LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München

Masterarbeit
in Bioinformatik

Extension of multi-level ontology analysis including adaption to continuous input values

Melanie Kopp

Aufgabensteller: Prof. Dr. Dr. Fabian Theis
Betreuer: Dr. Nikola Müller, Steffen Sass
Abgabedatum: 15. Oktober 2014

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

15. Oktober 2014 _____________________________

Melanie Kopp

Acknowledgement

At this point, I want to thank all those involved in this master thesis for their hints and advices, without whom this thesis would not have been realized.

First, I want to express my gratitude to Fabian Theis, who gave me the opportunity to do this master thesis at the Institute of Computational Biology, Helmholtz Center Munich.

A very special thank goes to Nikola Müller and Steffen Sass, my supervisors, who did not only spend a lot of time in this work, but also shared and developed new ideas and results of work with me.

In addition, I want to thank Dr. Florian Büttner for his support and advice. His consultations always were helpful.

Last, but not least, I would also like to express my sincere thanks to my family and my boyfriend Nick for their endless support, encouragement and trust in me during my bioinformatics study.

Abstract

The assessment of functional associations between gene sets from large-scale data is an important task in the field of systems biology. As high-throughput experimental techniques enable the investigation of biological functions across multiple omics, the model-based approach, named Multi-level ONtology Analysis (MONA), was recently introduced. MONA simultaneously identifies enriched functional ontology terms by integration of gene sets from multiple omics layers. (Omics levels include mRNA and protein expression data as well as knowledge about other molecular mechanisms, which contribute to the expression of gene products, for example DNA methylation or miRNA.) MONA models ontology term-to-gene relationships via a Bayesian network therewith accounting for term redundancies and multiple testing problems. Up to now, two gene set models are implemented in MONA, which are plugged to the base model. As the two models constitute only a subset of possible models, extending the variety of MONA models is essential for the inference of terms-to-gene relationships for truly multiple omics levels. In this thesis, we have implemented several extensions for the existing MONA model, including a flexible extension of the cooperative model to an arbitrary number of omics levels. In addition, a combined model was introduced fusing a two-level cooperative to the inhibitory model. All models were thoroughly evaluated with realistic synthetic data and outperformed related gene-set enrichment approaches as well as less complex MONA models. In addition, we applied MONA to a biological data set profiling adipocyte differentiation in order to reveal meaningful functional processes. To address the greatest drawback of MONA, we developed a working model for the MONA framework, which additionally infers term probabilities from p-value ranked gene sets instead of the current binary input values (differentially expressed / not differentially expressed), comprising the novel model cMONA. CMONA uses p-values as continuous observation for term inference and thus takes the strength of expression into account instead of requiring an arbitrary cutoff.

Zusammenfassung

Die Identifizierung funktionaler Eigenschaften von Genmengen, stellt eine wichtige Herausforderung im Bereich der Sytembiologie dar. Da experimentelle Hochdurchsatz-Methoden die Erforschung biologischer Funktionen über mehrere „-omics“-Level ermöglichen, wurde kürzlich ein neuer Ansatz namens Multi-Level Ontology Analysis (MONA) vorgestellt. MONA identifiziert parallel überrepräsentierte funktionelle Ontologie-Terme durch Integration von Genmengen mehrerer „omics“-Level. Diese können unter anderem mRNA- und Protein-Expressionsdaten sowie Vorwissen über andere molekulare Mechanismen enthalten, die zur Expression von Genprodukten beitragen, beispielsweise DNA-Methylierung oder microRNAs. MONA modelliert Term-zu-Gen-Beziehungen über ein Bayessches Netzwerk und handhabt damit Redundanzen und Probleme des multiplen Testens. Bis jetzt sind zwei Modelle für Genmengen in MONA implementiert, die an das Basismodel gekoppelt sind. Da diese Modelle jedoch nur eine Teilmenge möglicher Modelle darstellen, ist eine Erweiterung von MONA für die Inferenz von Term-zu-Gen-Beziehungen verschiedener „omics“-Level essentiell. Im Rahmen dieser Arbeit haben wir MONA erweitert. Dabei handelt es sich um eine flexible Erweiterung des kooperativen Modells auf eine beliebige Anzahl von „omics“-Ebenen. Darüber hinaus wurde ein kombiniertes Modell eingeführt, das ein Zwei-Ebenen-Modell mit dem inhibitorischen Modell fusioniert. Alle Modelle wurden gründlich mit realistischen synthetischen Daten evaluiert und übertrafen verwandte Ansätze sowie weniger komplexe MONA Modelle. Zusätzlich wurde MONA auf einen biologischen Datensatz angewandt, um bedeutsame funktionale Terme bezüglich der Differenzierung von Adipozyten. Der größte Nachteil von MONA ist die Nutzung binärer Eingabewerte (differentiell / nicht differentiell exprimiert). Um diesen Nachteil aufzuheben, haben wir einen Modellansatz namens cMONA entwickelt, der Term-Wahrscheinlichkeiten ausgehend von nach p-Wert sortierten Genmengen inferriert. CMONA nutzt p-Werte als kontinuierliche Beobachtungen für die Term-Inferenz und berücksichtigt folglich die Stärke der Expression, weshalb dieser Ansatz nicht von willkürlich festgelegten Signifikanzniveaus abhängig ist.

Overview

1 Introduction

Omics technologies measure biological systems on different molecular levels [1]. Measurements include for example mRNA or protein expression, DNA methylation or microRNA regulation [2]. These levels build complex functional networks of molecular interactions capturing cellular functions and pathways [3]. We typically employ omics technologies to understand and identify functional processes for the phenotypic condition under investigation [2]. This may facilitate studies on the mechanisms of likewise underlying diseases and thus can help developing corresponding drugs and treatments [4]. Different conditions can arise if cells adjust to certain signals coming from the environment, disease or even from mutations in the genome. Resulting alterations in gene expression generate a phenotypic state which is able to adjust to new conditions [5]. This can be referred to as "gene response" and is not only regulated by protein expression but also by a number of further regulatory mechanisms like mRNA expression, DNA methylation or post-transcriptional modification by microRNAs (miRNA).

The extraction of "knowledge" by investigation of gene responses poses an important challenge in the field of bioinformatics. Obtained gene expression profiles have to be interpreted to gain insights into biological mechanisms [6]. Functional annotation of genes or gene products are described in ontologies such as Gene Ontology (GO) or KEGG pathways, which represent structured vocabularies that are referred to as terms, representing biological functions, pathways, etc. for gene products [7]. For gene set analysis, a wide range of methods has been developed, including Fisher's exact test and Gene Set Enrichment Analysis (GSEA), which make use of such ontologies [8]. In doing so, gene products are mapped to their biological functions and it is determined, which gene sets are enriched between the different conditions. Existing methods have some drawbacks: On the one hand, these approaches do not consider the hierarchical GO structure, which finally implicates a large number of redundant inferred terms [9]. On the other hand, due to the GO term hierarchy, there terms cannot be statistically tested independently from each other. Therefore multiple testing corrections have to be performed retaining a prescribed family-wise error rate [10]. These issues can be solved by using model-based approaches, which try to identify a minimal set of groups that best explain the data [9]. To overcome the problem of integrated functional analysis across multiple omics, a framework was developed, named Multi-level Ontology Analysis (MONA) [2]. MONA uses a Bayesian network based approach which models a term-to-gene relationship for searching a minimal set of 'active' gene sets while taking into account measured information from multiple species [2].

Up to now, MONA has only two models implemented. The cooperative and inhibitory model, both represent only selected relationships between molecular species (e.g. miRNA post-transcriptional target inhibition) and constituting just a subset of possible models. To obtain an appropriate framework for gene set analysis, we extended the functionality of MONA in order to handle any number of omics levels for term probability inference. In addition, we introduce a concept for another extension of MONA going beyond binary values (differentially / not differentially expressed) now enabling us to use continuous gene observations, for example p-values.

Purpose of this thesis

The aim of this thesis was to systematically extend and test the functionality and flexibility of MONA in order to set up an appropriate framework for term inference. Evaluations of MONA were therefore performed not only using realistic synthetic data sets in comparison to other methods but also applied to biological data.

This thesis comprises three new features and functions for MONA: First, an extension for the cooperative model to allow for a user-set number of species to be integrated, which provides more flexibility for the integration of various gene response data. Second, both already existing models were fused into a combined model, which is called cooperative-inhibitory model, enabling the user to combine independently observed data across different molecular species, for example mRNA and methylation expression data, as well as inhibitory layer for post-transcriptional regulation of mRNA expression by miRNA.

Finally and most importantly this thesis comprises a valuable extension of MONA, named cMONA. We developed a working model for continuous model-based ontology analysis, since many methods, including MONA itself, perform gene set analysis only using binary values in gene sets. Our cMONA is the first step towards a model-based functional analysis incorporating statistical significances.

Each developed and implemented model was systematically evaluated using synthetically generated data and additionally applied to experimental data of adipocyte differentiation.

In chapter 2, we provide background information, which defines concepts and specifies methods used throughout this thesis. Chapter 3 comprises formal definitions of all developed and implemented MONA extensions including information on parameterization strategy. Chapter 4 describes biological materials and contains detailed description of the employed testing strategies. Chapter 5 gives a description of results and discussion. Chapter 6 concludes this thesis.

2 Background

2.1 Omics

The Suffix "-omics" is used to distinguish respective genome-wide studies on each molecular level. For example, genomics was the first "omics" derived to describe the genome, which encompasses studies of the entire DNA of an organism and was firstly mentioned by Hans Winkler in 1920 [11]. *Genomics* is referred to the study of genomes of organisms. Vast advances in developing methods for large-scale profiling in the past decades enabled the assessment of different molecular species. Consequently, "-omics" terms were also introduced for other subfields such as proteomics, transcriptomics or methylomics [12]. *Proteomics* comprises the study of the function of all expressed proteins, named proteome, involving for example protein-protein interactions, protein activity patterns and profiles in cancer patients [13]. Transcripts mirror the sequence of the DNA from which it was transcribed. Besides messenger RNA (mRNA), various other types of transcribed RNA exist, that are not further translated into proteins (non-coding RNA), for example transfer RNA (tRNA) and ribosomal RNA (rRNA) which are both involved in the translation process of protein biosynthesis [14] or microRNAs (miRNA), which control gene expression in plants and animals post-transcriptionally [15]. The transcriptome involves all transcripts present in a given cell and represents just a very small percentage of the genome, below 5% of the genome in humans, because only a very small part of the entire DNA is transcribed. In research, *transcriptomics* is used for example if one wants to determine when, where and how strong certain genes are turned on or off in cells or tissues. This information can for example indicate the amount of gene activity in both health and disease and thus lead to a deeper understanding of the contribution of gene activity to disease. Another mechanism contributing to gene expression is represented by DNA methylation [16]. DNA methylation involves the transfer of a methyl group to the C-5 position of the cytosine ring of DNA by the enzyme methyltransferase and states an important epigenetic layer that is involved in cellular differentiation processes and control of transcriptional potential. Methylation is a stable modification of genomic DNA and thus can be inherited. It dynamically changes during lifespan of cells and tissues and is susceptible to diet and environmental influences [1]. Thus, the methylome containing all DNA methylation sites for an organism can provide insights into the evolutionary history of DNA methylation as well as its dysregulation in certain disease states.

2.2 Ontologies

Gene ontologies describe relationships of each respective protein typically in a hierarchical and species-independent manner. Individual genes are then mapped to an ontology term in an ontology to better classify its functions. The most commonly used ontology is the Gene Ontology.

Gene Ontology (GO)

GO comprises three different vocabularies (ontologies), namely biological process, molecular function and cellular component, to describe features and properties of gene products by so called GO terms [7]. Each GO term within the ontology has its own name, unique identifier and definition indicating the category to which it belongs. GO terms are structured in a hierarchical manner; child terms being more specific and parent terms being more general. The structure can be described by a directed acyclic graph of which each term is represented by a node and a relationship between two terms by a directed edge. GO is not strict hierarchy because a term node can have more than one parent node.

KEGG PATHWAY

The KEGG PATHWAY database contains manually curated pathway maps which represent molecular interaction and reaction networks including metabolism, cellular signaling processes, organismal systems and human diseases [17]. It represents the biological system and its components on different levels like genes, proteins and chemical substances in combination with information about their relationships. Information about diseases and drugs are also provided in this database.

WikiPathways

WikiPathways is an open platform comprising biological pathways, which provide views of interactions of underlying processes [18]. Each pathway has its own dedicated wiki page containing useful information including the current diagram, description and references. Moreover, the collections of pathways can be browsed with combinations of species names and ontology-based categories.

2.3 Gene expression analysis

Different environmental conditions or disease states can influence the development of phenotypes by alterations in gene expression on different molecular levels like mRNA expression, methylation states or even by post-transcriptional modification by microRNAs [2]. The quantitative analysis of gene expression has become an integral part of most modern biological investigations. A technique often applied in gene expression analysis is microarray technology [19]. Microarrays are able to simultaneously capture variations in gene sequence or expression by hybridization of labeled DNA targets to a very large set of oligonucleotide probes [20]. This enables description of genome-wide expression changes. A microarray expression analysis typically results in a long list of differentially expressed genes, which is the starting point of further functional analysis, including gene set enrichment, which it is aimed to find patterns for differential expression [21].

2.4 Statistics

Moderated t-statistic with limma

On the assumption that gene expression is altered under e.g. different environmental conditions, information obtained from microarray experiments is derived in order to score genes according to their strength of differential expression [22]. For this purpose, Smyth developed the moderated t statistic, which uses local regression to determine significance for each gene with respect to its expression by fitting a linear model to the expression data. The level of different expression of each gene is represented by its respective log-fold-change. The more the respective log-fold-change differs from zero, the more significant the magnitude of different expression for a gene. Relative to a minimum log-fold-change cutoff, a p-value can be computed for each gene. P-values and/or log-fold-change can be then used as arbitrary cutoffs to determine differentially expressed genes.

Receiver operating characteristics (ROC) and Area under the receiver operator curve (AUC)

A receiver operating characteristic (ROC) curve visualizes the performance of a benchmark test by plotting the true positive rate (sensitivity) against the false positive rate (sensitivity) across varying cutoffs [23]. It illustrates the tradeoff between sensitivity and specificity. The overall accuracy is measured by the area under the ROC curve (AUC). It summarizes the entire location of the ROC curve and is an effective measure of sensitivity and specificity that describes the inherent validity of diagnostic tests. The closer its curve follows the upper left-hand corner of its ROC space, the more accurate the test and thus the higher its

AUC. If the curve lies on the diagonal of the ROC space, the diagnostic test gives random guesses. ROC curves are appropriate for comparing two or more alternative statistic tests applied to the same data or finding the optimal cut off values.

Goodness of Fit test

Starting from observed data, a Goodness of Fit test measures the "distance" between the data and the distribution or model, which is tested. The resulting distance is then compared to some threshold value. If the distance is below this threshold, which is also referred to as the critical threshold, the fit is considered as good. The Chi Square test is a famous example of a goodness of fit test which can be applied to any univariate, discrete distribution for which the cumulative distribution function can be calculated [24]. The probability curve of a chi-square distribution is an asymmetric curve and has only one parameter, k, which is a positive integer specifying the number of degrees of freedom. The null hypothesis, for which the test is defined, assumes that the data follow a specified distribution. The opposite holds for the alternative hypothesis.

2.5 Bayesian networks

Bayesian networks are directed acyclic graphs (DAG), containing nodes being random variables and directed edges representing probabilistic dependencies among the random variables [25]. A node without parents follows an unconditional probability; otherwise it follows a conditional probabilistic distribution, which is determined by its parent nodes. Using an assumption of conditional independence, it is able to efficiently infer posterior probabilities of variables by computing the joint distribution over a set of random variables. Consequently, it contains information to compute any probability of interest. This feature makes Bayesian networks a favorite tool for areas such as machine learning or text mining.

2.6 Functional Analysis

2.6.1 Gene set enrichment by Fisher's exact test

Fisher's exact test constitutes an application for functional analysis of large gene lists derived from high-throughput experiments [26] with respect to sets of genes rather than to individual genes [27]. Starting from a gene list, a measure of differential expression is calculated for each gene, usually a p-value from a t-test. This measure is used as cutoff for separating the gene list into differentially expressed and non-differentially expressed genes. Fisher's exact test makes use of the hypergeometric distribution and takes the size of the overlap between the gene set and the list of differentially expressed genes using a 2 x 2 contingency table. The table simply counts the number of genes on the microarray with every possible combination of the binary attributes 'differentially expressed' and 'in the gene set' [21]. Fisher's exact test examines the relationship between the two dimensions of the table by calculating the p-value for overrepresentation of the gene set among the differentially expressed genes using a test for independence. The null hypothesis formulates that a term is not active whereas the alternative hypothesis states that a term is active. Fisher's exact test determines the null distribution by randomly reassigning genes to the labels for being in the gene set and for being differentially expressed [21]. If the out coming p-value is small, the null hypothesis of the respective term being off is rejected [28].

2.6.2 Model-based enrichment analysis

Bayesian networks can be used in model-based approaches of gene set enrichment analysis to model the data with all categories simultaneously for identifying biological categories, which are overrepresented. Model-based Gene Set Analysis (MGSA) is an example for such an approach [29]. It aims to identify a set which comprises a minimal number of categories which is overrepresented in the given data by mapping genes to their respective categories within a Bayesian network.

By including prior knowledge, this approach is able to infer posterior probabilities of biological categories and parameters in order to identify ‘active’ groups. Using a model-based approach instead of hypothesis testing on each category separately (as for example done by Fisher’s exact test), it overcomes multiple testing problems and additionally avoids the prediction of term redundancies. Thus, this model-based approach provides a more accurate method for the analysis of high-throughput data.

2.6.3 Multi-level Ontology Analysis (MONA)

MONA is a model-based framework using a Bayesian network to infer term probabilities by integrating data from different “omics” [2]. While other methods are just capable of integrating a single species for the inference of term probabilities, MONA is able to combine multiple species response data. Simultaneously, it overcomes issues like the multiple testing problem and handling redundant term predictions which could be problematic using GO as ontology. MONA integrates multi-level omics data into a base model and also handles any combination of molecular levels.

Base model

The base model of MONA (see Figure 1, a) comprises a Bayesian network, which consists of a term layer (blue) representing the ontology terms and a hidden layer (green), which stands for the hidden gene response [2]. As displayed in the figure, the nodes of the term layer are mapped to one or more nodes of the hidden layer as, for example, defined by GO. In addition, the base model is defined by conditional probabilities: The term layer includes Boolean nodes which are Bernoulli- distributed and modeled by a probability p to be on.

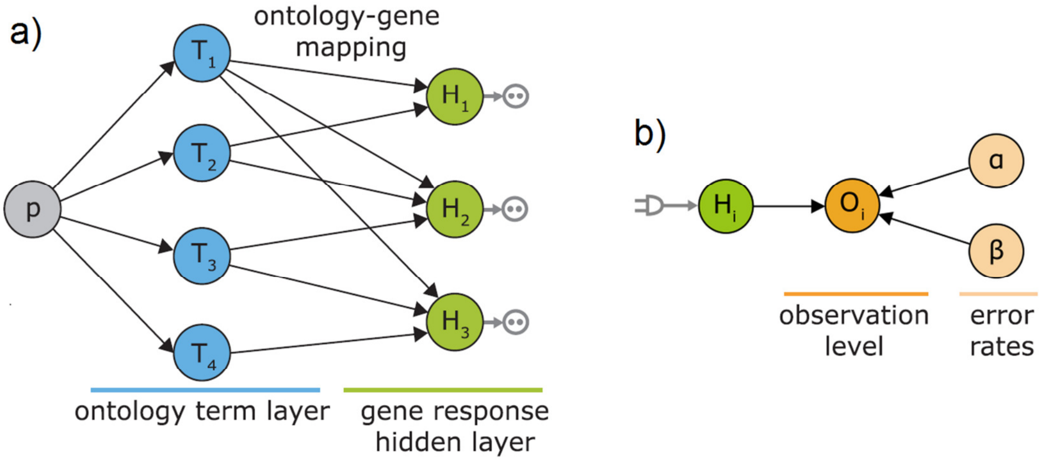


Figure 1: a) The base model: Hidden nodes (green) of the gene response which are observed by noisy measurements are attached by nodes representing the ontologies (blue). A prior p defines the probability for the terms to be on or off. b) Single-species model: observations O_i (orange) in terms of measurements from a single species are connected to exactly a node H_i of the hidden gene response. Furthermore, each observation node is attached by a false positive error rate (α_i) and false negative error rate (β_i). The figure was adopted from [2].

The hidden terms are also Boolean nodes and defined to be on if at least one of the term nodes, to which they are connected, is on. Let T be the used ontology terms, H_i a certain gene and $T(H_i)$ a set of terms which are connected to H_i. Then the following holds:

$$P(H_i|T) = \begin{cases} 1 & if\ \exists T_j\ \in T(H_i) : T_j = 1 \\ 0 & otherwise \end{cases}$$

The single species model, which is displayed in Figure 1 b), allows a single species to be observed. Observations in terms of noisy measurements are contained in the observation layer. Each observation node O_i is connected to exactly one node H_i of the hidden gene response, respectively. Moreover, each observation node possesses its own error rates α and β which represent false positive and false negative error rates. Let O_i be an observed gene measurement and H_i the corresponding gene response. Then the following holds:

$$P(O_i = 1|H_i) = \begin{cases} 1 - \alpha & if\ H_i = 1\ (true\ positive) \\ \alpha & if\ H_i = 0\ (false\ positive) \end{cases}$$

$$P(O_i = 0|H_i) = \begin{cases} 1 - \beta & if\ H_i = 0\ (true\ negative) \\ \beta & if\ H_i = 1\ (false\ negative) \end{cases}$$

Cooperative and inhibitory model

In addition to the single-level model, a cooperative and an inhibitory model are implemented in MONA. This enables the integration of measurements from different species in an independent or dependent manner [2].

In contrast to the single-species model, the cooperative model allows for two species, which can be considered as independent measurements of a common underlying gene response (see Figure 2), which might be for example changes in gene expression or DNA methylation [2]. Each observation node of the species is connected to exactly one hidden node, like it is the case for the single species model. As the measurements are noisy, each observation node has got its own false positive and false negative rate (α^I, α^{II} and $\beta^I,\ \beta^{II}$ for species *I* and *II*). The error rates for each of the species are defined as described in the base model.

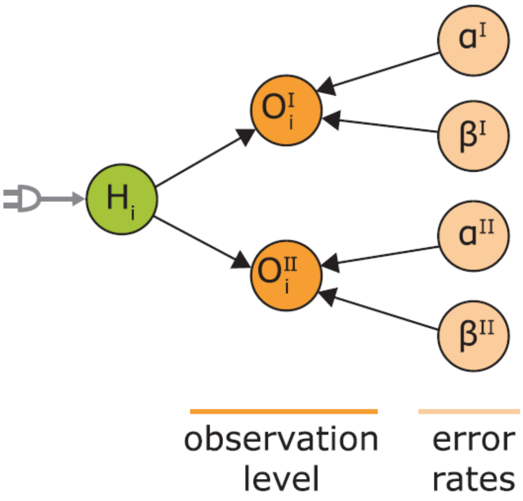


Figure 2: The cooperative model of MONA: The observation layer comprises independent noisy measurements from two species, for example mRNA and methylation. Each observed species has its respective false positive (α^{I}, α^{II}) and false negative error rate (β^{I}, β^{II}). The figure was adopted from [2].

The inhibitory model can be used in case of measurements, where one species is considered to act as an inhibitor of another species [2]. An example for this is represented by posttranscriptional modulation of mRNA by miRNA. According to that, one of the observed species is referred to as inhibited species, the other as inhibitor species. For the inhibitory observations, an additional hidden node $H_i^{I,inh}$ is introduced for every gene response H_i separately, like described by Figure 3. $H_i^{I,inh}$ Is a Boolean random variable which describes the state of the inhibitor (for example miRNA): If the inhibitor is active, then $H_i^{I,inh} = 1$, otherwise $H_i^{I,inh} = 0$. p^{inh} is the probability for $H_i^{I,inh}$ to be active. $O_i^{I,inh}$ states the observation of $H_i^{I,inh}$ and has its own false positive and false negative error rates. While the error rates for the inhibitor species $O_i^{I,inh}$ are defined as described above in the base model, following error rates hold for the inhibited species O_i^{I}:

$$P\left(O_i^I = 1 \middle| H_i^{I,inh}, H_i\right) = \begin{cases} 1-\alpha^I & if\,(H_i^{I,inh} = 0 \wedge H_i = 1) \vee \left(H_i^{I,inh} = 1 \wedge H_i = 0\right)(TP) \\ \alpha^I & if\,(H_i^{I,inh} = 1 \wedge H_i = 1) \vee (H_i^{I,inh} = 0 \wedge H_i = 0)(FP) \end{cases}$$

$$P\left(O_i^I = 0 \middle| H_i^{I,inh}, H_i\right) = \begin{cases} 1-\beta^I & if\,(H_i^{I,inh} = 1 \wedge H_i = 1) \vee \left(H_i^{I,inh} = 0 \wedge H_i = 0\right)(TN) \\ \beta^I & if\,(H_i^{I,inh} = 0 \wedge H_i = 1) \vee (H_i^{I,inh} = 0 \wedge H_i = 1)(FN) \end{cases}$$

A true gene response is reflected by an active inhibitor and an inactive inhibited species or by an inactive inhibitor and an active (which means non-inhibited) species.

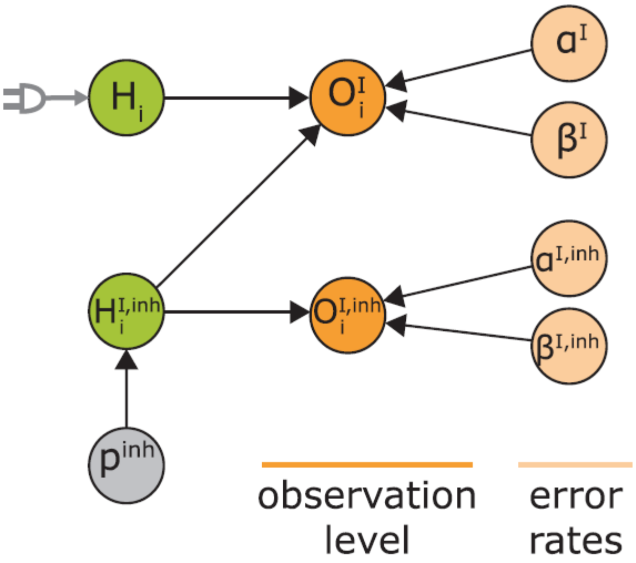


Figure 3: The inhibitory model of MONA: Two observed species which cannot be considered as independent measurements, for example post-translational modification of mRNA by miRNA. In this case, O_i^I represents the observed mRNA and $O_i^{I,inh}$ represents the observed miRNA. The observed mRNA node O_i^I is not only connected to both the hidden gene response of mRNA H_i but also to the hidden gene response of its inhibitor $H_i^{I,inh}$ which indicates that the expression of miRNA influences mRNA expression. p^{inh} is the probability for $H_i^{I,inh}$ to be active Apart from that $O_i^{I,inh}$ is

connected to its hidden gene response $H_i^{I,inh}$. Both the inhibitor and inhibited observed species have their own false positive and false negative error rates. The figure was adopted from [2].

2.6.4 2D enrichment integrating quantitative high-throughput data

The method 2D enrichment, a method, which was presented by Cox et al., aims to identify categories from genes, which show a consistent behavior between any two different omics data [30]. In this context, for each category it is tested whether its numerical value (for example p-value or fold-change), which was previously determined by a non-parametric test such as a multivariate analysis of variance (MANOVA), significantly deviates from the general global distribution of the data. This is done in order to be independent of the shape of the distribution from the numerical values. To test if a particular population tends to have larger values than another, the MANOVA test statistic compares the means of several groups, as showed in (1), regarding two groups in two dimensions:

$$\frac{s_{xx}d_y^2 + s_{yy}d_x^2 - 2s_{xy}d_xd_y}{s_{xx}s_{yy} - s_{xy}^2} \qquad (1)$$

where $d_x = \bar{x}_1 - \bar{x}_2\ and\ d_y = \bar{y}_1 - \bar{y}_2$ are the differences of the group means between group 1 and 2 in the x and y coordinated, respectively and s_{xx}, s_{yy} and s_{xy} are the summed squares of the deviations from the group means for x, y and mixed coordinates [30]. The result of the MANOVA test is defined as the 2D annotation enrichment p-values. The difference of average ranks of the significantly deviating annotations is characterized by a s-score, which is represented by a tupel consisting of two numbers,(s_x, s_y), where $-1 \leq s_x \leq 1$ and $-1 \leq s_y \leq 1$. Consequently, significant terms will avoid a circular region around the origin, as Figure 4 displays. The remaining parts can be divided into correlating, non-correlating and anti-correlating regions. In summary, it can be stated that the 2D enrichment provides a ‘no-cutoff’ method handling two ‘omics’ dimensions simultaneously, meaning that it is not necessary to define sets of regulated genes or proteins in advance, thereby reducing arbitrary factors. In contrast to other enrichment method, the 2D enrichment does not need a reference set to calculate scores regarding a set of genes or proteins relative to all genes in the genome and thus, for instance, avoids use of biased data in consequence of incomplete databases.

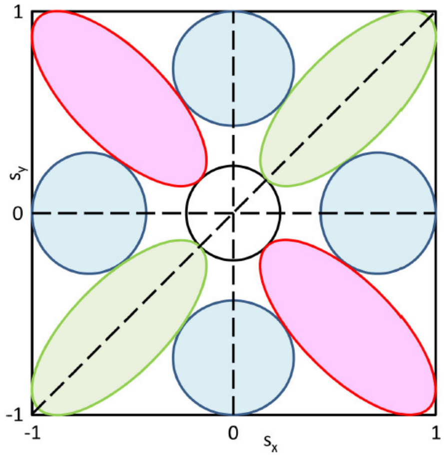


Figure 4: Schematic representation of the 2D annotation enrichment score, which is represented by a number pair inside the displayed rectangle. Significant terms will avoid the circular region around the origin. The green regions correspond to concordant up or down regulation. The blue regions correspond to terms that are up or down regulated in one direction, but not in the other, while the terms in the red regions show anti-correlating behavior. The figure was adopted from [30].

2.7 MiRNA-target relationship

Beside other regulatory mechanisms, gene expression in plants and animals is post-transcriptionally controlled by non-coding RNAs. Short non-coding RNAs are for example miRNAs, which interact with complementary sites of the mRNA’s 3’ untranslated region leading to an induced cleavage or to repression of productive translation [15]. While it has been shown that miRNAs are involved in stress signaling and diseases [31], their impact on distinct biological pathways and phenotypes remains largely unknown [32].

2.7.1 Identifying miRNA-target relationships

The identification of miRNA targets relies strongly on *in silico* predictions of miRNA seed regions and target sequences [33]. Public databases of validated miRNA-target pairs benefit from recent technological advances [34]. Several tools for miRNA-target prediction are presented in this section.

TargetScan

TargetScan predicts miRNA targets by looking for 8mer and 7mer sites matching the seed region of each miRNA [35]. Nonconserved sites as well as sites with mismatches in the seed region can optionally be identified and predicted. Predicted targets of mammalian miRNAs are ranked by predicted efficacy of targeting using context+ scores of the site, which is defined as the sum of the contribution of overall six features: Site-type contribution, 3' pairing contribution, local AU contribution, position contribution, target site abundance contribution and seed-pairing contribution. If a single miRNA is chosen as the representative miRNA for a specific targeted gene, all other miRNAs belonging to the same family are also predicted to target the same gene.

StarBase

Beside a variation of features like decoding Pan-Cancer and Interaction Networks, the identification of miRNA target interactions is another main feature of StarBase [36]. For this purpose, miRNA cleavage sites are predicted from CLIP-Seq and Degradome-Seq data from six organisms. To identify target clusters in animal and plants, respectively, six miRNA target prediction tools are used, including TargetScan, PicTar and miRanda. To discover new miRNA target sites from CLIP-Seq and Degradome-Seq, two web servers were provided.

MiRanda

MiRanda is an algorithm for detecting potential miRNA target sites in genomic sequences [37]. It is able to read RNA sequences and genomic DNA sequences from files in FASTA format for identifying potential target sites using a dynamic programming local alignment between query miRNA sequence and reference sequence. However, this algorithm uses scores which are based on sequence complementarity instead of sequence identity. Using the resulting alignment, the second part of the MiRanda algorithm estimates the thermodynamic stability of RNA duplexes. A fictional single-stranded RNA composed of the query sequence, a linker and the reference sequence is generated and is then used to calculate its minimum free energy.

MiRTarBase

MiRTarBase is an experimentally validated miRNA-target database containing interactions which were collected by data mining and manually survey of pertinent literature articles related to functional studies of miRNAs [38]. The collected miRNA-target interactions are validated experimentally using reporter assay, western blot, microarray and next-generation sequencing experiments. The database provides the most updated collection by comparing with other similar previously developed databases.

2.7.2 MiRlastic

As sequence-based prediction approaches predict a high number of false positives, it would be reasonable to combine them with expression data to obtain more accurate miRNA target relationships. To overcome this issue, a method called MiRlastic was developed which enables the construction of miRNA-mRNA regulatory networks representing potential regulatory relationships between mRNAs and miRNAs. It is assumed that miRNAs which are highly correlated in expression data tend to be functionally related and thus are coexpressed [39].

Thus, MiRlastic aims to retain groups of miRNAs that are more likely correlated with each other as compared to randomly sampled miRNAs [40]. Consequently, it includes biological properties in order to predict miRNA target relationships using a regression-based feature selection and an elastic net penalty which identifies all associations which are explained by measured expression values. The Elastic net penalty is based on a combined penalty of the least absolute shrinkage and selection operator (lasso) and ridge regression penalties [41]. It simultaneously performs automatic variable selection and continuous shrinkage. In some situations, the lasso approach produces unsatisfactory results and thus does not constitute a reliable method as it does not retain correlated variables. In contrast to lasso, elastic net is able to select groups of correlated variables, meaning that strongly correlated predictors tend to be in or out of the model, and thus prevents a loss of information [41]. In addition, ridge expression provides no proper solution since it performs no feature selection on the data. For modeling miRNA-mRNA relationships, elastic net is used as regression model to keep meaningful correlated miRNA predictors while excluding miRNAs with insufficient effect on the mRNA response. Using a balance between the lasso and ridge regression penalties, MiRlastic outperforms other multivariate regression models in multiple genomic datasets and suits biological understanding of miRNA-mRNA interactions [41].

2.8 Software

C# and Infer.net

MONA is implemented in C# using the Infer.NET framework (http://research.microsoft.com/infernet). Infer.NET can be used for a number of different problems like machine learning approaches, classifications or clustering. It also allows for running Bayesian inference on graphical models. All extensions of MONA were implemented based on C# and Infer.net.

R

The generation of the synthetic data sets, the evaluation, ROC curves and AUC, as well as all visualizations (including heatmaps, trees, volcano plots) for this thesis were performed using the R language [42]. R is a free software environment for statistical computing and graphics.

3 Methods: Extending MONA

For the purpose of obtaining an extended functionality and flexibility of multi-level ontology analysis, following three models were implemented in MONA:

1. The cooperative model was adjusted to handle any number of species for ontology term inference.
2. A combination of the cooperative and the inhibitory models was implemented to allow for the investigation of measurements that can be interpreted as both independent and dependent.
3. A working model was derived handling continuous observations in terms of values, which are obtained from differential expression analysis. Essential work for the analysis of p-value distribution was subsequently performed.

The following sections provide motivation and specification of each implemented extension. Evaluation strategy and results are described subsequently in Chapters 3 and 4, respectively.

3.1 Extended cooperative model

Up to this point, the cooperative model allowed just for two species as observations, for example combined measurements of mRNA expression and DNA methylation. In case of more than two species, the cooperative model could not be used as "plug and play" because the model first had to be adapted by modifying the implementation and then had to be recompiled before applying MONA to the data. In order to overcome these costly procedures the cooperative model was adjusted to allow for optional numbers. The resulting extended cooperative model is illustrated in Figure 5. An arbitrary number of N independent species can be handled which may be regarded as noisy measurements of an underlying gene response. Given N species, each observation node O_i^j $(j = 1, \dots, N)$ is connected to its respective hidden node H_i of the gene response and has its own false positive and false negative rates α^j and β^j, which are defined below.

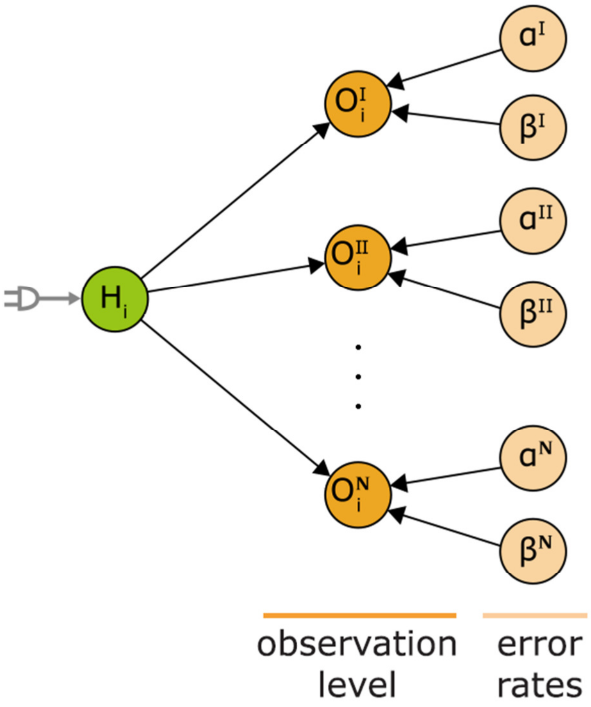


Figure 5: Extended Cooperative Model: Observations of an optional number N of species, which are coupled to an underlying hidden gene response node H_i. Each species j has its own false positive and false negative rates $\alpha_{\,i}$ and $\beta_{\,i}$. Note that each node H_i of the hidden layer is connected to an ontology term. The figure was reproduced from [2].

For any hidden node H_i, the species j is modeled as an independent observation $O_i^j$of gene response, which again depends on the respective false positive (α^j) and false negative (β^j) error rates. The error rates are then defined as:

$$P\left(O_i^j = 1 \middle| H_i\right) = \begin{cases} 1-\alpha^j & if\ H_i = 1 \\ \alpha^j & if\ H_i = 0 \end{cases}$$

$$P\left(O_i^j = 0 \middle| H_i\right) = \begin{cases} 1-\beta^j & if\ H_i = 0 \\ \beta^j & if\ H_i = 1 \end{cases}$$

For parameterization, uniform priors were chosen meaning all values in the domain of the distribution have equal density. Thus the shape parameters α^j and β^j of the Beta priors for the terms being active as well as for the false positive and false negative rates were set to 1.

In addition, Fisher's exact test was performed on the given synthetic dataset. Note that Fisher's exact test is just applicable for a single species. To compare the results of the different approaches, a receiver-operating-characteristic analysis (ROC curve) based on the p-values for all datasets together with the area under the curve (AUC) was performed.

3.2 Combined cooperative inhibitory model

The cooperative and the inhibitory model of MONA allow for the separate gene set analysis of independent or dependent measurements, respectively. Both models were combined, resulting in a cooperative inhibitory model, achieving an inference using both types of measurements. Figure 6 illustrates the cooperative-inhibitory model which connects a cooperative part (blue) to an inhibitory part (orange). The nodes of the observation layer consist of measurements from the first species (O_i^I), second species (O_i^{II}) and a species ($O_i^{I,inh}$), which is able to regulate the first species by inhibition. The first species is connected to both, the gene response H_i, which is modeled by independent measurement, and the gene response $H_i^{I,inh}$ of the inhibitory part, which is not regarded as independent. Furthermore, the second species is connected to H_i and the inhibitory species to $H_i^{I,inh}$, which has a probability p^{inh} of being active. Likewise the false positive and false negative error rates are chosen as for the cooperative and the inhibitory model described above.

The cooperative-inhibitory model can for example be applied to combined mRNA expression, DNA methylation and miRNA expression data.

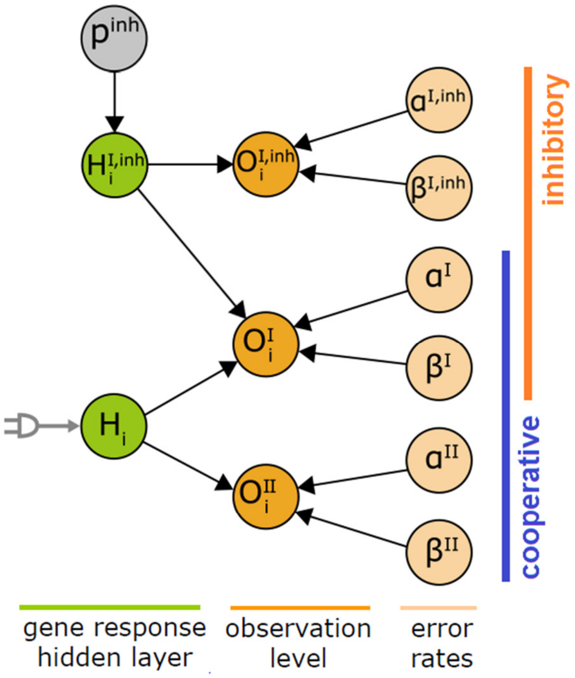


Figure 6: The cooperative inhibitory model fuses a cooperative part (blue), which allows the observation of two independent measurements with an inhibitory part (orange), which comprises the measurement of an inhibitor and an inhibited species. The second observed node O_i^{II} of the cooperative part is just connected to the hidden node H_i while the observed node O_i^{I} of the first species is connected to both H_i of the cooperative part and to $H_i^{I,inh}$ of the inhibitory part. The inhibitor species $O_i^{I,inh}$ is only connected to the node of its own hidden gene response which has the probability p^{inh} to be active. All nodes of the observation layer have their own corresponding false positive and false negative error rates. The figure was reproduced from [2].

In order to determine appropriate shape parameters a and b of the Beta prior probability p for the terms to be active, different values for a and b were used for evaluating the cooperative-inhibitory model. Results are described in Section 5.2.

3.3 cMONA - Continuous MONA for single species

Up to now, MONA infers ontology terms by using binary data as observations, which indicate whether the respective genes are differentially expressed or not. The aim is to implement a model, which allows using continuous input values and thereby considering the strength of differential expression by avoiding arbitrary thresholding of significance values. Meaningful continuous values are represented by p-values obtained from significance testing, e.g. by a moderated t statistic, meaning that each measurement has a respective p-value.

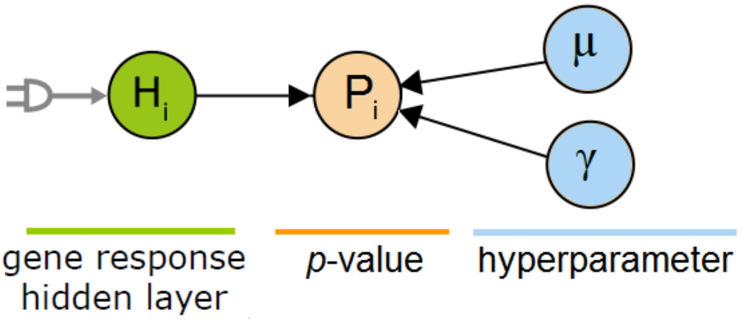


Figure 7: Current working model of cMONA allowing for continuous input values, represented by p-values. A p-value P_i is directly coupled to exactly one node H_i of the hidden gene response. μ and γ are hyperparameters specifying the error rates for Pi by representing mean and precision of the p-value distribution.

According to these requirements for a continuous MONA (cMONA), we propose a working model for single species (Figure 7) as follows. P-values are introduced as observations P_i (orange) from a single species and are connected to exactly a node H_i of the hidden gene response. The P_i observation level is replacing the binary $O_i$observation level from MONA. As p-values contain false positives and false negative errors, corresponding error rates have to be included. (Binary) MONA uses two error rates, which are coupled to its observation nodes O_i, considering false positive and false negative measurements and are represented by two Beta distributions α and β. However, in contrast to α and β, these error rates are not Beta distributed, and thus are replaced by the global hyperparameters μ

and γ, which represent the mean as well as the precision of the p-value distribution. Therefore, we have to derive distribution assumption for the p-values, which is a non-trivial task. Trimmer et al. presented a method for the determination of differentially expressed genes from set of genes without requiring any p-value threshold [43]. According to that, it was aimed to estimate the number of differentially expressed genes based on the p-value distribution, which was divided into a uniform distribution and an unknown, alternative distribution.

Consequently, we assume that p-values obtained from differentially expressed genes are log-log-normally distributed, otherwise uniformly distributed. If the hidden gene response $H_i = 1$, we expect that the respective p-value P_i comes from a lognormal distribution, if $H_i = 0$, we expect P_i to be uniformly distributed:

$$P(P_i|H_i) \sim \begin{cases} Log - Log - N(\mu, \gamma) & if\ H_i = 1 \\ Uniform & if\ H_i = 0 \end{cases}$$

In the context of this thesis, the model could not be fully implemented in MONA. Instead, a generic model for p-value distribution was implemented in Infer.net, which estimates mean and variance of p-values in a transformed log-log-space.

Analysis towards a generic model for p-value distribution

A fundamental step towards implementation of the generic model was to fit a mixture model estimating the distribution of given p-values which indicate the significance of differential expression. Concerning this, a set of genes can be divided into:

$$H0: No\ difference\ in\ gene\ expression$$

$$H1: Difference\ in\ gene\ expression$$

P-values under the null hypothesis which are obtained by multiple significance tests follow a uniform distribution [43]. In contrast, p-values under H1 are expected to have a very low p-value, for example 0.1 or 0.05 and thus accumulate around zero. Such p-values follow an alternative distribution, which is assumed to be lognormal. Consequently, the p-values follow a mixture distribution comprising uniform and lognormal.

A prerequisite for generating realistic synthetic datasets for the evaluation of cMONA is represented by modeling the p-value distribution. Therefore, a generic model was implemented in Infer.net which infers the parameters for the lognormal distribution given the fraction of genes under the null hypothesis. The model takes raw p-values as input parameter and transforms them into a log-log-space as it is expected that the transformed p-values follow a normal distribution. Based on the proportion of null p-values which are also given as input, the model infers mean and precision of the p-value distribution. The inferred parameters are then used to sample the lognormal-uniform mixture distribution. The generic model was applied to p-values obtained from mRNA expression data (chapter 4.2) as well as p-values obtained from ten independent data sets from the Gene Expression Omnibus.

4 Materials and Testing Strategy

The following chapter comprises data sets which were used to evaluate and compare results of the newly implemented models as well as their application for ontology term inference. Furthermore, it describes methods of functional analysis performed as preprocessing step for multi-level ontology analysis in order to identify differentially expressed genes.

4.1 Synthetic Data Generation

Both the new implemented and already existing models of MONA were evaluated using synthetically generated data using the GO. For this purpose, a number of non-redundant GO biological process terms were randomly sampled containing between five and hundred genes. Non-redundant means, that for every term in a set, it holds that none of the remaining terms is either a parent or an offspring of this term. Afterwards, the sampled terms were mapped to their corresponding genes in order to create the gene sets for the evaluation. This was done according to the different models respectively and is described in the corresponding sections. Furthermore, false positives and false negatives were introduced in the data with a probability of 25% respectively. Overall twenty synthetic data sets were generated for the evaluation of the extended cooperative model, the cooperative-inhibitory model and cMONA.

MONA and cMONA were systematically evaluated with synthetic and biological data sets. Performance was quantified consistently with Receiver Operating Curves (ROC) and respective area under the ROC curve (AUC). A term was considered to be active, if its posterior probability is above 0.5.

4.2 Biological Dataset

Besides synthetic data, expression measurements from mRNA, miRNA and methylation levels are used in this thesis for multi-level inference of biological processes involved in adipocyte differentiation. These datasets were generated by isolating preadipocytes from obese patients (with a BMI between 43 and 70) at the group of Harald Staiger (Universitätsklinikum Tübingen) kindly provided by Steffen Sass for this thesis. After isolation, the cells were differentiated to adipocytes *in vitro* within 20 days. Measurements were taken of the preadipocytes (day 0) and the fully differentiated mature adipocytes (day 20). For mRNA profiling the Affymetrix GeneChip Human Gene 1.0 ST Array, for miRNA profiling the Affymetrix GeneChip miRNA 2.0 Array and for DNA methylation the Illumina Infinium HumanMethylation450 Array was used. All these experiments were conducted in the group of Johannes Beckers (Institute of Experimental Genetics at the Helmholtz Center Munich).

4.3 Statistical Analysis of mRNA expression, methylation and miRNA expression data

The biological data sets were preprocessed meaning that active gene sets had to be determined. For this purpose, the Bioconductor library *limma* was used which provides functionalities with respect to analysis of gene expression microarray data, especially for assessment of differential expression.

For the determination of both mRNAs and methylation sites that are differentially expressed between preadipocytes and mature adipocytes, a moderated t statistic was first performed, which indicates the significance of differential expression. Figure 8 shows the heatmaps for the corresponding samples derived from preadipocytes and adipocytes of mRNA and methylation site measurements, respectively. The columns of samples derived from preadipocytes are marked blue while columns of samples from adipocytes are marked orange. The coloring of each heatmap indicates the row scaled expression values from row expression (white) over medium expression (light blue) to high expression (dark blue). It is observable that both species show differential expression patterns. The differences between the expression values between preadipocytes and adipocytes are stronger for measured mRNAs than for methylation sites.

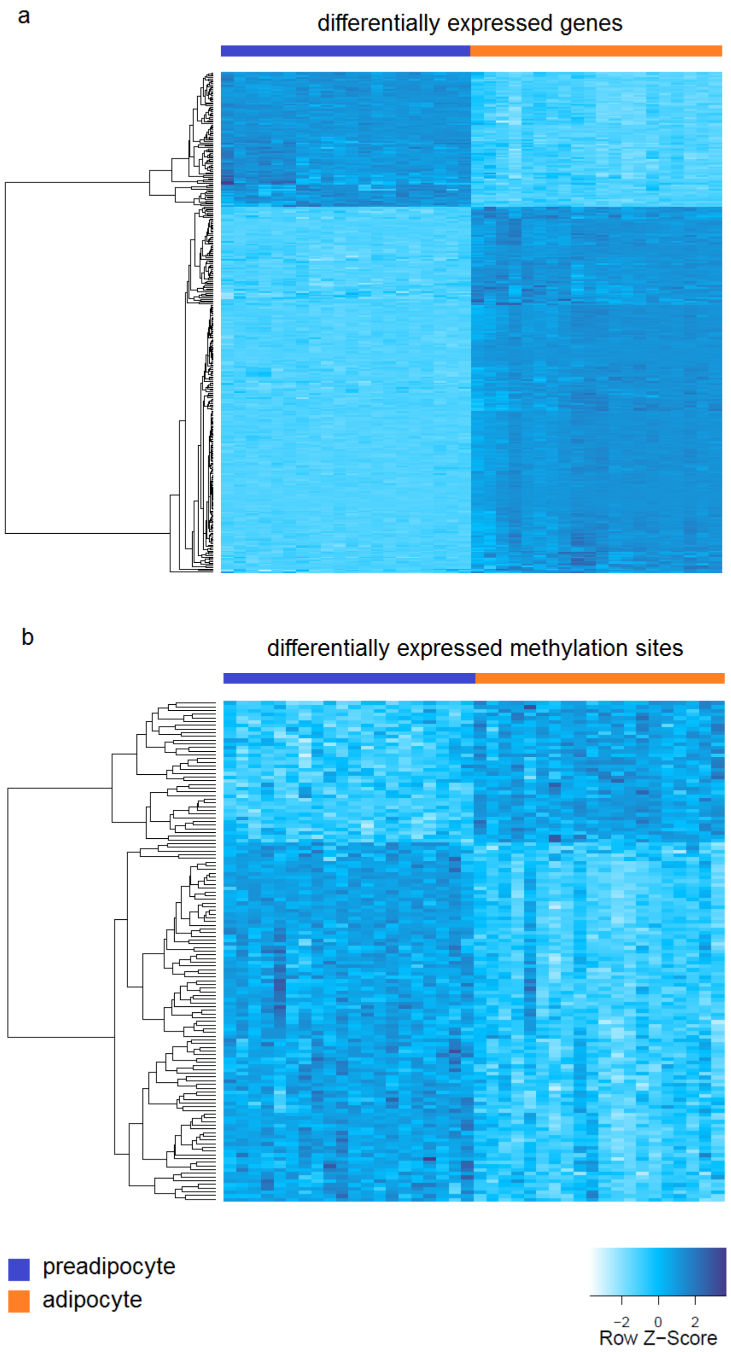


Figure 8: Heatmap showing (a) differentially expressed genes and (b) differentially expressed methylation sites between preadipocytes (blue column) and adipocytes (orange column). The gene expression and methylation values were standardized row-wise, respectively. Low values are indicated in white, middle in light blue whereas high values are colored in darker blue. The clustering of mRNAs is based on the correlation of mRNA and sample expression profiles, respectively. The same holds true for the methylation sites.

To determine genes which are considered as differentially expressed, thresholds had to be defined. For this purpose, volcano plots, showed in Figure 9, were created and used to identify changes in mRNA and methylation measurements. Each plot shows significance as a function of log-change. Based on these plots, it was decided to choose both an adjusted p-value cutoff and a log fold change cutoff and for methylation data, due to the number of data points with low p-values and high log fold changes, only an adjusted p-value cutoff.

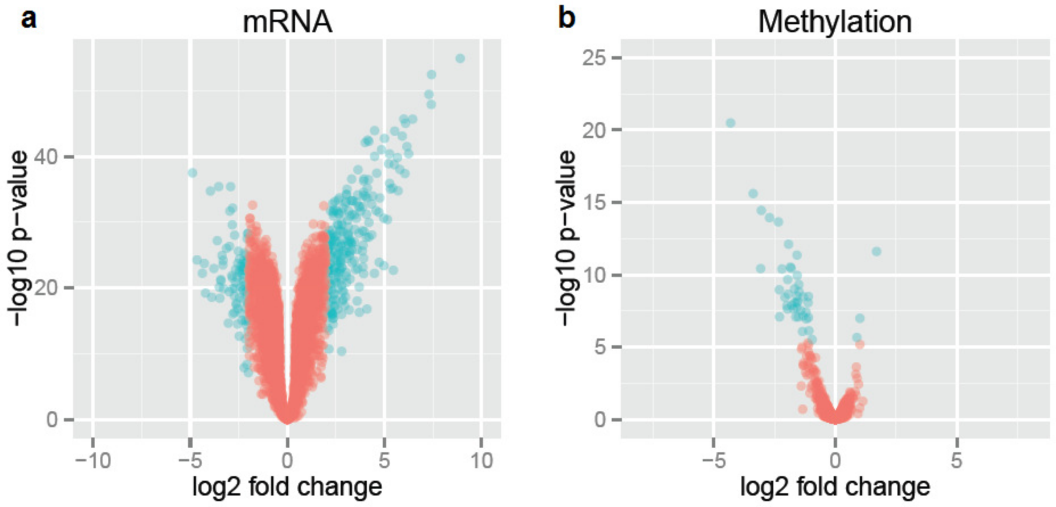


Figure 9: Volcano plots showing the chosen thresholds for defining gene sets for mRNA and methylation measurements: (a) 336 (cyan) of overall 19878 (rose) mRNAs with both an adjusted p-value below 0.05 and a log fold change above 2 are considered as significantly differentially expressed. (b) 66 (cyan) of overall 9278 (rose) methylation sites with an adjusted p-value below 0.05 are considered as significantly differentially expressed.

Consequently, a measured mRNA was defined to be considered significantly differentially expressed if its adjusted p-value was below 0.05 and its mean expression (log fold change) was 2-fold up- or down-regulated. Methylation sites were considered as significantly differentially expressed if their adjusted p-value was below 0.05. Starting from these chosen cutoffs, 336 mRNAs and 66 methylation sites were defined to be differentially expressed.

The third level of the biological data set, consisting of measured miRNAs, was preprocessed using the MiRlastic approach (described in 2.7). In order to determine significantly differentially expressed miRNAs, a moderated t statistic was first calculated as it was done for the mRNA and methylation sites data. The resulting p-values were adjusted by applying Bonferroni p-value correction for multiple testing. Based on these statistics, a miRNA was assumed to be differentially expressed if its adjusted p-value was below 0.05 and its mean expression was at least 4-fold up- or down-regulated between preadipocyte and adipocyte measurements. Next miRlastic was applied to the miRNA and also the mRNA measurements to assign potential targeted genes using miRNA-target predictions from TargetScan. Out of 9941 overall miRNA targets, 1395 were predicted to be significantly differentially expressed.

5 Results and discussion

This chapter contains the results with respect to evaluation and application of extended MONA. Section 5.1 comprises the evaluation results of the extended cooperative model, 5.2 describes both the evaluation of the cooperative-inhibitory model as well as its application to experimental datasets. Section 5.3 illustrates the results of a generative model for p-value distribution with respect to cMONA. Afterwards, an additional section will show evaluation results of MONA for miRNA alone enrichment.

5.1 Extension of Cooperative Model

The performances of the cooperative inhibitory models for both two and three levels and for Fisher's exact test (only applicable to single level) are displayed in Figure 10.

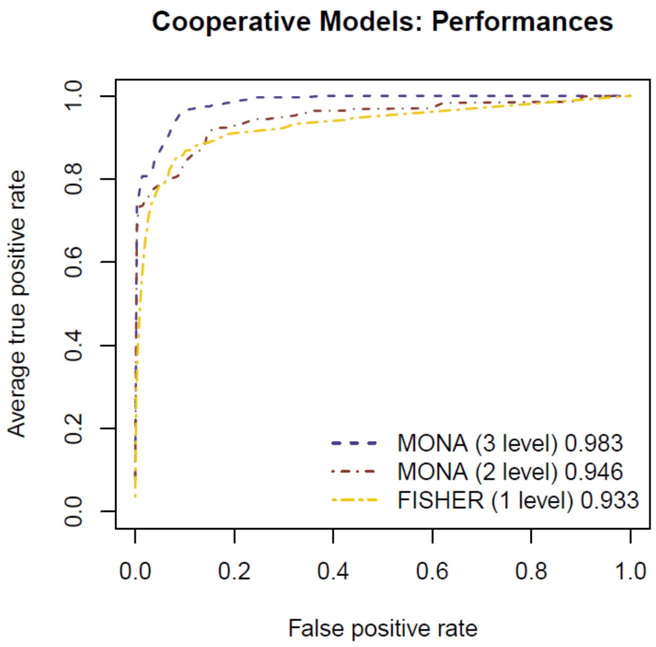


Figure 10: Performance of the cooperative model applied to synthetic data of two species (purple) and three species (blue) respectively. The inference of Fisher's exact test (yellow) is based on one species only. AUC values are listed in the respective figure legends.

The AUC value for the cooperative model with three levels was 0.983, for two levels 0.946 and for Fishers exact test 0.933. The ROC curve generated by the three species model differed from the curve generated by the two level model of MONA and indicated that a higher number of measurements correspond to more accurate results.

We have implemented the cooperative model to be used with more than three levels, which was not evaluated further. We anticipate even higher AUC values following the trend of one- to three-level MONA results.

5.2 Cooperative Inhibitory Model

5.2.1 Evaluation on synthetic data

In the following, optimal shape parameters of the Beta prior p for the cooperative-model were determined by evaluation. The heatmap in Figure 11 shows different priors and the corresponding AUC values. While the parameter b was chosen to be 1 or 2, a larger range was decided to be chosen for a because this prior restricts the number of predicted terms. The resulting AUC values for these priors range from 0.9750583 (corresponds to a = 1 and b = 2) to 0.9954245 (corresponds to a = 15 and b = 1). Altogether, all AUC values, regardless of whether a was chosen to be high or low, are relatively near to 1 which implies an accurate performance of the cooperative-inhibitory model regardless of the selected priors. We therefore selected a=15 and b=1 for further analysis.

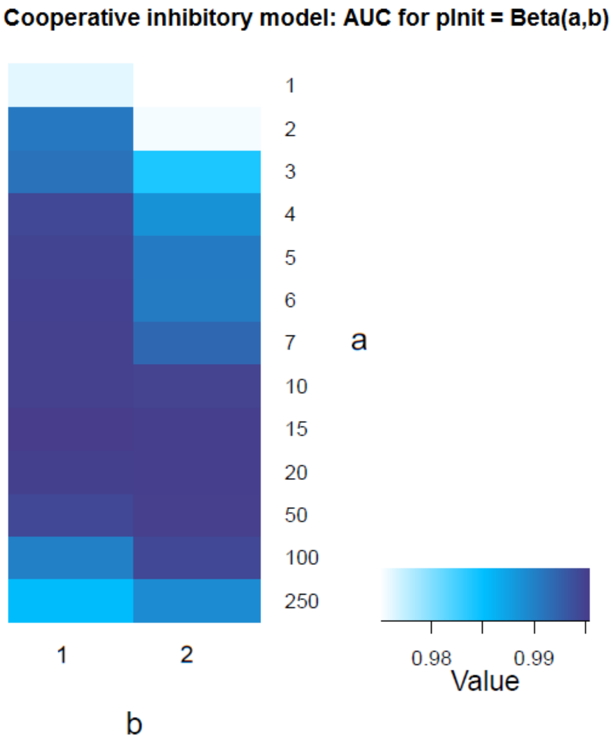


Figure 11: Evaluation of the cooperative-inhibitory model displaying AUC values using different shape parameters a and b of the Beta prior probability for the terms to be active. The model was tested using different values for a and b, respectively.

Evaluation

Figure 12 shows the results of the evaluation based on the synthetically generated data for the cooperative-inhibitory model (black), the cooperative model (yellow), the inhibitory model (blue) as well as for Fishers exact test on tree single species representing those species, which are used for the cooperative-inhibitory model. Regarding Figure 6, species 1 (violet) represents O^{I}, species 2 (brown) represents $O^{Inh,I}$ and species 3 (cyan) O^{II}.

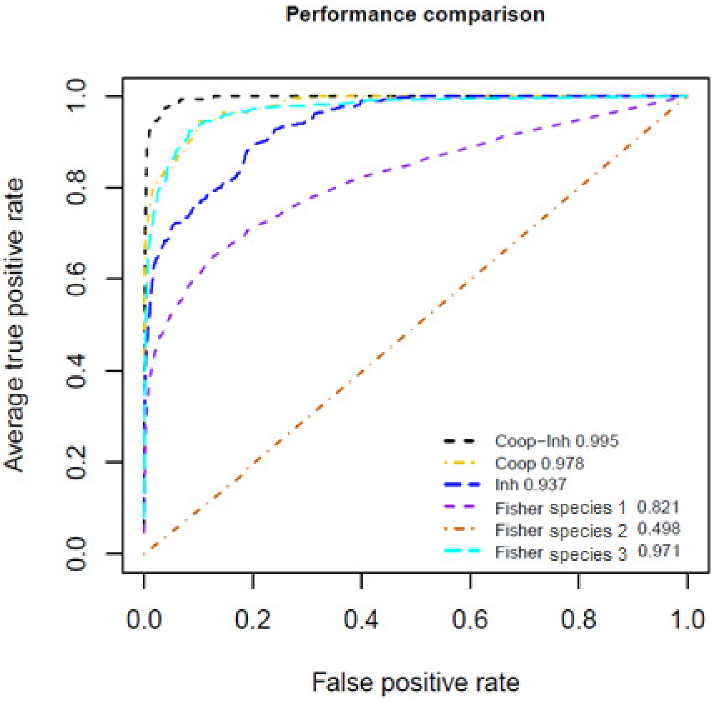


Figure 12: Evaluation of the combined, the cooperative and the inhibitory model of MONA as well as Fisher's exact test using synthetic data sets. ROC curves are shown in black for the combined model, yellow for the cooperative model, blue for the inhibitory model and violet, brown and cyan for single species performed by Fisher’s exact test.

It is observable that the cooperative-inhibitory outperforms all other approaches with an AUC value of 0.995. This supports the assumption, that using more information for the inference has a positive effect on the performance. The AUC values for the cooperative and the inhibitory model were 0.978 and 0.937, respectively. ROC curves generated by Fisher's exact test on the three species differ from the ROC curve generated by the cooperative-inhibitory model. Notably, results of miRNA alone enrichment with the Fishers exact test follow a random distribution. This will be further investigated in Section 5.4. As expected, the fusion of knowledge from the cooperative and the inhibitory model results in higher performance which is shown by a significantly better ROC curve in contrast to less complex MONA models as well as Fisher's exact test on single species.

5.2.2 Application to biological data

In order to assign the genes to functional categories, annotations from GO*, KEGG PATHWAYS* and *WikiPathways* were retrieved. Table 1 lists the obtained gene sets for these ontologies. For GO, a set of 14539 genes was obtained, containing 304 genes, whose mRNA was up- or down-regulated, 51 genes with an associated hypomethylated CpG site and 1205 genes which were considered to be targeted by differentially expressed miRNAs. The respective numbers for *KEGG PATHWAYS* and *WikiPathways* are also listed in Table 1 In total, 9457 functional categories of GO, 216 from *KEGG PATHWAYS* and 323 from *WikiPathways* were used as input for MONA together with the respective gene sets.

Table 1: Obtained gene sets for three different ontologies: GO, KEGG Pathways and WikiPathways. The columns show the number of determined genes whose mRNA was up- or down-regulated, genes with an associated hypomethylated CpG site, genes which were considered to be targeted by up- or down-regulated miRNA and the total number of genes in the respective gene set.

	mRNA	methylation	miRNA	total size
GO	304	51	1205	14539
KEGG Pathways	177	27	505	6576
WikiPathways	127	20	477	5288

The heatmaps displayed in Figures 13-15 show the predicted categories from the adipocyte differentiation dataset using GO, *KEGG PATHWAYS* and *WikiPathways*, respectively. To have a comparison with respect to other models of MONA, the cooperative, the inhibitory and the single species model were additionally applied.

For GO, 12 terms were obtained with a probability of above 0.5. Comparing the resulting terms of the cooperative-inhibitory model to the remaining models, it is obvious that there is not much overlap of predicted terms between the cooperative-inhibitory models and the other ones. In contrast, it is observable that the cooperative and the single mRNA model clearly overlap meaning that mRNA alone has a great impact on the cooperative model. Note that the single model applied to the methylation data did not predict any term to be active. This may be due to the small number of genes with an associated hypomethylated CpG site. Active terms predicted by the inhibitory model, were neither predicted by the miRNA single model, nor by the mRNA single model. The cooperative and the inhibitory model have no active terms in common, whereas the combination of both models reveals further active terms. The most interesting term revealed is *positive regulation of fat cell differentiation* which implies the maturation of preadipocytes into adipocytes [44]. Further terms which could only be inferred by the cooperative-inhibitory model and which are shown to be altered in adiposity, are *vitamin metabolic process* [45], *fatty acid metabolic process* [46], *response to toxic substance* [47], *one-carbon metabolic process* [48] and *selenium compound metabolic process* [49]. These results show that the cooperative model is able to infer meaningful functional GO terms which are altered in adipocyte differentiation and therefore are enriched between the gene sets. In addition, it also predicts *positive regulation of fat cell differentiation*, a term, which is referred to as adipocyte differentiation. Consequently, the outcome of the new model shows that the combined information of both, the cooperative and the inhibitory model, produces a model which enables the identification of meaningful GO terms with respect to adipogenesis.

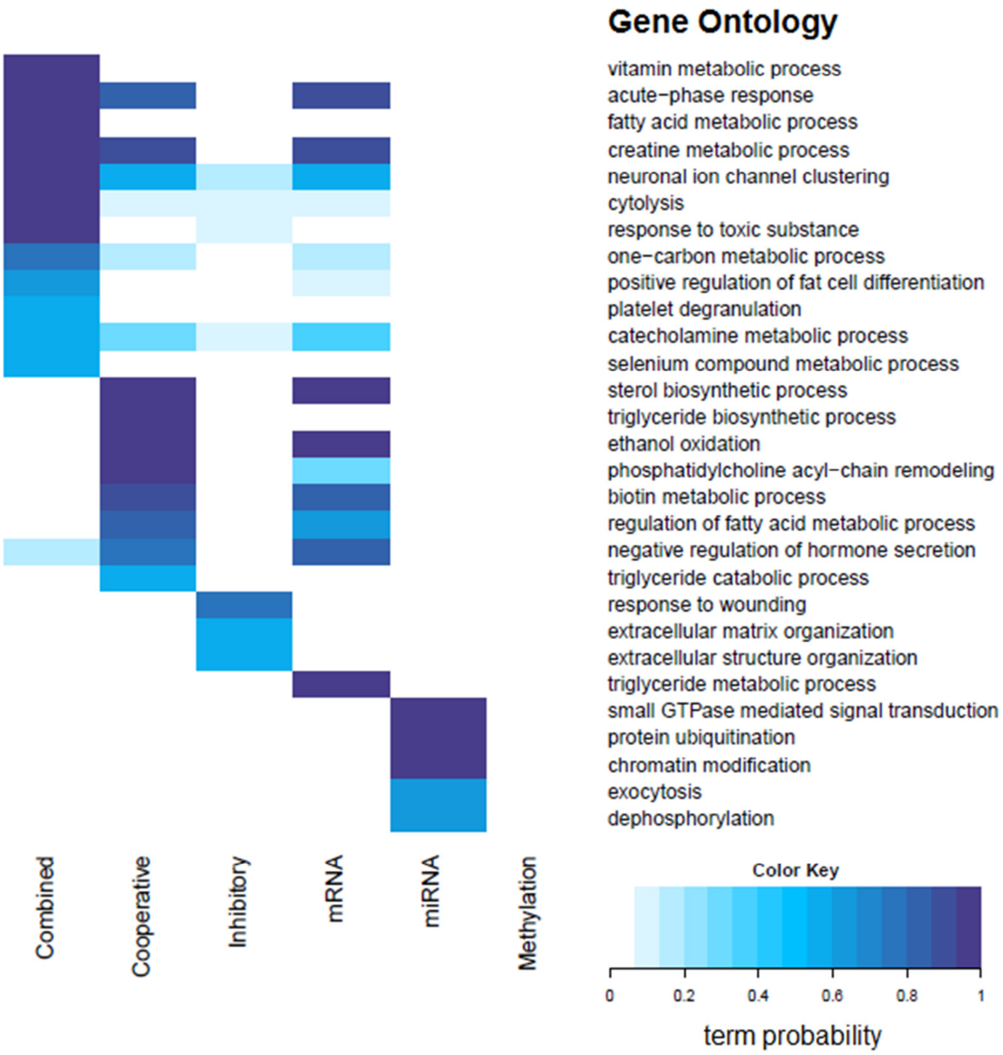


Figure 13: Heat map showing the biological processes of adipocyte differentiation predicted by the cooperative-inhibitory, cooperative and inhibitory models as well as single model for mRNA, miRNA and methylation, using GO. The color indicates the functional category in the respective run. No terms could be predicted by the single model for methylation data.

Regarding the outcomes predicted by the cooperative-inhibitory model and the remaining models using *KEGG PATHWAYS*, Figure 14 shows that the overlap of terms predicted by cooperative-inhibitory model and the remaining models is higher than their overlap using GO terms showed in Figure 13. Nevertheless the number of predicted KEGG terms of the cooperative-inhibitory is definitely higher compared to the other models. There are some terms predicted only by the new model. The first one is *Metabolism of xenobiotics by cytochrome p450*. As for GO, the single species model applied to methylation data did not infer any terms. Xenobiotics are harmful, lipid-soluble chemical substances that are foreign to the human body meaning neither naturally produced by nor expected to be present within that organism [50]. It has been proposed that there is a link between exposure to certain xenobiotics and obesity by showing that some xenobiotic-metabolizing forms of P450 were expressed in white adipose tissue [51]. Another term is *Phenylalanine metabolism* which is also associated with human obesity [52]. Further terms known to be changed between preadipocytes and adipocytes are *Vitamin digestion and absorption* [53], *Glycolysis/Gluconeogenesis* [54], *Biosynthesis of unsaturated fatty acids* [55], *taurine and hypotaurine metabolism* [56], *Retinol metabolism* [57] and *ABC transporters* [58]. All these terms were revealed only by the combined model and are either shown or proposed to be altered in adipose tissue.

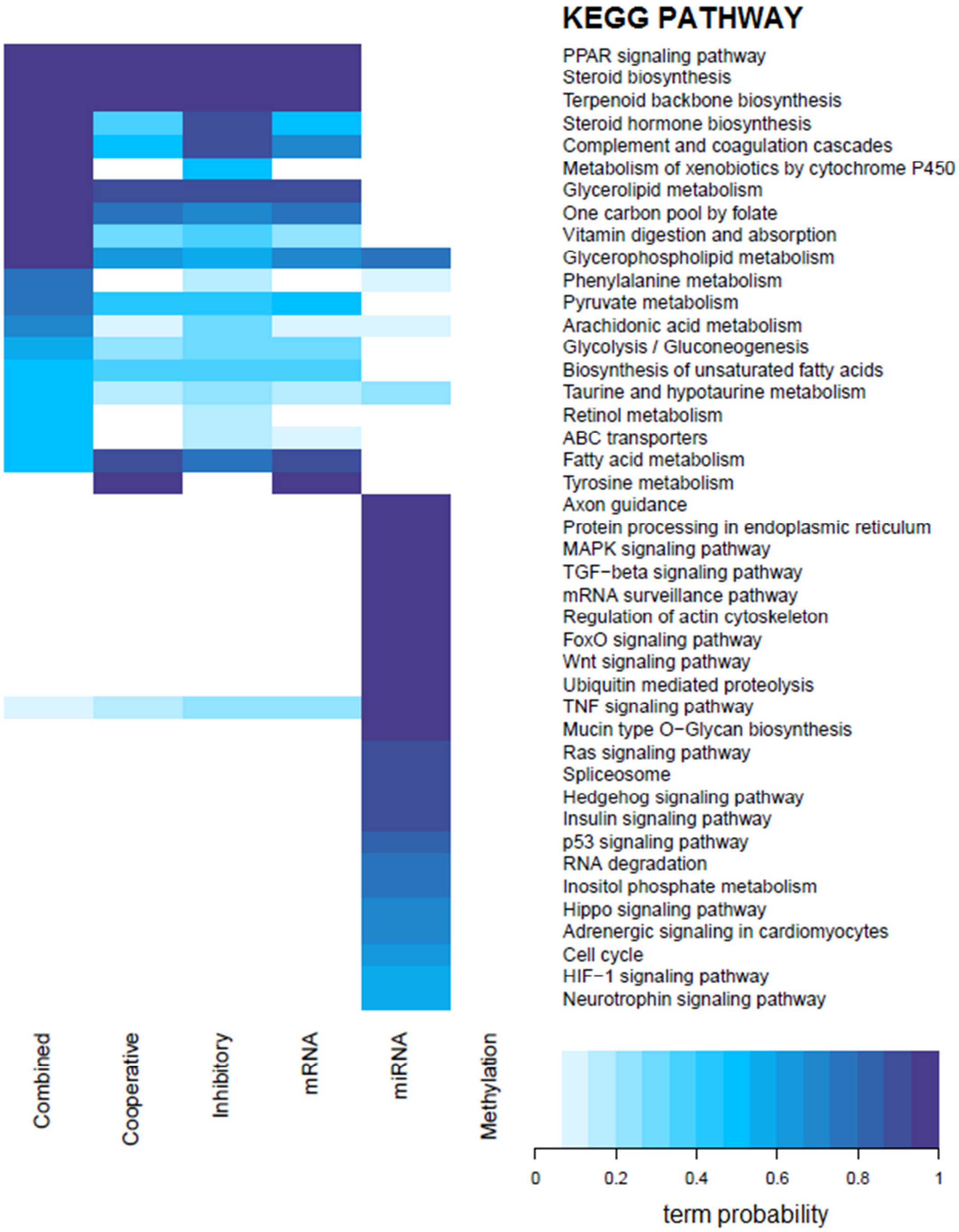


Figure 14: Heat map showing the biological processes of adipocyte differentiation predicted by the cooperative-inhibitory, cooperative and inhibitory models as well as single model for mRNA, miRNA and methylation, using *KEGG PATHWAYS* as ontology. The color indicates the functional category in the respective run. No terms were predicted by the single model for methylation data.

Considering the inferred terms of WikiPathways which are displayed in Figure 15, some aspects are becoming apparent. First, in contrast to other models handling more than one species in parallel, the cooperative model inferred a significantly smaller number of terms. Furthermore, there are more terms identified to be active by the cooperative-inhibitory model then by the remaining models (except the miRNA single species model). There are also terms arising which were shown to be altered in adiposity: *Irinotecan pathway* [59], *blood clotting cascade* [60], *the alpha-linoleic acid metabolism* [61], *Fatty Acid Beta Oxidation* [62] and *Arachidonic acid metabolism* [63].

Concerning three different ontologies, the application of the cooperative-inhibitory model, on the one hand, enhances the probabilities of terms which were also predicted separately by the other models, and on the other hand, it additionally infers some meaningful terms which were not inferred by the other models. Using biological process terms from the *GO*, the cooperative-inhibitory model infers many terms which could not be identified by several other models of MONA, regardless of whether cooperative, inhibitory or single species. These retrieved terms are possible processes which are altered during adipogenesis. Regarding *KEGG PATHWAYS* and *WikiPathways*, it is observable, that the cooperative model, the inhibitory model as well as the single species model applied to mRNA data had a strong impact on the results of the cooperative-inhibitory model. The single species model for miRNA predominantly predicted terms originating from signaling and DNA modification pathways whereas the single species model for methylation data did not infer any categories, maybe due to the low number of differentially methylated CpG sites. However, the combined integration of mRNA, methylation and miRNA data in MONA enables the determination of functional categories exceeding the ability of the separate models of MONA.

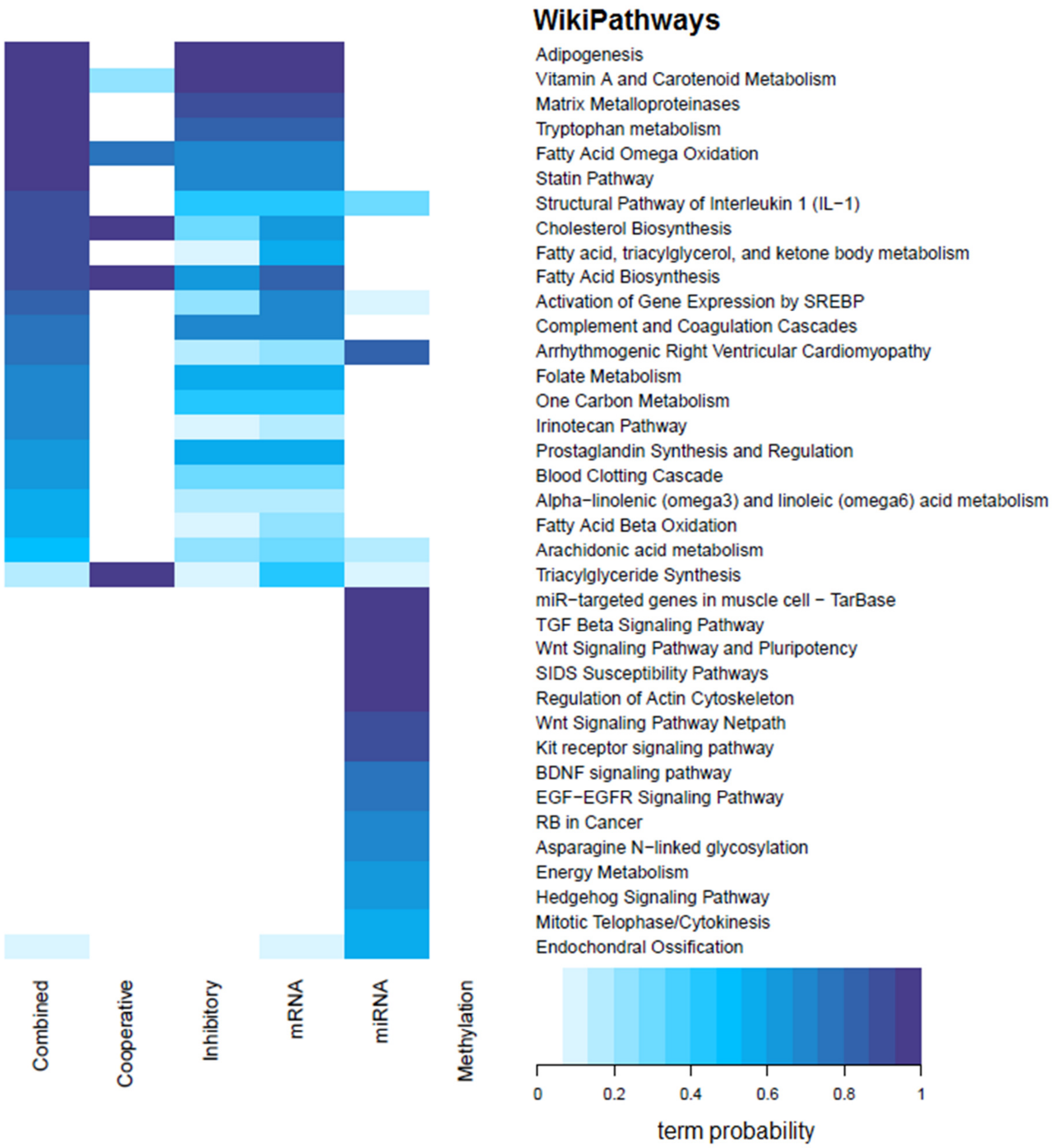


Figure 15: Heat map showing the biological processes of adipocyte differentiation predicted by the cooperative-inhibitory, cooperative and inhibitory models as well as single model for mRNA, miRNA and methylation, using *WikiPathways* as ontology. The color indicates the functional category in the respective run. No terms could be predicted by the single model for methylation data.

In order to set up a framework for the application on data sets which comprise measurements from mRNA, methylation and miRNA levels, the cooperative and the inhibitory model were combined. The cooperative-inhibitory model was shown to outperform the cooperative and the inhibitory model as well as Fisher's exact test on synthetic data sets. In addition, the application of the combined model to mRNA, methylation and miRNA data from adipocyte differentiation using categories from GO, *KEGG PATHWAYS* and *WikiPathways*, revealed meaningful functional terms which could not be inferred by the cooperative and/or the inhibitory models alone. Figure 16 shows a part of the GO tree containing those nodes predicted by the cooperative-inhibitory model (yellow) and Fisher's exact test (cyan). Nodes predicted by both approaches are colored blue. It shows that the newly implemented model overcomes term redundancies while Fisher's exact test predicts a huge number of redundant terms.

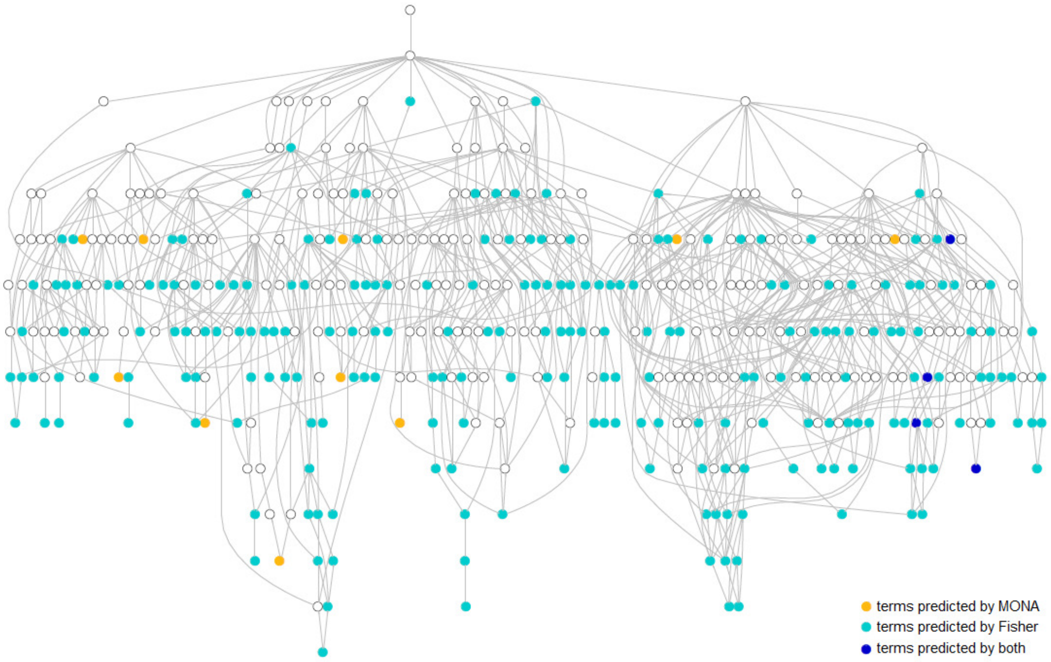

● terms predicted by MONA
● terms predicted by Fisher
● terms predicted by both

Figure 16: Visualization of a cutout of the GO tree containing biological processes. The colored nodes display biological processes which were predicted to be active by MONA's cooperative-inhibitory model (yellow) and Fisher's exact test (cyan). Terms identified by both methods are shown in blue.

5.3 CMONA

The drawback of (binary) MONA is that it makes use of binary input values requiring an arbitrary threshold. To overcome this, we derived a working model, named cMONA, which allows continuous input values in terms of p-values. CMONA takes into account the expression strength and is the first step towards a model-based functional analysis incorporating statistical significances without an arbitrary threshold.

5.3.1 Evaluation of generic model for p-value distribution sampling

The generic model was applied to p-values which were previously obtained from the experimental data set comprising mRNA expression measurements from adipocyte differentiation. Using the R package *fdrtool*, which can be used for estimation of local false discovery rates [64], the proportion of p-values under H0 was estimated, meaning the proportion of p-values of not differentially expressed genes. Within the mRNA expression data set, about two-thirds of the mRNAs were estimated to be differentially expressed. Mean and precision of the log-log-normal distribution were predicted to be about 0.325 and 1.79. These parameters were then used for p-value sampling: If a gene was considered to be differentially expressed, its transformed log-log p-value was expected to come from a normal distribution with the previously inferred mean and precision parameters, otherwise from a uniform distribution. These assumptions were used to sample p-values starting from the gene set which was previously given as input for mean and precision estimation. Figure 17 displays the distribution of the p-values obtained from the data set (blue) as well as the sampled p-value distribution (pink) in both, the transformed space (left) and the original space (right).

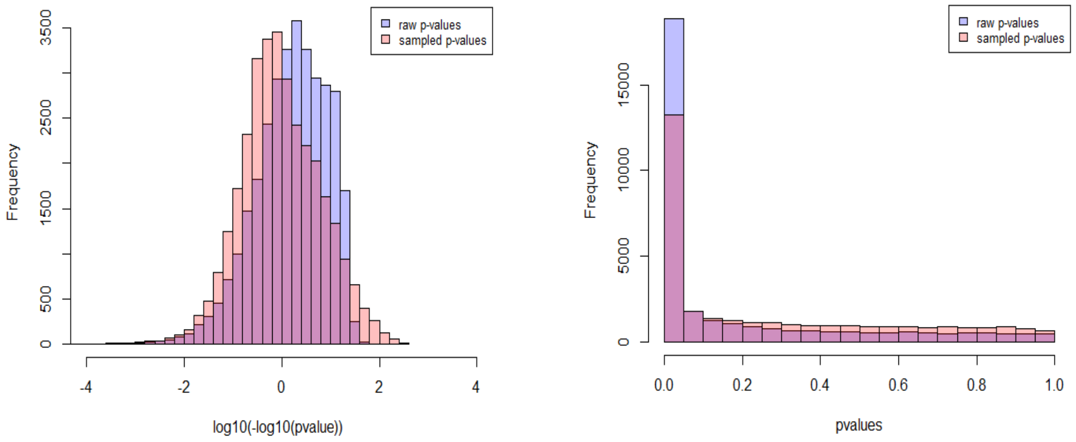


Figure 17: Histograms of p-value distribution of the experimental mRNA expression data set obtained from the moderated t statistic (blue) and sampled p-values (pink). Left: distributions in the transformed space (log-log-space). Right: distributions in the original data space.

Since the experimental data set contains two-thirds of differentially expressed genes, the distribution of transformed p-values is slightly right-shifted. But as we expected a normal distribution of transformed log-log p-values, further 10 data sets from the Gene Expression Omnibus were evaluated. Both the obtained and the sampled p-value distributions of these data sets are displayed in the Supplementary Section (Figure A).

5.4 Excursion: Application of MONA to miRNA enrichment

To gain a deeper insight into the regulatory mechanisms, gene set enrichment using miRNAs as observations became more important. Starting from miRNAs as observed species, the characterization of enriched pathways of targeted genes poses a difficult challenge, since functional properties of miRNAs are characterized by their targets [65]. To identify enriched pathways for measured miRNAs, enrichment methods make use of miRNA-target binding predictions [66]. Methods for such predictions are known to produce many false positive relationships [67]. In addition, there is a many-to-many relationship between miRNAs and multiple genes [66]. If these relationships are taken into account for correlation analyses which are based on one-to-one relationships, the biological context of cell signals are ignored [66]. Considering these issue, MONA was evaluated by application to miRNA alone enrichment. This evaluation was performed on synthetic data sets which were generated based on miRNA-target matrices of TargetScan [35], StarBase [36], MiRanda [37] and MiRTarBase [38] as well as on combinations of them. In addition, Fisher's exact test was equally evaluated for performance comparison.

The generation of the data is displayed by Figure 18. For each data set, five non-redundant GO terms containing five to hundred genes were sampled, then mapped to their corresponding genes followed by mapping these genes to their targeted miRNAs for each miRNA-target matrix. Consequently, twenty gene-sets were generated per miRNA-target matrix for the evaluation. As MONA uses a category-to-gene relationship for the inference, the resulting miRNAs could not directly be used as input and thus had to be mapped back to their corresponding targeted genes. If the expression of these miRNAs is considered to be altered, then their targeted mRNAs and thus the resulting gene products are also considered to be altered. However, miRNA-target predictions may infer large amounts of false positive mRNA targets for individual miRNA regulators. Thus different selection criteria were chosen which should keep the number of the gene set reduced. It was decided not to take all miRNAs for the last mapping step. Instead, the occurrence of all mapped miRNAs was determined and it was decided to take the most common one to the most common five miRNAs for the last mapping step. The same was done for the rarest miRNAs. After their creation, MONA's single model and Fisher's exact test were applied to all data sets followed by generating ROC curves.

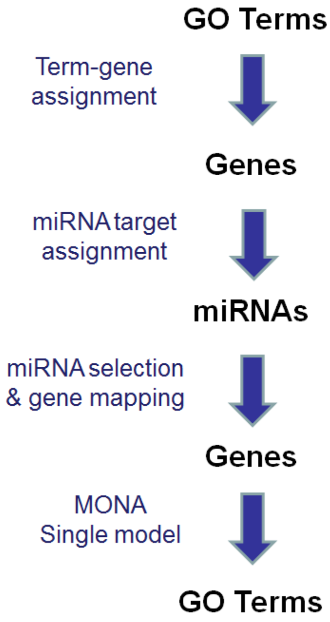
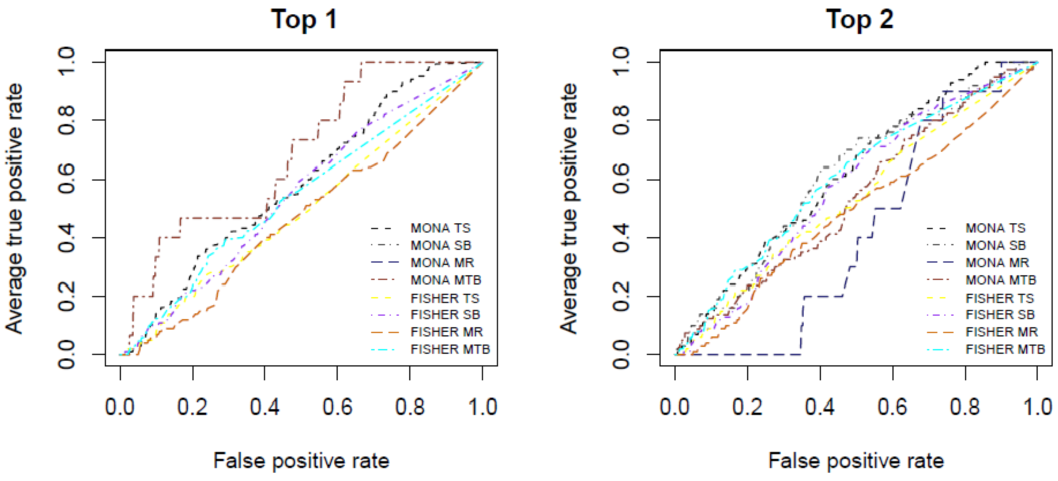
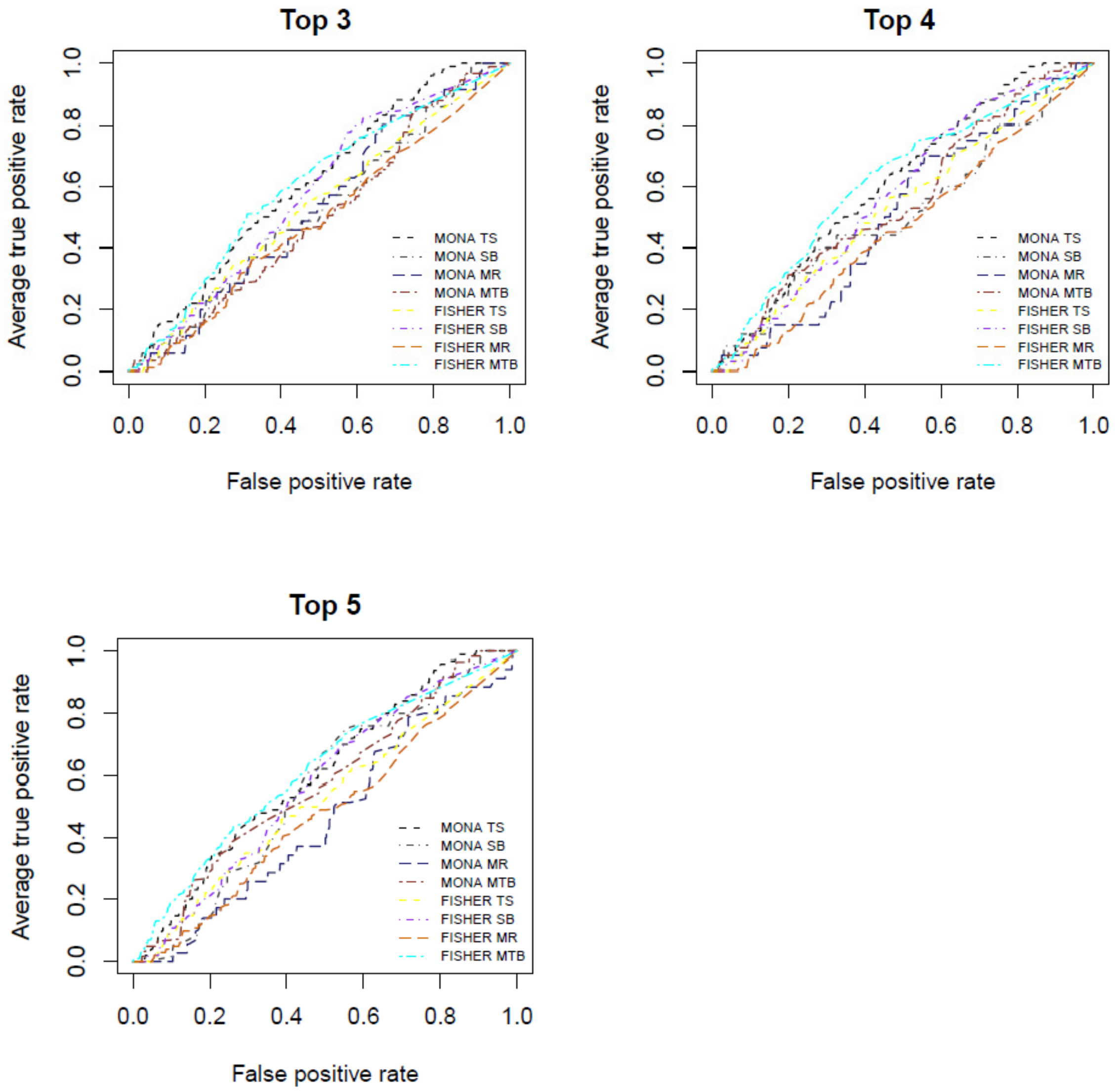


Figure 18: Schematic representation displaying the generation of synthetic data sets for the miRNA alone enrichment. First, non-redundant GO terms were sampled which were mapped to the corresponding genes. Secondly, genes were mapped to miRNAs by which they are targeted, using different miRNA-target predictions. In the next step, a selection criterion is applied to the resulting miRNAs to reduce their number. Afterwards, the obtained miRNAs were mapped back to genes which were then considered as differentially expressed and used for the single enrichment by MONA (and Fisher's exact test).

Figure 19 and 20 show the results of the inference of MONA and Fisher's exact test, for the most 1 to 5 most common as well as rarest obtained miRNAs, respectively. Taken together, one can say that the performance of MONA is near to random, which is shown by ROC curves near the diagonal of the ROC space. Nevertheless, there are exceptions regarding the performance of MONA's miRNA alone enrichment, for example taking in account only the most common miRNA predicted by MONA and using MiRTarBase for miRNA-target relationship (Figure 19, Top 1). It shows a ROC curve clearly better than random, but on the other site, the shape of the curve also shows that MONA was not able to perform a successful inference for all data sets. An inference was considered to be successful, if the mean of the inferred Beta distribution of the shape parameter for the term probabilities not to be 'active', was above 0.8, since it was expected that only a small fraction of terms is active. Another performance of MONA not being random is shown in Figure 20 using the Top 2 rarest miRNAs for generating the data by MiRanda. It appears that miRNA-target relationships which are experimentally validated, like in the case of MiRTarBase, are more suitable for miRNA enrichment. An argument is that *ab initio* and sequence-based miRNA-target predictions may infer false positives. This can, for example, be avoided by using experimentally validated. Figures displaying the performances using combinations of miRNA-target predictions can be found in the Supplementary Section (Figure B1 and B2).

Most common miRNAs





Top 3

Top 4

Top 5

Figure 19: Plots showing the performances of bot MONA and Fisher's exact test for the most common (Top 1) to the five most common (Top 5) miRNAs for single miRNA enrichment. . The data sets were generated using following miRNA-target predictions: TargetScan (TS), StarBase (SB), MiRanda (MR) and MiRTarBase (MTB).

Rarest miRNAs

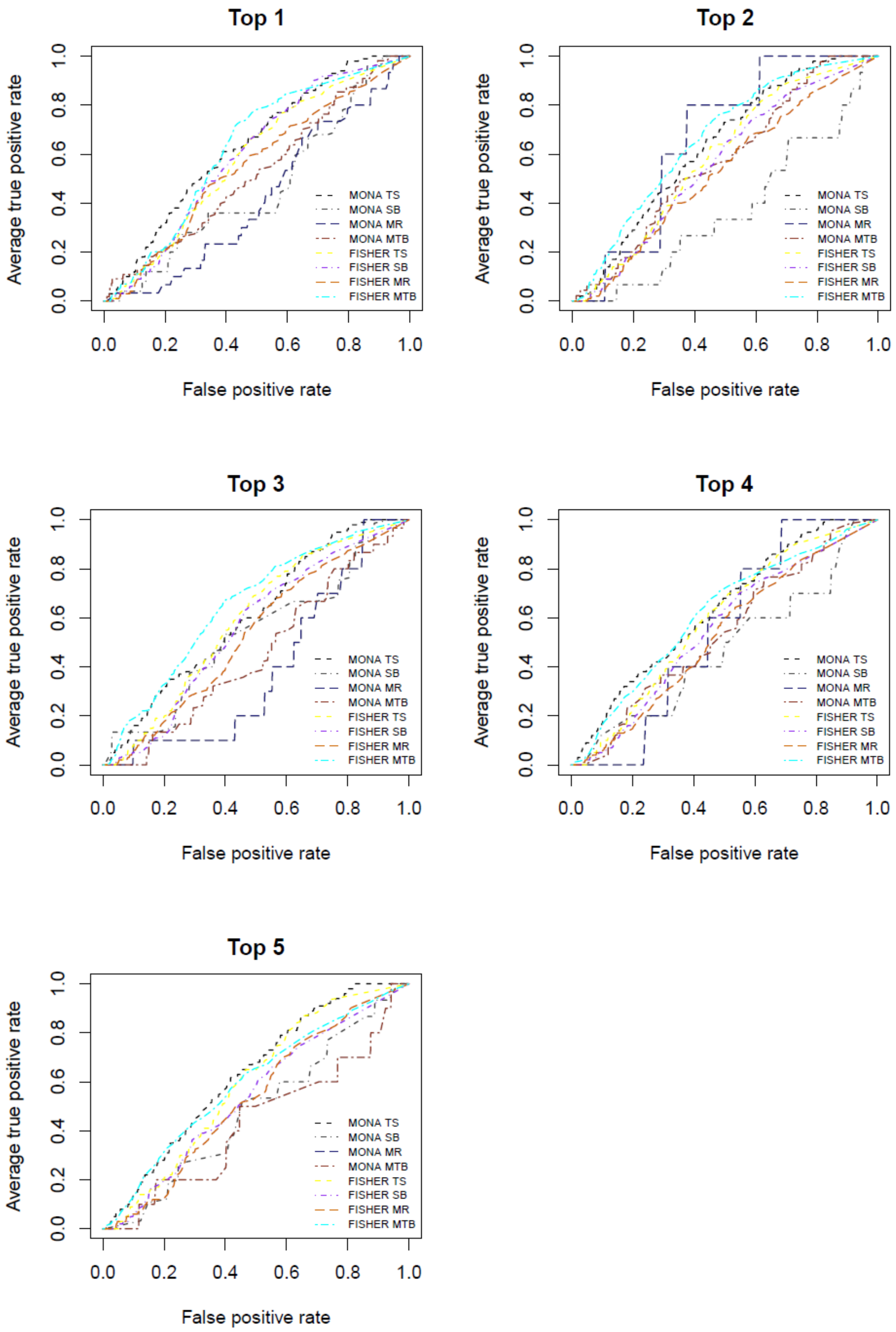


Figure 20: Plots showing the performances of bot MONA and Fisher's exact test for the rarest (Top 1) to the five rarest (Top 5) miRNAs for single miRNA enrichment. The data

sets were generated using following miRNA-target predictions: TargetScan (TS), StarBase (SB), MiRanda (MR) and MiRTarBase (MTB).

6 Summary and Outlook

Multi-level Ontology Analysis (MONA) provides a model-based framework, which is capable of integrating results across multiple molecular species for the assessment of functional gene responses. This thesis comprises extensions of MONA regarding functionality and flexibility. Therefore, three additional models were implemented.

First, the cooperative model was adjusted in order to handle any number of species, which can be considered as independent observations, for example mRNA expression, DNA methylation and protein expression. The extended model showed to perform more accurately by the integration of three instead of two species.

Second, the cooperative and the inhibitory model were fused into one combined model allowing for the investigation of measurements that can be interpreted as both independent and dependent. It performs better than less complex MONA models and Fisher’s exact test. The application of the cooperative-inhibitory model to experimental data (comprising mRNA, DNA methylation and miRNA expression levels) of adipocyte differentiation revealed meaningful functional terms.

Finally, a working model, named cMONA, was developed to overcome MONA’s drawback by allowing for p-values continuous observed input values instead of binary values. In the context of this thesis, the model could not be fully implemented but essential work interrogating p-value distribution assumptions. Up to now, a generic mixture model was implemented, which estimates p-value distribution. It was applied to mRNA expression data as well as to ten randomly selected data sets from the Gene Expression Omnibus database.

Future implementations of cMONA should include the implementation of cMONA for single species with the here determined distribution assumptions for (not) significantly regulated genes. Furthermore, other models for the assessment of gene responses could be implemented, including an “activating” model which could be applied e.g. to protein phosphorylation measurements as many enzymes are “activated” by phosphorylation events.

7 References

[1] Pelizzola M et al., "The DNA methylome," *FEBS Letters,* no. 585(13), pp. 1994-2000, Jul 2011.

[2] Sass S et al., "A modular framework for gene set analysis integrating multilevel omics data," *Nucleic Acids Research,* no. 41(21), pp. 9622-9633, Nov 2013.

[3] Hyduke DR et al., "Analysis of omics data with genome-scale models of metabolism," *Molecular Biosystems,* vol. 9(2), pp. 167-74, Feb 2013.

[4] Thang J, Xing Z, Ma M, Wang N, Cai YD, Chen L, Xu, "Gene Ontology and KEGG Enrichment Analyses of Genes Related to Age-Related Macular Degeneration," *BioMed Research International,* p. 10, 08 2014.

[5] A. Kashiwagi, I. Urabe, K. Kaneko and T. Yomo, "Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced Attractor Selection," *PLoS ONE,* p. 10, 12 2006.

[6] Subramanian A et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences,* vol. 102(43), pp. 15545-50, Oct 2005.

[7] Ashburner M et. al, "Gene Ontology: tool for the unification of biology," *Nature Genetics,* no. 25(1), pp. 25-29, May 2000.

[8] Abatangelo L et al., "Comparative study of gene set enrichment methods," *BMC Bioinformatics,* 2009.

[9] Bauer S et al., "Model-based gene set analysis for Bioconductor," *Bioinformatic,* pp. 1882-1883, 2011.

[10] Yi N et al., "Multiple comparisons in genetic association studies: a

hierarchical modeling approach," *Statistical Applications in Genetics and Molecular Biology,* pp. 35-48, 2014.

[11] Hans Winkler, Verbreitung und Ursache der Pathogenesis im Pflanzen- und Tierreiche, Jena: Fischer, 1920.

[12] Yadav SP, "The wholeness in suffix -omics, -omes, and the word om," *Journal of Biomolecular Techniques,* p. 277, Dec 2007.

[13] Tyers M et al., "From genomics to proteomics," *Nature,* pp. 193-197, Mar 2003.

[14] Lodish H et al., "The Three Roles of RNA in Protein Synthesis," in *Molecular Cell Biology. 4th edition.*, New York, W. H. Freeman, 200.

[15] Brennecke J et al., "Principles of MicroRNA-Target Recognition," *PLoS Biology,* p. 85, Mar 2005.

[16] Liu Y et al., "Methylomics of gene expression in human monocytes," *Human Molecular Genetics,* pp. 5065-74, Dec 2013.

[17] Kanehisa M et al., "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research,* pp. 199-205, 2014.

[18] Kelder T et al., "WikiPathways: building research communities on biological pathways," *Oxford University Press,* 2011.

[19] Yaari G et al., "Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations," *Nucleic Acid Research,* 2013.

[20] King HC et al., "Gene expression profile analysis by DNA microarrays: promise and pitfalls," *JAMA,* 2001.

[21] Goeman JJ et al., "Analyzing gene expression data in terms of gene sets:

methodological issues," *Bioinformatics,* pp. 980-987, 2007.

[22] Smyth GK, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology,* 2004.

[23] Hajian-Tilaki K, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine,* no. 4(2), pp. 627-35, 2013.

[24] Witt PL et al., "Nonparametric testing using the chi-square distribution," *Physical Therapy,* no. 66(2), pp. 266-268, Feb 1986.

[25] F. Ruggeri, R. Kenett and F. Faltin, "Bayesian Networks," in *Encyclopedia of Statistics in Quality and Reliability*, Wiley, 2007.

[26] Huang da W et al., "Bioinformatics enrichment tools: paths towards the comprehensive functional analysis of large gene lists," *Nucleic Acids Research,* pp. 1-13, 2009.

[27] Abatangelo L et al., "Comparative study of gene set enrichment methods," *BMC Bioinformatics,* Sep 2009.

[28] Ryman N et al., "Statistical power when testing for genetic differentiation," *Molecular Ecology,* pp. 2361-73, Oct 2001.

[29] Bauer S et al., "GOing Bayesian: a model-based gene set analysis of genome-scale data," *Nucleic Acids Research,* pp. 3523-32, Jun 2010.

[30] Cox J et al., "1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data," *BMC Bioinformatics,* 2012.

[31] Mendell JT et al., "MicroRNAs in stress signaling and human disease," *Cell,* pp. 1172-87, Mar 2012.

[32] Nam S et al., "MicroRNA and mRNA integarted analysis (MMIA): a web tool for examining biological functions of microRNA expression," *Nucleic Acids Research,* pp. W356-62, Jul 2009.

[33] Harsh Dweep et al., "In-Silico Algorithms for the Screening of Possible microRNA Binding Sites and Their Interactions," *Current Genomics,* no. 14(2), pp. 127-136, Apr 2013.

[34] Hao Z et al., "Advances in the Techniques for the Prediction of microRNA Targets," *International Journal of Molecular Sciences,* no. 14(4), pp. 8179-8187, Apr 2013.

[35] Lewis BP, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell,* vol. 120(1), pp. 15-20, Jan 2014.

[36] Li JH et al., "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data," *Nucleic Acids Research,* pp. D92-D97, Nov 2013.

[37] Enright AJ et al., "MicroRNA targets in Drosophila," *Genome Biology,* 2003.

[38] Hsu SD et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research,* pp. D78-85, Jan 2014.

[39] Sass S et al., "MicroRNAs coordinately regulate protein complexes," *BMC Systems Biology,* no. 5(1), Aug 2011.

[40] Sass S, "Integration of multiple omics levels for the analysis of adipocyte differentiation," *unp. Diss.,* 2014.

[41] Zou H et al., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B,* pp. 301-320, Sep 2004.

[42] R. Core Team, "R: A language and environment for statistical computing," 2012.

[43] BARTHOLOMÉ K et al., "Estimation of Gene Induction Enables a Relevance-Based Ranking of Gene Sets," *Journal of Computational Biology,* no. 16(7), pp. 959-967, Jul 2009.

[44] Gregoire FM et al., "Understanding adipocyte differentiation," *Physiological Reviews,* no. 78(3), pp. 783-809, Jul 1998.

[45] Wood RJ, "Vitamin D and adipogenesis: new molecular insights," *Nutritial Reviews,* no. 66(1), pp. 40-46, Jan 2008.

[46] Frayn KN et al., "Fatty acid metabolism in adipose tissue, muscle and liver in health and disease," *Essays in Biochemistry,* no. 42, pp. 89-103, 2006.

[47] Medina-Gomez G et al., "Adipogenesis and lipotoxicity: role of peroxisome proliferator-activated receptor gamma (PPARgamma) and PPARgammacoactivator-1 (PGC1)," *Public Health Nutrition,* no. 10(10A), pp. 1132-1137, Oct 2007.

[48] Verbrugghe A et al., "Peculiarities of one-carbon metabolism in the strict carnivorous cat and the role in feline hepatic lipidosis," *Nutrients,* no. 5(7), pp. 2811-2835, jul 2013.

[49] Hassan A et al., "Selenium promotes adipogenic determination and differentiation of chicken embryonic fibroblasts with regulation of genes involved in fatty acid uptake, triacylglycerol synthesis and lipolysis," *The Journal of Nutritional Biochemistry,* no. 25(8), pp. 858-867, Aug 2014.

[50] Lang M et al., "Metabolism of xenobiotics and chemical carcinogenesis.," *IARC Sci Publ,* no. 148, pp. 13-22, 1999.

[51] Ellero S et al., "Xenobiotic-Metabolizing Cytochromes P450 in Human White Adipose Tissue: Expression and Induction," *Drug Metabolism and*

Disposition, no. 38(4), pp. 679-686, Apr 2010.

[52] Swierczynski J et al., "Serum phenylalanine concentration as a marker of liver function in obese patients before and after bariatric surgery," *Obesity Surgery,* no. 19(7), pp. 883-889, Jul 2009.

[53] Nimitphong H et al., "25-Hydroxyvitamin D3 and 1,25-Dihydroxyvitamin D3 Promote the Differentiation of Human Subcutaneous Preadipocytes," *PloS one,* no. 7(12), 2012.

[54] Houde VP et al., "Chronic rapamycin treatment causes glucose intolerance and hyperlipidemia by upregulating hepatic gluconeogenesis and impairing lipid deposition in adipose tissue," *Diabetes,* no. 59(6), pp. 1338-1348, Jun 2010.

[55] Veltri BC et al., "Adipose fatty acid composition and rate of incorporation of alpha-linolenic acid differ between normal and lipoprotein lipase-deficient cats," *Journal of Nutrition,* no. 136(12), pp. 2980-2986, Dec 2006.

[56] Ueki I et al., "3T3-L1 Adipocytes and Rat Adipose Tissue Have a High Capacity for Taurine Synthesis by the Cysteine Dioxygenase/Cysteinesulfinate Decarboxylase and Cysteamine Dioxygenase Pathways," *Journal of Nutrition,* no. 139(2), pp. 207-214, Feb 2009.

[57] Schupp M et al., "Retinol saturase promotes adipogenesis and is downregulated in obesity," *Proceedings of the National Academy of Sciences,* no. 106(4), pp. 1105-1110, Jan 2009.

[58] Leslie W. Chinn, "ABC transporters in adipose tissue," in *Pharmacodynamic Effects of Xenobiotic ABC Transporters in Peripheral Tissues*, Proquest, 2011.

[59] Shah P et al., "Altered irinotecan pharmacokinetics in diet-induced obesity," *The FASEB Journal,* no. 28, Apr 2014.

[60] De Pergola G et al. , "Coagulation and fibrinolysis abnormalities in obesity," *Journal of Endocrinological Investigation,* no. 25(10), pp. 899-904, Nov 2002.

[61] Casado-Diaz A et al. , "The omega-6 arachidonic fatty acid, but not the omega-3 fatty acids, inhibits osteoblastogenesis and induces adipogenesis of human mesenchymal stem cells: potential implication in osteoporosis," *Osteoporosis International,* no. 24(5), pp. 1647-1661, May 2013.

[62] Ellis JM et al., "Adipose acyl-CoA synthetase-1 directs fatty acids toward beta-oxidation and is required for cold thermogenesis," *Cell Metabolism,* vol. 12(1), pp. 53-64, Jul 2010.

[63] Sawa SC et al., "Association of adipose tissue arachidonic acid content with BMI and overweight status in children from Cyprus and Crete," *British Journal of Nutrition,* no. 91(4), pp. 643-649, Apr 2004.

[64] Strimmer K, "fdrtool: a versatile R package for estimating local and tail area-based false discovery rates," *Bioinformatics,* no. 24(12), pp. 1461-1462, Jun 2008.

[65] Wang B et al., "Challenges for MicroRNA Mircorarray Data Analysis," *Mircoarrays,* no. 2(2), Jun 2013.

[66] Calura E et al., "Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles," *Nucleic Acids Research,* no. 42(11), 2014.

[67] Mendoza MR et al., "RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier," *PLoS One,* no. 8(7), Jul 2013.

[68] Tarca AL et al., "Analysis of microarray experiments of gene expression profiling," *American Journal of Obstetrics & Gynecology,* pp. 373-388, 2006.

[69] Muniatequi A et al., "Joint analysis of miRNA and mRNA expression data," *briefings in Bioinformatics,* pp. 263-78, May 2013.

[70] Wang X et al., "Gene set enrichment analysis for multiple continuous phenotypes," *BMC Bioinformatics,* Aug 2014.

[71] Iriazzy RA et al., "Gene set enrichment analysis made simple," *Statistical methods in medical research,* vol. 18(6), pp. 565-575, Dec 2009.

[72] Massa MS et al., "Gene set analysis exploiting the topology of a pathway," *BMC Systems Biology,* Sep 2010.

[73] Rao DC, "Phenotype choices/Options," in *Genetic Dissection of Complex Traits, 2nd Edition*, Academic Press, 2008, p. 300.

[74] Edgar R et al., "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research,* no. 30(1), pp. 207-210, Jan 2002.

8 Supplementary

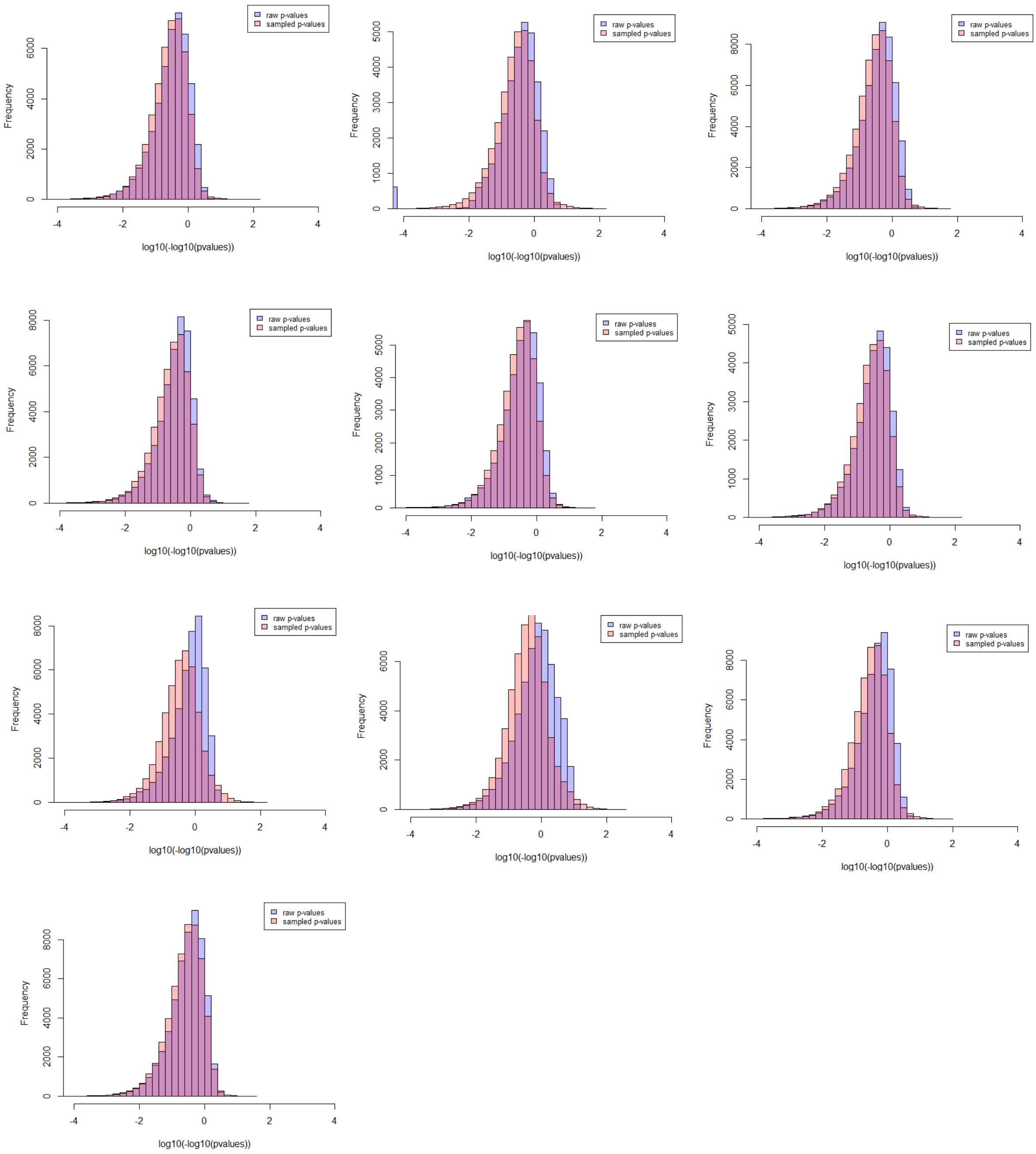
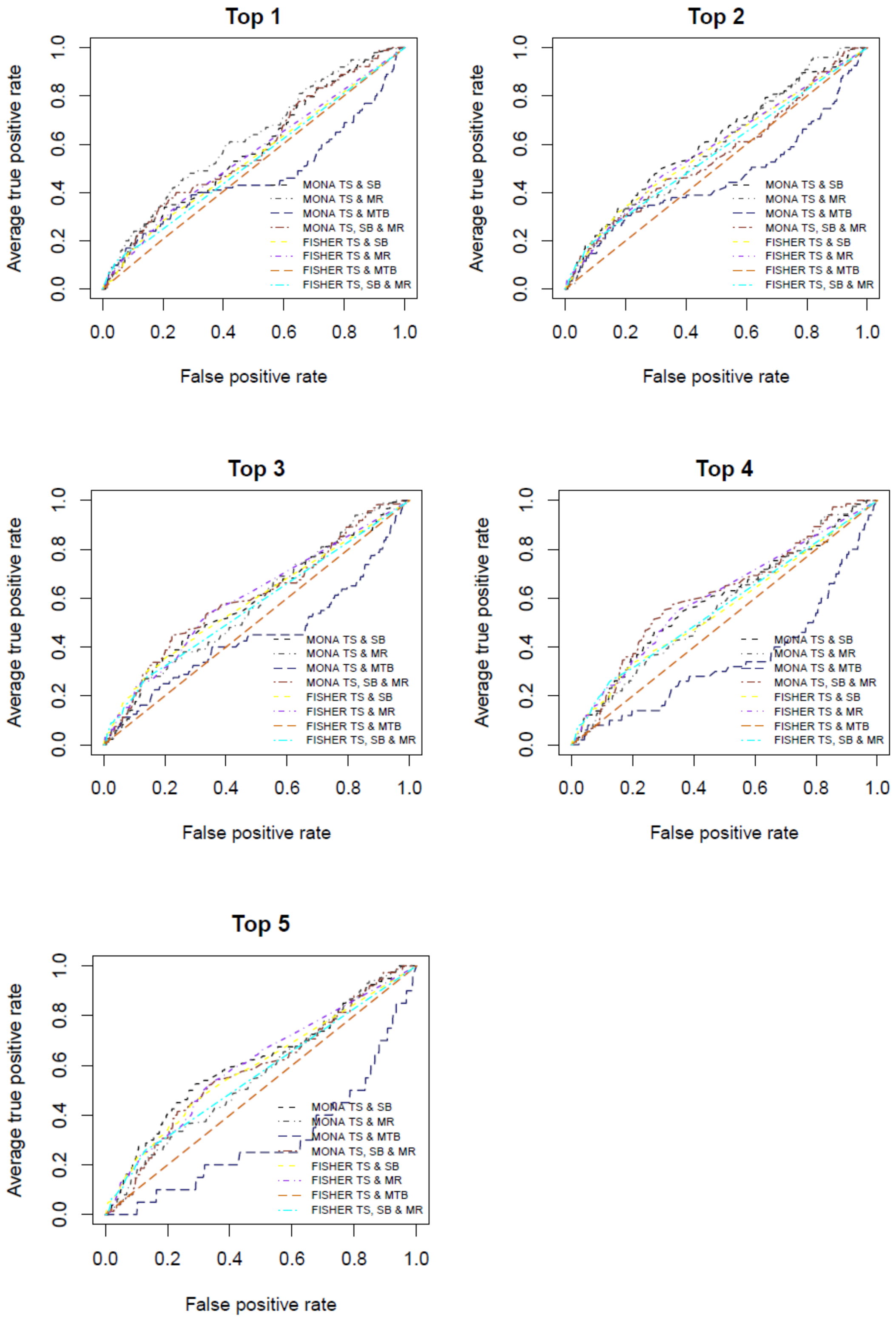


Figure A: Distributions of experimental (blue) and sampled p-values (pink) in the transformed log-log-space. The experimental data of the 10 sets was randomly taken from the Gene Expression Omnibus Database. The sampled p-value distribution was generated by the generic model of cMONA.


Top 1

Top 2

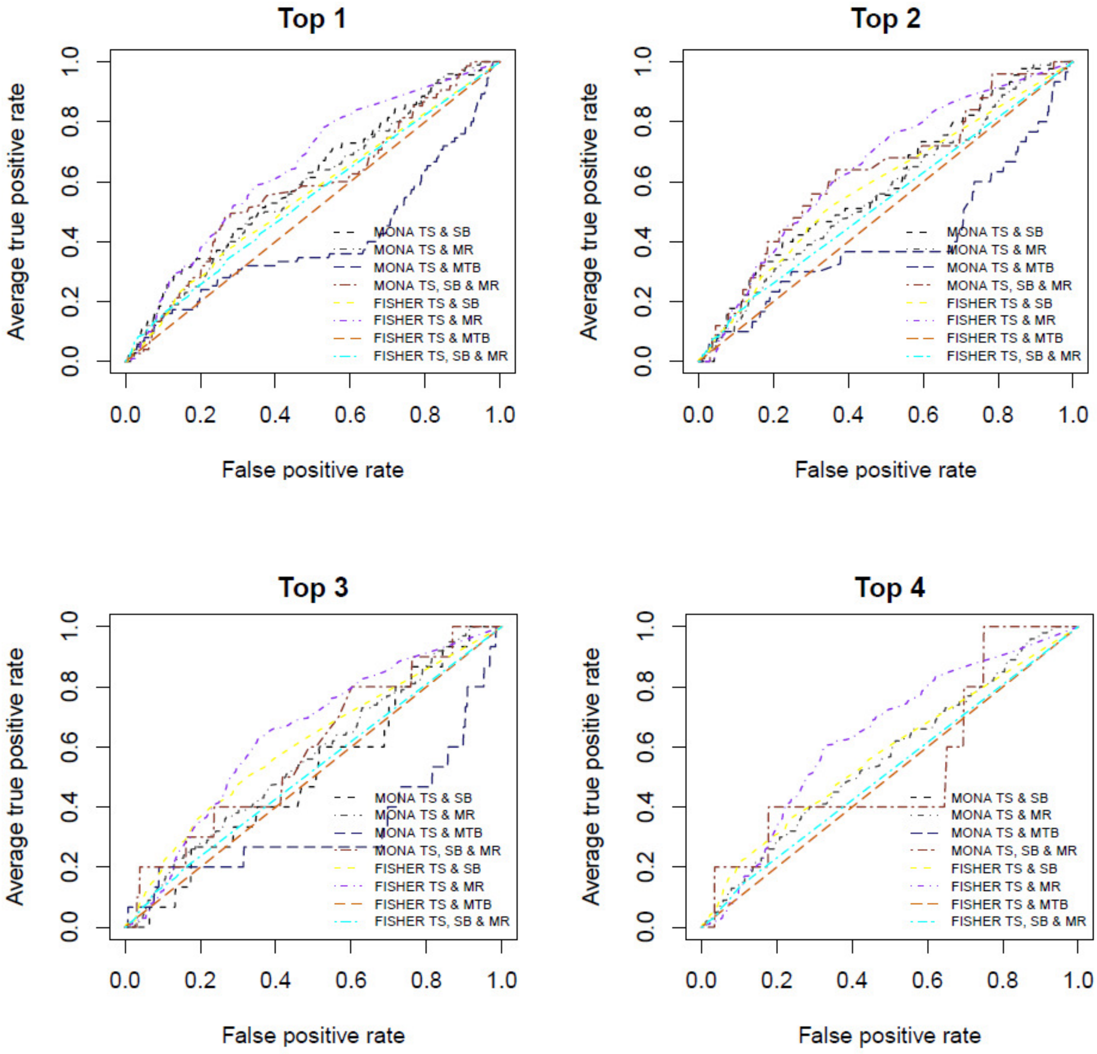
Top 3

Top 4

Top 5

Figure B1: Plots showing the performances of bot MONA and Fisher's exact test for the most common (Top 1) to the five most common (Top 5) miRNAs for

single miRNA enrichment. The data sets were generated using combinations of certain miRNA-target predictions: TargetScan and StarBase (TS & SB), TargetScan and MiRanda (TS & MR), TargetScan and MiRTarBase (TS & MTB) as well as combination of TargetScan, StarBase and MiRTarBase (TS, SB & MTB).



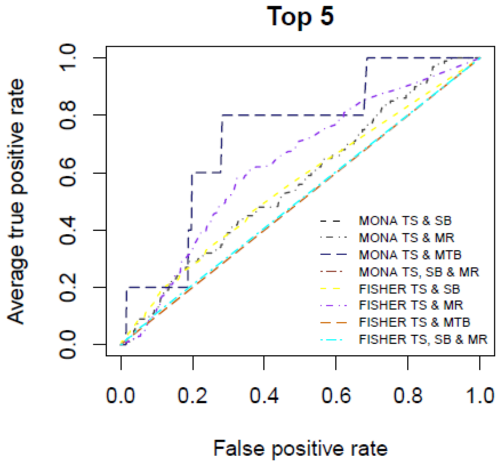


Figure B2: Plots showing the performances of bot MONA and Fisher's exact test for the rarest (Top 1) to the five rarest (Top 5) miRNAs for single miRNA enrichment. The data sets were generated using combinations of certain miRNA-target predictions: TargetScan and StarBase (TS & SB), TargetScan and MiRanda (TS & MR), TargetScan and MiRTarBase (TS & MTB) as well as combination of TargetScan, StarBase and MiRTarBase (TS, SB & MTB).