same data type, e.g. Entrez IDs or gene symbols. In this work we will describe how we improved MONA such that it allows predictions based on an input of mRNA and Metabolon names, trying to abuse the correlation between gene expression and metabolite concentration patterns in order to link abnormalities in expression levels directly to corresponding metabolic pathways. This was accomplished by pre-processing the input in an appropriate way, such that the cooperative model of MONA could be used for predictions (see Fig.1).

3.2 Data collection

In order to perform gene set enrichment MONA accesses data mappings, e.g. genes and their corresponding GO-Terms, from its associated database. In order to be able to integrate metabolomics data for MONA's cooperative model two kinds of information had to be added to the database: 1) Associations between metabolites and their corresponding catalytic enzymes. 2) Associations between metabolite-enzyme-reactions and the metabolic pathways they belong to. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [20] [19] contains metabolic pathways in the form of networks, where metabolites are nodes which are connected through edges corresponding to metabolic enzymes. These pathways are available in an XML-like format. We used the KEGGGraph R package [40] available from Bioconductor [15] to parse the XML-files for all metabolic pathways known in the mouse and extracted the information required in order to perform gene set enrichment with MONA. The algorithm for extracting this information works the following way: For every pathway, first a list of corresponding reactions is determined. After this the algorithm identifies the corresponding substrates and products for each reactions and creates a table with unique reaction to metabolite relationships. The same happens for reaction to enzyme relationships. Finally these two tables are merged by the reaction name column leading to the desired compound to enzyme relations. These mapping steps are required as the KEGG XML files do not provide metabolite to enzyme relationships directly. In order to map KEGG compound IDs to Metabolon names which are used by MONA we used the data provided by the Reconstruction of The HUman Genome (Recon x) project [12] . MONA uses Metabolon names as they provide a stable set of name terms for each metabolite. This is necessary in order to avoid confusion based off different metabolite name notations.

There was a number of metabolites about which i could not derive any information regarding their association with an enzyme from the KEGG database. As it is not intended that these metabolites are ignored in MONA predictions we determined direct metabolite to pathway relationships from the Recon project data and transferred these relations for the mouse. This makes biological sense, as the close evolutionary relationship between human and mice indicates strong similarities between the metabolic pathways and metabolic reactions between human and mouse [30].

As we integrated metabolomics data for MONA not only for mouse but also for human input data, we also created the required mappings for human genes. For enzyme to pathway relationships we used the KEGG.db R package [10] and the Recon project data to create the required mappings.

3.3 MONA metabolite

In order to make the usage of MONA with metabolite data as user friendly an easy as possible, we added an extra metabolite model option to the surface of the web application. The input that is required for the prediction is very straight forward. There are four lists that have to be provided for the application. The first list contains a set of differentially expressed mRNAs in the form of Entrez IDs or gene symbols. The second list is a background list of all the mRNAs that have been measured in the experiment. The third list contains a set of metabolites with differential concentrations and the fourth and last list contains the background of all the metabolites measured. Metabolon names are used for metabolites to eliminate the risk of unintentional ignoring of some input because of different metabolite name notations in the input and the database.

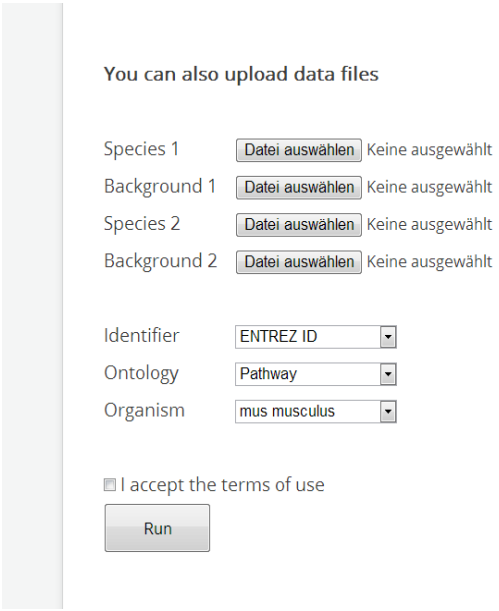

Figure 2: Input list upload

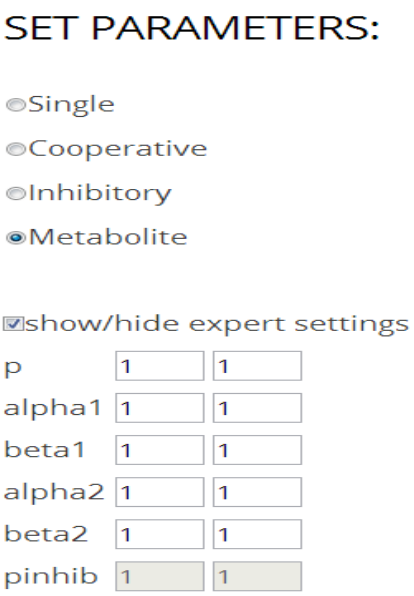

Figure 3: Model selection

3.3.1 Data pre-processing

After metabolite data has been submitted the data is pre-processed in a specific way such that it can be used for predictions by the MONA core application. First of all, a number of metabolites from the input differentially expressed metabolites is determined, which cannot be mapped to an enzyme. These metabolites are taken away from the set of differentially expressed metabolites and are added to the set of differentially expressed mRNAs as well as to their corresponding background. After that the differentially expressed metabolites are replaced by their corresponding enzymes. Finally those mRNAs in the mRNA background which could not be mapped to any metabolite, as well as the metabolites that could not be mapped to an mRNA, are added to a list of missing entities. The fact that the metabolites which couldn't be mapped to an enzyme are switched to the list of differentially expressed mRNAs is a necessary step as the cooperative model of MONA is only able to deal with missing values in the second input species. After the input has been transformed like this, it can be converted to the standard MONA input and predictions can be done using the cooperative model of the MONA core application. This input pre-processing is very quick and doesn't increase the computational time of the application in a noticeable way. This implies that in the model, mRNA and metabolites are regarded as independent noisy observations with a common enzymatic response, meaning that corresponding mRNA and metabolites are expected to follow similar patterns of differential expression and concentration, taking estimated false positive and false negative ratios into account. These mRNA and metabolite observations link to a hidden node in the model, which representing the common enzymatic response, marking it as active or not, depending on the observations. A result term in the model can only be active if at least one corresponding hidden node is active (see Figure 4).

3.4 performance assessment

To assess the performance of the integrated metabolite model for MONA we compared prediction results of gene set enrichment tools for a large number of randomly generated

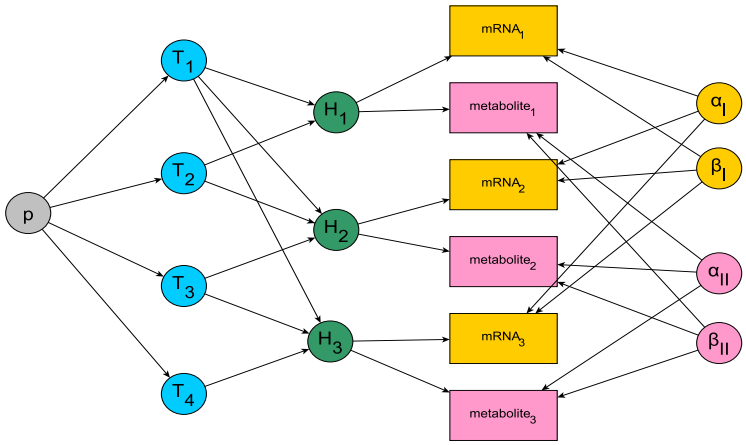

Figure 4: The MONA metabolite model. Each hidden enzymatic response is linked to a mRNA and a corresponding metabolite observation, taking false positive rate alpha and false negative rate beta into account. The hidden enzymatic response nodes are linked to metabolic pathway terms which are linked to a probability for them to be active.

test datasets. we compared three methods: 1) MONA with metabolite data integration, 2) gene set enrichment using Fisher's exact test [17] [7] and 3) gene set enrichment using Model-based Gene set Analysis (MGSA) with the corresponding R package used for implementation [4] . Both, Fisher's exact test and MGSA are only capable of using a single level input set for their predictions. The random test sets were created the following way: First we created a mRNA assignment matrix. This assignment matrix has a column for every KEGG-pathway in the MONA database and a row for every mouse gene that is associated with at least pathway. If a pathway is associated with a gene the respective field of the matrix contains 1 else 0. A second matrix of similar kind was built for all mouse metabolites in the database that are associated to at least one pathway. After that a sample of three to seven pathways were randomly selected from the total set of pathways and were marked as activated. Now all genes corresponding to the activated pathways were estimated from the gene assignment matrix. After adding false negatives and false positives to this set of genes, the set was used as set of differentially expressed genes serving as input for the prediction methods. For MONA an additional set of differentially expressed metabolites was created in a similar fashion, serving as second input data set for the MONA metabolite model. Finally all three prediction methods were run using the generated input. MONA was used on both, metabolite and gene level simultaneously, while MGSA anf Fisher's exact test were only used on gene level. The prediction results were evaluated. For the evaluation different ratios of false positives and false negatives for the test sets were used.

3.4.1 MONA metabolite

The MONA-predictions were made using the MONA metabolite model we developed in the exact same way as described in chapter "MONA metabolite".

3.4.2 MGSA

As already mentioned, the MGSA method was implemented using the "mgsa" R package provided by Bioconductor [4] [15] . The prediction method provided by the said package gets two lists as input. The first list is a boolean typed list of the same length as the mRNA background, indicating for each mRNA whether it is differentially expressed or not. The second list is an assignment containing information about which mRNA belongs to which pathway.

3.4.3 Fisher's exact test

The prediction method using Fisher's exact test [7] [17] was implemented using the "fisher.test" method which comes with R by default. For each metabolic pathway, the prediction method gets a two times two matrix as input, containing information about how many genes corresponding to this pathway are differentially expressed, about how many of them are not differentially expressed, and about how many of the genes not corresponding to the pathway are bzw. are not differentially expressed. Thus, a p-value which signals, whether a pathway is active or not is calculated for every pathway. The total number of these p-values represents the gene set enrichment analysis resullt using Fisher's exact test.

3.4.4 Performance assessment process

The whole process of assessing and comparing the performance of MONA metabolite is directly controlled by a .NET console application, calling the required R-scripts in the appropriate order. The application can be provided with the number of test runs supposed to be made and false negative and false positive rates to be used for the test set generation. Then, for each run, first the test data sets are generated. After that the predictions for each method are made. The prediction results are stored as temporary data files on the hard disk. After the last run is finished, all the predictions are evaluated and compared by a final R-script, using the ROCR R package [31] to plot results.

3.5 M 200 extraction of differentially expressed genes

In Order to use MONA metabolite to perform gene set enrichment analysis for the M 200 dataset, we had to extract lists of differentially expressed genes between pairs of two classes of mice.

The following classes were compared:

- Wild type with type II diabetes untreated
- Diabetes untreated with diabetes treated with metformin
- Diabetes untreated with diabetes treated with drug x
- Diabetes untreated with diabetes treated with a comibnation of metformin and drug x

3.5.1 Determination of differential expressed genes and metabolites

The data came in form of Microsoft Excel spreadsheet files. In order to use MONA metabolite for predictions, the data was pre-processed as follows. First of all the spreadsheet files were read into R using the "read.xls" function from the gdata R package. [39] . The data contains p-values describing the differential expression the measured mRNAs. These p-values were determined by a linear modeling approach implemented in the R package "limma" [32]. Multiple testing correction was performed for the p-values. A p-value of 0.05 was set as the threshold for a gene to be regarded as differentially expressed.

Differential metabolite concentrations were determined using a x test approach. Afer the that, the determined differentially expressed metabolites had to be filtered such, that only metabolites were taken into consideration, for which a Metabolon name mapping exists, as this is obligatory for MONA metabolite to handle them as input. The resulting sets of differentially expressed genes and metabolites, along with their respective backgrounds for each comparison of two of the introduced mouse groups were used as input for gene set enrichment analysis with MONA metabolite.

3.6 Using MONA metabolite to analyze the generated differential expression data from the M 200 dataet

The enrichment analysis was performed, using the MONA web application with the MONA metabolite model. The lists of differentially expressed genes and metabolite were uploaded manually along with their respective genetic and metabolic backgrounds.

For the comparison of wild type and untreated diabetic mice the priors were altered. As the number of differential genes and metabolites between these classes is very large. The prior probabilities for a pathway to be inactive which MONA uses, follow a beta distribution. By default, no assumption is made about how many pathways are active. In order to only get the most significant hits for the comparison of wild type and untreated diabetic mice, the parameters for the beta distribution of pInit were set to $\alpha = 1000, \beta = 1$ (Figures 5 and 6). For all the other enrichments, the standard settings of MONA were used.

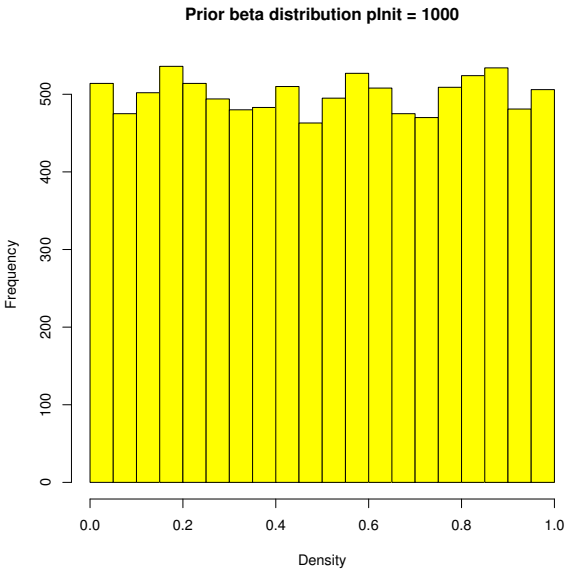

Figure 5: Default

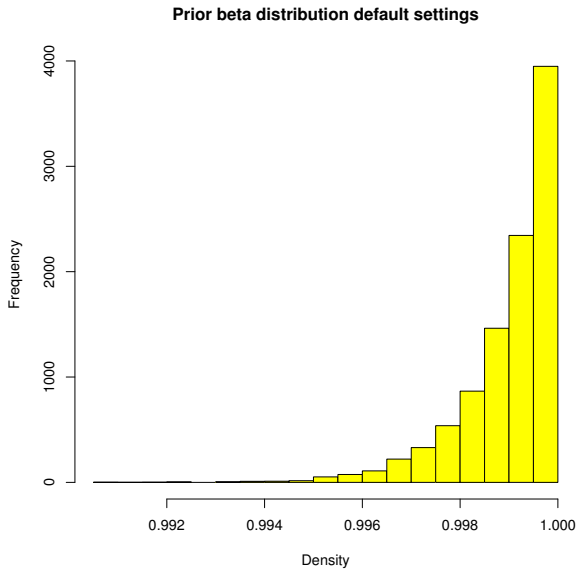

Figure 6: pInit = 1000

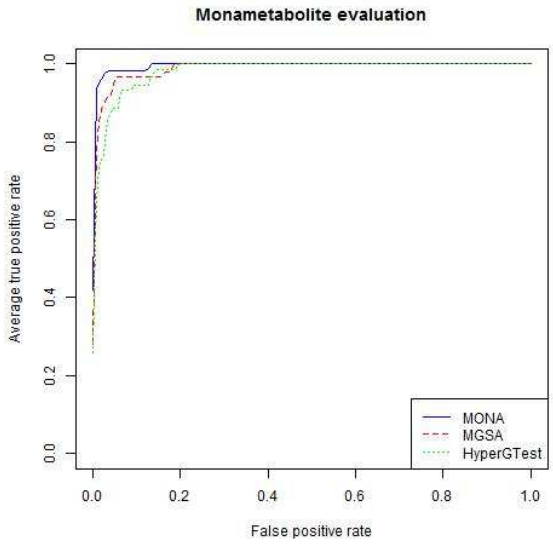

Figure 7: ROC curve for the performance comparison of MONA metabolite with MGSA and Fisher's exact test for false positive rate $\alpha = 0.25$ and false negative rate $\beta = 0.35$

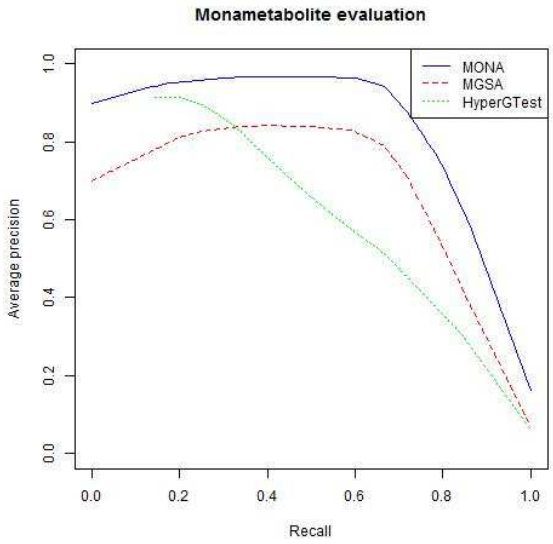

Figure 8: Precision-recall-curve for the performance comparison of MONA metabolite with MGSA and Fisher's exact test for false positive rate $\alpha = 0.25$ and false negative rate $\beta = 0.35$

4 Results

4.1 Generated data and mappings

For mice, all in all we were able to extract 17147 unique gene to pathway relationships including 6660 genes and 225 pathways. Furthermore 9232 metabolite to enzyme relations with 222 metabolites were derived. Moreover, additional 83 direct metabolite to pathway relationships were determined.

For humans, a number of 4592 unique metabolite to enzyme relationships including 153 metabolites and 546 genes, as well as 326 direct metabolite to pathway relationships could be found.

4.2 Results performance assessment and comparison

To assess the performance of MONA we ran predictions for self-generated test data-sets and compared the results of MONA metabolite to MGSA and Fisher's exact test.
Figures 7 and 8 show the performance of MONA compared to MGSA anf Fisher's exact test. It is clearly visible, that MONA metabolite outperforms the other methods in terms of false positive and and false negative ratios.

4.3 M 200 extraction of differentially expressed genes

Table 1 shows, which input-sets for MONA were generated from the Mouse 200 data-set, along with their respective sizes. There are large differences between the input sets for wild-type against diabetic mice. Not quite as many differentially expressed genes and metabolites are contained in the input sets for the comparison of diabetic mice with and without treatment. The only exception here is the comparison of diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x were a decent amount of genes were regarded as differentially expressed.

groups compared	# diff. genes	# genes in background	# diff. metabolites	# metabolites in background
wild-type vs. diabetic	3697	25,649	206	279
diabetic vs. metformin-treated	1	25,649	5	279
diabetic vs. drug-x-treated	110	25,649	99	279
diabetic vs. combination treatment	208	25,649	5	280
wild-type vs. diabetic after two weeks	3856	25,649	184	274
diabetic vs. metformin-treated after two weeks	10	25,649	7	273
diabetic vs. drug-x-treated after two weeks	0	25,649	0	273
diabetic vs. combination treatment after two weeks	651	25,649	9	274

Table 1: Overview about the sizes of the generated input files for MONA from the described mouse liver data-set. The first column contains the two compared classes and the other columns contain the numbers of differentially expressed genes, genes in background, differentially concentrated metabolites and and metabolites in background from left to right.

4.4 Results M 200 gene set enrichment analysis

For each comparison of two different classes of mice from the Mouse 200 study, gene set enrichment was performed using the integrative model MONA metabolite in order to determine consequences of type II diabetes and different approaches of medical treatment for the metabolism of mice. We try to evaluate the MONA metabolite output by having a close look at what kind of pathways are predicted to be significant and which of their corresponding genes and metabolites are differential. We also compare the results with those, other methods like MGSA and Fisher's exact test would have achieved for the data.

4.4.1 Wild-type mice vs. diabetic mice four hours after disease outbreak

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
1.00000	03010	Ribosome	119	62	33	0	0	0	0
1.00000	04146	Peroxisome	80	75	42	0	0	0	0
1.00000	00982	Drug metabolism - cytochrome P450	87	76	39	5	0	0	5
1.00000	04610	Complement and coagulation cascades	76	71	34	1	0	0	0
0.99970	04142	Lysosome	123	115	42	0	0	0	0
0.98771	03040	Spliceosome	138	86	28	0	0	0	0
0.98603	00020	Citrate cycle (TCA cycle)	31	30	15	8	5	3	8
0.98307	03018	RNA degradation	76	61	19	0	0	0	0
0.97745	00010	Glycolysis / Gluconeogenesis	62	52	21	2	1	1	2
0.97018	03008	Ribosome biogenesis in eukaryotes	86	66	22	0	0	0	0
0.95866	00520	Amino sugar and nucleotide sugar metabolism	48	44	17	2	1	1	2
0.94799	00280	Valine, leucine and isoleucine degradation	50	46	22	11	3	3	11
0.94722	00270	Cysteine and methionine metabolism	39	29	13	7	3	3	7
0.94537	00260	Glycine, serine and threonine metabolism	34	33	14	12	10	9	12
0.93458	00100	Steroid biosynthesis	18	15	8	4	2	1	4

Table 2: The table shows the results of MONA metabolite for the comparison of diabetic and wild-type mice. The columns show (from left to right) the predicted probabilty of a pathway, the pathway identifier, the name of the pathway, how many genes correspond to that pathway, how many of those are differentially expressed within the dataset, how many metabolites correspond to the pathway, how many of them are differentially concentrated in the Mouse 200 data and how many of the metabolites which correspond to the pathway could be mapped to a gene product

Table 2 shows the results of MONA metabolite for differentially expressed genes and metabolites with differential concentrations between diabetic mice and healthy wild-type mice. Having a quick look at the table it quickly becomes clear that there are a lot of metabolic pathways predicted to be different between the two compared types of mice. The table shows the most significant results, all of which are predicted to be differential with a very high probability of over ninety percent. The results suggest changes in basic cellular compartments and mechanisms like the peroxisome or the splicosome as well as changes in basic metabolic pathways like the citrate cycle.

Figure 9 compares the results for the comparison between diabetic and wild-type mice produced by MONA metabolite and the single level model of MONA. The figure shows that the single level model predicts the exact same pathways to be differential as MONA metabolite. Also the probabilities for the results are close to the same.
Regarding the overall topic of the data-set, some of the pathways predicted to be different seem to be especially interesting. One of those is the citrate cycle as one of the most basic and important parts of eukaryotic metabolism. Figure 10 shows the pathway and its changes in gene expression and metabolite concentration caused by type II diabetes. A majority of all the enzymes being part of the pathway are under-expressed. There is a higher concentration of malate and fumarate and a lower concentration of cis-aconitate, than expected. Another pathway which is strongly suggested to be differential by MONA and which is of special interest is the drug metabolism pathway. Looking at Figure 11 it can be observed, that close to all enzymes which are responsible for the processing of a number of popular drugs are under-expressed with the exception of the CYP1A2 gene which encodes cytochrome P450, family 1, subfamily A, polypeptide 2 and which is over-expressed.

To compare the results of MONA metabolite for the Mouse 200 data-set with the results of existing approaches we had a look at the results of MGSA when run for the mRNA set and the metabolite set respectively. Figure 12 shows these results in comparison with the results of MONA metabolite. It can be seen that MONA metabolite agrees with terms predicted to be differential by MGSA for the mRNA data-set only and for both the mRNA and the metabolite dataset. However the terms only predicted to be differential by MGSA for the metabolite data-set only do not show up in the results of MONA metabolite.

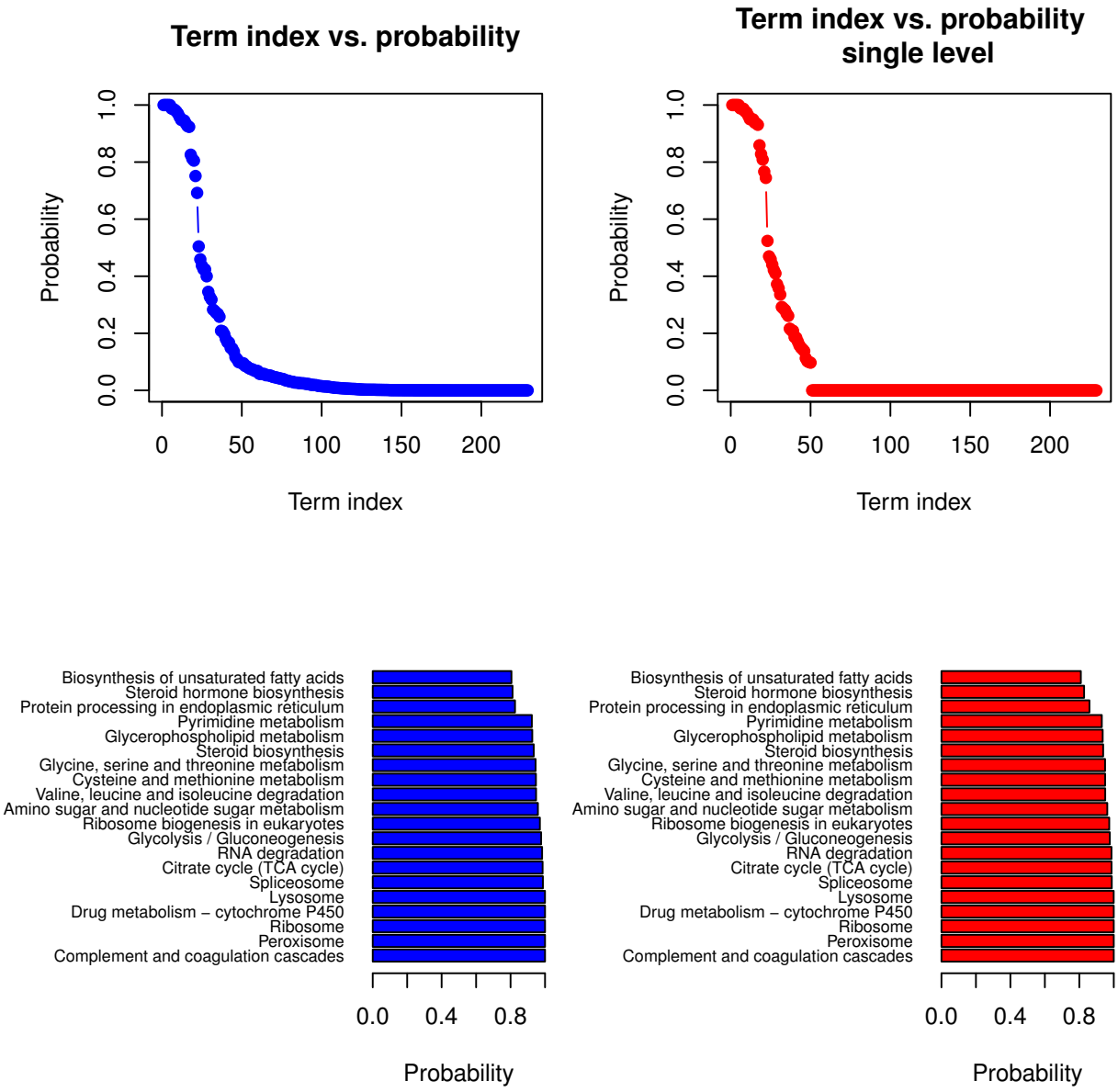

Figure 9: Comparison of the results of MONA metabolite with the single level model of MONA. The single level results are colored red and the MONA metabolite results are shown in the plots with blue color. The top two plots show each term with its probability as a curve showing the overall distribution of term probabilities. The bar-plots on the bottom show the top result pathways. The length of the bar describes how likely they are to be differential.

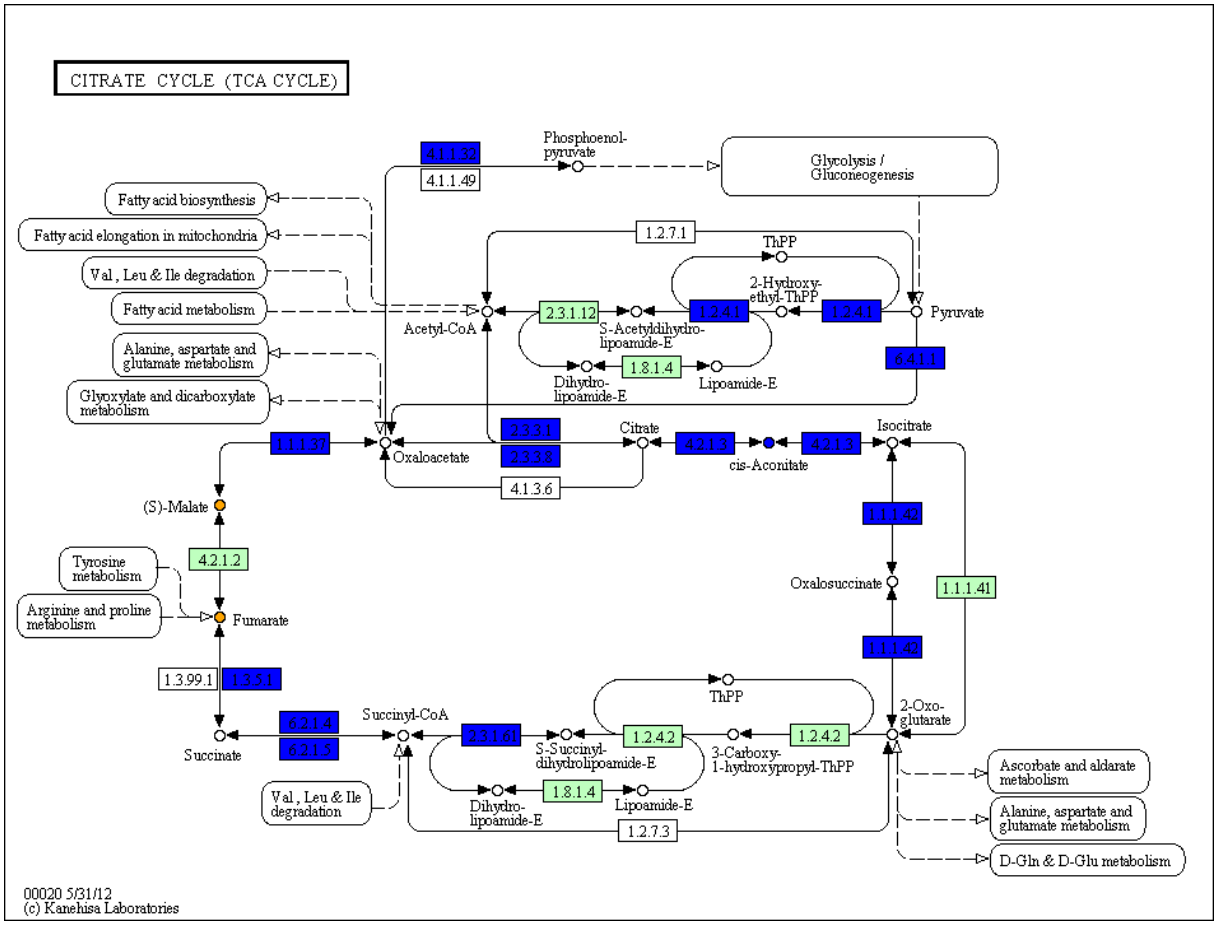

Figure 10: Citrate cycle with differences between wild-type mice and diabetic mice four hours after disease outbreak. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

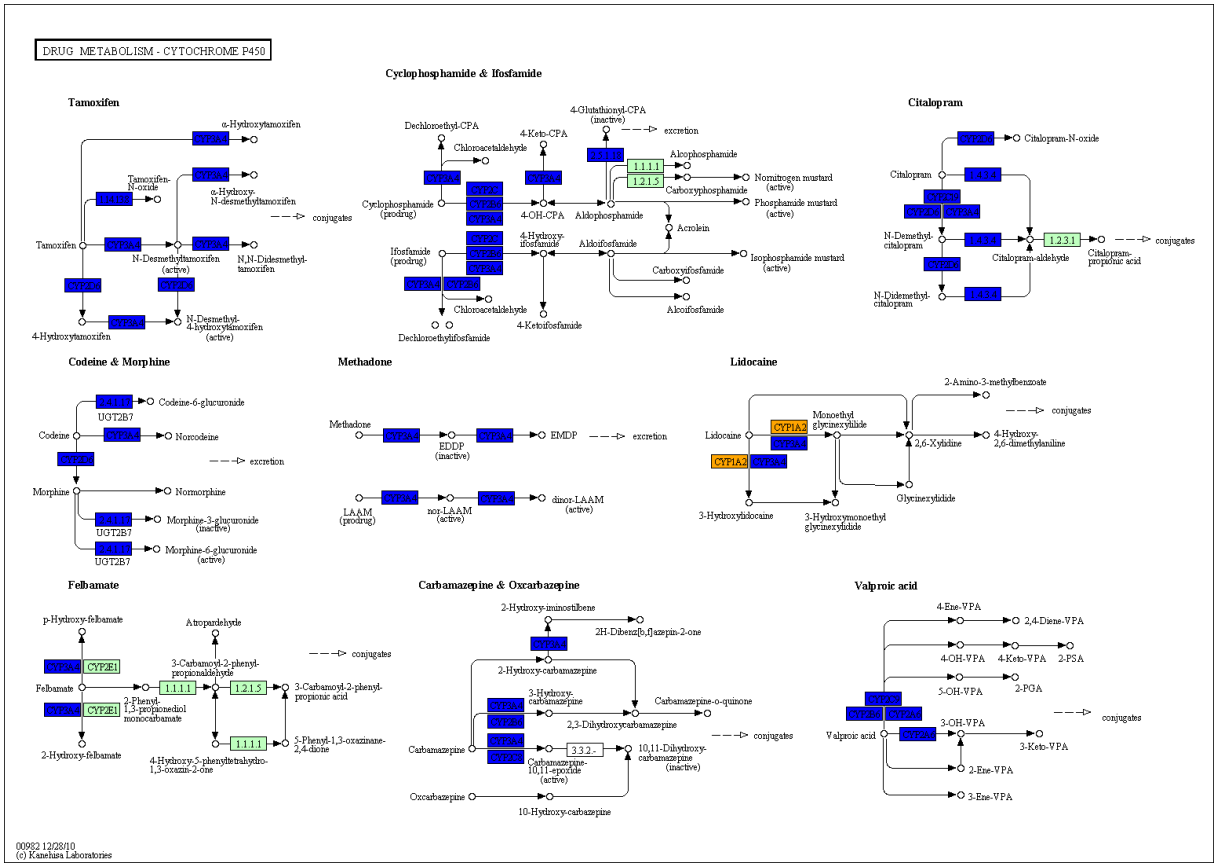

Figure 11: Drug metabolism pathway with differences between wild-type mice and diabetic mice four hours after disease outbreak. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

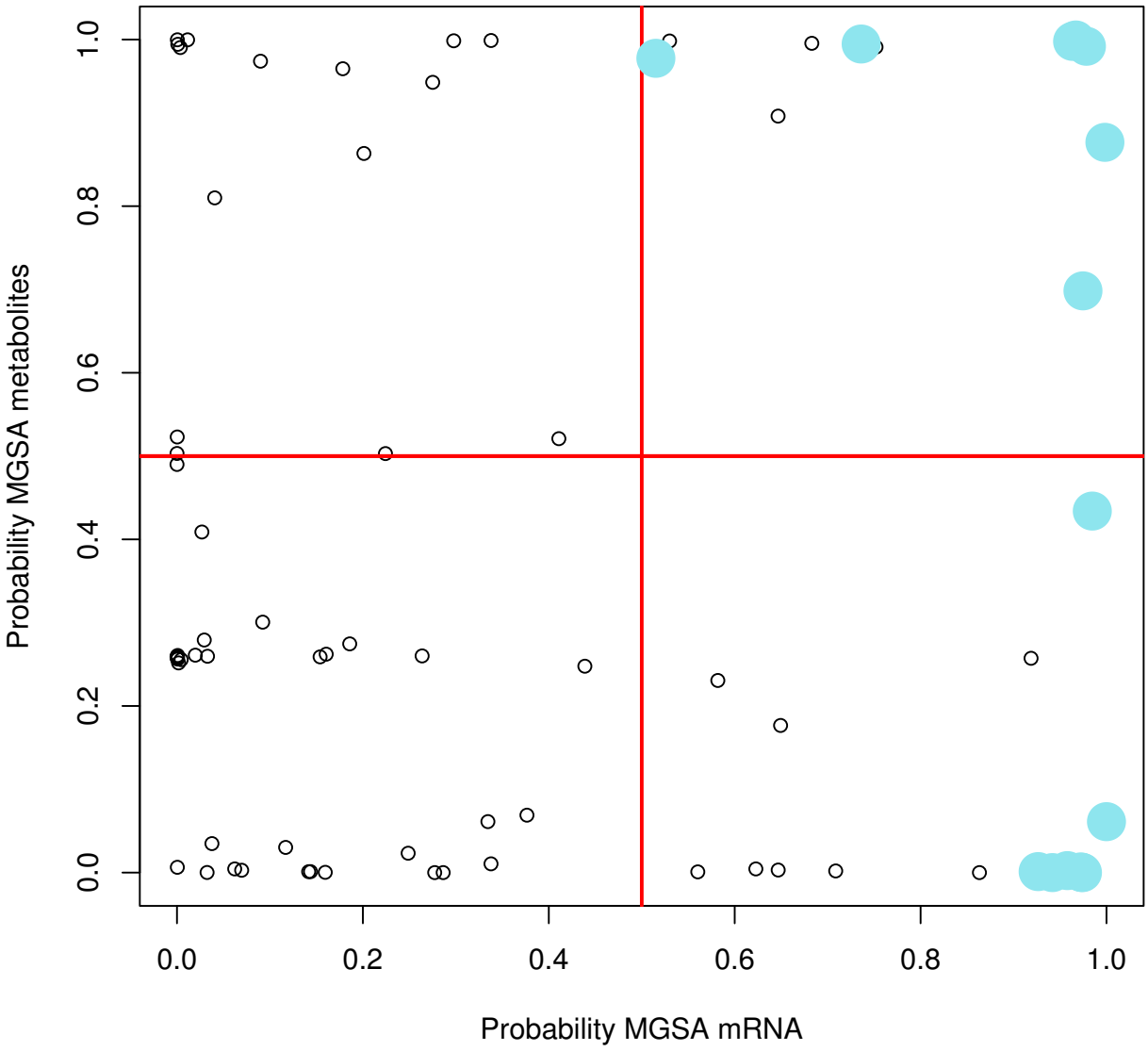

Figure 12: Comparison of MGSA predictions for the wild-type versus diabetic mice comparison from the Mouse 200 dataset. For each pathway it was predicted how likely it is for the pathway to be differential by MGSA, providing mRNA data (x-axis) or metabolite data (y-axis) as input. The red lines mark the significance level of 0.5. The colored dots mark the pathways which are also predicted to be differential by MONA metabolite with a probability of more than 0.5

4.4.2 Wild-type mice vs. diabetic mice two weeks after disease outbreak

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
1.00000	00982	Drug metabolism - cytochrome P450	87	76	41	5	0	0	5
1.00000	04610	Complement and coagulation cascades	76	71	36	1	0	0	0
1.00000	04146	Peroxisome	80	75	36	0	0	0	0
0.99856	00190	Oxidative phosphorylation	147	104	37	1	1	1	0
0.96765	00071	Fatty acid metabolism	48	41	19	3	1	1	3
0.94054	04142	Lysosome	123	115	37	0	0	0	0
0.84829	05219	Bladder cancer	43	38	17	0	0	0	0
0.83514	02010	ABC transporters	45	42	16	2	1	1	0
0.83437	00100	Steroid biosynthesis	18	15	9	4	2	1	4
0.56781	00600	Sphingolipid metabolism	41	34	14	3	3	2	3
0.50137	03050	Proteasome	45	41	15	0	0	0	0
0.49988	00010	Glycolysis / Gluconeogenesis	62	52	18	2	1	1	2
0.44479	04130	SNARE interactions in vesicular transport	35	31	12	0	0	0	0
0.42681	01040	Biosynthesis of unsaturated fatty acids	25	23	10	26	11	11	0
0.39889	04120	Ubiquitin mediated proteolysis	140	121	37	0	0	0	0

Table 3: The table contains the results of MONA metabolite for the comparison between wild-type and diabetic mice two weeks after disease outbreak. The columns contain the same information as the columns of Table 2

The investigation of the data for the comparison of wild-type and diabetic mice after two weeks using MONA metabolite also shows a number of high probability predictions, even though not quite as many as the comparison of the mentioned classes of mice after four hours. Again some basic metabolic pathways are contained within the results, such as glycolysis and fatty acid metabolism. Moreover the results suggest discrepancies in basic cell compounds and in intra-cellular transportation as shown in Table 3 by high probabilities for results like ABC-transporters.

The results of MONA metabolite for the wild-type vs diabetic after two weeks data-set holds some contrast to the predictions of MONA's single model as indicated by Figure 13. The single model generally predicts slightly higher probabilities for pathways to be altered. Some pathways which get high probabilities from the single model approach are particularly less good of a result for MONA metabolite, e.g. RNA transport (03013) which has a probability of roughly 80% in the single level results and only around 20% in the MONA metabolite outcome. Other pathways are slightly more probable for MONA metabolite though, like for example bladder cancer (05219)
A pathway which is contained in the prediction results and which is of special interest for examining the effects of type II diabetes on the metabolism of the mouse is the steroid biosynthesis pathway. Figure 14 shows the pathway and how it is altered within the diabetic mice. A large number of enzymes in this pathway are over-expressed. The metabolite cholesterol has a higher concentration than expected. Another set of genes of interest here are the ABC-transporters subfamilies. The changes are shown in Figure 15. A number of genes are differential in all the abc-transporter subfamilies. Some of them are over-expressed and some under-expressed with about equal distribution.

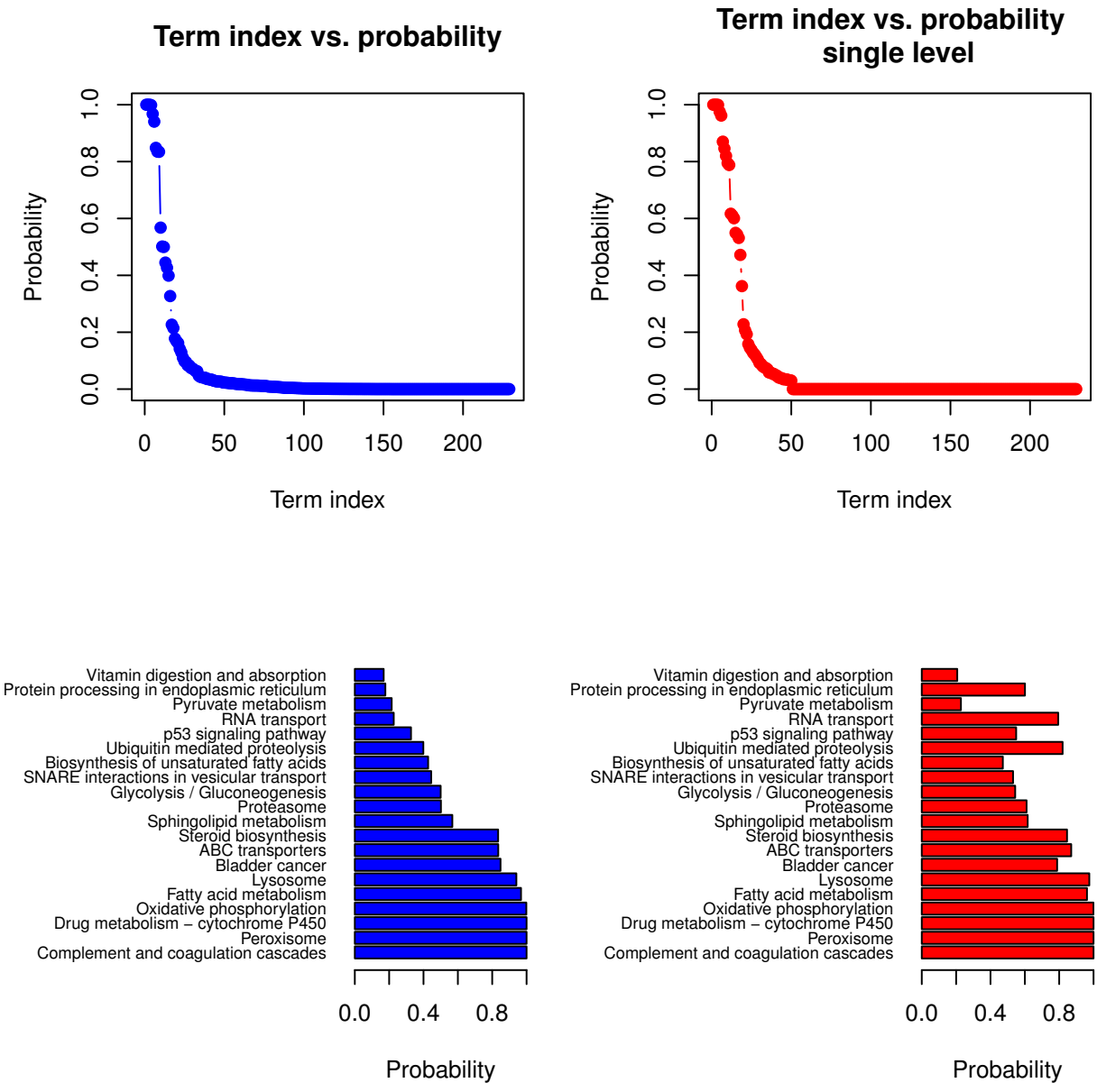

Figure 13: Comparison of MONA metabolite results and MONA single level results

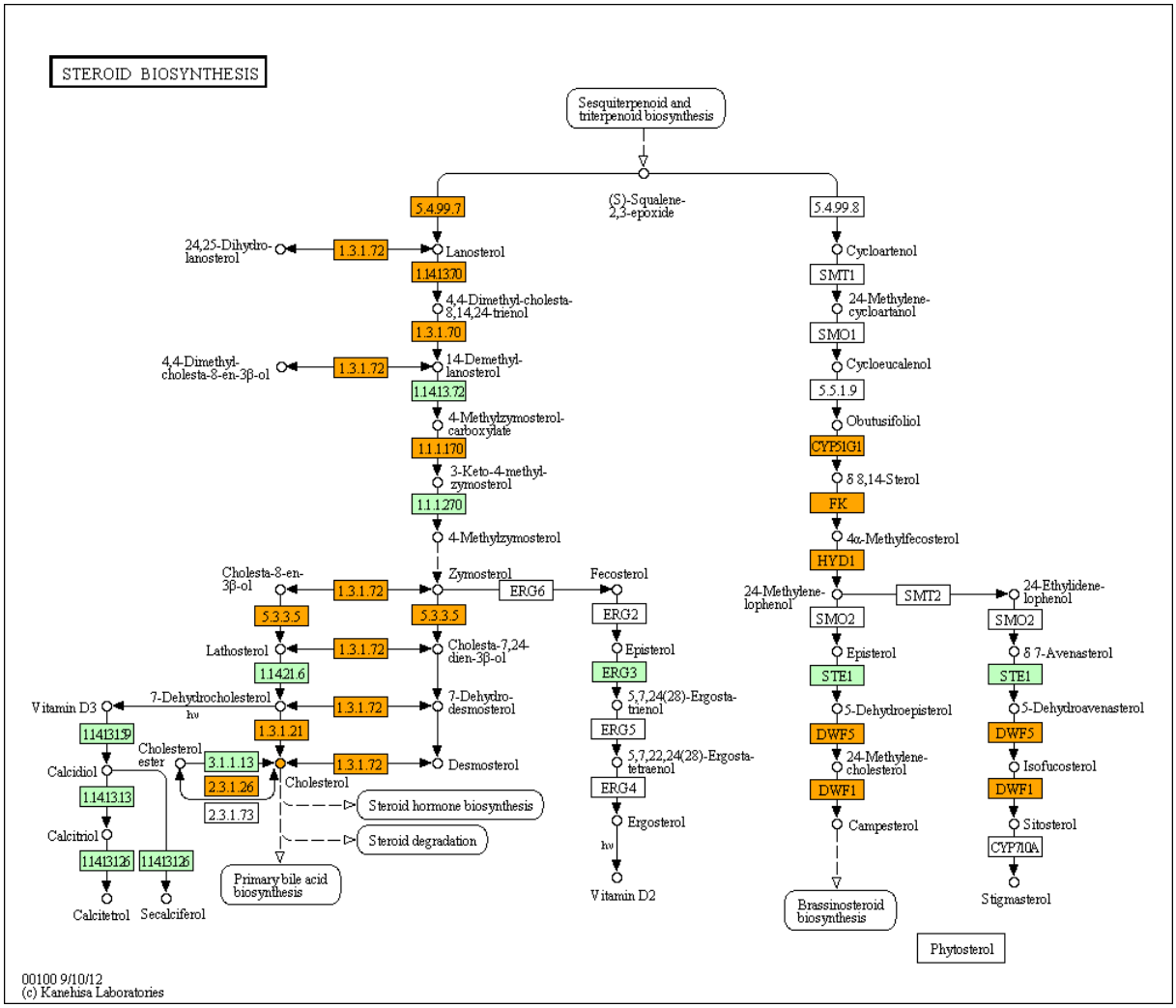

Figure 14: Steroid biosynthesis pathway with differences between wild-type mice and diabetic mice two weeks after disease outbreak. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

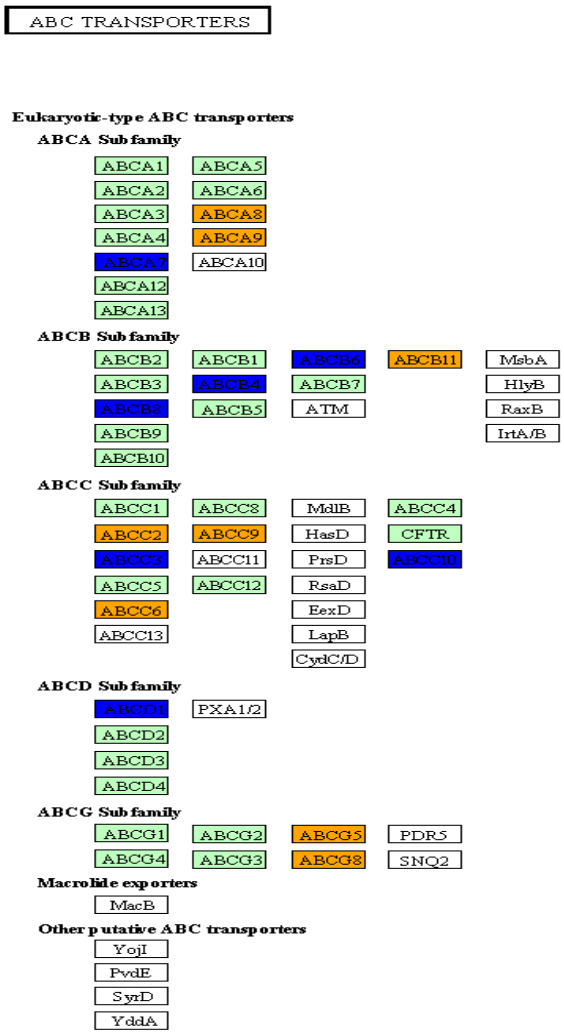

Figure 15: ABC-transporters subfamilies with differences between wild-type mice and diabetic mice two weeks after disease outbreak. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

4.4.3 Untreated diabetic mice vs. diabetic mice after different medical treatments

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
0.00226	00195	Photosynthesis	0	0	0	1	1	0	0
0.00226	04111	Cell cycle - yeast	0	0	0	1	1	0	0
0.00226	00550	Peptidoglycan biosynthesis	0	0	0	1	1	0	0
0.00226	02020	Two-component system	0	0	0	1	1	0	0
0.00225	00471	D-Glutamine and D-glutamate metabolism	3	1	0	4	3	0	4
0.00225	00121	Secondary bile acid biosynthesis	0	0	0	3	1	0	0
0.00225	00785	Lipoic acid metabolism	3	1	0	0	0	0	0
0.00114	00780	Biotin metabolism	2	2	0	1	1	0	1
0.00092	00120	Primary bile acid biosynthesis	15	15	1	10	6	0	10
0.00058	00730	Thiamine metabolism	4	3	0	1	1	0	1
0.00058	00300	Lysine biosynthesis	3	3	0	1	1	0	1
0.00058	00524	Butirosin and neomycin biosynthesis	4	3	0	1	1	0	1
0.00058	00061	Fatty acid biosynthesis	6	3	0	12	5	0	0
0.00015	00400	Phenylalanine, tyrosine and tryptophan biosynthesis	7	5	0	4	2	0	4
0.00008	00460	Cyanoamino acid metabolism	6	6	0	2	2	0	2

Table 4: MONA metabolite results for the comparison of diabetic mice and mice that have been treated with metformin for four hours.

Running MONA on the data-set, generated from differential metabolites and genes between diabetic mice and mice and metformin-treated mice after four hours of therapy, did not produce any good results, as revealed in Table 4. The probabilities for changes in any pathways are overall really low.
Two weeks of treatment did also not reveal any new findings regarding the changes in metabolism invoked through metformin (Tab. 5).

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
0.00225	00195	Photosynthesis	0	0	0	1	1	0	0
0.00225	00550	Peptidoglycan biosynthesis	0	0	0	1	1	0	0
0.00225	04111	Cell cycle - yeast	0	0	0	1	1	0	0
0.00225	02020	Two-component system	0	0	0	1	1	0	0
0.00224	00471	D-Glutamine and D-glutamate metabolism	3	1	0	4	3	0	4
0.00224	00121	Secondary bile acid biosynthesis	0	0	0	3	1	0	0
0.00224	00785	Lipoic acid metabolism	3	1	0	0	0	0	0
0.00113	00780	Biotin metabolism	2	2	0	1	1	0	1
0.00057	00730	Thiamine metabolism	4	3	0	1	1	0	1
0.00057	00524	Butirosin and neomycin biosynthesis	4	3	0	1	1	0	1
0.00057	00300	Lysine biosynthesis	3	3	0	1	1	0	1
0.00057	00061	Fatty acid biosynthesis	6	3	0	12	5	0	0
0.00015	00400	Phenylalanine, tyrosine and tryptophan biosynthesis	7	5	0	4	2	0	4
0.00007	00460	Cyanoamino acid metabolism	6	6	0	2	2	0	2
0.00004	04122	Sulfur relay system	10	7	0	0	0	0	0

Table 5: Results for MONA metabolite for the comparison of diabetic mice and diabetic mice after two weeks of treatment with metformin

This does also apply for the comparisons of diabetic mice and mice after four hours of treatment with drug x (Tab. 6), diabetic mice and mice after two weeks of treatment with drug x and diabetic mice and mice after four hours of treatment with a combination of both medicaments (Tab. 7).

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
0.19022	04111	Cell cycle - yeast	0	0	0	1	1	1	0
0.19022	02020	Two-component system	0	0	0	1	1	1	0
0.19022	00550	Peptidoglycan biosynthesis	0	0	0	1	1	1	0
0.19022	00195	Photosynthesis	0	0	0	1	1	1	0
0.00293	00471	D-Glutamine and D-glutamate metabolism	3	1	0	4	3	1	4
0.00293	00785	Lipoic acid metabolism	3	1	0	0	0	0	0
0.00293	00121	Secondary bile acid biosynthesis	0	0	0	3	1	0	0
0.00111	00780	Biotin metabolism	2	2	0	1	1	1	1
0.00042	00730	Thiamine metabolism	4	3	0	1	1	0	1
0.00042	00300	Lysine biosynthesis	3	3	0	1	1	1	1
0.00042	00524	Butirosin and neomycin biosynthesis	4	3	0	1	1	1	1
0.00042	00061	Fatty acid biosynthesis	6	3	0	12	5	3	0
0.00006	00400	Phenylalanine, tyrosine and tryptophan biosynthesis	7	5	0	4	2	2	4
0.00002	00460	Cyanoamino acid metabolism	6	6	0	2	2	2	2
0.00001	04122	Sulfur relay system	10	7	0	0	0	0	0

Table 6: Results of MONA metabolite for the comparison of diabetic mice and diabetic mice after four hours of treatment with drug x

Neither MONA metabolite nor the single level model of MONA predicted any significant metabolic changes for these classes .

Probability	ID	Pathway name	# corr. genes	# c.g. in backgr.	# c.g. in diff.	# corr. meta	# c.m. in backgr.	# c.m. in diff.	# mapped meta
0.05476	00730	Thiamine metabolism	4	3	1	1	1	0	1
0.00248	00232	Caffeine metabolism	9	8	1	16	0	0	16
0.00246	02020	Two-component system	0	0	0	1	1	0	0
0.00246	04111	Cell cycle - yeast	0	0	0	1	1	0	0
0.00246	00550	Peptidoglycan biosynthesis	0	0	0	1	1	0	0
0.00246	00195	Photosynthesis	0	0	0	1	1	0	0
0.00245	00471	D-Glutamine and D-glutamate metabolism	3	1	0	4	3	0	4
0.00245	00785	Lipoic acid metabolism	3	1	0	0	0	0	0
0.00245	00121	Secondary bile acid biosynthesis	0	0	0	3	1	0	0
0.00128	00780	Biotin metabolism	2	2	0	1	1	0	1
0.00067	00300	Lysine biosynthesis	3	3	0	1	1	0	1
0.00067	00524	Butirosin and neomycin biosynthesis	4	3	0	1	1	0	1
0.00067	00061	Fatty acid biosynthesis	6	3	0	12	5	0	0
0.00036	04977	Vitamin digestion and absorption	24	18	2	1	0	0	0
0.00019	03450	Non-homologous end-joining	12	12	1	0	0	0	0

Table 7: Results of MONA metabolite for the comparison of diabetic mice and diabetic mice after four hours of treatment with a combination of metformin and drug x

4.4.4 Untreated diabetic mice vs. diabetic mice after two weeks of treatment with a combination of metformin and drug x

In contrast to previously described attempts of medical treatment, analyzing the comparison of diabetic mice and mice after two weeks of treatment with a combination of metformin and drug x, did reveal a couple of changes. MONA metabolite predicted six pathways to be altered with a probability of more than 0.5. There seem to be significant changes in arginine and proline metabolism, taruine and hypotaurine metabolism and glutathione metabolism. Moreover, there are alterations in the pathway related to type I diabetes mellitus (Fig. 8).

Prob	ID	Pathway name	# corr.genes	# c.g. inbackgr.	# c.g. indiff.	# corr.meta	# c.m. inbackgr.	# c.m. indiff.	# mappedmeta
1.00000	00330	Arginine and proline metabolism	54	46	2	18	10	2	18
0.99935	00480	Glutathione metabolism	54	51	1	6	5	1	6
0.87412	04141	Protein processing in endoplasmic reticulum	169	141	13	0	0	0	0
0.69025	03008	Ribosome biogenesis in eukaryotes	86	66	8	0	0	0	0
0.56860	00430	Taurine and hypotaurine metabolism	10	10	0	2	2	1	2
0.52761	04940	Type I diabetes mellitus	59	52	3	0	0	0	0
0.41745	03013	RNA transport	168	128	10	0	0	0	0
0.34719	04146	Peroxisome	80	75	6	0	0	0	0
0.33055	00910	Nitrogen metabolism	23	18	1	3	3	1	3
0.32533	00670	One carbon pool by folate	19	16	0	0	0	0	0
0.31208	04976	Bile secretion	71	63	5	3	0	0	0
0.23066	04150	mTOR signaling pathway	53	44	4	0	0	0	0
0.21330	04722	Neurotrophin signaling pathway	131	119	7	0	0	0	0
0.21221	03015	mRNA surveillance pathway	93	69	5	0	0	0	0
0.19534	04622	RIG-I-like receptor signaling pathway	69	62	5	0	0	0	0

Table 8: Results of MONA metabolite for the comparison of diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x

It is also to notice, that MONA's single level model using only the mRNA expression data as input did not predict any pathway to be altered, thus including the metabolite data leads to very different findings (Fig. 16).

The arginine and proline metabolism pathway shows an under-expression of the Gls gene which codes the glutaminase enzyme. Moreover, an under-expression of the Nos1 gene coding for the nitric-oxide synthase is observable. Besides these enzymes there are also two metabolites which exist in lower concentrations in the medically treated mice than in the untreated diabetic mice, namely glutamate and citruline (Fig. 17).

Another pathway which also yields some changes of interest, is the glutathione metabolism pathway. In this pathway the Gene that codes the glutathione S-transferase is overexpressed. The metabolite L-Glutamate shows lower concentration values than expected.

Besides the two pathways just mentioned, the differences in the type I diabetes mellitus pathway might also be a matter of concern. Here, GroEl, a molecular chaperone is overexpressed and perforin 1, encoded by the PRF1 gene is also over-expressed.

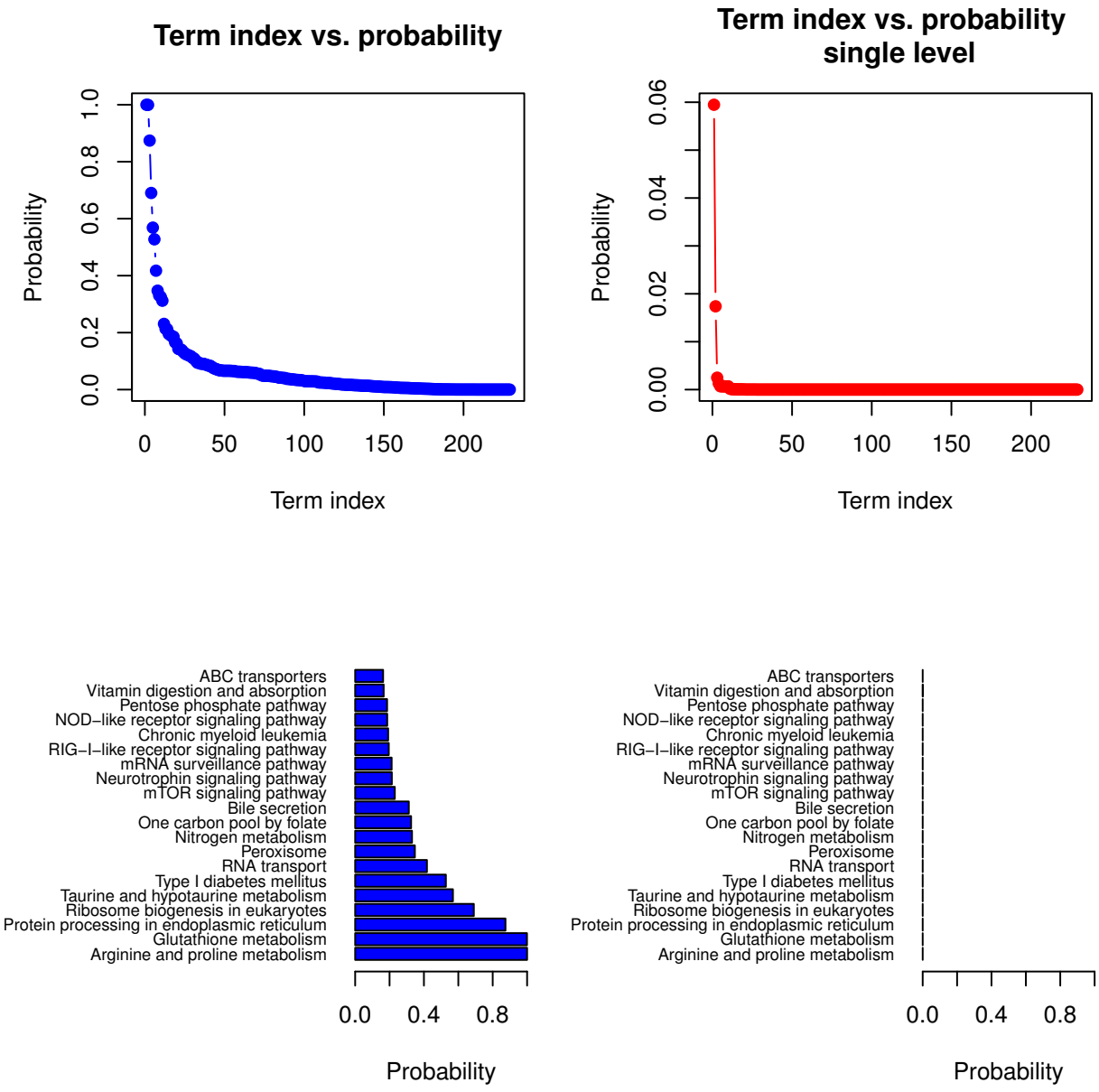

Figure 16: Comparison of the results for MONA metabolite and MONA's single level model for the comparison between diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x.

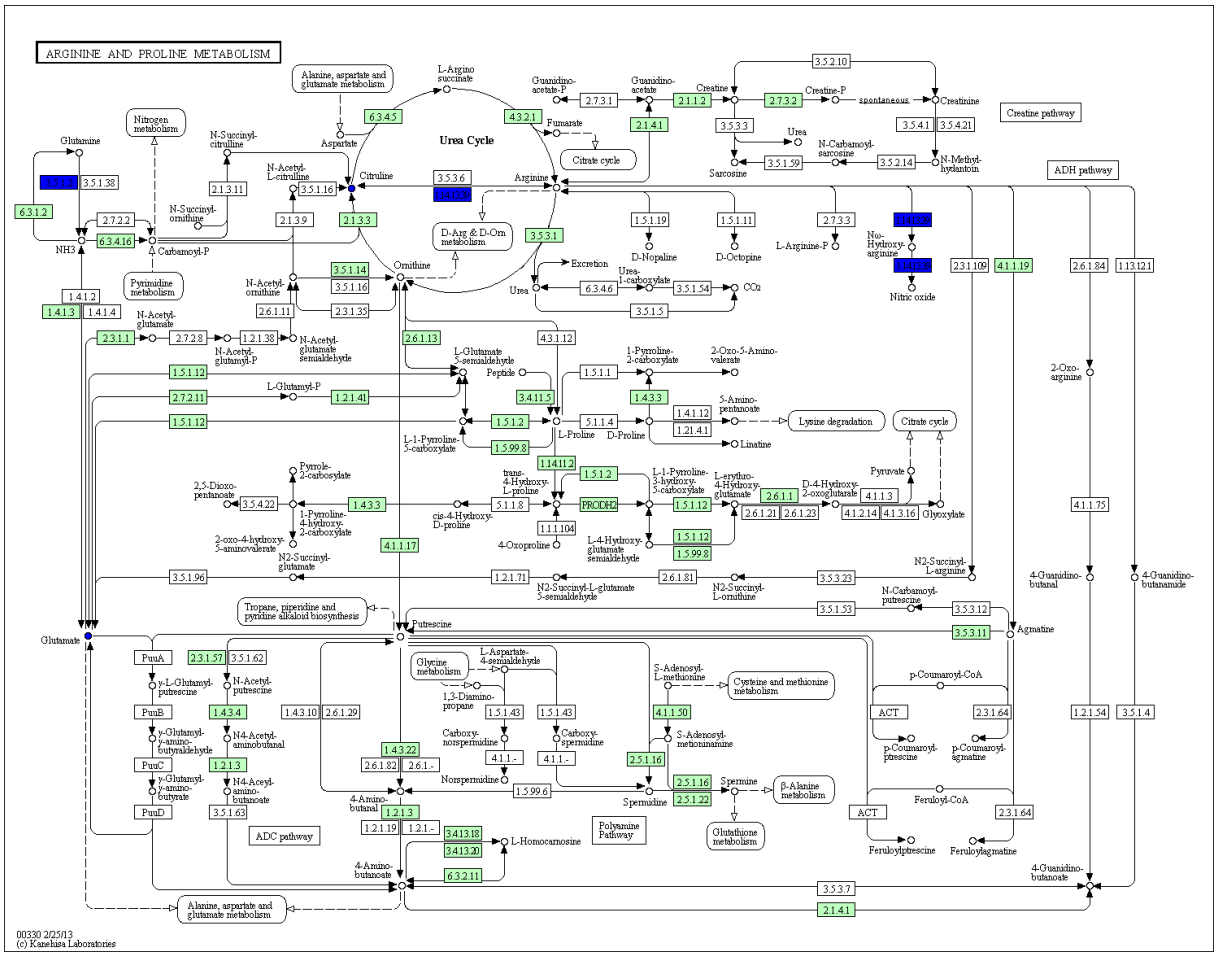

Figure 17: Arginine and proline metabolism pathway with differences between diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

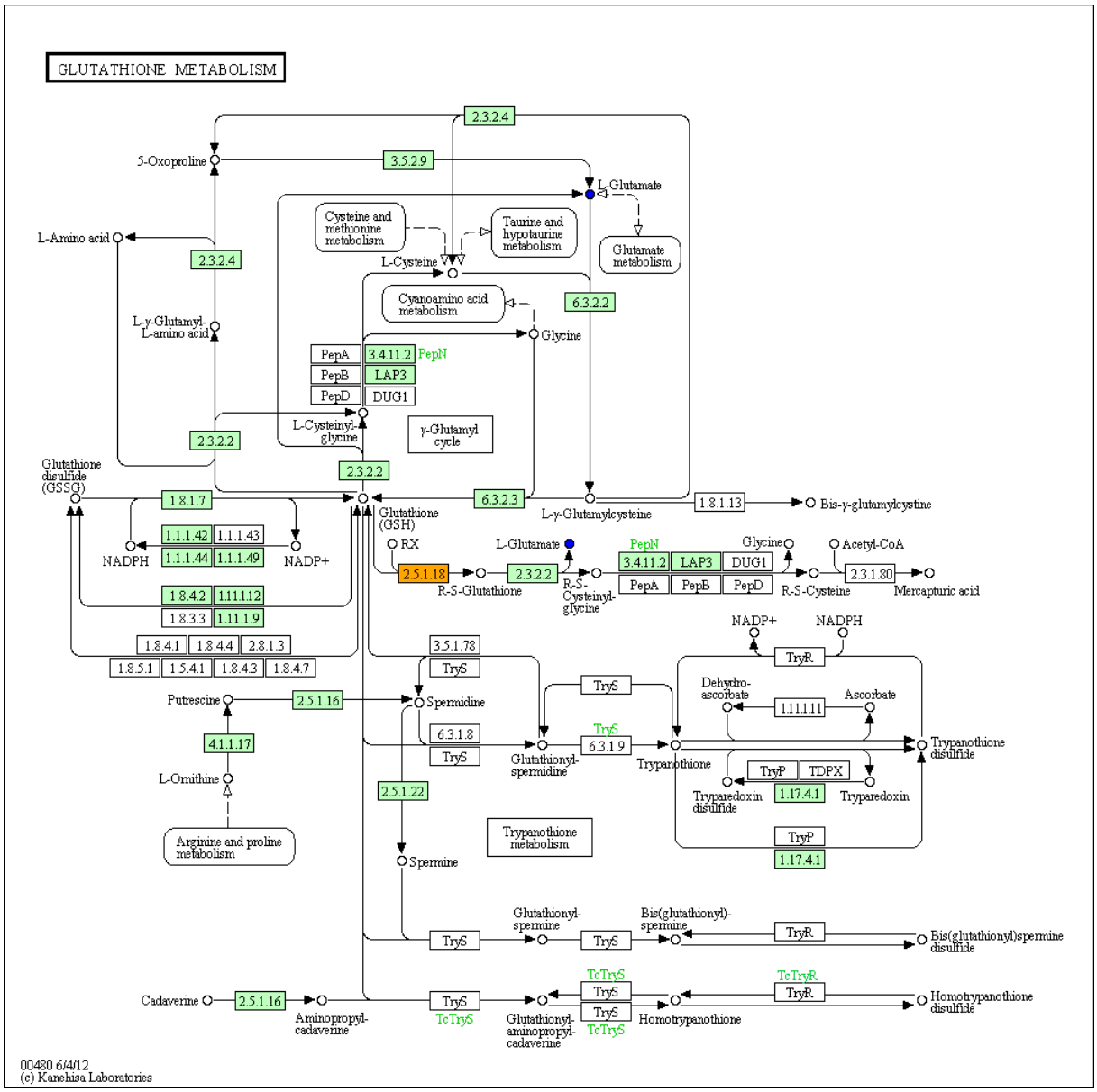

Figure 18: Glutathione metabolism pathway with differences between diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

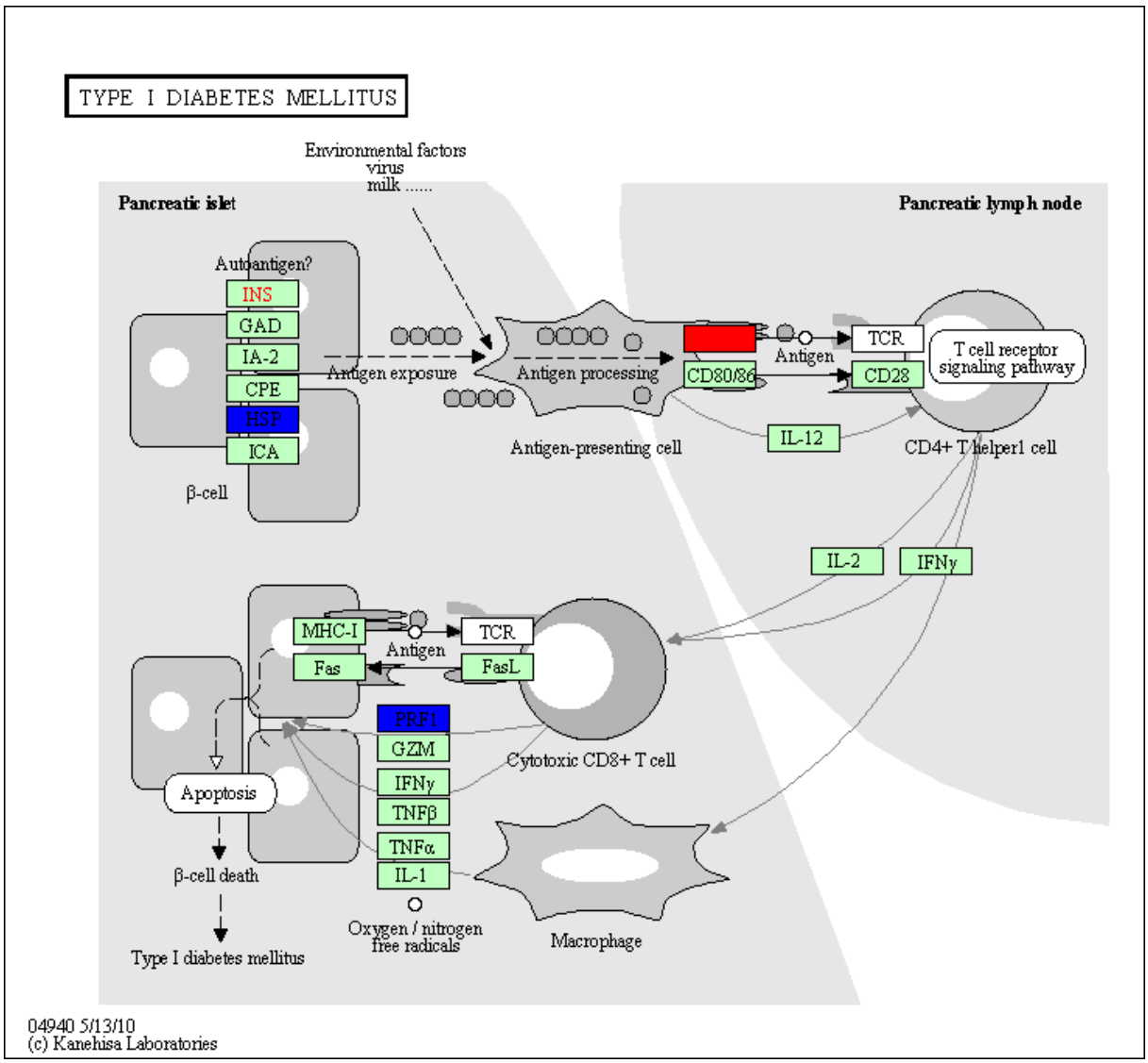

Figure 19: Type I diabetes mellitus pathway with differences between diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x. Green signals those genes which take part course of the metabolic pathway in the mouse. Genes and metabolites marked blue are under-expressed and orange marked genes and metabolites are over-expressed within the medically treated mice in comparison to the untreated diabetic mice.

5 Discussion

5.1 Generated data and mappings

A comprehensive set of mappings could be created from the KEGG-annotations for the pathways of *mus musculus*. One problem that arises, however, is that the KEGG-mappings consist of KEGG identifiers for genes and metabolites, while MONA uses the unique metabolon identifications to assign metabolites to pathways in order to avoid complications which might occur from different ways of spelling etc.. Although every KEGG-gene-identifier can in general be mapped to a Entrez-ID or a gene symbol, not all KGG-compound-identifiers can be mapped to a metabolon term. From this follows that known metabolite to pathway relationships cannot be taken into account by MONA because of missing metabolon name assignment. By regularly updating the MONA database though, this problem will get less serious over time, as current annotations and assignments are always going to improve. Improving the quality of metabolite to pathway mappings will inevitable also cause a performance improvement for MONA in terms of coverage.

5.2 Performance assessment and comparison

Comparing the performance of MONA metabolite to MGSA anf Fisher's exact test showed, that MONA metabolite clearly outperforms the other methods. This shows that the additional information gained by integrating transcriptomics and metabolomics data does indeed lead to better prediction results compared to just taking the transcriptomics data into account.

5.3 Extraction of differentially expressed genes and metabolites from the Mouse 200 data-set

Ignoring errors in measurement the process of extracting differentially expressed genes and metabolites is pretty much free of complications. Two things have to be considered, though.

The first problem that occurs is the same as for the generation of the metabolite to pathway mappings. Only a part of the metabolites marked as differentially expressed by statistical tests can be considered for the gene set enrichment process as a metabolon annotation is only available for this part of metabolites. As already mentioned this flaw is going to improve with better assignments and annotations being available in the future.

The other fact that has to be considered extracting differential mRNAs and metabolites is the choice of the p-value offset. We chose a p-value of 0.05 which is a very popular and commonly accepted offset. However, dependent on the case and different conditions, it might be reasonable to chose another offset in order to improve results.

5.4 Mouse 200 gene set enrichment analysis

In the following, we will discuss the results of the gene set enrichment analysis performed with MONA metabolite for the input-sets generated from the Mouse 200 data-set. We will discuss the possible reasons for changes in metabolism caused by type II diabetes mellitus and different approaches of medical treatment, as well as the consequences for the organism that result from these changes.

5.5 Wild-type mice vs. diabetic mice four hours after disease outbreak

The gene set enrichment analysis of the differential mRNAs and metabolites from the comparison of wild-type mice and diabetic mice four hours after disease outbreak suggested many alterations in quite a large number of pathways including pathways responsible for very basic cellular and metabolic processes, such as the citrate cycle and glycolysis (See Fig. 2). First of all it is to note here, that the results very much agree with the predicted changes by the single level model of MONA (See Fig. 9). This comes from the huge number of genes assigned as differential leading to a number of changes in pathways which is so big, that including metabolite information seems to not be necessary for good enrichment results. If one has a closer look at the changes in phenotype and overall health that type II diabetes mellitus brings with itself, it appears reasonable that such a big part of the gene pool is altered within diseased individuals.

There are two pathways within the pathways suggested to be altered by the gene set enrichment analysis, we will discuss more closely. We will look at changes for certain corresponding mRNAs and metabolite concentration in detail and draw conclusions on how these changes might affect the organism and thus lead to diabetes-related symptoms.

At first we will focus on the changes in the citrate cycle. It is clearly visible that most enzymes along the path are under-expressed. For the metabolites, there is an under-expression of cis-Aconitate and an over-expression of malate and fumarate. As commonly known, type II diabetes causes defective hepatic glucose production [3], which strongly correlates with our findings of low activity of enzymes involved in the oxidation of carbohydrates. The fact that malate and fumarate are over-expressed leads from the fact that fumarate also comes from arginine and proline metabolism which is also predicted as active by the gene set enrichment. The high concentration of fumarate and the normal activity of the fumarase explain the high concentration of malate.

Another pathway that appears to be of special interest is the drug metabolism pathway. It is clearly visible that the enzymes taking part in the metabolism of a number of popular drugs are under-expressed to a big part (See Fig. 11). This correlates with already performed studies where the inhibition of drug metabolism in liver samples of rats was observed [1].

5.6 Wild-type mice vs. diabetic mice two weeks after disease outbreak

The gene set enrichment analysis of the comparison of wild-type mice with diabetic mice two weeks after disease outbreak also shows a larger number of results suggesting changes in a number of pathways. The results contain a number of pathways which were already expected to be corresponding to type II diabetes, like fatty acid metabolism and steroid biosynthesis, which leads to the conclusion that the results yield biological sense. One thing that attracts attention here, however, is the fact that there are some changes between the predictions of MONA metabolite and MONA's single level model. There are a number of pathways which are marked as active by the single level model but not by MONA metabolite. One example is the biosynthesis of unsaturated fatty acids. One of the core strengths of MONA metabolite is, that it is able to deal with redundancies between the terms on a multiple input scale. As the term fatty acid metabolism is already listed as a high probability result, and unsaturated fatty acid biosynthesis is a part of the whole

fatty acid metabolism, biosynthesis of unsaturated fatty acids, could be regarded as a term redundancy which is filtered out by the MONA metabolite model. For the other terms like protein processing in endoplasmatic reticulum such an assumption cannot be made. There could be different reasons for the high probability of these terms, for example they could represent false positive results, as MONA's cooperative model shows better performance in terms of false positives as the single level model [29].

One of the pathways marked as active and which we want to have a closer look at is the steroid biosynthesis pathway. It is commonly known that steroid levels are directly related to diabetes [2]. Having a look at Figure 14 it is distinctly visible that almost all the enzymes, which are part of the steroid biosynthesis are over-expressed. This leads to a higher concentration of the metabolite cholesterol. High cholesterol levels are directly related to a higher risk of getting type II diabetes [8].

The second change on metabolism which is predicted to be associated to type II diabetes by MONA metabolite is the change in ABC-transporter subfamilies. It has already been shown that diabetes has effects on the expression of hepatic ABC-transporters [37]. This leads to another correlation of our gene set enrichment predictions and already existing biological research which suggests that the predictions of MONA metabolite are biologically meaningful.

5.7 Untreated diabetic mice vs. diabetic mice after different medical treatments

As the Tables 6, 4, 5 and 7 show, not pathways were predicted as differential for the mice treated with metformin, drug x and the mice treated with a combination of both for four hours. This is not very surprising regarding the number of differential genes and metabolites which were inferred from the Mouse 200 data-set for the comparison of untreated diabetic mice and the medically treated mice just described (See Tab. 1). Regarding these numbers of differential genes and metabolites are this low compared with the size of the measured genetic background, no results for the gene set enrichment could have been expected.

5.8 Untreated diabetic mice vs. diabetic mice after two weeks of treatment with a cominaation of metformin and drug x

The most interesting results were achieved from the comparison of untreated diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x. As to see in Table 8 a number of pathways have been marked as active by MONA metabolite. First of all it is to mention that the predictions of MONA metabolite and MONA's single level model differ greatly from each other for this comparison. While MONA metabolite suggests six active pathways, the single level model was not able to achieve any results. This shows that taking metabolite concentrations into account can in fact make a big impact on the enrichment results. Besides this it also has to be noted that some pathways were marked as active by MONA metabolite, but not by the single level model, even though the pathways do not have any metabolites associated to them. One example for such a case is the type I diabetes mellitus pathway. This results appears to be very reasonable regarding the Mouse 200 data-set, and thus, it seems confusing that MONA metabolite would mark this pathway as active, but the single level model not. The reason for this is the fact that for MONA metabolite the

whole picture of gene and metabolite to pathway associations is considered. Even though there are no metabolites directly linked to the type I diabetes mellitus pathway, the fact that the additional information about metabolites is there general, has a positive effect on the performance of the inference process of the bayesian network, thus leading to better results. This shows another core advantage of MONA metabolite, leading to results which are not that obvious in the first place and which can not be achieved by single level approaches.

As the number of genes and metabolites marked as differential for the comparison of diabetic mice and diabetic mice after two weeks of treatment with a combination of metformin and drug x is significantly lower than for the comparison of wild-type and diabetic mice it becomes especially interesting to have a closer look at the predicted pathways and their alterations.

The first active pathway we will have a closer look at here, is the arginine proline metabolism (See Fig. 17). The first thing to catch the eye is the under-expression of the nitric-oxide synthase along with its product, the metabolite citruline. The fact, that both enzyme and product are under-expressed together again suggests that the prediction is biologically reasonable. Besides this, there also an under-expression of both the metabolite glutamate and the glutaminase, of which glutamate is a product. This enforces the point just made.

It seems surprising that a pathway is predicted to be altered with a high probability although only two genes out of over two hundred differentially expressed genes correspond to it. An explanation for that lies in the integration of transcriptomics and metabolomics. As there is big number of differentially expressed genes, the inference probably assumes a big number of false positives within the mRNA input set. In contrast to that, there are only nine metabolites with differential concentrations. Like this, great importance is linked to the occurrence of these metabolites in combination with already a low number of genes.

The next pathway we examine is the glutathione metabolism. Here, an under-expression of L-glutamate occurs together with an over-expression of glutathione-S-transferase. Although L-glutamate is neither a substrate nor a product of the glutathione-S-transferase they still lie in direct neighborhood to each other in the pathway. Nevertheless a direct relationship between the over-expression of the glutathione-S-transferase and the under-expression of L-glutamate can not be inferred. This may be a starting point for further examination in order to gain new biological insights.

The last pathway to look at is the type I diabetes mellitus pathway. Here the chaperonin GroEL and perforin I are under-expressed. According to KEGG perforin I is only related to type I diabetes, which raises the question why it is differential as the observed data was gained from mice diseased with type II diabetes. However recent studies have shown that perforin also plays a role in inflammation which is characteristic for type I and type II diabetes [34]. This shows that the role of perforin is not yet fully known for both types of diabetes. As perforin, the chaperonin GroEL does also not seem to have anything to do with type II diabetes but only seems to be related to type I diabetes. To the best of our knowledge no relation ship between GroEl and type II diabetes has been determined yet. Nevertheless there might be such a relation which just has not been detected yet.

6 Conclusion

Metabolic pathways are the building blocks of an organism's metabolism. In order to respond to different environmental conditions, these pathways can be activated. Each pathway consists of a set of biochemical reactions in which metabolites are converted. It is well known that the reactions responsible for the transmutations of metabolites are catalyzed by enzymes. The activation of a pathway leads to different concentrations of corresponding metabolites. As the metabolites and enzymes have a direct relationship to each other, it follows that different levels in enzyme activity cohere with the concentrations of their substrates and products. Therefore we expect that if a pathway is active, there are different levels of the activity of its enzymes and different concentrations of their corresponding metabolites. A possibility to measure enzyme activity is to measure the amount of mRNA which translates to the enzyme. The method introduced in this thesis aims to make statements about the activity of metabolic pathways by analyzing differential mRNAs and metabolites in a data-set by using the Bayesian Network model MONA metabolite. We also showed that MONA metabolite outperforms existing methods

We used this method to find active pathways in the Mouse 200 data-set and were able to find a lot of correlations with already existing research result which proves the proves that the method provides biologically meaningful results. Moreover we found out, how the combination of metformin and a new drug called drug x in this thesis affects the metabolism of type II diabetic mice. These findings could serve as starting points to further investigate the effectiveness of this new kind of medical treatment for type II diabetes mellitus. Finally we found some proteins whose mRNA was differentially expressed within the Mouse 200 data but which have no association to type II diabetes as of now, illustrating that a lot of information about type II diabetes has yet to be uncovered.

References

[1] DENNIS M Ackerman and KENNETH C Leibman. Effect of experimental diabetes on drug metabolism in the rat. *Drug Metabolism and Disposition*, 5(4):405–410, 1977.

[2] Peter Arner, Rolf Gunnarsson, Sven Blomdahl, and Carl-Gustav Groth. Some characteristics of steroid diabetes: a study in renal-transplant recipients receiving high-dose corticosteroid therapy. *Diabetes care*, 6(1):23–25, 1983.

[3] James R. Bain, Robert D. Stevens, Brett R. Wenner, Olga Ilkayeva, Deborah M. Muoio, and Christopher B. Newgard. Metabolomics applied to diabetes research: moving from information to knowledge. *Diabetes*, 58(11):2429–2443, Nov 2009.

[4] Sebastian Bauer, Peter N. Robinson, and Julien Gagneur. Model-based gene set analysis for bioconductor. *Bioinformatics*, 27(13):1882–1883, Jul 2011.

[5] Christopher M. Bishop. Model-based machine learning. *Philos Trans A Math Phys Eng Sci*, 371(1984):20120222, Feb 2013.

[6] Nathan Blow. Metabolomics: Biochemistry's new look. *Nature*, 455(7213):697–700, Oct 2008.

[7] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec 2004.

[8] Liam R Brunham, Janine K Kruit, Terry D Pape, Jenelle M Timmins, Anne Q Reuwer, Zainisha Vasanji, Brad J Marsh, Brian Rodrigues, James D Johnson, John S Parks, et al. β-cell abca1 influences insulin secretion, glucose homeostasis and response to thiazolidinedione treatment. *Nature medicine*, 13(3):340–347, 2007.

[9] Max Bylesjö, Daniel Eriksson, Miyako Kusano, Thomas Moritz, and Johan Trygg. Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. *Plant J*, 52(6):1181–1191, Dec 2007.

[10] Marc Carlson. *KEGG.db: A set of annotation maps for KEGG.*

[11] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[12] Natalie C. Duarte, Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard O. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777–1782, Feb 2007.

[13] E. Dudley, M. Yousef, Y. Wang, and W. J. Griffiths. Targeted metabolomics and mass spectrometry. *Adv Protein Chem Struct Biol*, 80:45–83, 2010.

[14] Oliver Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol*, 48(1-2):155–171, Jan 2002.

[15] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H. Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[16] V. S. Gomase, S. S. Changbhale, S. A. Patil, and K. V. Kale. Metabolomics. *Curr Drug Metab*, 9(1):89–98, Jan 2008.

[17] Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics*, 23(22):3024–3031, Nov 2007.

[18] Michael Inouye, Johannes Kettunen, Pasi Soininen, Kaisa Silander, Samuli Ripatti, Linda S. Kumpula, Eija Hämäläinen, Pekka Jousilahti, Antti J. Kangas, Satu Männistö, Markku J. Savolainen, Antti Jula, Jaana Leiviskä, Aarno Palotie, Veikko Salomaa, Markus Perola, Mika Ala-Korpela, and Leena Peltonen. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol*, 6:441, Dec 2010.

[19] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[20] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, Jan 2012.

[21] Gabi Kastenmüller, Werner Römisch-Margl, Brigitte Wägele, Elisabeth Altmaier, and Karsten Suhre. metap-server: a web-based metabolomics data analysis tool. *J Biomed Biotechnol*, 2011, 2011.

[22] Jan Krumsiek, Karsten Suhre, Anne M. Evans, Matthew W. Mitchell, Robert P. Mohney, Michael V. Milburn, Brigitte Wagele, Werner Romisch-Margl, Thomas Illig, Jerzy Adamski, Christian Gieger, Fabian J. Theis, and Gabi Kastenmuller. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet*, 8:e1003005, 2012.

[23] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5:21, 2011.

[24] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *J Proteome Res*, 11(8):4120–4131, Aug 2012.

[25] John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):34, 2011.

[26] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends Biotechnol*, 16(9):373–378, Sep 1998.

[27] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, Apr 2012.

[28] Henning Redestig and Ivan G. Costa. Detection and interpretation of metabolite-transcript coresponses using combined profiling data. *Bioinformatics*, 27(13):i357–i365, Jul 2011.

[29] Steffen Sass, Florian Buettner, N. S. Mueller, and F. J. Theis. A modular framework for gene ontology analysis integrating multilevel omics data. *Nucleic Acids Research*, 37:1–11, 2009.

[30] Martin I. Sigurdsson, Neema Jamshidi, Eirikur Steingrimsson, Ines Thiele, and Bernhard O. Palsson. A detailed genome-wide reconstruction of mouse metabolism based on human recon 1. *BMC Syst Biol*, 4:140, 2010.

[31] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, Oct 2005.

[32] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[33] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[34] HE Thomas, JA Trapani, and TWH Kay. The role of perforin and granzymes in diabetes. *Cell Death & Differentiation*, 17(4):577–585, 2009.

[35] Johan Trygg. O2-pls for qualitative and quantitative analysis in multivariate calibration. *Journal of chemometrics*, 16(6):283–293, 2002.

[36] Johan Trygg and Svante Wold. O2-pls, a two-block (x–y) latent variable regression (lvr) method with an integral osc filter. *Journal of Chemometrics*, 17(1):53–64, 2003.

[37] Willie M van Waarde, Henkjan J Verkade, Henk Wolters, Rick Havinga, Juul Baller, Vincent Bloks, Michael Müller, Pieter JJ Sauer, and Folkert Kuipers. Differential effects of streptozotocin-induced diabetes on expression of hepatic abc-transporters in rats. *Gastroenterology*, 122(7):1842–1852, 2002.

[38] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[39] Gregory R Warnes and G Gorjanc. gdata: Various r programming tools for data manipulation. *R package version*, 2(2), 2008.

[40] Jitao David Zhang and Stefan Wiemann. Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470–1471, Jun 2009.