LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Masterarbeit
in Bioinformatik

Model-driven analysis of metabolic changes in adipocytes during differentiation

Benjamin Drexler

Aufgabensteller:	Prof. Fabian Theis
Betreuer:	Ferdinand Stückler
Abgabedatum:	22.01.2013

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

22.01.2013 _____

Benjamin Drexler

Abstract

Obesity is a major health risk factor in the Western world and developing countries. It describes the excess of white adipose tissue, which consists primarily of adipocytes. The differentiation of preadipocytes to mature adipocytes is called adipogenesis. In this thesis, we investigated this differentiation process based on invitro experiments. The differentiation to mature adipocytes was observed in three independent experiments using the human Simpson-Golabi-Behmel Syndrome cell strain. Over a timespan of 28 days, the intra- and extracellular concentrations of 188 metabolites were measured by mass-spectrometry. This metablomics data was analysed applying the following methods to obtain a better knowledge about adipogenesis. To get a first overview of the dynamics in the metabolite dataset, a k-means clustering was performed. These results indicated an increase of concentration of several phosphatidylcholines and a decrease of concentration of amino acids. Therefore, an enrichment analysis was performed to assess statistically significant concentration changes during differentiation. At first, biochemical class information was used for the metabolite set definition. This revealed a down regulation of intracellular amino acid concentrations during adipogenesis. To include a pathway feature in the metabolite set definition, we utilised a Gaussian graphical model based on the population data of the Kooperative Gesundheitsforschung in der Region Augsburg study. This allowed for the identification of several phosphatidylcholines. Their concentrations are highly altered during the experiment and are most likely in connection with the formation of lipid droplets. In addition, data about intra- and extracellular metabolite concentrations enabled us to investigate the exchange of metabolites between these two environments. The exchanges of metabolites might be in connection with signaling by metabolic intermediates or helps to answer questions regarding metabolite concentrations in blood. There are several time courses of metabolite concentrations, that could be due to an exchange between the two environments. Additionally, the exchange of metabolites might be also only during a short timespan. Therefore, we applied several methods which are based on the correlation coefficient or the t-value of the Student's t-test. The results indicate that several methods are needed to cover the variety of the biological scenarios which involve the exchange of metabolites. Further research in form of tracer experiments is needed to confirm these results. The results of this thesis can be used as a starting point for further investigations to increase the knowlegdge about the adipogenesis and the role of adipocytes as an endocrine cell. This knowledge is essential to reduce the impact of obesity and the associated diseases as major health risk factor.

Zusammenfassung

Adipositas ist ein schwerwiegendes Gesundheitsrisiko in der Westlichen Welt und Entwicklungsländern. Es ist definiert als der Überschuss an weißem Fettgewebe, welches hauptsächlich aus Adipozyten besteht. Adipogenesis bezeichnet die Differenzierung von Präadipozyten zu Adipozyten. In dieser Thesis wurde dieser Prozess basierend auf in-vitro Experimenten untersucht. Die Differenzierung zu Adipozyten wurde in drei unabhängingen Experimenten beobachtet, welche mit der menschlichen Simpson-Golabi-Behmel Syndrom Zelllinie durchgeführt wurde. Die intra- und extrazellulären Konzentrationen von 188 Metabolite wurden über eine Zeitspanne von 28 Tagen mittels Massenspektrometrie gemessen. Um ein besseres Verständnis über den biologischen Prozess Adipogenesis zu gewinnen, wurden diese Daten mit den folgenden Methoden analysiert. Ein k-means Clustering wurde durchgeführt, damit ein erster Überblick über die Daten ermöglicht wurde. Die Ergebnisse des Clusterings deuteten auf einen Anstieg der Konzentrationen von mehrere Phosphatidylcholinen, sowie eine Verringerung der Aminosäure Konzentrationen. Daher wurde eine Enrichment Analyse angewendet, um statistisch signifikante Änderungen der Konzentrationen festzustellen. Zunächst wurden Informationen über die biochemischen Klassen zur Definition von Metabolit-Gruppen verwendet. Mit Hilfe dieser wurde eine signifikante Runterregulierung der intrazellulären Aminosäure Konzentrationen festgestellt. Damit die Definition der Metbolit-Gruppen zusätzlich Pathway Informationen enthält, wurde ein Gaussches graphisches Modell basierend auf den Populationsdaten der KORA Studie erstellt. Diese Metabolit-Gruppen erlaubten die Identifikation von Phosphatidylcholinen, deren Konzentrationen stark verändert waren, welche wahrscheinlich im Zusammenhang mit der Entstehung von Lipid-Tröpfchen stehen. Durch die Messung von intra- und extrazelluläre Konzentrationen war es möglich den Austausch von Metaboliten zwischen diesen beiden Umgebungen zu untersuchen. Dieser Austausch könnte im Zusammenhang mit der Signalgebung von metabolischen Zwischenprodukten sein oder Fragen bezügliche der Metabolit-Konzentrationen im Blut beantworten. Es gibt mehrere Konzentrationsverläufe, welche dem Austausch von Metaboliten zu Grunde liegen könnten. Zusätzlich kann der Austausch nur während einer kleinen Zeitspanne statt finden. Daher wurden mehrere Methoden zur Untersuchung des Austauschs angewendet. Die Resultate verdeutlichen, dass mehrere Methoden notwendig sind, um die Vielfalt der biologischen Szenarien abzudecken. Die Ergebnisse dieser Untersuchungen müssen jedoch mit weiteren experimentellen Methoden bestätigt werden. Die Ergebnisse dieser Thesis können als Startpunkt für tiefergehende Untersuchungen dienen, damit das Wissen über Adipogenesis und die Rolle von Adipozyten als endokrine Zelle erweitert wird. Dieses Wissen is essentiell, um den Einfluss von Adipositas und die damit verbundenen Krankheiten als schwerwiegendes Gesundheitsrisiko zu verringern.

Acknowledgements

Thanks to Ferdinand Stückler for being a great supervisior with inspiring ideas and constant assistance, to Helmut Laumen and Kerstin Ehlers for their feedback throughout the whole work of this thesis and their improvements on a first manuscript, to Fabian Theis for his feedback during the thesis and notes to a first manuscript, to Prof. Hans-Werner Mewes for providing the possibility to write this thesis and to the whole CMB group for a great working environment.

Contents

Chapter 1

Introduction

In this thesis, we investigate the differentiation of preadipocytes to adipocytes. This process was observed in three independent in-vitro experiments over a time-span of 28 days. The intra- and extracellular concentrations of 188 metabolites were measured. These metabolomics data was analysed with several methods to indentify metabolites or metabolite sets that show significant changes in concentration during the experiment. Additionally, the exchange of metabolites between the intra- and extracellular environment is examined. In this chapter, we explain the biochemical background of the differentiation process and its meaning for the organism. In the beginning of Chapter 2, we explain the experimental design and the dataset. Afterwards, the applied methods are introduced. The results of these methods are shown in Chapter 3 and their biological meaning is discussed. At the end, we give a summary of this thesis and present an outlook in Chapter 4. Additional Figures and Tables, which are mainly of the results, are included in the Appendix.

1.1 Adipogenesis

The differentiation of mesenchymal stem cells to preadipocytes and then the terminal differentiation to adipocytes is called adipogenesis. Adipocytes are primarily embedded in white adipose tissue (WAT). Obesity describes the excess of WAT. Since it is associated with cardiovascular diseases and diabetes type 2, it is a major health risk factor in the Western world and developing countries like China [39] or India [14]. The WAT consists mainly of adipocytes, preadipocytes, fibroblasts, nerves and diverse immune cells [7]. It regulates the energy homeostasis of the organism, thus it is not only a passive energy depot. Rather, the WAT fulfills the function of an endocrine organ, which secretes several factors that play a central role in insulin sensitivity, lipid metabolism and satiety [10], immunological

responses and cardiovascular diseases [16, 21]. The size of the WAT is largely dependent on the number and size of the adipocytes. The number of adipocytes can increase during childhood and puberty [33]. In adulthood, there is an anually turnover of adipocytes of approximately 10%.

In this thesis, we investigate the differentiation step from preadipocyte to a mature adipocyte. There are several factors which are well established to influence the adipogensis. Both, the canonical and non-canonical WNT signaling inhibit the adipogenesis of a committed preadipocyte [8]. However, the canonical WNT signaling seems to be important for the survival of committed preadipocytes [28]. TGFβ is another signaling pathway that regulates adipogenesis. Depending on the experimental design and the cell state, the TGFβ signaling may inhibit or promote adipogenesis [8]. The stiffness and tension of the extracellular matrix (ECM) is another factor, which contributes to the regulation of adipogenesis [8]. Changes of the ECM and cytoskeletal components induce an alteration in cell shape from fibroplast-like to a spherical shape [11].

After a preadipocyte gets to a committed state, the major regulator of the adipogenesis are the transcription factors CCAAT/enhancer binding protein-β (CEBPβ), CEBPδ and peroxisome proliferator-activated receptor-γ (PPARγ) as they induce a transcriptional cascade (see Figure 1.1). Without PPARγ or CEBPβ and CEBPδ, the cell is not able to differentiate to a mature adipocyte. These major regulators induces the expression and synthesis of proteins which are associated with the adipocyte phenotype. These include, inter alia, fatty acid-binding protein 4 (FABP4), glucose transporter 4 (GLUT4), leptin, adiponectin, adipsin and resistin [8, 11]. Some of these help to fulfill the role as an endocrine cell with key parts in various physiological processes. Additionally, mRNA levels of enzymes that are involved in tricylglycerol synthesis and degradation are increased [11]. The cell accumulates lipids, which are stored in lipid droplets (LDs). These are cellular organelles that consist of a phospholipid monolayer as the membrane and have a hydrophobic core. Phosphatidylcholines are the most prevalent phospholipid of the membrane with up to 60% [24]. The biogenesis of LDs is not yet fully understood, but the current model invovles a budding of a nascent lipid droplet nearby the endoplasmatic reticulum. The main task of the phospholipids is the formation of a boundary to store triglycerols and other lipids, but they may also be involved in differentially recruitment of LD proteins [24]. The mechanism behind regulation of adipogenesis, biogenesis of LDs and the role of adipocytes as endocrine cells are yet to be completely understood and leave room for further research. A better understanding of these biological processes may help to increase the knowledge about obesity and may reduce the impact of the associated health risk factors in the Western world and developing countries.

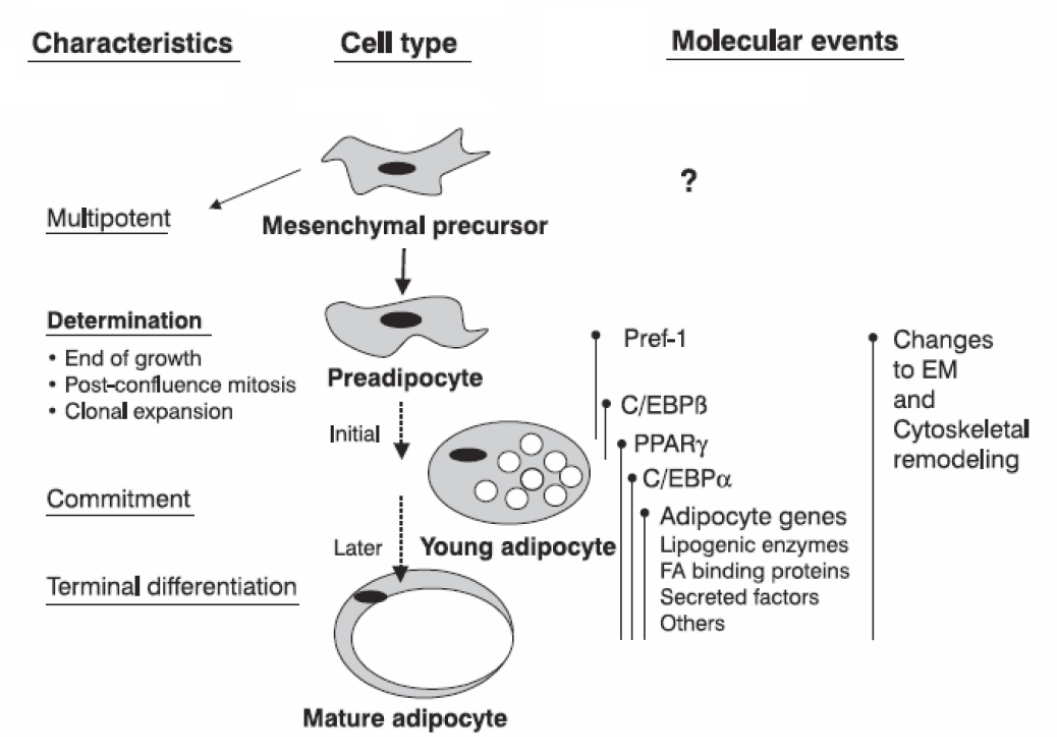

Figure 1.1: Scheme of the differentiation from a preadipocyte to a mature adipocyte. Characteristics and molecular events, e.g. the expression of transcription factors, that occur during the adipogenesis are denoted on the side. The Figure is adapted from [25].

1.2 Simpson-Golabi-Behmel Syndrome Cell Strain

The Simpson-Golabi-Behmel Syndrome (SGBS) cell strain can be used to investigate the differentiation of human preadipocytes to adipocytes and were the used cell strain in this thesis. The cell were prepared from an adipose tissue specimen of a patient with SGBS [9]. SGBS is a genetic disorder which is inherited in a X-linked recessive fashion. It is characterized by pre- and postnatal overgrowth, coarse facies, congenital heart defects, and other congenital abnormalities [40]. The mutation which leads to the disease is not detected in this cell strain.

Preadipocytes isolated from regular adipose tissue donors are well established for studies of adipogenesis. However, they have several drawbracks due to issues of the availability and variability of different donors. In addition, they have a limited life span. In contrast, the SGBS cell strain can proliferate up to 50 generations and it retains the capacity to differentiate to adipocytes. The analysis of gene expression of SGBS cells showed that they behave like adipocytes of regular adipose tissue [36].

The SGBS cells are not transformed or immortalized. They can differentiate under serum- and albumin-free culture conditions. The differentiation is induced by exposure to a mixture of insulin, triiodothyronine, cortisol and a PPARγ agonist. Wihthin a few days, an accumulation of lipids ain the cell is visible as small lipid droplets. Overall, using these cells one can reach up to >90% adipogenic differentiationrate [9].

The SGBS cell strain has been used in several studies analysing adipose differentiation, adipocyte glucose uptake or lipolysis. In general, cell lines or strains can be used to study the effects of single factors or hormones on specific cell types in vitro. However, these studies need to be verified in vivo, since the tissues are part of a complex organism and communicate with other cells and tissues.

1.3 Metabolomics

The metabolome describes the abundance of metabolites in a cell, tissue or organism. Metabolites are the products and intermediates of metabolism. The metabolome can change within seconds in contrast to the genome or proteome [5]. All cellular processes of the the transcriptome and proteome end in metabolites. Therefore, the metabolome is considered to be an indicator of an organism's phenotype endpoint. At the moment, around 40,000 metabolites are known of the human organism (Human Metabolome Database [38]). Additional metabolites are

taken up through the environment, which increases the number of metabolites up to 100,000 [5].

Metabolomics refers to the comprehensive study of these metabolites and their reactions. There are several methods to determine the metabolites and their concentrations. The three major determination methods are chromatography, nuclear magnetic resonance and mass-spectrometry. All of these methods include several submethods with each of them has their assets and drawbacks. Therefore, most of them are limited to a specific class of metabolites with specific biochemical properties. A composition of methods can be used to cover a wide range of metabolites.

Metabolomics is used for various applications in research. For instance, the ratio of phenylalanine and tyrosine can be used as a biomarker of the disease phenylketonuria (PKU) for newborn [6]. Another study examined the human response to glucose challenging [32]. Illig et al. [15] used metabolomics in a genome wide association study to show that there is association between metabolite frequencies and genetic variations. There is also one study that used metabolomics to investigate the adipocyte differentiation of the 3T3-L1 murine cell line [27]. Metabolomics can be combined with other *omics* like transcriptomics or genomics to gain further knowledge about biological processes.

1.4 Motivation

As described in Section 1.1, the biological process adipogenesis is not yet completely understood. In this thesis, we have the possibility to analyse metabolomics data of a human cell strain that differentiates from preadipocytes to adipocytes. We want to identify metabolites or pathways of which the concentration is highly altered during the experiment. These metabolites or pathways are most likely to be regulated during adipogenesis and their influence could be essential to comprehend the differentiation process. In addition, the extracellular measurements enable us to investigate the exchange of metabolites between the intra- and extracellular environment during adipogenesis. These results could be set in a connection with the metabolite concentrations in blood samples or help to understand the signaling of metabolic intermediates. Overall, the results of this thesis might be able to increase the understanding of adipogenesis or reveal interesting aspects of these metabolomics data which could be followed up by further research.

Chapter 2

Materials and Methods

This chapter deals with the applied methods and the used materials in this thesis. It explains the used dataset for the analysis, necessary preprocessing steps and the performed analysis methods. All the results are then shown and discussed in the Chapter 3.

2.1 Experimental Data of Adipocyte Differentiation

The dataset contains the data of the biological experiment which was conducted by the workgroup of Dr. Helmut Laumen (Technische Universität München). In the following sections, the experimental design and further information about the measured metabolites are explained.

2.1.1 Experimental Design

The experiment was conducted in order to analyze the adipogenesis of the human Simpson-Golabi-Behmel Syndrome (SGBS, see Section 1.2) cell strain. The cells were cultivated for 32 days (from day -4 to day 28). The growth medium was replaced every two days. The differentiation of the cells is induced by a change of growth medium from *proliferation medium* to *induction medium* at day 0. Another switch of growth medium to *feeding medium* occured at day 4. The composition of induction medium and feeding medium is slightly different, whereas the proliferation medium is the only one which contains fetal calf serum (FCS).

The cells were harvested before and at eight days after induction of differentiation. An overview of the harvesting days and the changing growth medium is displayed in 2.1. Additionally, a sample of the supernatant was taken and

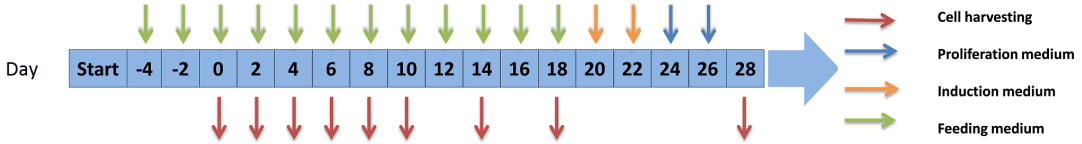

Figure 2.1: Overview of the experimental design. The harvesting points and the change of growth medium are marked.

the metabolite concentrations in cells and supernatant were measured with mass-spectrometry. The measurements of the cells will be referred to as *intracellular* and supernatant measurements will be referred to as *extracellular*. The medium has an influence on the extracellular metabolite measurements. This is especially the case for the proliferation medium, since it is the only medium with FCS. Some of these metabolites are measured in the supernatant and an effect of the medium is measured in the supernatant. Proliferation medium was used from day -2 and day 0. Hence, day 0 is left out for analysis which deal with extracellular measurements.

The experiment was conducted three times (experiment #1 to #3) which we refer to as *biological replicates*. The concentration of an intracellular metabolite sample was also measured three times which are referred to as *technical replicates*. There are no technical replicates for the supernatant samples. To sum up, the number of intracellular measurements per metabolite is 81 (number of biological replicates × number of harvesting points × number of technical replicates) and the number of extracellular ones is 27, since there are no technical replicates. Hence, the overall number of datapoints should be 108 per metabolite. This is not the case, because the determined concentration for all metabolites at experiment #3, day 6, technical replicate #1 and for all technical replicates of experiment #3, day 10 are missing. So, there are only 104 datapoints per metabolite available.

2.1.2 Measurement of Metabolites

The metabolite concentrations were determined with the Biocrates AbsoluteIDQ^{TM} p180 kit. It is based on electrospray ionization tandem mass spectrometry (ESI-MS/MS) and it measures the concentration of 188 metabolites. To give a brief overview about the panel of measured metabolites, it contains 40 acylcarnitines, 21 amino acids, 21 biogenic amines, 90 glycerophosholipids, 15 sphingomyelins and 1 sugar. Table A.1 lists all 188 metabolites of this dataset.

Since not all samples fit onto one plate, the concentrations were determined in two batches. The first batch contains the intracellular samples of all nine harvesting points and technical replicates for experiment #1 and experiment #2. Experiment #3 is present with the first five harvesting points (day 0 to day 8). The remaining

harvesting points of experiment #3 are measured on batch #2. All extracellular measurements were also part of batch #2.

Additionally, *zero samples* were measured for every metabolite besides the experiment samples. These are samples that only contain extracting solvent (80% methanol) and are used to determine the noise in the measurement. Every batch contained eight zero samples per metabolite. So overall, there are sixteen zero samples.

2.1.3 Differentiation Marker

As described in Section 1.1, there are several transcription factors and other proteins that are expressed during the differentiation to adipocytes. Some of these markers were measured during the experiment to assess the differentiation status of the cells. The following markers were measured:

- Lipid accumulation: measured by Oil-Red-O staining
- Glycerol-3-phosphate dehydrogenase (GPDH): enzyme activity measured
- Peroxisome proliferator-activated receptor-γ (PPARγ): mRNA expression measured
- Leptin: mRNA expression measured

The data of these differentiation markers is shown in the Figures A.1, A.2 and A.3. Dr. Helmut Laumen provided an analysis of these differentiation marker to assign different phases of the differentiation to the measuring days of the experiment. All differentiation markers indicate a preadipocyte state at day 0 and 2, since there is no activity or they are not expressed. At day 4, PPARγ and Leptin are slightly expressed, but there is no lipid accumulation. From day 4 to 14 an increase of GPDH enzyme activity, PPARγ mRNA levels and the accumulated lipids is observeable. This indicates that the cells were accumulating triglycerides. These three markers reached the maximum level at day 18 to 28. Additionally, the expression of leptin increases at 18 and reaches its maximum at day 28, which confirms that the cells are differentiated to adipocytes. Based on these observations, day 0 and 2 are considered to be the early phase of differentiation, day 4 to 14 the middle phase and day 18 and 28 the late phase.

2.2 Quality Control

Since metabolomics data is always affected by noise, it is crucial to perform a quality control, because the analysis is highly dependent on the quality of the data.

The follwing criteria were used to perform the quality control.

Missing Values

Metabolites contain a varying number of missing values, so that there is no measured concentration. This is dependent on the difficulty to measure this certain metabolite. Since an analysis is only meaningful with a sufficient amount of datapoints, we defined the threshold such that the fraction of missing values should be below 75% of a metabolite. This criteria was applied independently for intra- and extracellular, so that a metabolite with a high fraction of missing values in one environment is not necessarily left out in the other environment.

Coefficient of Variation

The coefficient of variation (CV) is a normalized measure to evaluate the dispersion of a distribution. It is defined as follows:

$$CV := \frac{\sigma}{\mu}$$

where σ denotes the standard deviation and μ the mean of the technical replicates for a metabolite at a certain experiment and measuring day. Since there are no technical replicates for extracellular measurements, this was only possible for the intracellular measurements. Afterwards, the intra- and extracellular measurements of metabolites were excluded in which over 80% of the datapoints had a CV greater than 0.25.

Limit of Detection

The mean of the zero samples were used to calculate the specific limit of detection (LOD) for every metabolite. Because the zero samples only contain extracting solvent, they are used to evaluate the noise in the measurement of this metabolite and a measurement below the LOD indicates that this measurement is mostly affected by noise.

Two unique LODs per metabolite were calculated, since two batches were used in the measurement process. For the intracellular case, metabolites will be exlcuded, if 70% of the measurements are below the LOD. The threshold for extracellular measurements was reduced to 60%. The distinct thresholds were used to take the varying number of datapoints in the two environment into account. A more

conservative threshold was used, since the number of datapoints is much smaller for the extracellular case.

2.3 Preprocessing of Data for Analysis

Due to the characteristics of the dataset (see Section 2.1), a preprocessing of the data was necessary to proceed with the analysis. This implied a correction of a batch effect and imputing of the missing values.

2.3.1 Correction of batch effect

Since one part of experiment #3 was measured on batch 1 and the other part on batch 2 (see Section 2.1.2), it was necessary to correct for this batch effect. That is, the measured concentration of an identical sample results in different concentrations when measured on different batches.

The affected measurements of the batch effect are the intracellular values of experiment #3 at the measuring days 14, 18 and 28. The correction of the batch effect was performed as follows. Every metabolite is treated individually and two means are calculated. One is the mean of experiment #3 of the intracellular values of day 14 to day 28. The other mean is based on the same measuring days, but it consists of the values of experiment #1 and #2. The correction value for every metabolite is then the difference between the mean of experiment #3 and the mean of experiment #1 and #2. This value is then subtracted of every value which was described to be a target of the batchcorrection earlier in this section. It could be the case that this correction methods leads to negative values, i.e. measured value $<$ correction value. Such values are set to a missing value and might be imputed in a later step of the preprocessing.

2.3.2 Imputing of Missing Values

Some metabolites contain missing values. These missing values are imputed by a process which consists of two steps that are sequentially performed.

Imputing of Missing Values Due to Low Concentrations

The first imputing step is based on the assumption that some values are missing due to a low concentration of the metabolite. Thus, there is no signal in the mass-spectrometry measurement. These values are imputed with the minimal measured concentration of this metabolite over all experiments and measuring days. But only missing values were qualified, which had a technical replicate (same measuring day

and experiment) with a concentration that was below the LOD. This indicates that the missing value was due to a low concentration for this sample and this imputing method is applicable.

Imputing with Linear Interpolation

For the second step, there are two variants to impute a missing value. The first uses the information of the same experiment. The mean over the technical replicates of the measuring day before and after the missing value are calculated and the missing values is imputed with the mean of these two values. If this approach is not possible (i.e. missing values at the measuring days before and after), the mean of the other two experiments at the measuring day of the missing value is calculated. Imputed values of this step were not used for imputing of further values. This method was also used to impute missing values of time course metabolomics in the HuMet study [17].

2.4 Statistical Analysis

This section deals with fundamental statistical analysis. They were used as a part of other applied methods or they provide a general analysis of the data.

2.4.1 Normalization of the Data

The concentration of the metabolites can vary in great ranges, but these differences do not necessarily indicate the biological relevance of the metabolite. Therefore, normalization of the metabolite concentrations was performed depending on the analysis or applied method.

Calculation of the Fold Change

The fold change allows an easy evaluation whether the concentration of a metabolite was in- or decreasing over the time in respect of the starting point of the experiment. For the calculation of the fold change, the mean of all values at a certain measuring day (up to 9 for intracellular and up to 3 for extracellular) was calculated. Afterwards, the ratio from the measuring day to the first measuring day is calculated. We also logarithmised the fold changes to base 2, so that a value of 1 corresponds to a doubled concentration and a value of -1 to a halved one. Some concentrations are measured with a 0 and the logarithm results in a missing value. These values are replaced with the lowest value of the metabolite. The log(fold change) of a metabolite will be referred to as logFC.

Fold Change of Grouped Measuring Days

Some of the methods or analysis were not only applied for the measuring days, but also with grouped measuring days to account for differentiationphases. The calculation of the logFCs had to be adapted for that. First of all, the value of a differentiationphase is the mean of all measuring points (all technical replicates and all experiments) of all measuring days within the certain differentiationphase. When comparing two differentiationphases (e.g. early to middle or middle to late), the fold change is then calculated in such way, that it is relative to the first mentioned differentiationphase.

Z-Score

The z-score is another way of normalizing the concentrations of a metabolite. With this normalization, the mean of the values is 0 and the standard deviation is 1. The z-score is defined as follows:

$$z := \frac{x - \mu}{\sigma}$$

with μ being the mean and σ being the standard deviation of the considered measuring points. These are typically all measuring points (all experiments, all measuring days and all technical replicates) of a metabolite of one of the two environments.

2.4.2 Distribution of Data

The knowledge about the distribution of the data is very important, since some statistical tests or methods are based on the assumption that the dataset underlies a normal distribution. Earlier investigations of metabolomics data revealed that they usually follow a log-normal distribution [17, 18]. Figure A.4 shows the distribution over all measuring points of the dataset, when the values are logarithmised.

We performed a visual examination of our dataset with Q-Q plots to validate these observations. A Q-Q plot displays the quantiles of two distributions [37]. In this case, we compare the quantiles of the sample (i.e. metabolite concentration) against the theoretical quantiles of a normal distribution. The plot will be close to linear, if the sample underlies a normal distribution. Overall, a majority of the metabolites were closer to a log-normal distribution than to a normal distribution. Thus, if not stated otherwise, the log(concentrations) will be used for further analysis. Some metabolites have measured concentrations of 0, for which the logarithm is not defined. Hence, before the concentrations are logarithmised, these 0-values are replaced with the minimum value of this metabolite which is not zero.

2.4.3 Student's t-Test

A Student's t-test can be applied to evaluate whether the means of two populations are equal. To apply this test, the populations have to follow a normal distribution and the variance has to be equal. As a part of the test, the t-value is calculated. Afterwards, the t-distribution is used to calculate the confidence interval to evaluate the H_0 hypothesis which assumes that the means of the two populations are equal. The H_0 hypothesis is rejected at a significance level α of 0.05.

In this thesis, we are dealing with log-normal distributed data. We can not ensure that the two used populations of the t-test always have the same variance, but we assume it and we have to keep that in mind when evaluating the results.

2.4.4 Multiple Testing Correction

Multiple testing describes the simultaneously application of a statistical test, which typically leads to an increase of a type I error (or false positives). This is, the H_0 hypothesis is falsely rejected. Several methods have been proposed to correct for multiple testing. In this thesis, we applied a method which is based on the control of the false discovery rate (FDR) [2]. The FDR is defined as the expected proportion of false positives to the number of hypothesis which are declared to be significant. This method is considered to be less stringent, but more powerful, than methods which are based on the family wise error rate such as Bonferroni correction.

2.4.5 Principal Component Analysis

The principal component analysis (PCA) is a statistical multivariate technique to extract the important information from the data [1]. A principal component is a linear combination of several variables, e.g. metabolites or measuring days. The explained variance of a principal component decreases with an ascending numbering of the component, so that the first one has the largest possible variance. Additionally, the calculation of a principal component is under the constraint to be orthogonal to the preceding principal component. In this thesis, the PCA is used to give an brief overview of the data. For this, the concentrations were normalized with the z-score (see Section 2.4.1).

2.5 Clustering of Metabolite Time Courses

Several algorithms were applied to cluster the time courses of the metabolites. Their results are used to give an overview of the dataset with an visual representation.

2.5.1 Distance Measures

A distance measure is used to calculate the distance between two observations, e.g. fold changes of metabolites. Two of these distance measures were used in this thesis.

Euclidean Distance

The Euclidean distance between two vectors x and y (e.g. fold changes of metabolites) is defined as follows:

$$d(x, y) := \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.1}$$

- x, y: vectors with the length of n
- x_i, y_i: the i-th element of the vector, e.g. the concentration at measuring day 4

Pearson Product-moment Correlation Coefficient

The Pearson product-moment correlation coefficient (PCC) describes the linear dependence between two variables. The PCC r for two samples x and y is defined as follows:

$$r := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.2}$$

- x_i: the i-th obversation of x
- $\bar{x}$: the mean of x

The PCC r ranges from -1 and 1. A value above 0 indicates a positive correlation and a value below 0 a negative correlation. At a value of 0, there is no linear dependence between the two variables.

2.5.2 Hierachical Clustering

The hierachical clustering tries to build a hierachy of clusters. For that, the distance between all samples is calculated with a metric. Afterwards, the samples or clusters with the smallest distance are joined in further clusters. This bottom-up approach is also called agglomerative hierachical clustering. At the end, there is

one cluster which cointains all samples. The hierachy of clusters can be visualized by a dendogramm.

For the initial distance calculation between the samples, the distance measures Euclidean distance. The distance between two clusters is calculated with the complete linkage criteria:

$$d(X, Y) := max\{d(x, y) : x \in X, y \in Y\}$$

To cluster the data, the mean of the technical replicates and the mean of the three experiment was calculated. Afterwards, the z-score (see Section 2.4.1) over the measuring days of a metabolite was computed to normalize the data.

2.5.3 k-means Clustering

The k-means clustering algorithm assigns the observations (i.e. metabolite concentrations) to k cluster. It consists out of two steps which are iterated. The algorithm stops when the assignment of the samples do not change anymore. The two steps are called assignment and update step. In the assignment step, the samples are allocated to the cluster with the closest mean or centroid, respectiveley. We decided to use the Euclidean distance and the PCC as a distance measure. A re-calculation of the mean or centroid is performed in the update step. k obversations are choosen randomly as the initial centroids. Since it is not ensured that the k-means clustering algorithm converges to the global minimum, we repeated the calculation 500 times.

Determine the Number of Clusters via Silhouette

The silhoutte aids with the visual investigation of a cluster analysis [29]. It can be used to evaluate whether an observation lies within or merely at the outside of its cluster. For every observation i, the value s_i is calculated which is defined as follows:

- i: the i-th observation, i.e. metabolite concentrations
- a_i: average distance to all the other observations in the cluster A
- b_i: average distance to all the observations to the closest neighbor cluster B

$$s_i := \frac{b_i - a_i}{max\{a_i, b_i\}}$$

so that s_i ranges from -1 to 1. The s_i values are displayed in a plot at which the s_i values are grouped in the corresponding clusters and in a descending order.

This plot can be used to evaluate the result of a cluster analysis. In this thesis, we mainly used it to determine a k, the numbers of clusters, for the k-means clustering.

2.6 Enrichment Analysis of Metabolite Changes

Enrichment analysis are widely used to evaluate whether there is a certain gene annotation (e.g. biological or molecular function) enriched in a gene set. In this thesis, we used an enrichment analysis to assess the changes in concentration of metabolite sets.

2.6.1 Definition of Sets

A metabolite set is a group of metabolites that have a similarity or connection in respect of certain characteristics. The following sections describe the applied methods to define the metabolite sets.

Biochemical Classes

This definition of the metabolite sets is based on the biochemical classes and properties of the metabolites, e.g. sphingomyelins or amino acids.

GGM-based Classification

This method consists of two steps. The first one is the calculation of a Gaussian graphical model (GGM) and the second one is the detection of communities within this GGM.

The GGM can be represented as a network in which the nodes are metabolites and the edges are based on partial correlation between the metabolites. The partial correlation describes the linear dependency between two random variables, i.e. the concentrations of two metabolites. Furthermore, the partial correlation is conditioned against the whole set of random variables in contrast to the Pearson correlation coefficient. This leads to the outcome that indirect effects of correlation are removed as it can be seen in an example in Figure 2.2.

At first, we estimated a GGM with the values of this dataset. The GeneNet package was used to calculate the partial correlation [31]. The outcome was not satisfying due to the low number of samples in contrast to the number of variables, i.e. metabolites. Therefore, we used the GGM which was already calculated by Krumsiek et al. [18]. This GGM is based on the metabolite concentrations of the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) study [13, 15]. Even though, our dataset is based on cells that undergo adipogenesis

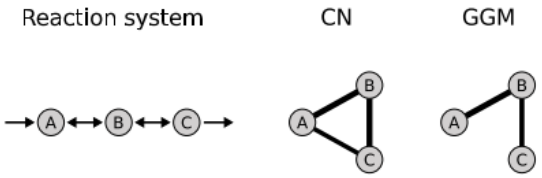

Figure 2.2: This example illustrates that indirect interactions between metabolites are eliminated in a Gaussian graphical model (GGM) in contrast to a correlation network (CN). Figure is adapted from Krumsiek et al. [18].

and the KORA study has samples of blood serum, we decided that we can use the KORA GGM, because both datasets are based on human samples and the fundamental metabolic pathways should be the same. Krumsiek et al. calculated a partial correlation coefficient above 0.1619 to be significant. We decided to use a cutoff of 0.18 for edges between the metabolites to follow a more conservative approach. The corresponding network is called a GGM which was used to detect modules within.

The next step is the detection of communities within the GGM. A community is a group of nodes in the network that show a high density of edges within the community, but sparse connections to nodes outside the community [22]. There are several approaches to detect such a community structure. We applied an algorithm that is based on the maximization of modularity. The modularity is defined as the number of edges within the groups minus the expected number of edges in an equivalent random network [22]. The applied algorithm is a heuristic approach to optimize the modularity [4]. It reiterates two phases in which the number of communities decreases until there is no further improvement of the modularity. The algorithm was provided as the brain connectivity toolbox in MATLAB (The Mathworks Inc.) [30].

Finally, a manual curation step of the community classification was necessary because of two reasons. First, not all metabolites in our dataset were measured in the KORA study. Second, there a metabolites in the GGM which do not have any edges to other metabolites. Hence, they are considered as an own community by the detection algorithm.

2.6.2 Enrichment Analysis

We implemented two methods to perform an enrichment analysis which are based on different approaches. One uses t-tests to evaluate the changes of a metabolite

set and the other one compares the distribution of the set metabolites against the distribution of all metabolites.

Hypergeometric Test (t-Test based)

First of all, a t-test of the log(concentrations) (see Section 2.4.2) between all datapoints of the the first measuring day and another one is performed. This is done for all measuring dayss and metabolites. Thus, the overall number of t-tests is number of metabolites $\times$ (number of measuring days -1). The H_0 hypothesis that the two samples have the same mean is rejected at a significance level α of 0.05. Thus, the log(concentrations) of the metabolite are considered to be significantly different between these two measuring days. To account for multiple testing, p-values were corrected using FDR (see Section 2.4.4).

Afterwards, the metabolite set information is used to perform a hypergeometric test for every metabolite set and every assessed measuring days combination. The hypergeometric test is based on the hypergeometric distribution. It describes the probability of x successes in N draws (without replacement) from a finite population of size M containing K elements with the desired characteristic. These variables have the following meaning in our test:

- x: number of metabolites of one specific set which show a significant difference
- N: overall number of metabolites which show a significant difference
- M: overall number of metabolites
- K: number of metabolites in the corresponding set

The probability density function is defined as follows:

$$f(x|M,K,N) := \frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

which is used in the cumulative distribution function to calculate the p-value as follows:

$$\text{p-value} = 1 - \sum_{i=0}^{x} \frac{\binom{K}{i-1}\binom{M-K}{N-i}}{\binom{M}{N}}$$

This test is identical to the one-tailed Fisher's exact test [26]. This approach reveals which metabolite sets show an enriched fraction of metabolites with a significant difference at a certain measuring day.

Distribution-based Tests

We used two tests to evaluate whether the distribution of the $\log_2$(fold changes) of a metabolite set is different than the distribution of all metabolites at a certain measuring day. Hence, the number of tests is number of metabolite sets $\times$ number of measuring days.

Kolmogornov-Smirnov Test

The Kolmogornov-Smirnov (K-S) two-sample test is a nonparametric test that compares the underlying probability density function of two populations [41], i.e. $\log_2$(fold changes) of the metabolite set and all metabolites. For that, it quantifies the distance between the distributions. The H_0 hypothesis is that both distributions are equal and is rejected at a significance level α of 0.05.

Wilcoxon Rank Sum Test

The nonparametric Wilcoxon rank sum test is used to evaluate whether the mean ranks of two populations differ. The test relies on the assumption that the two samples are independent, but in contrast to the t-test, there is no need for an underlying normal distribution. The H_0 hypothesis states that the mean difference is zero and is rejected at a significance level α of 0.05.

Weighted Enrichment Analysis

The weighted enrichment analysis determines whether the values of a metabolite set are statistically overrepresented. That is, the values are higher than the values of the other metabolites. The logFCs or t-values of a t-test were used as the values of the metabolites, so that they are comparable between the metabolites. For the weighted enrichment analysis, the sum of the values for each metabolite set is calculated. Therefore, this test is performed with non-negative values, i.e. the absolute-values of the logFCs or t-values. This sum is denoted as e. To assess the statistical overrepresenation, the set assignments of the metabolites are randomly shuffled. Afterwards, the sum e_r of this random metabolite sets are calculated and compared to e. This process was repeated $r = 10^6$ times and the number of times when $e_r > e$ was counted as f. The empirical p-value is then calculated as follows: $p = f/r$. This method was recently applied on metabolomics data [19].

2.7 Analysis of the Intra- and Extracellular Metabolite Dependency

The following sections describe methods which were used to assess the exchange of a metabolite between the intra- and extracellular environment. These methods are based on the correlation of the concentration or use the t-statistic to evaluate the changes in concentrations.

2.7.1 Global Correlation

This methods uses the Spearman's rank correlation coefficient (SCC) to evaluate whether there is an association between the intra- and extracellular log(concentrations) of a metabolite. To account for the distribution of the metabolite data, we applied the rank correlation coefficient. The SCC ρ is defined as follows:

$$\rho := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- x, y: intra- and extracelluar concentrations
- n: number of considered measuring days
- x_i, y_i: rank of the i-th concentration
- $\bar{x}$: mean rank of x

The log(concentrations) of all three experiments are used. For that, the mean of the technical replicates of the intracellular measurements is calculated. As explained in Section 2.1.1, the extracellular concentrations of some metabolites is affected by the growth medium at the first measuring day. Therefore, day 0 is left out of the calculation of ρ. So overall, 24 concentrations are used for the calculation (number of experiments $\times$ (number of measuring days - 1) $= 3 \times 8 = 24$). The probability that ρ is significantly different from 0 is calculated with a permutation test. A metabolite is considered to be correlated or anti-correlated when the p-value is below 0.05.

2.7.2 Window-based Correlation

In addition to the global correlation (see Section 2.7.1), we applied a method that can be considered as a local approach. The Spearman's correlation coefficient ρ is calculated with the concentrations that are in between a certain time window. The number of considered measuring days is the window size which is denotated

as ws. Hence, a metabolite does not have a single ρ. Instead there are $8 - ws + 1$ ρ's per metabolite. The calculation of ρ is slightly altered:

$$\rho := \frac{\sum_{i=s}^{e}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=s}^{e}(x_i - \bar{x})^2 \sum_{i=s}^{e}(y_i - \bar{y})^2}}$$

- x, y: intra- and extracelluar concentrations
- s: start index of the window
- e: end index of the window
- x_i, y_i: rank of the i-th concentration
- $\bar{x}$: mean rank of x

2.7.3 t-Value Based

This method seeks for remarkable changes between measuring days of the intra- and extracellular log(concentrations). For that, a t-test is applied for all adjacent measuring days starting at day 2 (see Section 2.1.1 for the reason), e.g. day 2 and day 4, day 4 and day 6 and so on. Dependent on the environment, the two populations consist of nine (intracellular) or three (extracellular) log(concentrations) for a single t-test. Due to the low number of samples per measuring day, it is unlikely to observe significant changes. Hence, we did not use the p-value to seek for significant changes, but instead the t-value of the t-test (see Section 2.4.3) was used to evaluate the change between the measuring days. A certain threshold was defined, so that $|t| > treshold$ indicates a timespan (e.g. day 4 to day 6) with a remarkable change. This *threshold* was set to 1.25. After the assessment of the timespans for intra- and extracellular was done, we were looking for metabolites, in which there were at least two timespans that showed a remarkable change for both environments, i.e. intra- and extracellular.

Chapter 3

Results and Discussion

In this chapter, we show the results of the different analysis methods and discuss their biological meaning. First of all, results of the data preprocessing (e.g. quality control or handling of the batch effect) are shown. The next section deals with the analysis of metabolite time course of a single environment. In the end, the exchange between intra- and extracellular is investigated.

3.1 Metabolomics Data Analysis and Preprocessing

Before further analysis could be performed, a preprocessing of the data was necessary. This includes a quality control of the metabolites, correction of a possible batch effect and imputing of missing values. The results of these methods are shown in the first two sections. The last section gives a brief overview of the data.

3.1.1 Quality Control

The quality control was individually carried out for the intra- and extracellular environment. At the end there are three sets of metabolites that passed the quality control. One for intracellular, one for extracelluar and the last one is the intersection of the two sets before. The workflow and the numbers for the intersection of intra- and extracellular are shown in Figure 3.1. Table 3.1 and Table 3.2 give an overview about the criteria and the metabolites which did not pass the certain criteria.

The quality control consists of two phases. The first phase was the elimination of metabolites with a fraction of missing values equal or higher than 75%. 10 (intracellular) or 6 (extracellular) metabolites were excluded due to this criteria (see Figure 3.2). The other two criteria, coefficient of variation (CV) and limit

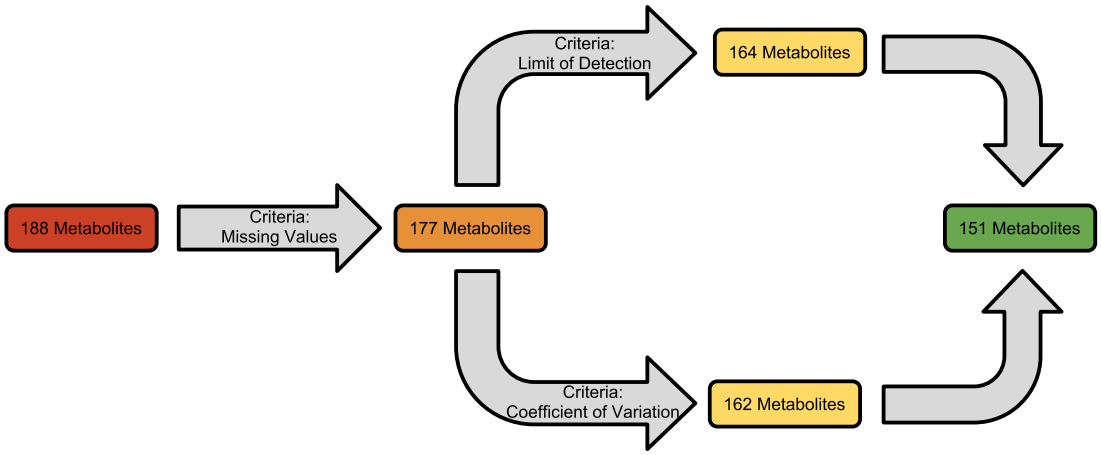

Figure 3.1: The workflow of the quality control. The first phase was the elimination of metabolites with a high number of missing value. The resulting set of metabolites was the basis for the application of the other criteria, which were carried out independently. At the end, the intersection of these two sets was build. The given numbers are for the combination of the intra- and extracellular environment.

of detection (LOD) (see Section 2.2), were independently applied in the second phase. We wanted to evaluate the consensus between these two criterias, because metabolites that do not pass the limit of detection criteria, could be more affected by noise in the measurement and this could also be exposed in a high CV.

The LOD criteria was not passed by 11 (intracellular) or 4 (extracellular) metabolites, respectively. This discrepancy can be explained with the influence of the growth medium, which in general, leads to higher concentrations in the extracellular environment. Figure 3.3 displays the distribution of the fractions below the LOD. The evaluation of the CV criteria was only performed on the intracellular measurements, because there are no extracellular technical replicates. Since no evaluation for the extracellular measurements was possible, we decided to exclude the 15 metabolites, that did not pass this criteria (see Figure 3.4), not only for the intracellular environment, but also for the extracellular one. The consensus between the CV and LOD criteria is 0 at the intracellular environment, which was surprising. Hence, the combination of these two criteria leads to a further reduction of the set of metabolites.

At the end, there are 152 metabolites in the intracellular set and 167 metabolites in the extracellular set. The intersection of these two sets contains 151 metabolites.

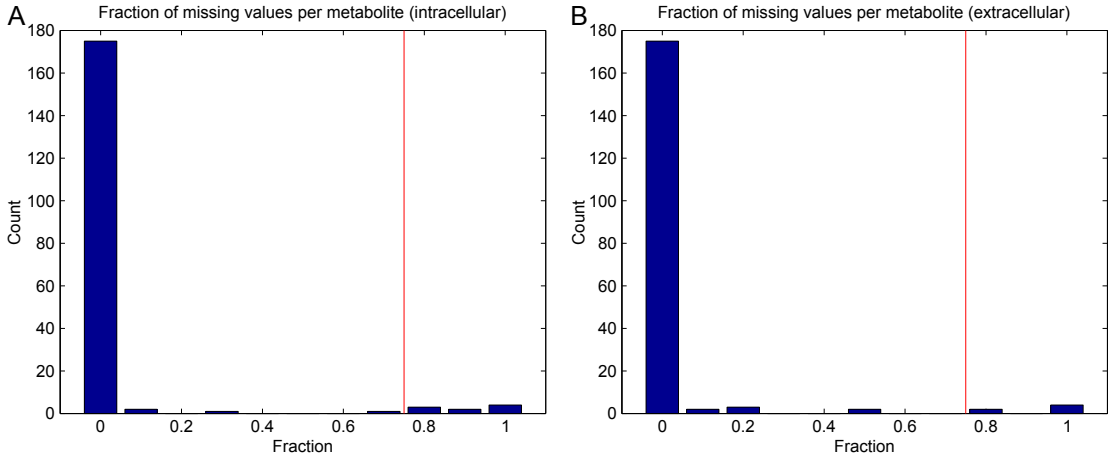

Figure 3.2: The fraction of missing values per metabolite is shown in a histogram representation for the intracellular (A) and extracellular (B) measurements. All metabolites with a fraction above 0.8 (red line) were excluded.

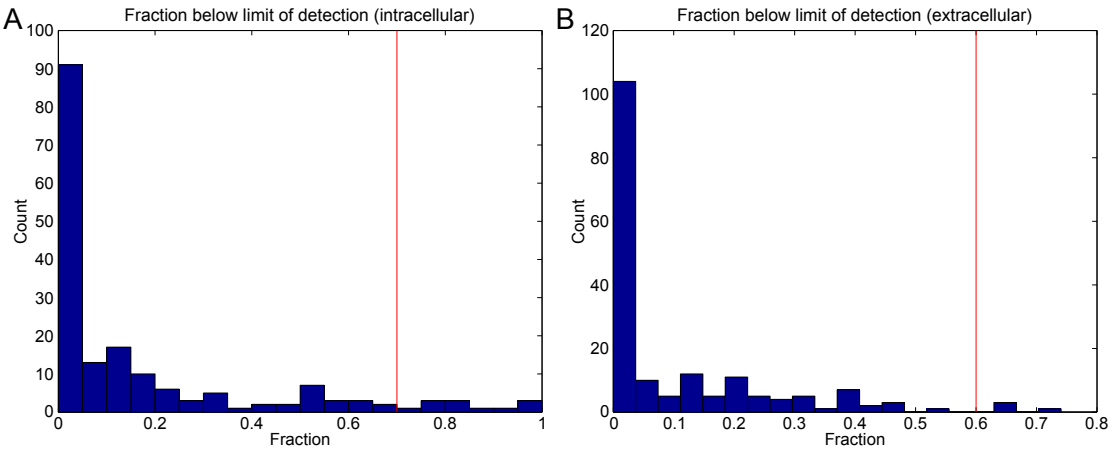

Figure 3.3: The fraction of measuring points below the limit of detection (LOD) per metabolite is shown in a histogram representation for the intracellular (A) and extracellular (B) measurements. The applied threshold of the quality control is indicated with the vertical red line. The threshold is 0.7 for intracellular and 0.6 for extracellular. The distinct thresholds were used to take the varying number of datapoints in the two environment into account.

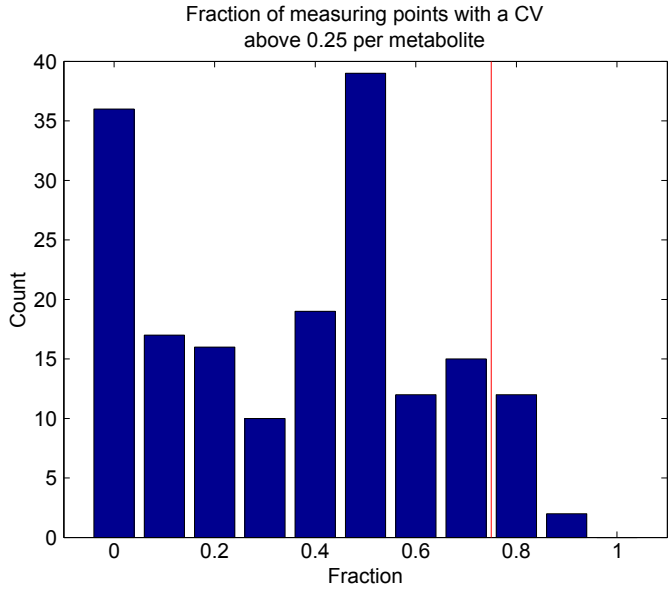

Figure 3.4: The coefficient of variation (CV) was calculated for every metabolite, every experiment and every measuring day. The fraction of CVs above 0.25 per metabolite is shown in this histogram. Metabolites with a fraction above 0.8 were excluded in the quality control.

Intracellular

Missing Values	Coefficient of Variation	Limit of Detection
Carnosine, DOPA, Dopamine, Nitro-Tyr, OH-Pro, SDMA, Serotonin, total DMA, ADMA, Histamine	Arg, Putrescine, Sarcosine, Taurine, PC aa C40:5, PC aa C42:1, PC aa C42:4, PC aa C42:5, PC ae C40:3, PC ae C42:2, SM (OH) C22:1, SM (OH) C22:2, SM C20:2, SM C26:0, SM C26:1	C14:1, C14:2, C16:1, C16:1-OH, C7-DC, Cit, Creatinine, PEA, PC aa C40:1, PC aa C42:0, PC aa C42:6

Table 3.1: List of the metabolites that were excluded due to the certain criteria for the intracellular environment.

Extracellular

Missing Values	Coefficient of Variation	Limit of Detection
Dopamine, Nitro-Tyr, OH-Pro, SDMA, Spermine, Carnosine	Arg, Putrescine, Sarcosine, Taurine, PC aa C40:5, PC aa C42:1, PC aa C42:4, PC aa C42:5, PC ae C40:3, PC ae C42:2, SM (OH) C22:1, SM (OH) C22:2, SM C20:2, SM C26:0, SM C26:1	PC aa C40:5, PC aa C42:1, SM (OH) C22:2, SM C18:0

Table 3.2: List of the metabolites that were excluded due to the certain criteria for the extracellular environment.

3.1.2 Preprocessing of Data for Analysis

This section deals with the results of methods which are applied before further analysis of the dataset could be carried out. This includes the correction of a batch effect in the measurements and also the imputing of missing values. The effects of these methods are also illustrated.

Correction of Batch Effect

The intracellular concentrations of all metabolites were measured on a different batch (batch 2) for day 14, 18 and 28 of experiment #3 (see Section 2.1.2). The concentrations at these measuring days showed a notable difference to the other two experiments at these measuring days. Depending on the biochemical class of the metabolite, the concentrations of batch 2 were higher or lower, whereas the three experiments showed a similar course at the first measuring days. An example is shown in Figure 3.5A. Therefore, we assumed that these differences are due to the different batches.

We performed a PCA of all intracellular measuring points (81 in total) to confirm this assumption. The first and second principle component are able to discriminate between the two batches for the most part of the measuring points (see Figure 3.6A). So, we decided to perform a correction of the batch effect as explained in Section 2.3.1. The PCA of the corrected concentrations is shown in Figure 3.6B. Based on these results, the corrections leads to an improvement, since the measuring points of batch 2 are not clearly separated anymore. There is still a small bias in the data, but this could also be explained due to the progression in

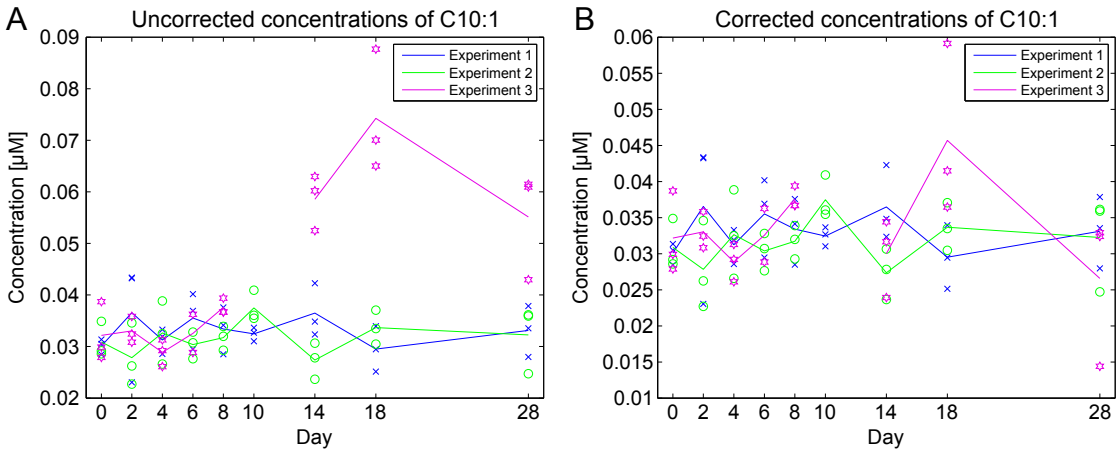

Figure 3.5: Intracellular concentrations of C10:1. The three replicates of each experiment are denoted with a unique symbol. The concentrations of experiment #3 at day 10 are missing. They will be imputed in a later stage. Panel A shows the concentrations without a correction of the batch effect and panel B the corrected concentrations. The affected measuring days are day 14 to day 28 of experiment #3.

the experiment (see the PCA in Section 3.1.3), since the corrected concentrations are in the late stage of the experiment. The effect of the correction were also manually evaluated. Figure 3.5B shows an example of the correction.

The batch correction can lead to negative values which are then set to a missing value. For the set of 151 metabolites, this is the case for 77 measuring points. These value will also be imputed in the next step.

Imputing of Missing Values

Using the set of metabolites of both environments, the dataset consist of 16,308 datapoints (151 metabolites $\times$ 108 measuring points), of which 894 are a missing value ($\sim$5.5%). 604 of these missing values are due an error in the conduction of the experiment (see Section 2.1.1), so that there are no samples at certain measuring points.

After Step 1 of the imputing method (see Section 2.3.2), there are 787 missing values left (107 values are imputed). So, only a small portion of the missing values was qualified for this imputing method. Step 2 (see Section 2.3.2) imputed 757 missing values, so that 30 missing values are left at the end of the imputing process. These values are part of two metabolites and could not be imputed, because the values were missing for all three experiments for consecutive measuring days. These metabolites might be left out in further analysis, when the applied method is not able to handle missing values. The corresponding numbers of the other two sets

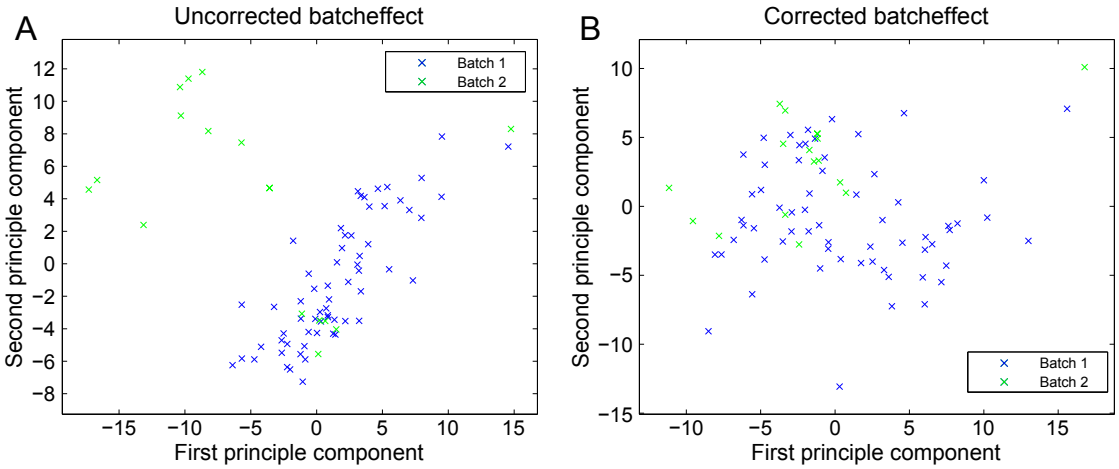

Figure 3.6: PCA of the intracellular measuring points. The effects of the batch effect (panel A) and the correction of it (panel B) are visualized.

	Intracellular	Extracellular	Both
Before imputing	899	1441	894
After step 1	792	1311	757
After step 2	30	183	30

Table 3.3: The number of missing values for the three sets of metabolites before and after the imputing steps.

of metabolites are listed in Table 3.3. An example of the imputing for C10:2 is shown in Figure 3.7.

3.1.3 Overview of the Data

At first, a brief overview over the dataset is given. Figure 3.8A shows the intracellular measurements which a normalized by the z-score (see Section 2.4.1). It is notable that a overwhelming fraction of the metabolites has a higher concentration in the early phase (day 0 to 4) than in the end stage (day 28). We performed a hierachical clustering (see Figure 3.8B) to further investigate the involved metabolites and to make this observation more comprehensible. The cluster which shows the clearest down regulation to the end of the experiment consist of 60 metabolites. A large part of the amino acids, some acylcarnitines (C12 to C18) and a various composition of glyercophospholipids are present in this cluster. There is also a fraction of metabolites that reaches its peak at day 10 to 18. This cluster consist

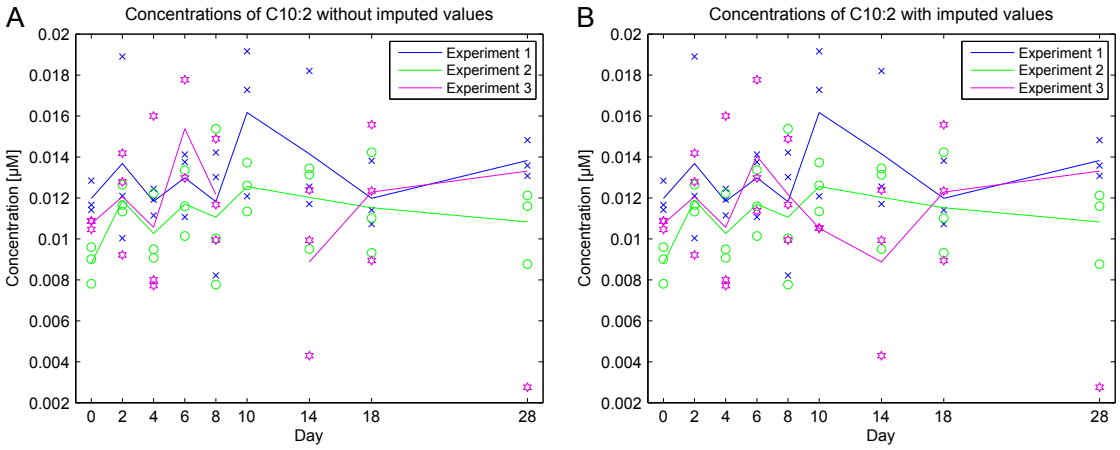

Figure 3.7: Intracellular concentrations of C10:2. The three replicates of each experiment are denoted with a unique symbol. (A) The concentrations of experiment #3 for one technical replicate at day 6 and for all three technical replicates at day 10 are missing. (B) The missing concentrations are imputed.

almost exclusively of phosphatidylcholines (PC aa and PC ae) and acylcarnitines (C3 to C10). See Section 3.2.1 for a more in depth cluster analysis.

In contrast to the intracellular measurements, the extracellular ones do not show such a clear trend towards a down regulation at the end of the experiment (see Figure 3.9). Additionally, Figure 3.10 displays the normalized value when day 0 is included. This represenation further illustrates the effect of the growth medium at day 0.

Figure B.1 and B.2 in the appendix show a similar overview, but in this case, the mean over the experiments was not computed, so that there are three measuring points per day and metabolite. For the intracellular measurements, the clustering looks very similar to the clustering with the mean over the experiments. Three major clusters, which have their peak in concentrations either in the early, middle or late phase of the experiment, are observeable. There is also a fourth cluster of which its metabolites show not such a clear peak at a certain timepoint. The results of the extracellular measurements are not that well separated and the result looks very inhomogenous.

Principle Component Analysis

We performed several principle component analysis (PCAs) to display different information. First, the PCA of all measuring points in both environments (108 measuring points in total) is shown in Figure 3.11. The first principle component is able to divide these two environments very good, as it can be seen due to the two

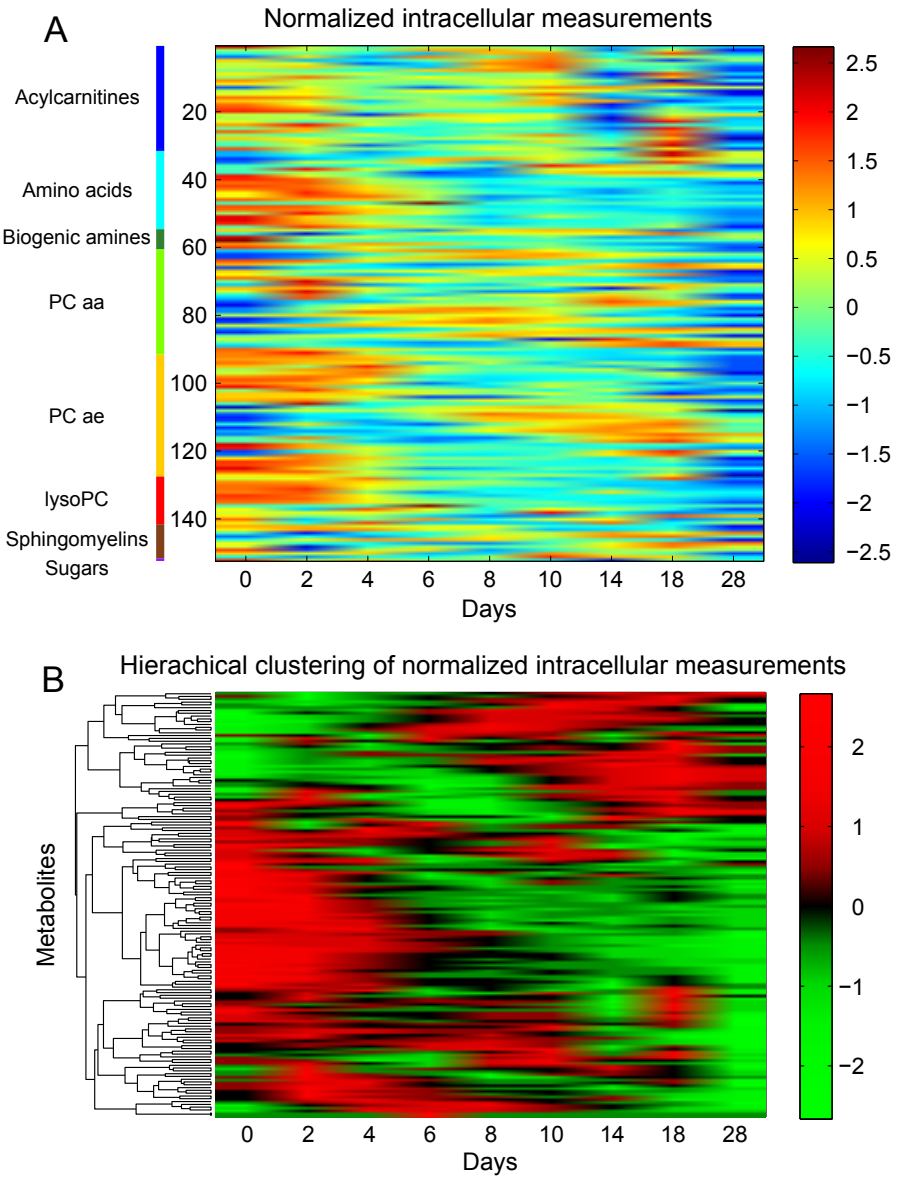

Figure 3.8: Represenation of the normalized intracellular measurements. The average over the technical replicates and experiments was computed, so that there is only one measuring point per day. (A) 152 metabolites are normalized with the z-score. The metabolites are ordered according to their biochemical classes. (B) A hierachical clustering was performed with Euclidean distance as metric and complete-linkage.

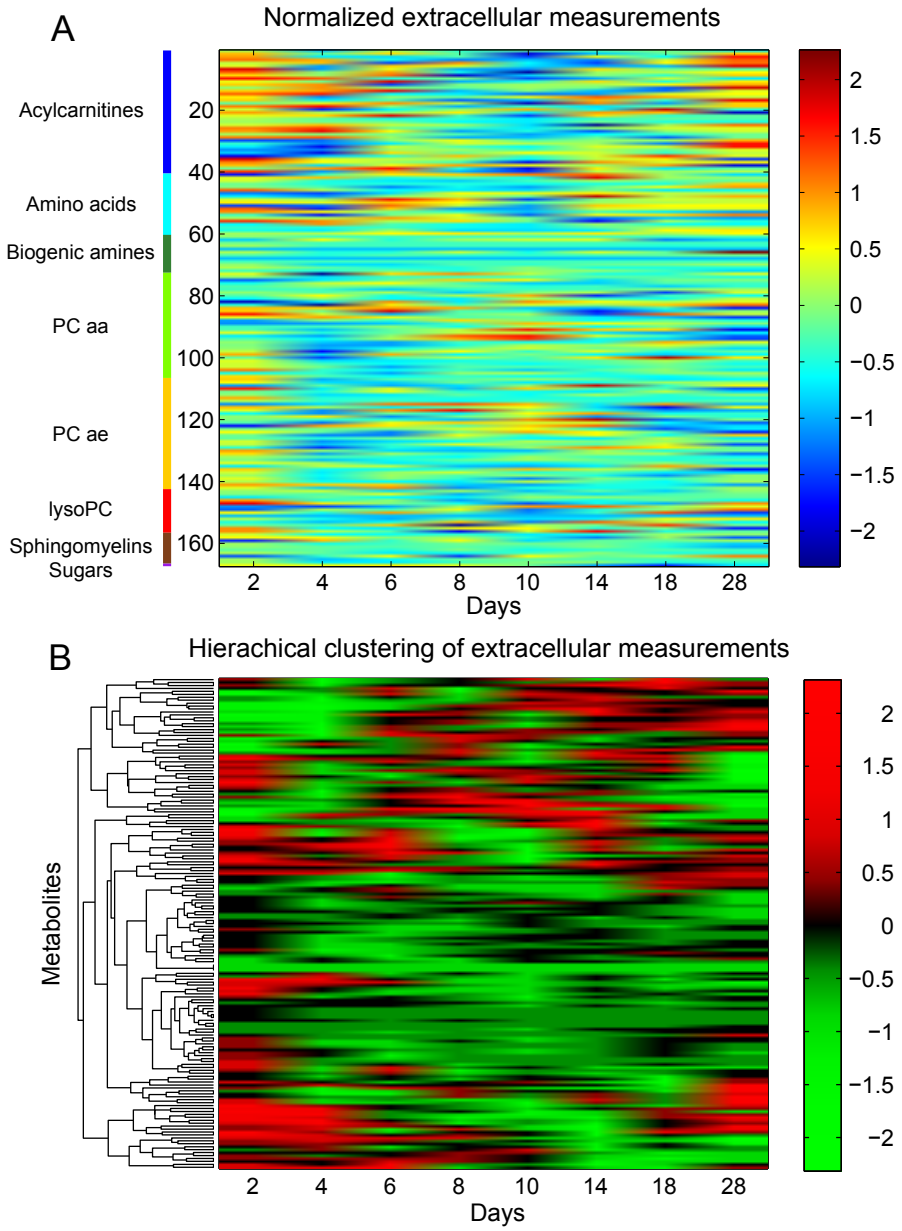

Figure 3.9: Represenation of the normalized extracellular measurements. The average over the technical replicates and experiments was computed, so that there is only one measuring point per day. (A) 167 metabolites are normalized with the z-score. The metabolites are ordered according to their biochemical classes. (B) A hierachical clustering was performed with Euclidean distance as metric and complete-linkage.

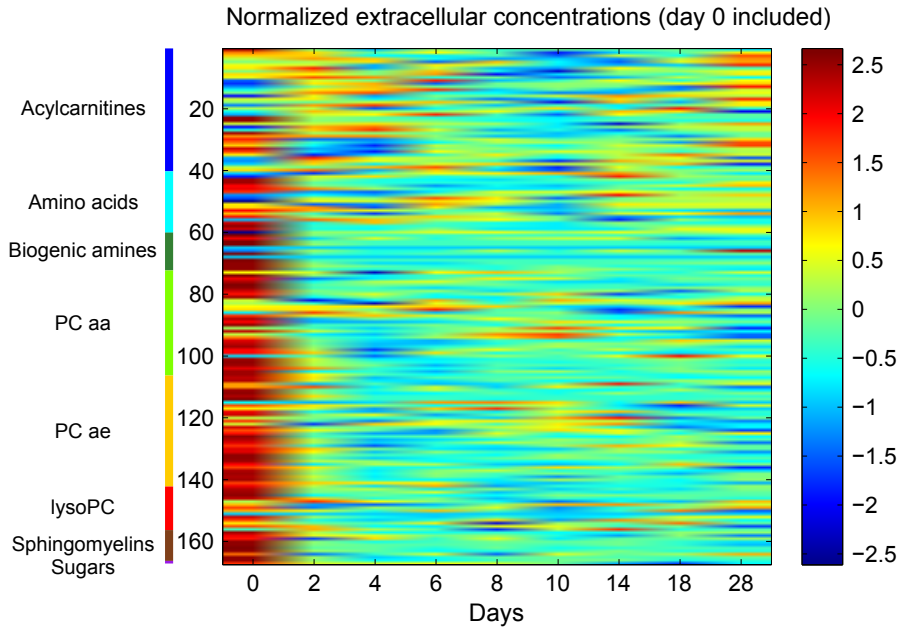

Figure 3.10: The normalized extracellular measurements of 167 metabolites from day 0 to day 28 are shown. The bias of the growth medium at day 0 is observeable.

clusters. The extracellular environment has less measuring points, because there are no technical replicates measured. The red circle highlights the extracellular measuring points of all three experiments at day 0. As in Section 2.1.1 explained, the composition of the growth medium at day 0 differs due to the addition of fetal calf serum. The second principle component was able to separate this measuring day which is affected by this medium effect.

Another PCA was performed to show the differences or changes between the days which is displayed in Figure 3.12. There are two representations. Figure 3.12A used the mean of the technical replicates and hence, there are three points per measuring day which represent the three experiments. The points of each measuring day have a tendency to cluster. To show this information more clearly, the mean of the three experiments was calculated in the PCA of Figure 3.12B. As a result, a trajectory of the measuring days is visible which represents the progression of the experiment over the timespan. The distance between the corresponding measuring days can already be used to get an idea about the different phases in the differentiation process, i.e. day 0 and 2 as an early, day 6, 8 and 10 as middle and day 14, 18 and 28 as the late phase. Day 4 is in between the early and middle phase. These results match with the information about the differentiation phases, which were provided by Helmut Laumen with an analysis of differentiation marker (see Section 2.1.3). We also performed a PCA for the extracellular measurements

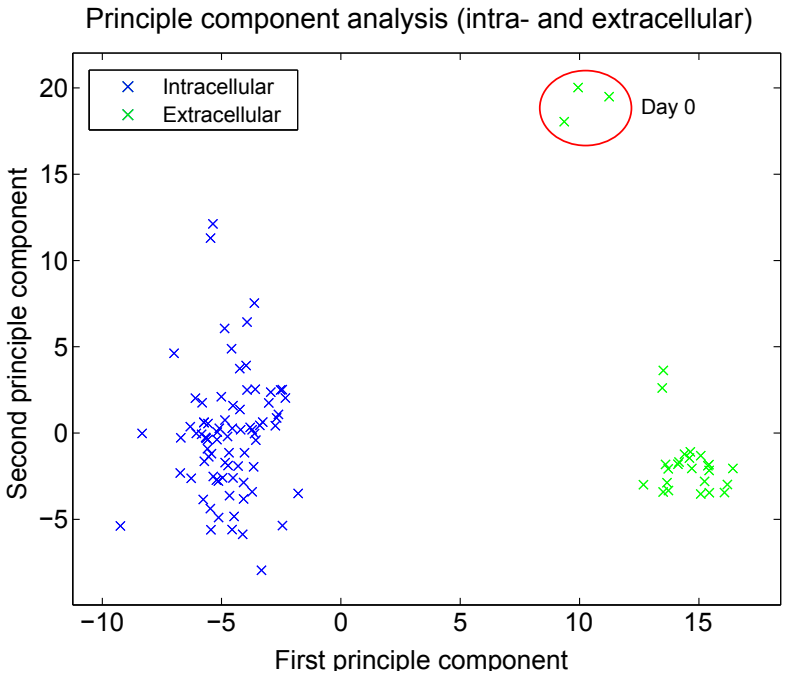

Figure 3.11: A PCA of all intra- and extracellular measuring points (108 in total). They are coloured according to their environment, i.e. intracellular in blue and extracellular in green. Day 0 of all three extracellular experiments are highlighted with a red circle. These measuring points are separated due to the different composition of the growth medium (see Section 2.1.1).

(see Figure B.3). The progression in the differention during the experiment is not so clearly visible.

Since a standard PCA is not able to deal with missing values, we performed this PCA with imputed values at first. There is also a Baysian PCA which is able to be applied on data with missing values [35]. We used the GP-LVM toolbox in MATLAB (The Mathworks Inc.) to perform a non-linear PCA without imputed values [20]. The results were very similar to the results of the standard PCA with imputed values (see Figure B.4).

3.1.4 Interpretation of the Extracellular Measurements

Before we continue with the clustering analysis of the intra- and extracelluler measurements, we discuss the interpretation and meaning of the extracellular concentrations. When the intracellular concentration of a metabolite decreases between two measuring days, there are two possible explanations. First, the metabolite has been catalyzed to another metabolite. Second, there was an exchange to the

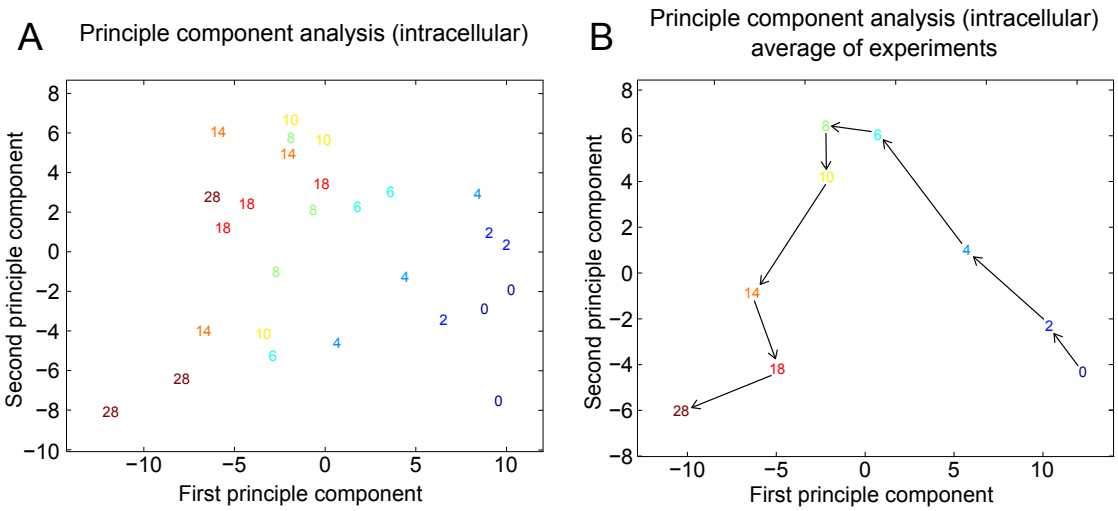

Figure 3.12: A principle component analysis of the intracellular measuring days. The concrete day of the point is denoted as a number in the figure. (A) The mean of the technical replicates was calculated, so that the three points per day represent the experiments. It is notable, that the three points of a measuring show a tendency to cluster. (B) Additionally, the mean of the experiments was computed. A trajectory dependend on the time point of differentiation can be observed.

extracellular environment. On the other hand, a decrease of the extracellular concentration can be explained with a third option. This is due to the fact that the extracellular concentrations are the measurement of the supernatant. Hence, the metabolite concentrations of the medium have an effect on the extracellular concentrations. This is the first problem in the interpretation of the extracellular concentrations. The second one is the medium change which occurs every two days. This is a reset of the extracellular metabolite concentrations. So for example, the concentration between measuring day 14 and 18 is the change between day 16 and 18 starting at the medium concentration. We tried several methods to recalculate the extra- and also the intracellular measurements, but we were not able to find a solution which made the meaning of the concentrations of the two environments entirely comparable.

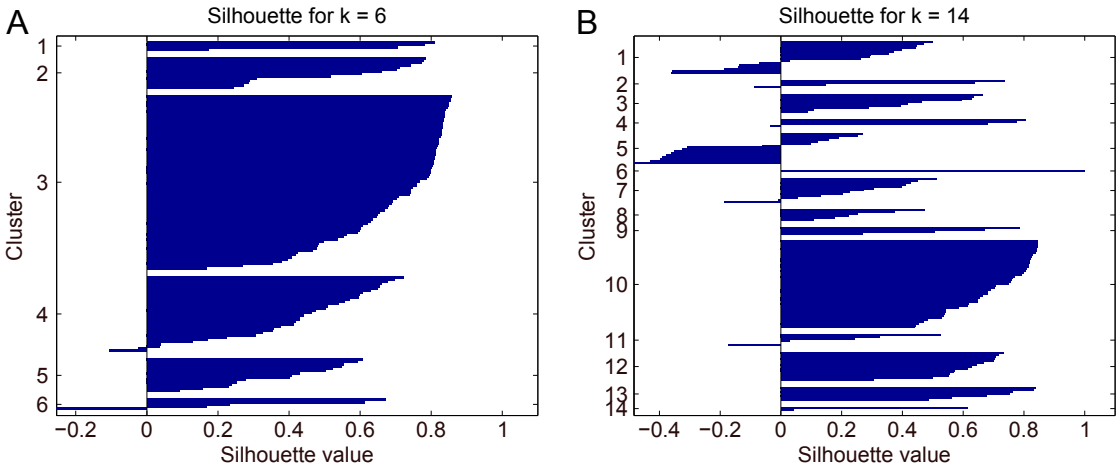

Figure 3.13: Silhouette of the clusters for $k = 6$ (panel A) and $k = 14$ (panel B). The silhouette values are much higher for the result of $k = 6$ and this representation illustrates the decrease of quality of the clustering result between these two k.

3.2 Analysis of Metabolite Time Courses

3.2.1 Clustering of Metabolite Time Courses

We performed a k-means clustering of the log(fold changes) (logFCs) to get a first overview of the data. The clustering was performed with the Euclidean distance and the Pearson product-moment correlation coefficient as distance measures. The assignments of the metabolites to their clusters for all the clustering results are listed in the tables in Appendix C.

Clustering Based on Euclidean Distance

A k-means clustering was performed for the logFCs of the metabolites with Euclidean distance as the metric. The fold changes normalize the measurements, so that the values of the metabolites are comparable, and a up or down regulation of the metabolite is visible. A range of 2 to 15 was used as k for the k-means clustering. The cluster results itself and the silhouette plots (see Section 2.5.3) were used to evaluate the cluster analysis. Figure 3.13 displays the silhouette values for $k = 6$ and $k = 14$. These results are based on the intracellular fold changes. They illustrate the differences between a good clustering result ($k = 6$) to a mediocre one. In general, the silhouette values are higher and less values are negative.

In fact, $k = 6$ obtained the best clustering result for the intracellular fold changes. The six clusters cover a range of elementary biological responses (see Figure 3.14). Cluster #1 and #2 are both up regulated metabolites, but cluster

#1 reaches its peak at a much earlier point. One can distinguish these two clusters as an early- and late-responder. Cluster #3 is the largest cluster and contains the metabolites that do not change much over the experiment. The fluctuations of these metabolites are most likely due to noise in the measurement. Cluster #4, #5 and #6 are all down regulated differ mainly due to the degree of down regulation. The assignment of metabolites to the clusters is given in the Table C.1 (Appendix A).

The k-means clustering of the extracellular fold changes does not show such a clear result. The silhoutte values of the clustering with $k = 13$ are displayed in Figure 3.15. This was the best result for the extracellular environment. Table C.2 (Appendix A) lists the clusters and their assigned metabolites. In comparison to the intracellular result, there are more metabolites with a negative silhouette value. Even though, the choosen k is much higher, the clusters do not look as homogenous (see Figure 3.16). The time courses of the metabolite fold changes show much more fluctuation in the extracellular environment than in the intracellular one. Additionally, there are metabolite time courses that look artificial with a flat line, i.e. cluster #13. These are metabolites with a lot of values that were measured with a concentration of 0. They were replaced with the minimum value of this metabolite (see Section 2.4.1).

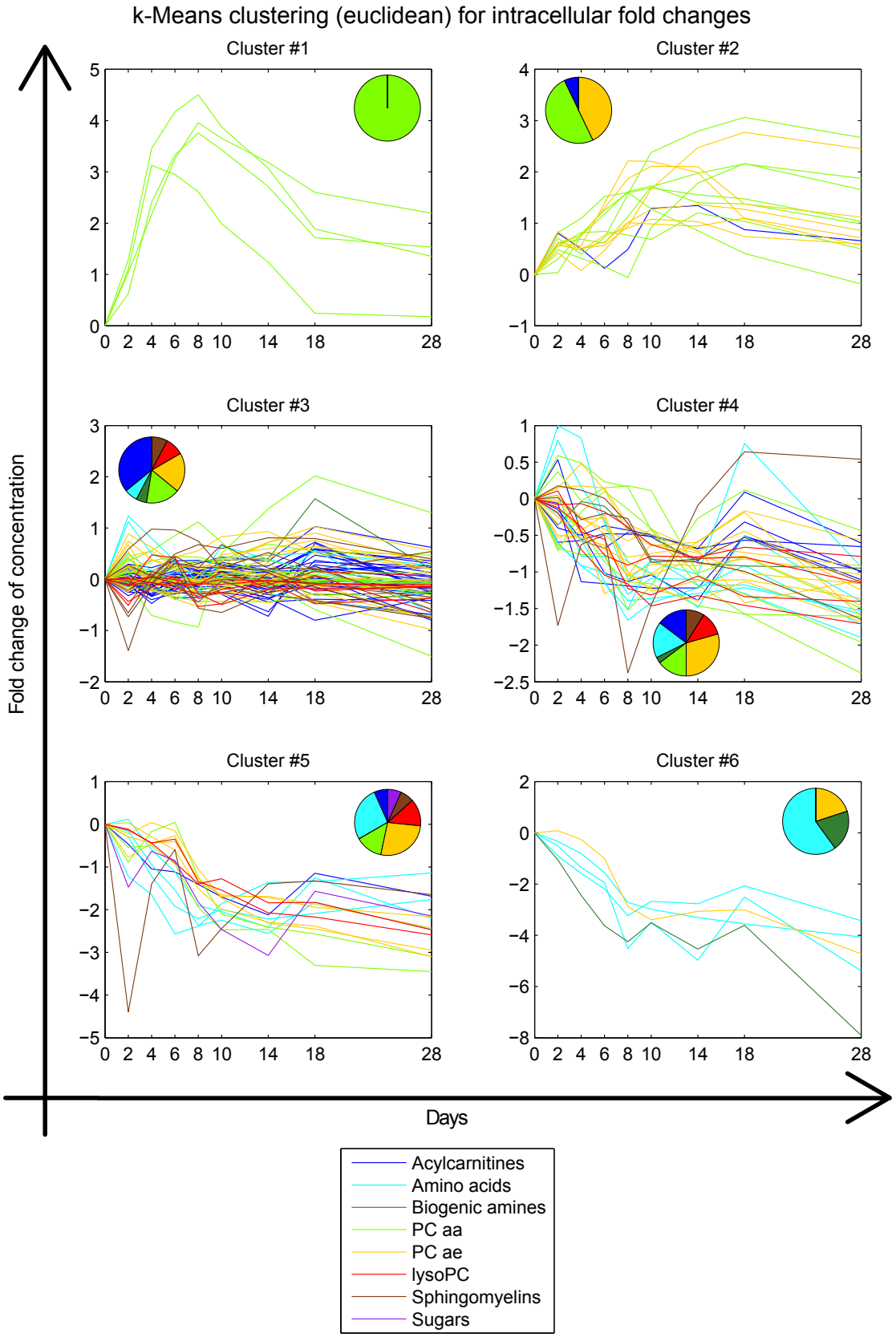

Figure 3.14: Overview of the k-means clustering with $k = 6$ for the intracellular logFCs based on Euclidean distance. The clusters are ordered in a way, so that there is a descending from a up to a down regulation. The piechart illustrates the distribution of biochemical classes within the cluster.

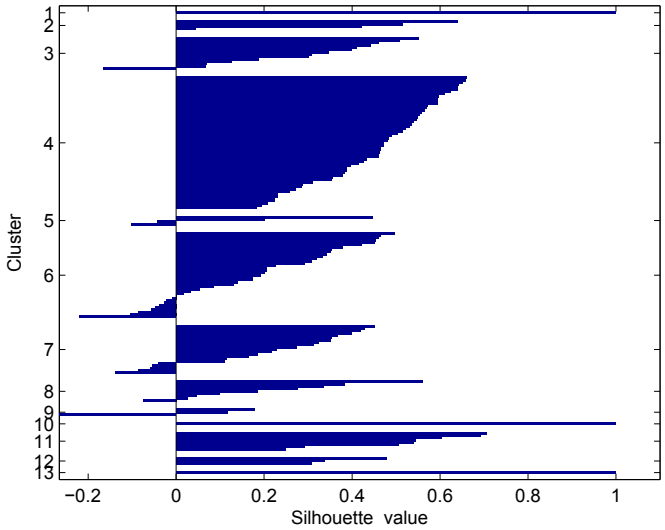

Figure 3.15: Silhouette values of the k-means clustering of extracellular logFCs for $k = 13$ and Euclidean distance.

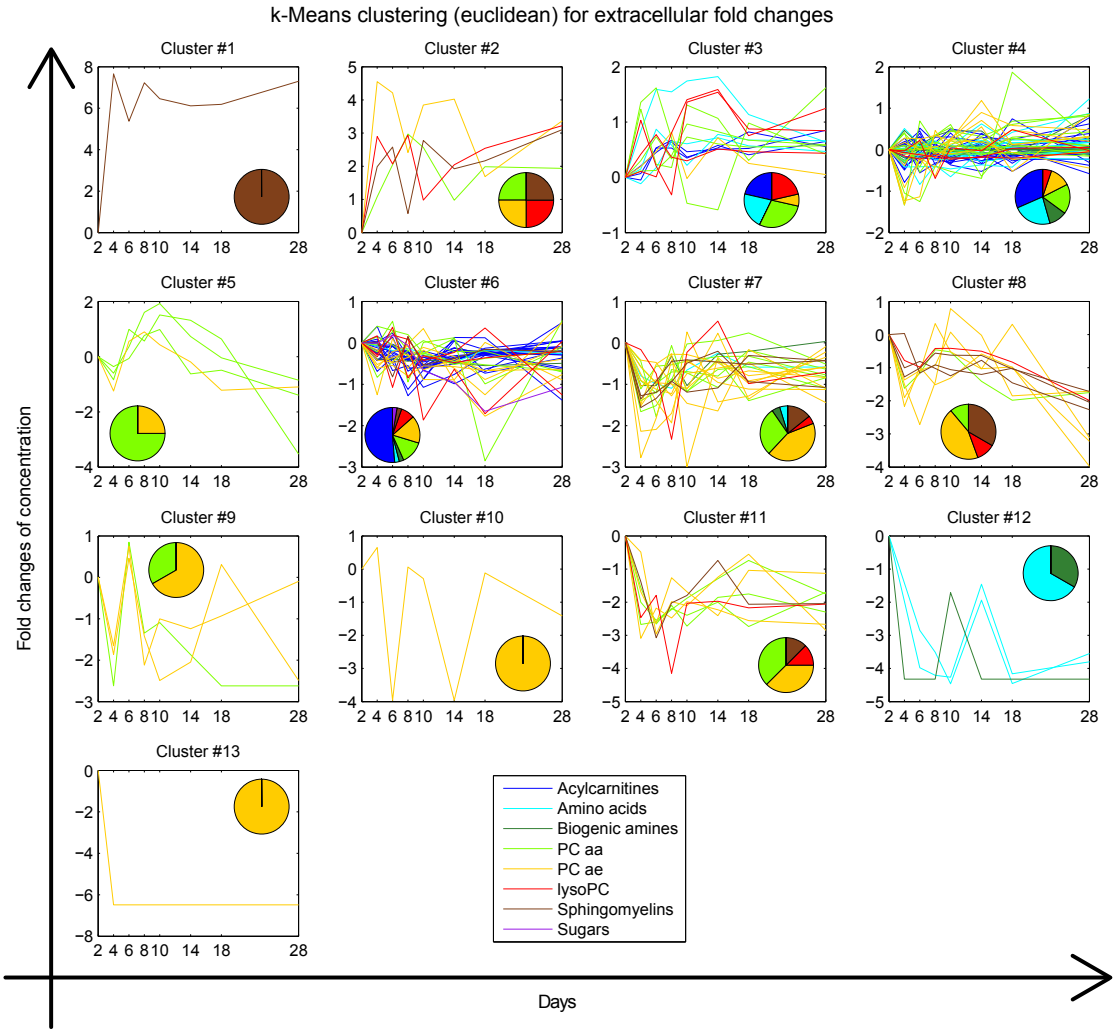

Figure 3.16: Overview of the k-means clustering with $k = 13$ for the extracellular logFCs based on Eucliden distance. The clusters are ordered in a way, so that there is a descending from a up to a down regulation. The piechart illustrates the distribution of biochemical classes within the cluster.

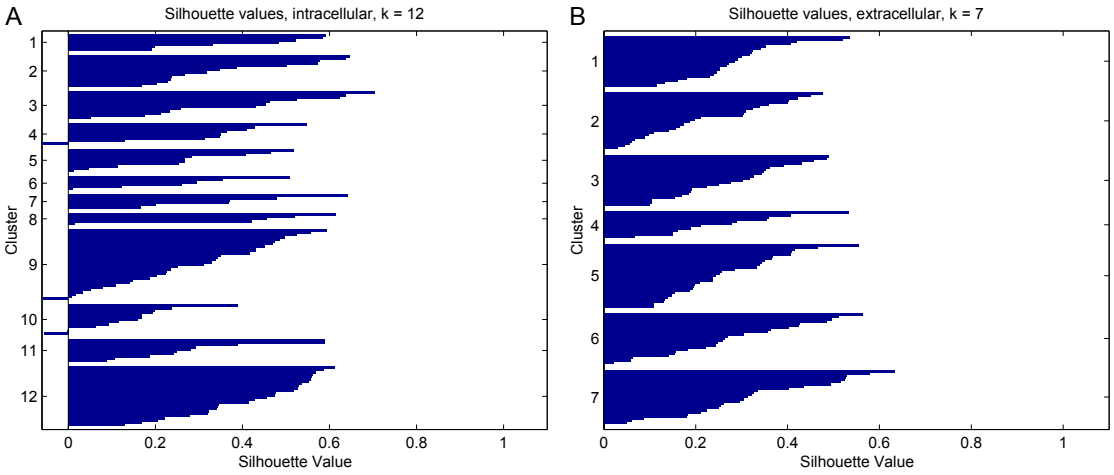

Figure 3.17: Silhouette values of the k-means clustering of intracellular (A) and extracellular (B) logFCs. Pearson product-moment correlation coefficient was used as distance measure.

Clustering Based on Correlation Coefficient

The clustering analysis of the logFCs was also performed with the Pearson product-moment correlation coefficient (PCC) (see Section 2.5.1) as distance measure. The range of k was between 2 and 15. The consideration of the silhouette values indicate that a clustering with $k = 12$ provides the best result for the intracellular measurements and $k = 7$ for the extracellular one. The silhouette values of these two are shown in Figure 3.17. There are almost no negative silhouette values, which indicates that the clustering of the metabolites is appropiate.

The clustering results for the intracellular logFCs are displayed in Figure 3.18. There are some biochemical classes of which the majority of its metabolites are only present in a few clusters. For example, the amino acids are only part of cluster #9 and #12 with the exception of two amino acids. These two clusters show a trend towards a down regulation. A similar observation can be done for the acylcarnitines, which are predominant in the clusters #5, #6 and #11. Figure 3.19 shows the clustering of the extracellular logFCs with $k = 7$. Even though the silhouette values of this clustering suggested a good clustering, its is a demanding task to evaluate this result. The time courses of the metabolites within a cluster do not look homogenous. The same applies for the composition of the clusters regarding its composition of biochemical classes. The two tables C.3 and C.4 list the assignment of metabolites to the clusters.

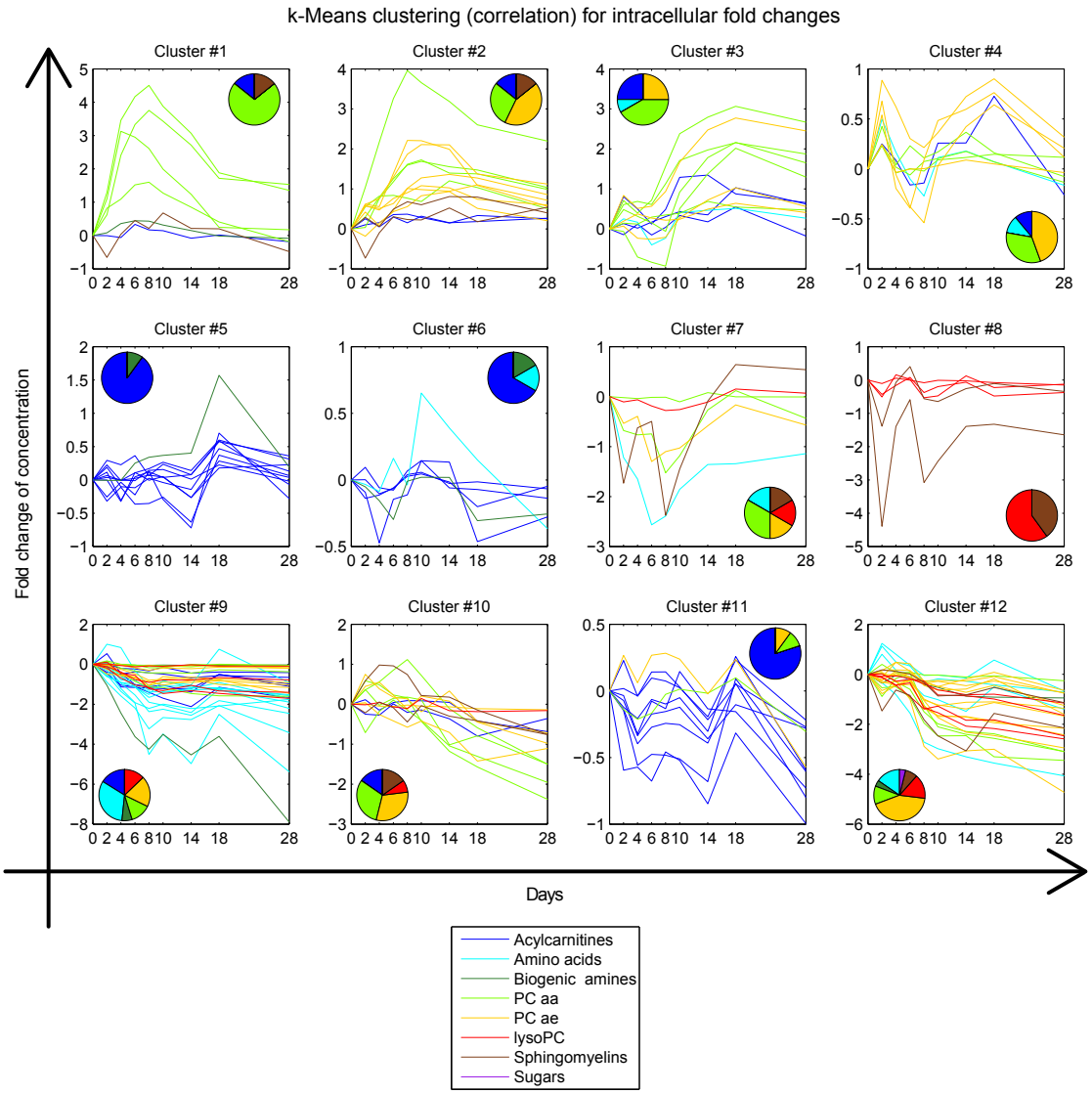

Figure 3.18: Overview of the k-means clustering with $k = 12$ for the intracellular logFCs based on correlation coefficient as the metric. The clusters are ordered in a way, so that there is a descending from a up to a down regulation. The piechart illustrates the distribution of biochemical classes within the cluster.

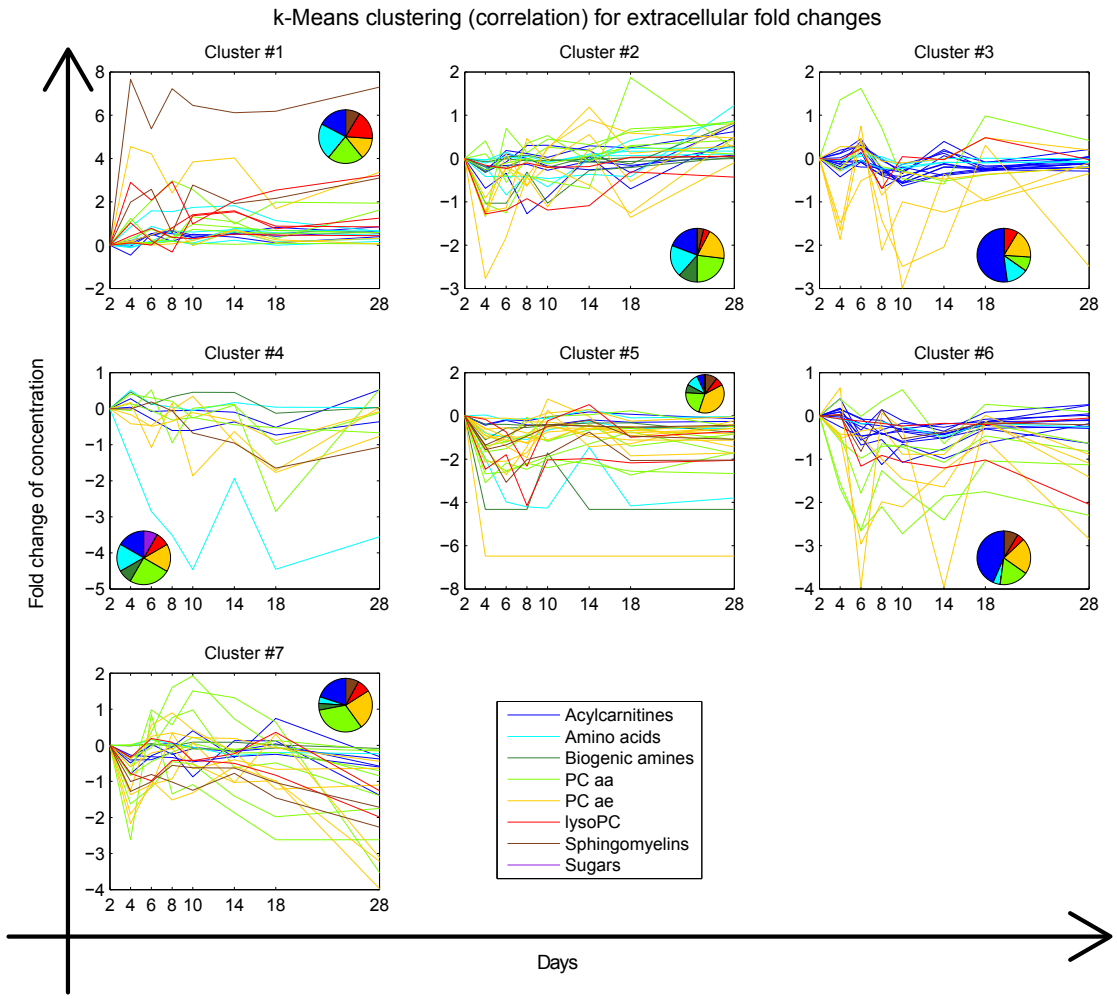

Figure 3.19: Overview of the k-means clustering with $k = 7$ for the extracellular logFCs based on correlation coefficient as the metric. The clusters are ordered in a way, so that there is a descending from a up to a down regulation. The piechart illustrates the distribution of biochemical classes within the cluster.

Conclusion of the Clustering Analysis

We performed the k-means clustering with two different distance measures, i.e. Euclidean distance and PCC. Regarding the intracellular measurements, the result with Euclidean distance enabled to distinguish between the metabolites considering their up or down regulation during the adipocyte experiment. There is one cluster (#3) of which its metabolites vary between a logFC of 1 to -1. The results of the PCC clustering was able to separate these metabolites and assign it to clusters of other metabolites. This is due to the differences of the distance measures. PCC takes the trend of a time course into account, whereas the Euclidean evaluates whether the time courses are in the same range of values. But it is most likely, that the fluctuations of these metabolites, that vary between 1 and -1, are due to noise in the measurements. Therefore, it might be reasonable to exclude metabolites that do not have an absolute logFC above 1 for future analysis.

The clustering of the extracellular measurements seems to be a demanding task. Both distance measures were not able to provide a clustering which enables a reasonable biological interpretation of the clusters. There could be two reasons for that. First, several metabolites have a measured concentrations of 0 at various measuring points. Our implementation of the logFC computation replaces these 0-values. Depending on the amount of 0-values, the time course of this metabolite can be artificial to some extent as already described earlier. Second, the meaning of the extracellular concentrations is not the same as for the intracellular ones. This was already discussed in Section 3.1.4.

However, first conclusions can be drawn from the results of the clustering analysis. There is a fraction of diacyl phosphatidylcholines (cluster #1 and #2 of the intracellular clustering result with Euclidean distance, see Figure 3.14) and also acyl-alkyl phosphatidylcholines that are strongly up regulated during the adipocyte experiment. On the other hand, the major part of the amino acids is down regulated as it can be seen in the clusters #9 and #12 of the intracellular result with PCC as a distance measure (see Figure 3.18). Before we discuss about the biological meaning of these metabolite concentration changes, we perform an enrichment analysis to further investigate the behaviour of the biochemical classes or pathways in the next section to assess whether there are significant changes of concentrations during the experiment.

Biochemical class	Number of metabolites
Acylcarnitines	35
Amino acids	19
Biogenic amines	5
Phosphatidylcholines (PC aa)	31
Phosphatidylcholines (PC ae)	36
lyso-phosphatidylcholines	14
Sphingomyelines	10

Table 3.4: List of the biochemical classes and the corresponding number of metabolites after quality control. These biochemical classes were used as metabolite sets for the enrichment analysis. Table A.1 lists the metabolites of the biochemical classes.

3.2.2 Enrichment Analysis of Metabolite Changes

The clustering analysis provided groups of metabolites that behaved similar in respect of a up or down regulation during the experiment. We follow another approach with the enrichment analysis. The results of this analysis are described in this section. The enrichment analysis is a method which investigates the behaviour of certain metabolite sets. These sets are defined in a way such that the metabolites of a set share certain characteristica, e.g. they belong to the same biochemical class. The enrichment analysis tries to identify metabolite sets that show a bevhaviour which is different in respect of the other metabolites. For this, several different test variants are applied.

Hypergeometric Test with Biochemical Classes as Metabolite Sets

The first variant of enrichment analysis was performed for the intracellular measurmenets as a hypergeometric test (see Section 2.6.2) with the biochemical classes (see Section 2.6.1) of the metabolites as sets. To enable a comparison between the intra- and extracellular environments, the group of 151 metabolites was used. This group was the intersection of intra- and extracellular that passed the quality control. Among these metabolites, H1 was the only sugar. No matter how H1 behaves during the experiment, this set with only one metabolite would have not shown a signal in an enrichment analysis. Therefore, it was also excluded and the enrichment analysis was performed with 150 metabolites. Table 3.4 gives an overview of the biochemical classes.

A t-test (see Section 2.4.3) for every metabolite was applied for a pair of two measuring days. The first measuring day was always day 0 and the second one was one of the remaining eight, e.g. day 0 and 2 or day 0 and 4. A number of 1,200 t-tests (8 measuring day pairs $\times$ 150 metabolites) was computed. To account

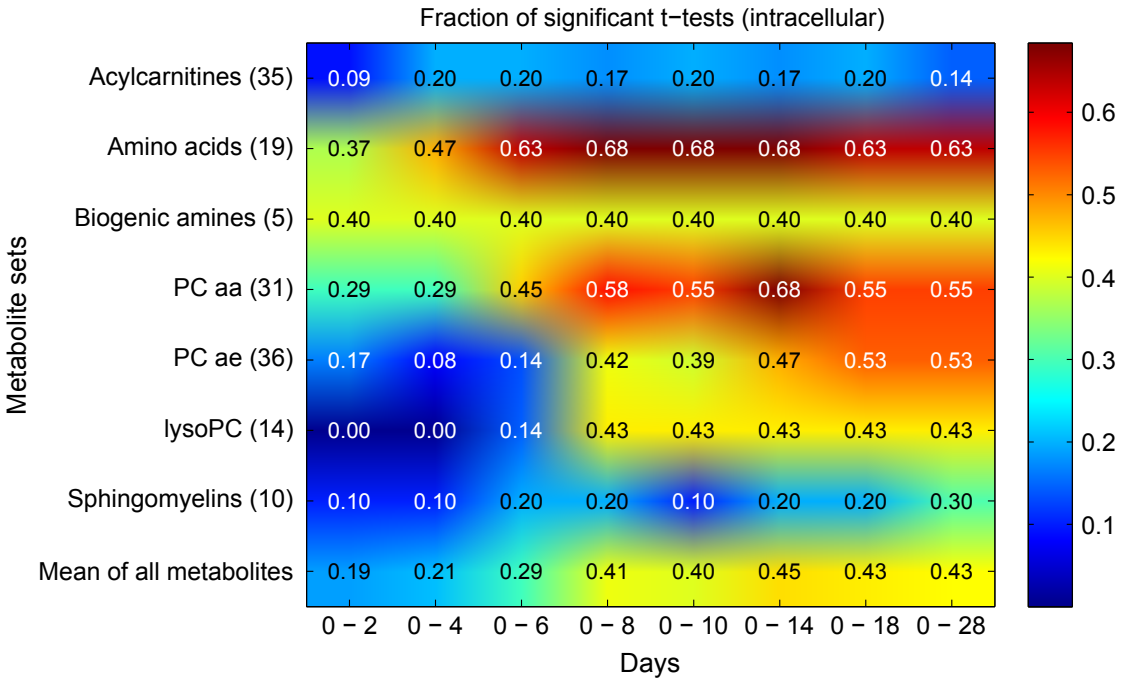

Figure 3.20: A t-test for every metabolite was performed between a pair of measuring days which is declared on the x-axis. The fraction of significant metabolites for each metabolite set is shown. The metabolite sets are based on the biochemical classes. Additionally, the fraction of all metabolites is shown (last row).

for multiple testing, the significance level α was adjusted with FDR (see Section 2.4.4) to 0.0154. Figure 3.20 displays the fraction of metabolites per set which showed a significant difference at the corresponding measuring day in comparison to day 0. The group of amino acids and phosphatidylcholines ($PC\ aa$) achieve the highest fractions. It is also noteworthy, that the fraction of all metabolites increases towards the end of the experiment.

The enrichment analysis itself was performed as a hypergeometric test (see Section 2.6.2). The p-values are shown in Figure 3.21A. We interpret these p-values as an indicator of the enrichment of a certain metabolite set. This first hint can then be followed with a further investigation of the metabolite set specific measurements. The results of the hypergeometric test show that the sets of amino acids and phosphatidylcholines ($PC\ aa$) show the strongest signal over all measuring days. This means, in comparison to all the other metabolites, an overwhelming part of the metabolites of these sets are up or downregulated. A weaker signal is observeable for the other group of phosphatidylcholines ($PC\ ae$) at day 18 and 28.

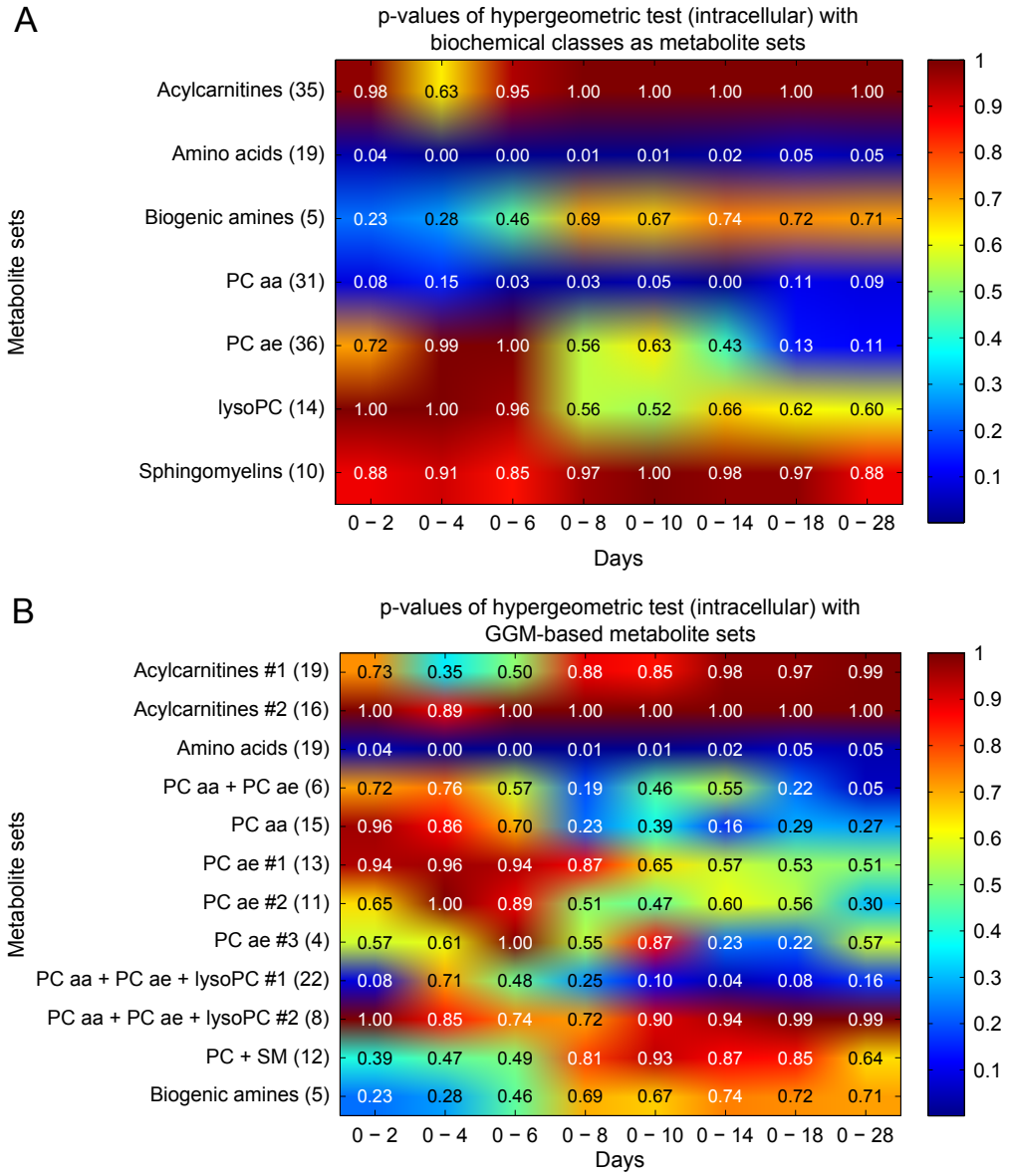

Figure 3.21: The p-values of a hypergeometric test. A t-test for every metabolite was performed between a pair of measuring days which is declared on the x-axis. Since FDR adjusted the signifcance level α to 0, we applied the Bonferroni correction [3] to account for multiple testing. (A) The biochemical classes were used as the metabolite sets. The adjusted signifance level α is 8.93×10^{-4}. (B) The hypergeometric test was performed with the GGM-based metabolite sets (see Section 3.2.2). The adjusted signifance level α is 5.21×10^{-4}.

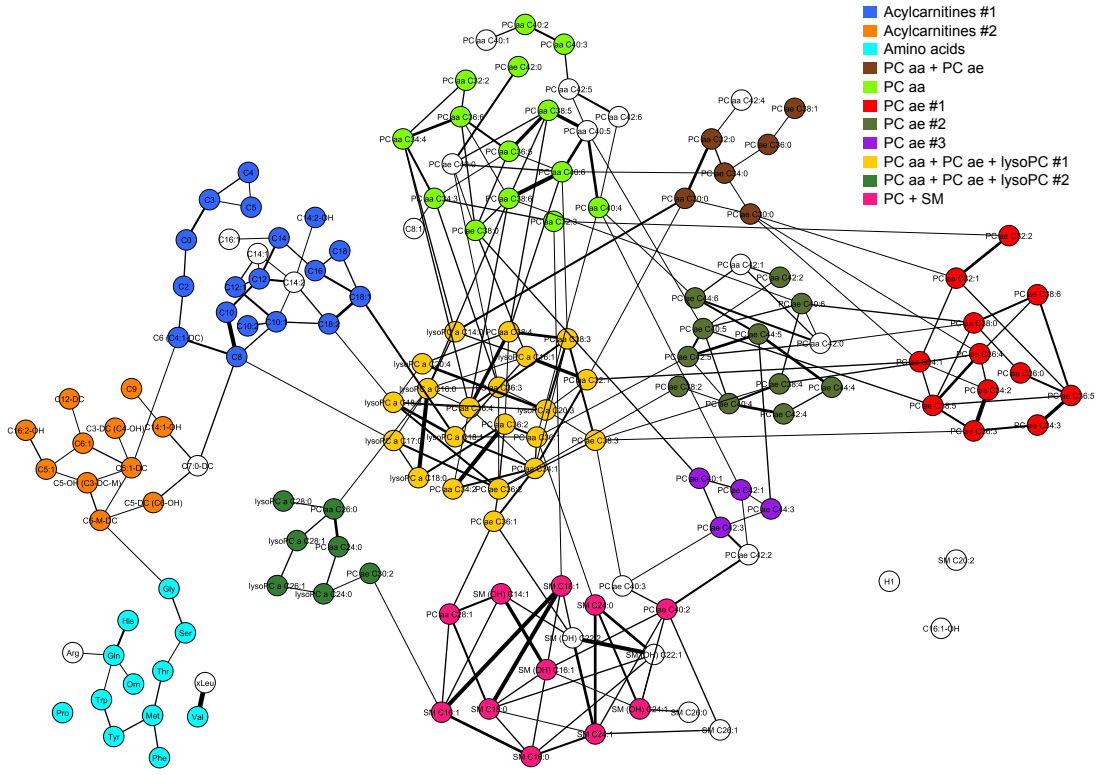

Figure 3.22: A Gaussian graphical model (GGM) of the metabolites based on the KORA study. An edge between metabolites denotes a partial correlation coefficient of 0.18 or greater. The metabolites are ordered and colored according to their metabolite set. The set definition is based on a community detection algorithm and a manual curation step. Metabolites that are not in the dataset of this experiment are colored white.

Hypergeometric Test with GGM-based Metabolite Sets

During adipogensis, metabolites of certain pathways might be up or downregulated. Therefore, we also wanted to include pathway features for the metabolite set definition. So far, this was not the case, because the biochemical classes only consider the biochemical properties of the metabolites. At the moment, it is a demanding task to obtain complete pathway annotations, because a mapping between the measured metabolites and resources with annotations are not accurate or annotations are missing. Hence, we decided to use Gaussian graphical modeling (GGM) as another approach, which has shown to be successful in reconstructing biological pathways from metabolomics data [18]. This was also done with a metabolite panel of Biocrates measurements.

The general method of the GGM-based metabolite set definition is described in Section 2.6.1. Here, we describe the intermediate results in the definition process.

The modularity clustering algorithm detected 15 communities in the GGM, which is based on the blood serum samples of the KORA study (KORA GGM). Even though, our dataset is based on cells that undergo adipogenesis and the KORA study has samples of blood serum, we decided that we can use the KORA GGM, because both datasets are based on human samples and the fundamental metabolic pathways should be the same. These 15 communities were then mapped to the 150 metabolites of the adipocyte experiment. Since not every metabolite of the KORA study is present in the group of the 150 metabolites due to a different Biocrates metabolite panel or excluded metabolites in the quality control, the mapping reduced the number of communities to 13. Vice versa, 23 of the 150 metabolites were not measured in the KORA study, e.g. some acylcarnitines, all biogenic amines, several amino acids and glycerophospholipids. Hence, these 23 metabolites are not present in the GGM and were not assigned to a community. Thus, a manual curation of these metabolites was performed. The last step was the merging of two communities, that consist of only one amino acid, with the community that contained all the other amino acids. At the end, there were 12 communities or metabolite sets, respectively. Table 3.5 gives an overview of the metabolite sets and its metabolites. Figure 3.22 shows the KORA GGM with a coloring of the nodes according to the metabolite sets. As it can be seen on the white nodes, it was not possible to map every of the 150 metabolites to the metabolites of the KORA study.

Similar to Section 3.2.2, the enrichment analysis was performed as a hypergeometric test with the 12 communities as the metabolite sets. The fractions with significant different metabolites per set are shown in Figure D.1 and the result of the enrichment analysis itself is displayed in Figure 3.21B. Since the metabolite set of the amino acids is excactly the same, the results do not differ from the result with the biochemical classes as sets. Besides the amino acids, two sets, that contain phosphatidylcholines (i.e. *PC aa + PC ae + lysoPC #1* and *PC aa + PC ae*) have the lowest p-values. This result is similar to the result with the biochemical classes, but a finer distinction in the composition is possible. Additionally, this set definition puts the results more in a pathway context.

Amino Acid Metabolism

The hypergeometric test with biochemical classes and the GGM-based definition of metabolite sets showed an enrichment of the metabolite set *amino acids*. Therefore, we take a look at the time courses of this metabolite set, which consists of 19 amino acids. Figure 3.23 displays the logFCs of the amino acids in comparison to the remaining metabolites. The set of amino acids can be separated into two main groups. The larger group is composed of amino acids, that have a logFC between -0.5 to -1.5 at measuring day 2 and 4. In the progression of the differentiation,

Metabolite set	Metabolites
Acylcarnitines #1	C0, C2, C3, C4, C5, C6, C8, C10, C10:1, C10:2, C12, C12:1, C14, C14:2-OH, C16, C16:2, C18, C18:1, C18:2
Acylcarnitines #2	C3-OH, C3-DC, C3:1, C4:1, C5-OH, C5-DC, C5-M-DC, C5:1, C5:1-DC, C6:1, C9, C12-DC, C14:1-OH, C16:2-OH, C16-OH, C18:1-OH
Amino acids	Gln, Gly, His, Met, Orn, Phe, Ser, Thr, Trp, Tyr, Ala, Asn, Asp, Glu, Ile, Leu, Lys, Pro, Val
PC aa + PC ae	PC aa C30:0, PC aa C32:0, PC ae C30:0, PC ae C34:0, PC ae C36:0, PC ae C38:1
PC aa	PC aa C30:2, PC aa C32:2, PC aa C32:3, PC aa C34:3, PC aa C34:4, PC aa C36:5, PC aa C36:6, PC aa C38:5, PC aa C38:6, PC aa C40:2,PC aa C40:3, PC aa C40:4, PC aa C40:6, PC ae C38:0, PC ae C42:0
PC ae #1	PC aa C36:0, PC aa C38:0, PC ae C30:1, PC ae C32:1, PC ae C32:2, PC ae C34:1, PC ae C34:2, PC ae C34:3, PC ae C36:3, PC ae C36:4, PC ae C36:5, PC ae C38:5, PC ae C38:6
PC ae #2	PC aa C42:2, PC ae C38:2, PC ae C38:4, PC ae C40:4, PC ae C40:5, PC ae C40:6, PC ae C42:4, PC ae C42:5, PC ae C44:4, PC ae C44:5, PC ae C44:6
PC ae #3	PC ae C40:1, PC ae C42:1, PC ae C42:3, PC ae C44:3
PC aa + PC ae + lysoPC #1	PC aa C32:1, PC aa C34:1, PC aa C34:2, PC aa C36:1, PC aa C36:2, PC aa C36:3, PC aa C36:4, PC aa C38:1, PC aa C38:3, PC aa C38:4, PC ae C36:1, PC ae C36:2, PC ae C38:3, lysoPC a C14:0, lysoPC a C16:0, lysoPC a C16:1, lysoPC a C17:0, lysoPC a C18:0, lysoPC a C18:1, lysoPC a C18:2, lysoPC a C20:3, lysoPC a C20:4
PC aa + PC ae + lysoPC #2	PC aa C24:0, PC aa C26:0, PC ae C30:2, lysoPC a C24:0, lysoPC a C26:0, lysoPC a C26:1, lysoPC a C28:0, lysoPC a C28:1
PC + SM	PC aa C28:1, PC ae C40:2, SM (OH) C14:1, SM (OH) C16:1, SM (OH) C24:1, SM C16:0, SM C16:1, SM C18:0, SM C18:1, SM C22:3, SM C24:0, SM C24:1
Biogenic Amines	Ac-Orn, Kynurenine, Met-SO, Spermidine, alpha-AAA

Table 3.5: Overview of the 12 GGM-based metabolite sets with 150 metabolites.

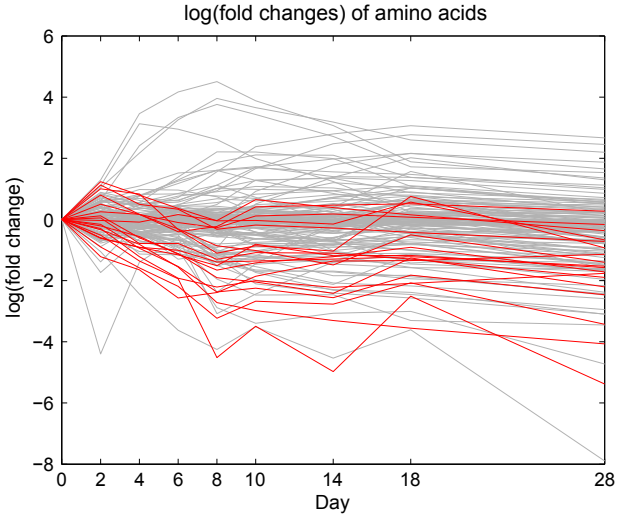

Figure 3.23: The intracellular logFCs of the amino acids (19 metabolites) are shown in red. The major part of this metabolite set is down regulated. The logFCs of the remaining metabolites is displayed in gray.

their logFCs decreases even further and the logFC varies between -2 to -4. The amino acids of the second group have a small peak at day 2 with a logFC around 1. Afterwards, their concentration descreases and the logFC mainly fluctuates around 0.

To summarize, a predominant part of the amino acids is down regulated over the whole time span of the experiment. Hence, the hypergeometric test indicates a significant enrichment of this metabolite set (p-values between 0.01 and 0.001 for certain measuring days, e.g day 4 to 10). The pathways of energy generation and oxidative stress are known to be induced during adipogenesis [25]. In her Diploma thesis, Dorothea Portius figured the role of amino acids in such pathways out and therefore, how the lowered concentrations of the amino acids can be explained due to the adipogenesis of the cell. Thus, the results of the amino acid metabolite set indicate that the enrichment analysis provides biological meaningful results.

Distribution-based Tests with GGM-based Metabolite Sets

So far, an enrichment analysis based on t-test results was performed. The problem with that kind of test is the hard cutoff whether there is a significant difference in the metabolite concentration between two measuring days. These results of the t-test are the basis of the hypergeometric test. It could be the case that several metabolites of a certain set scarcely miss this cutoff, but overall a major part of the metabolites of this set show a tendency towards a up or down regulation. This

metabolite set might be interesting, but the hypergeometric test would not show a signal for this metabolite set. Therefore, we applied two different tests which are based on the distribution of the logFCs (see Section 2.4.1 for the fold change calculation).

These two tests, Kolmogornov-Smirnov test (K-S test) and Wilcoxon rank sum test (see Section 2.6.2), evaluate whether the distribution of two populations differ. They were used to compare the logFCs of all metabolites against the logFCs of the metabolite set. This was done for the metabolite set and the measuring days. Day 0 was not included into the analysis, because the logFC of every metabolite is 0 per definition. The p-values that the underlying distributions of these two populations are the same are displayed in Figure 3.24.

The results of the two tests are quite similar, but the K-S test shows more often a signal for a certain metabolite set. This could be due to fact, that the rank sum test only compares the mean of the two underlying distributions, whereas the K-S test uses the distance between the two distributions. Hence, the K-S test could be more sensitive. In comparison to the hypergeometric test, both distribution-based tests also show a clear signal for the set of *amino acids* over all measuring days except day 2. Additionally, *PC + SM* at day 6 and *acylcarnitines #2* at day 18 obtain a low p-value. The latter are especially interesting, since they had very high p-values for the hypergeometric test.

Weighted Enrichment Analysis with GGM-based Metabolite Sets

The weighted enrichment analysis (see Section 2.6.2) is another method which overcomes the hard cutoff issue of the t-test based hypergeometric test. Since the values of the metabolites have to be comparable, we used the logFCs. Additionally, we also applied the weighted enrichment analysis with t-values of t-tests. The setup of the t-tests was similar to the one of the hypergeometric test (see Section 3.2.2), i.e. a t-test for a every metabolite and measuring day pairs with day 0. Since the values have to be non-negative for the weighted enrichment analysis, the absolute-values of the logFCs and t-values were used.

The results of the weighted enrichment analysis for both approaches, logFCs and t-values, are shown in Figure 3.25. Overall, the results of both variants are very similar. Like all the earlier applied tests, both approaches obtain low p-values for the set of *amino acids* over all measuring days. Besides that, the *PC aa + PC ae* set shows a signal for day 10 to 28. This is especially the case in the results of the t-values approach. There are more tests which obtained p-values in the region from 0.2 to 0.5. Thus, it seems like that weighted enrichment analysis with t-values enables a clearer distinction between a signal or no signal for a set than the logFC approach.

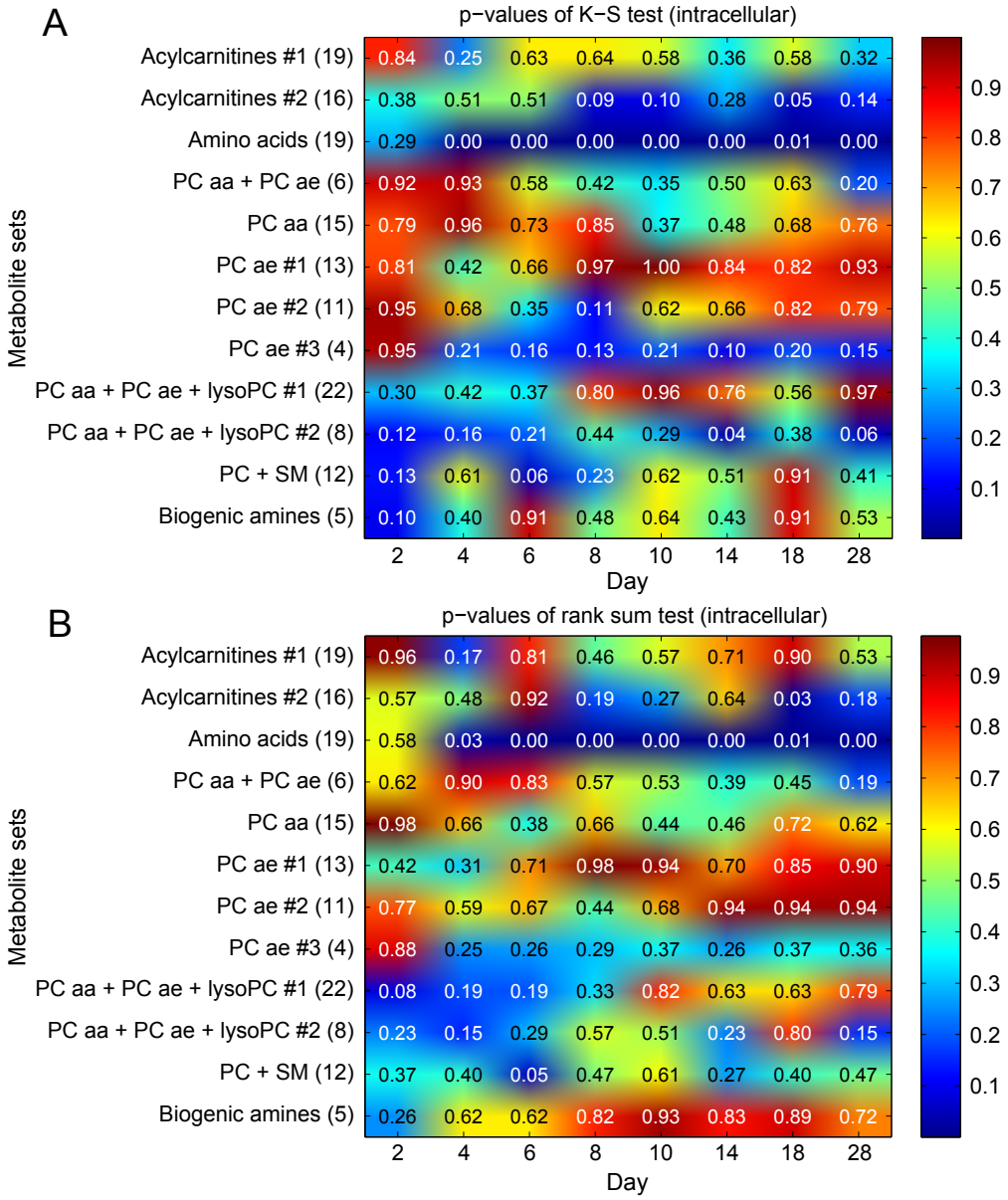

Figure 3.24: Results of an enrichment analysis with the GGM-based metabolite set definition (y-axis). The distribution of the log(fold changes) of all metabolites were compared to the distribution of log(fold changes) for the metabolites of a certain set. The p-value describes the probability that the underlying distribution of the two populations are equal. The Kolmogornov-Smirnov test (K-S test) (A) and the Wilcoxon rank sum test (B) were used. The adjusted significance level α is 0.0029 (K-S test) and 0.0015 (rank sum test) after FDR.

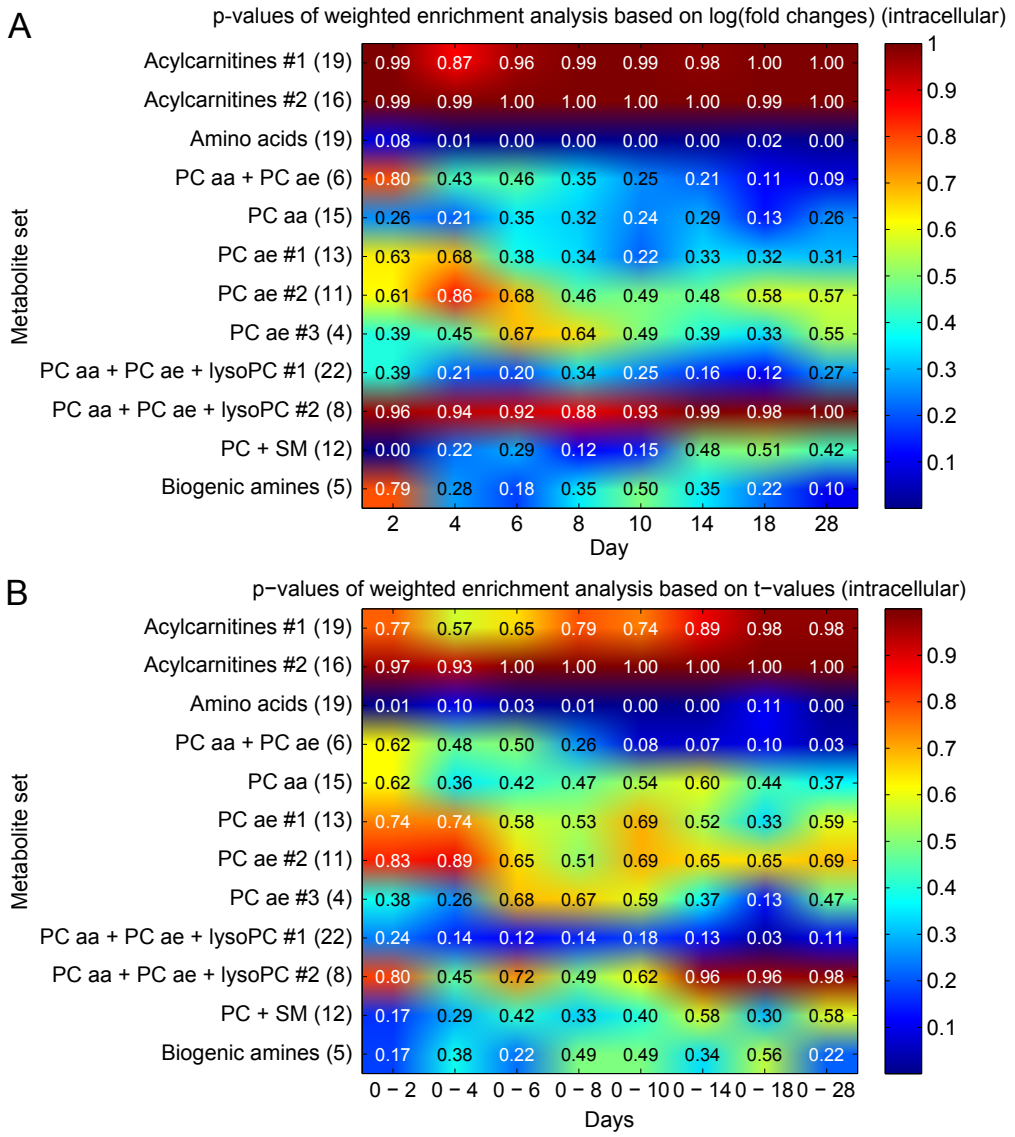

Figure 3.25: The p-values of a weighted enrichment anaylsis with a GGM-based metabolite set definition. (A) The intracellular logFCs of the metabolites were used as the values of the metabolites. The adjusted significance level α is 0.0015 after FDR. (B) A t-test of the intracellular concentrations was applied for every metabolite between the measuring day pair which is denoted on the x-asis. The t-values were then used for the weighted enrichment analysis. The adjusted significance level α is 4.4×10^{-4} after FDR.

Conclusion of the Enrichment Analysis for Intracellular Measurements

Overall, four different methods were applied as an enrichment analysis for the intracellular measurements. These methods are a hypergeometric test based on t-test results, two distribution-based tests with logFCs and the weighted enrichment analysis, which was performed with t-values and logFCs. Before we continue with two further enrichment analysis (i.e. extracellular measurements and grouping of measuring days according two differentiation phases), we want to recap the results of the already applied methods. Additionally, we discuss how the interpretation of the results can differ between the methods.

All performed methods showed the metabolite set *amino acids* to be enriched at almost every measuring day. In Section 3.2.2, a further investigation of this metabolite set and possible biological explanation was given. Besides the amino acids, there were several metabolite sets with phosphatidylcholines that obtained low p-values of the test statistics. The hypergeometric test with the biochemical classes as metabolite sets gave the first indication towards an enrichment of diacyl phosphatidcylcholines, since the *PC aa* metabolite set had low p-values (see Figure 3.21A). This metabolite set contains 31 diacyl phosphatidcylcholines and is one of the largest metabolite sets. The use of GGM-based metabolite sets allowed a finer distinction between the phosphatidylcholines and puts them also in a pathway context.

The metabolite sets *PC aa + PC ae + lysoPC #1* and *PC aa + PC ae* were two sets that obtained low p-values in several tests at different measuring days. The logFCs are displayed in Figure 3.26. A major fraction of the *PC aa + PC ae + lysoPC #1* set has an increased or decreased concentration. The up regulated metabolites have higher absolute logFCs. On the other hand, a larger part of the metabolites have a negative logFC. Overall, there is only a small fraction of this metabolite set which is not altered in concentration. This set has 22 metabolites and is the largest metabolite sets. It consists of 10 diacyl phosphatidylcholines, 3 alyl-alkyl phosphatidylcholines and 9 lyso-phosphatidylcholines. The length of the fatty acids of the lyso-phosphatidylcholines varies between 14 and 20 carbon atoms. The number of carbon-carbon double bonds ranges from 0 to 4 times. All lyso-phosphatidylcholines of this experiment that have a fatty acid with 20 or less carbon atoms are part of this set. The diacyl and acyl-alkyl phosphatidylcholines of this metabolite set have at least one carbon-carbon double bond and the summed number of carbon atoms for the two fatty acids is between 32 and 38.

The metabolite set *PC aa + PC ae* contains 6 metabolites, i.e. 2 diacyl and 4 acyl-alkyl phosphaticylcholines. The sumed number of carbon atoms of its metabolites varies between 30 and 38. For the exception of one metabolite (PC ae C38:1), they all have satured fatty acids. The hypergeometric test and the weighted enrichment analysis indicate an enrichment for the metabolite set *PC aa + PC ae* in

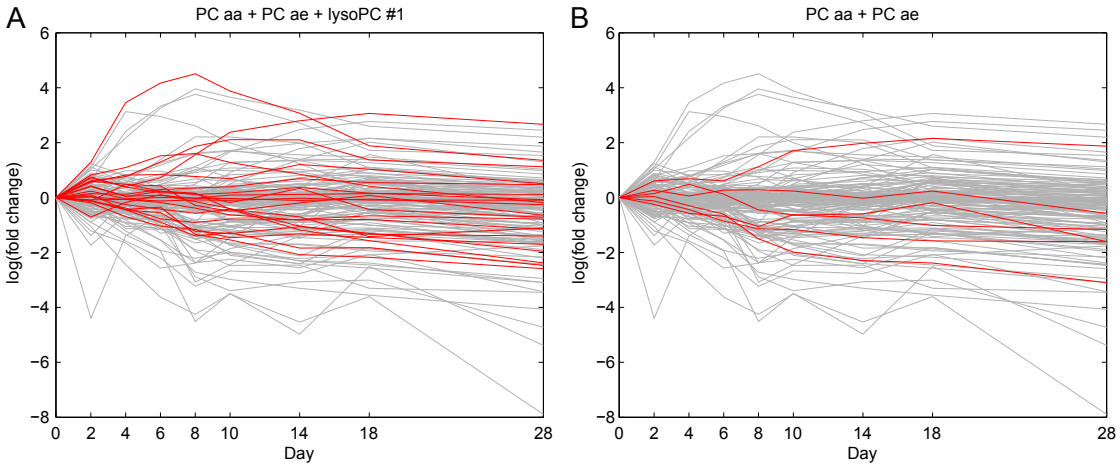

Figure 3.26: The intracellular logFCs of the metabolite sets *PC aa + PC ae + lysoPC #1* (A) and *PC aa + PC ae* (B) are shown in red. The logFCs of the remaining metabolites are shown in gray.

the late stage of the experiment. The inspection of the logFCs (see 3.26B) shows that four metabolites have a logFC below -1 at day 28 and one metabolite has a logFC around 2. The logFC of the remaining metabolite fluctuates around 0. So, the majority of this metabolite set is up or down regulated at the end of the experiment.

To summarize these results, the tests for the enrichment analysis indicated an enrichment of two sets which contain several phosphatidylcholines besides the amino acids. Phosphatidylcholines are part of the phospholipid class, which form all sorts of biological membranes within cells. The lipid composition determines characteristics of these membranes like fluidity or alters the binding of additional proteins, which can change the function of the membrane [34]. The lipid droplets (LD), which store triglycerides and are formed during adipogenesis, consist of phospholipids. The formation and size of these LDs may be influenced by the phosholipids [24]. Additionally, the differential recruitment of LD proteins may also be dependent on the phospholipid composition. Therefore, the changes of the phosphatidylcholines could be necessary adaptions of the lipid composition for the formation of LDs during adipogenesis.

The hypergeometric test is based on the t-tests which evaluate whether there is a significant difference of the metabolite concentrations between the investigated measuring day and the first measuring day. That is, the hypergeometric test only takes into account that there is a significant change and does not consider whether the change is positive or negative. Hence, a metabolite set which is enriched according to the hypergeometric test has an enriched fraction of metabolites with

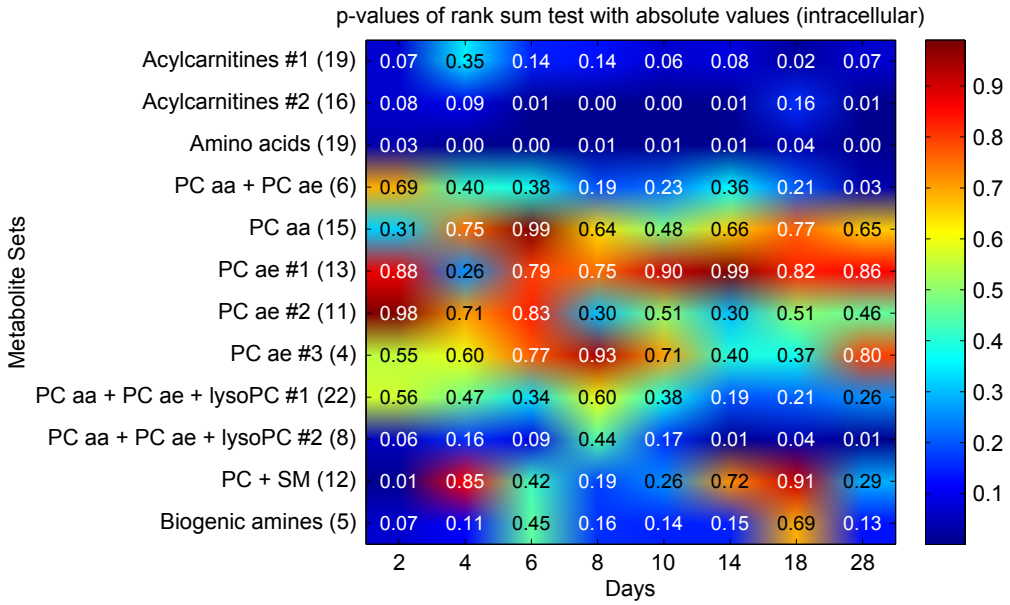

Figure 3.27: Results of an enrichment analysis with the GGM-based metabolite set definition (y-axis). The distribution of the absolute log(fold changes) of all metabolites were compared to the distribution of log(fold changes) for the metabolites of a certain set with the Wilcoxon rank sum test. The p-value describes the probability that the underlying distribution of the two populations are equal. These p-values are not corrected for multiple testing. The adjusted significance level α is 0.0021 after FDR.

a significant change. The weighted enrichment analysis uses absolute-values for the evaluation. In this context, an enriched set means that the metabolites of this set show a higher change than the remaining metabolites.

The results of the distribution-based methods have to be interpretated in different way. They compare the distribution of the set metabolites against the distribution of all metabolites. The outcome of such a test is that the metabolite set behaves different than the other metabolites. For this, the distribution-based tests try to distinguish between an in- or decrease of the metabolite concentration. This could be a problem for the Wilcox rank sum test. It uses the mean of the two distributions to evaluate the difference. It could be the case that the positive and negative logFCs balance each other, so that the metabolite set seems to behave like the other metabolites even though the metabolites of this set are highly up and down regulated. One could try to solve this issue with the use of absolute-values, but it appears to be that the Wilcox rank sum test becomes to sensitive (see Figure 3.27).

These differences in the test statements and interpretation can be illustrated with the following investigation of the metabolite set *acylcarnitines #2*. The

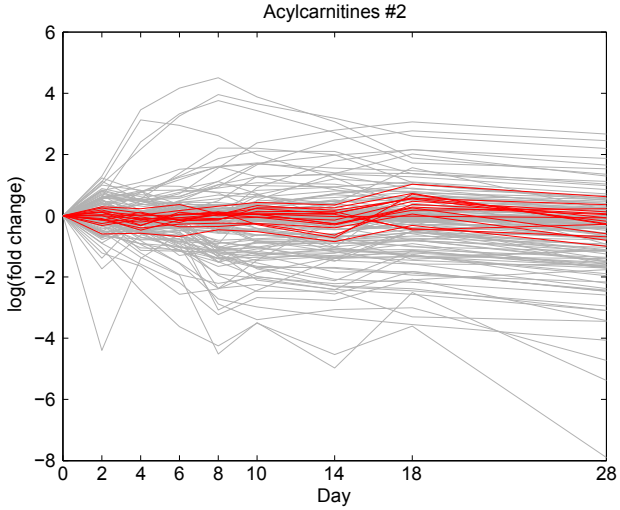

Figure 3.28: The intracellular logFCs of the metabolite set *acylcarnitines #2* are shown in red. The logFCs of the remaining metabolites are shown in gray.

distribution-based tests indicated an enrichment of this metabolite set. This was totally in contrast to the results of the other methods. The logFCs of the metabolites are displayed in Figure 3.28. As it can be seen, the logFCs of all metabolites fluctuate around 0 during the whole experiment. There are two possible explanations, why the distribution-based tests indicated an enrichment of this set. Both explanations rely on the fact that the logFCs of the set metabolites fluctuate around 0. First, the standard deviation of this distribution is lower than the one of the distribution of all metabolites. Second, the mean logFC of all metabolites is slightly below 0, whereas the mean logFC of this metabolite set is slightly above 0. Even though this metabolite set behaves different than the other metabolites, it is unlikely that this metabolite set is interesting for further research in respect of adipogenesis, since there is no really change of metabolite concentrations. Therefore, we are going to focus on the weighted enrichment analysis as the method in the following sections.

Enrichment Analysis of the Extracellular Concentrations

Thus far, the enrichment analysis was performed for the intracellular concentrations. In this section, we want to take a look at the results of the extracellular concentrations. The analysis was performed as a weighted enrichment of the logFCs with the GGM-based metabolite set definition. The results are shown in Figure 3.29. There is less accordance with the results of the intracellular environments. For example, there are two sets (*PC ae #1* and *PC + SM*) that obtain low p-

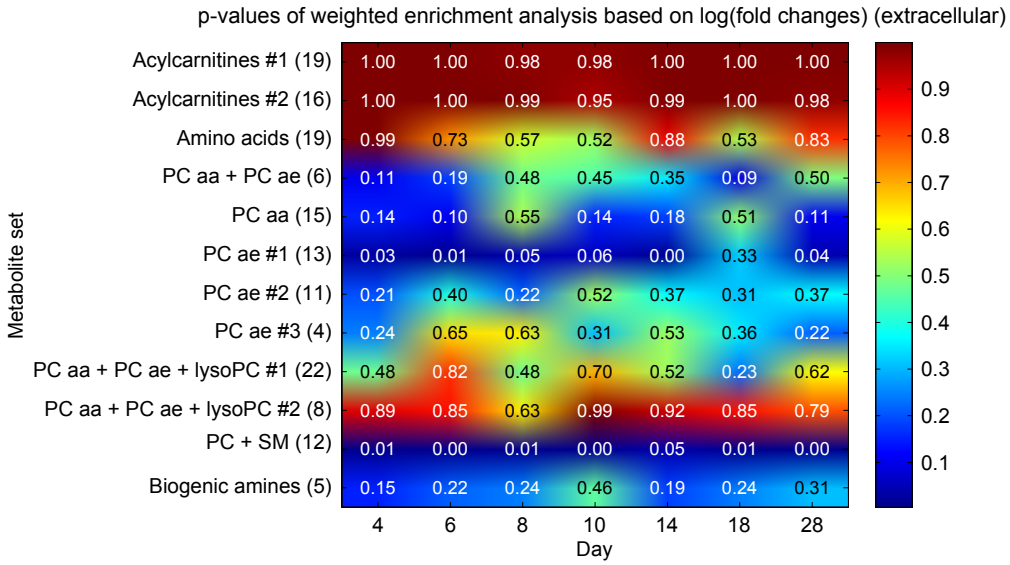

Figure 3.29: The p-values of the weighted enrichment analysis based on the extracellular log(fold changes) and GGM-based metabolite set definition.FDR adjusted the significance level α to 0. Therefore, we used Bonferroni correction to account for multiple testing [3]. The adjusted significance level α is 5.95×10^{-4}.

values for almost all measuring days. These sets did not have low p-values in an enrichment analyis for the intracellular concentrations. In contrast, the sets that showed up frequently in the intracellular results (e.g. *amino acids* or *PC aa + PC ae + lysoPC #1*) have high p-values at this enrichment analysis. So, it seems like that there is only a small connection of these two environments in respect of the metabolite set behaviour. This is further investigated in Section 3.3.

The logFCs of the two enriched sets are displayed in Figure 3.30. The metabolite set *PC + SM* contains all sphingomyelins of this dataset and two phosphatidylcholines (PC aa C28:1 and PC ae C40:2). Three of these metabolites show a strong upregulation at day 4 and maintain this level to the end of the experiment. These metabolites are SM (OH) C24:1, SM C22:3 and PC aa C28:1. The remaining metabolites have logFCs that mostly vary between -1 and -2. A sphingomyelin consists of sphingosine, a fatty acid and a phosphorylcholine as a head group. It was shown in mice that the experession of the sphingosine kinase is elevated during adipogenesis [12]. The sphingosine kinase facilitates the phosphorylation of sphingosine to sphingosine 1-phosphate (S1P). Thus, the breakdown of sphingomyelins to generate S1P could be an explanation of the lowered concentrations of sphingomyelins in the adipogenesis.

The *PC ae #1* set consists mainly of acyl-alkyl phosphatidylcholines (PC ae) with a combined chain length between 30 and 38 carbon atoms, which are unsat-

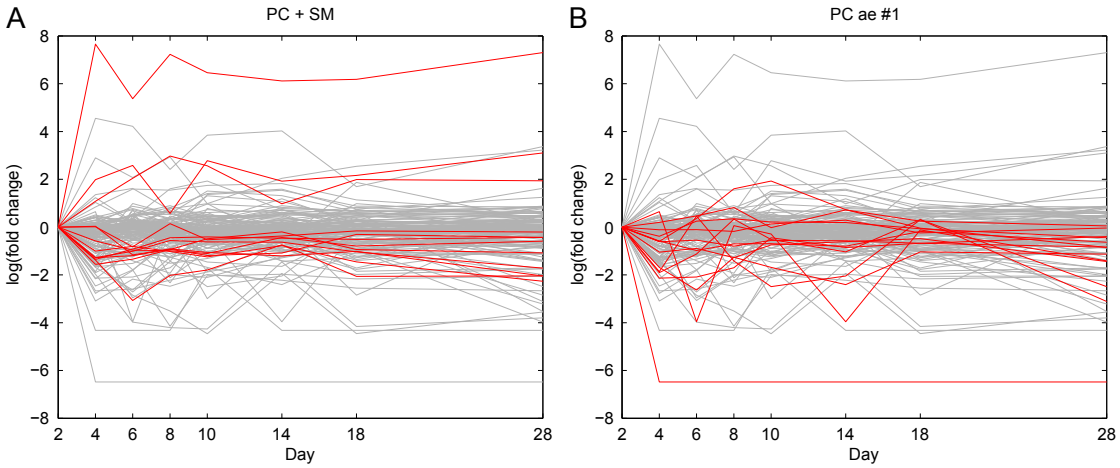

Figure 3.30: The extracellular logFCs of the metabolite sets *PC + SM* (A) and *PC ae #1* (B) are shown in red. The logFCs of the remaining metabolites are shown in gray.

urated between one and six times. This metabolite set also contains two diacyl phosphatidylcholines (PC aa C36:0 and PC aa C38:0). Overall, the logFCs of almost every metabolite is varying between 0 and -4, but there are two exceptions. First, there is PC aa C38:0 which reaches its peak with a logFC of 2 at day 10. The second exception is PC ae C34:1, which has the lowest logFCs of all metabolites of this experiment. This is due to the fact that the measured concentrations of PC ae C34:1 are 0 for all three experiments from day 6 to 28. This is also the case for experiment #2 and #3 at day 4, so that experiment #1 is the only measuring point with a value unequal 0. As described in Section 2.4.1, the measuring points with a value of 0 are replaced with the lowest value which is unequal 0. Hence, the logFC is equal for every measuring day starting at day 4 and the time course is constant. We wanted to ensure that the set *PC ae #1* is not enriched due to this metabolite. Thus, we repeated the weighted enrichment analysis and excluded PC ae C34:1. The metabolite set *PC ae#1* still obtained low p-values.

The extracellular enrichment analysis was also carried out with the other tests. The results for the weighted enrichment analysis with t-values, the K-S and the rank sum test are included in the appendix (see Figure D.2 D.3 and D.4). It is remarkable that there are almost no similarities between these four results, whereas the application of the tests for the intracellular measurements shared similar results to some extent. This could be an indicator that it is a demanding task to analyze the extracellular measurements. This could be due to the lack of technical replicates or metabolites with a high fraction of measured concentrations that are 0. Additionally, we also discussed the differences in the interpretation of the extracellular measurements in Section 3.1.4.

At the end, we also want to point out the results of the hypergeometric test based on t-test for each metabolite. There was not a single metabolite which obtained a p-value low enough that means of the two measuring days is considered to be significant different. This can be explained with the small population size, since there are no technical replicates for the extracellular measurements. Hence, the t-tests were performed with three values per population. Since there are no metabolites with a significant difference, the hypergeometric test did not show any results.

Grouping of Measuring Days According to differentiation phases

As described in Section 2.1.3, the adipogenesis can be divided into three phases. These phases can be distinguished by lipid accumulation within the cell, GPDH enzyme activity, mRNA expression levels of the key adipogenic transcription factor PPARγ and the mature adipocyte marker leptin. A metabolite ought to show a similar behaviour at the measuring days within a certain differentiation phase. Thus, we also performed a weighted enrichment analysis, for which the measuring days were grouped according to one of the three differentionphases. The grouping is as follows:

- Early phase: day 0 and 2 (early differentiation phase = no expression or activity of any assessed marker)
- Middle phase: day 4, 6, 8, 10 and 14 (differentiation and continously increasing lipid accumulation in lipid droplets = increase of PPARγ mRNA, GAPDH expression and lipid staining)
- Late phase: day 18 and 28 (mature adipocytes = lipid accumulation completed, GAPDH activity and PPARγ levels reached maximal levels, leptin expression detected)

The weighted enrichment analysis was performed with the GGM-based metabolite set definition and the intracellular logFCs. The calculation of the fold changes is adapted as described in Section 2.4.1. For example, the comparison between middle and late is based on the fold change of the middle to the late differentiation phase. For the three possible comparisons, the p-values of the weighted enrichment analysis are shown in Figure 3.31. As it could have been expected from the results for the ungrouped measuring days, the metabolite set *amino acids* has low p-values for the two comparison with the early phase. The fold changes between the early and middle differentiation phase are visualized in a GGM (see 3.32A). The comparison between the middle and late phase does not obtain a low p-value. This is due to the fact that the concentration of the amino acids shows the strongest

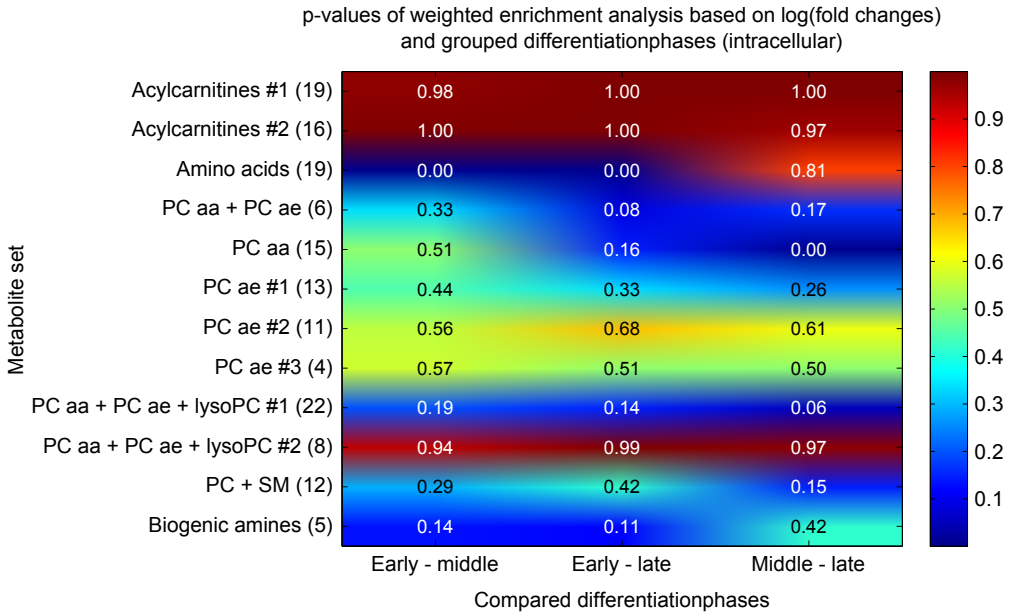

Figure 3.31: The p-values of the weighted enrichment analysis with grouped measuring days according to their differentiation phase. The logFCs were calculated as described in Section 2.4.1, so that the fold change describes the change between the first named differentiation phase to the second named on the x-axis. The grouping of the measuring days is as follows: day 0 and 2 are early, day 4, 6, 8, 10 and 14 are middle and day 18 and 28 are late. The metabolite sets are based on the GGM and denoted on the y-axis. The adjusted significance level α is 0.004.

decrease in the first days (day 2 to 8). After that, there is only a slight decrease of concentration to the end of the experiment (see Figure 3.23). The majority of the amino acids has a logFC between 0 and -1 (see 3.32B), which does not lead to an enrichment by the analysis. These results allow the conclusion that the switch of the amino acid metabolism (see Section 3.2.2) occurs in the transition from the early to the middle phase and is then maintained in the late phase.

The metabolite set *PC aa* has a low p-value for the comparison between the middle and late phase. This set consist of 13 diacyl phosphatidylcholines with combined chain length between 30 and 40 carbon atoms with 2 to 6 double bonds per fatty acid chain. Additionally, this set contains two acyl-alkyl phosphatidylcholines (PC ae C38:0 and PC ae C42:0). As it can be seen in the fold change colored GGM, these two acyl-alkyl phosphatidylcholines are slightly lowered in concentrations, whereas four of the diacyl phopsphatidylcholines have a logFC of -1.5 or lower. In contrast to that, PC aa C40:3 and PC aa C40:4 show an increase of concentration. Since PC aa C40:6 is one of the PCs which is lowered in concentrations, there could be a connection. Overall, there is enough change in both directions (i.e. increase and decrease of concentrations), so that this metabolite set is enriched according to this analysis. As described earlier (see Section 3.2.2), the change of PC concentrations may be explained with the remodelling within the cell during the adipogenesis.

The representation of the fold changes between the phases in the GGM represents a lot of information very cleary. It can be used to find interesting group of metabolites and as a starting point for further investigations. For example, the GGM with the fold changes of the early to the late phase (see Figure 3.33) exposed an interesting pair of metabolites, PC aa C38:3 and PC aa C38:4. They have a very high partial correlation of 0.48, but PC aa C38:3 has a very low logFC of -2.5, whereas PC aa C38:4 has a high logFC around 2.25 (see Figure 3.34). This anti-correlation of these metabolites starts at day 4. A desaturation of one of the two fatty acid chains of PC aa C38:3 could explain these time courses. This hypothesis could be verified with another experiment which assesses the activity of the corresponding enzyme(s).

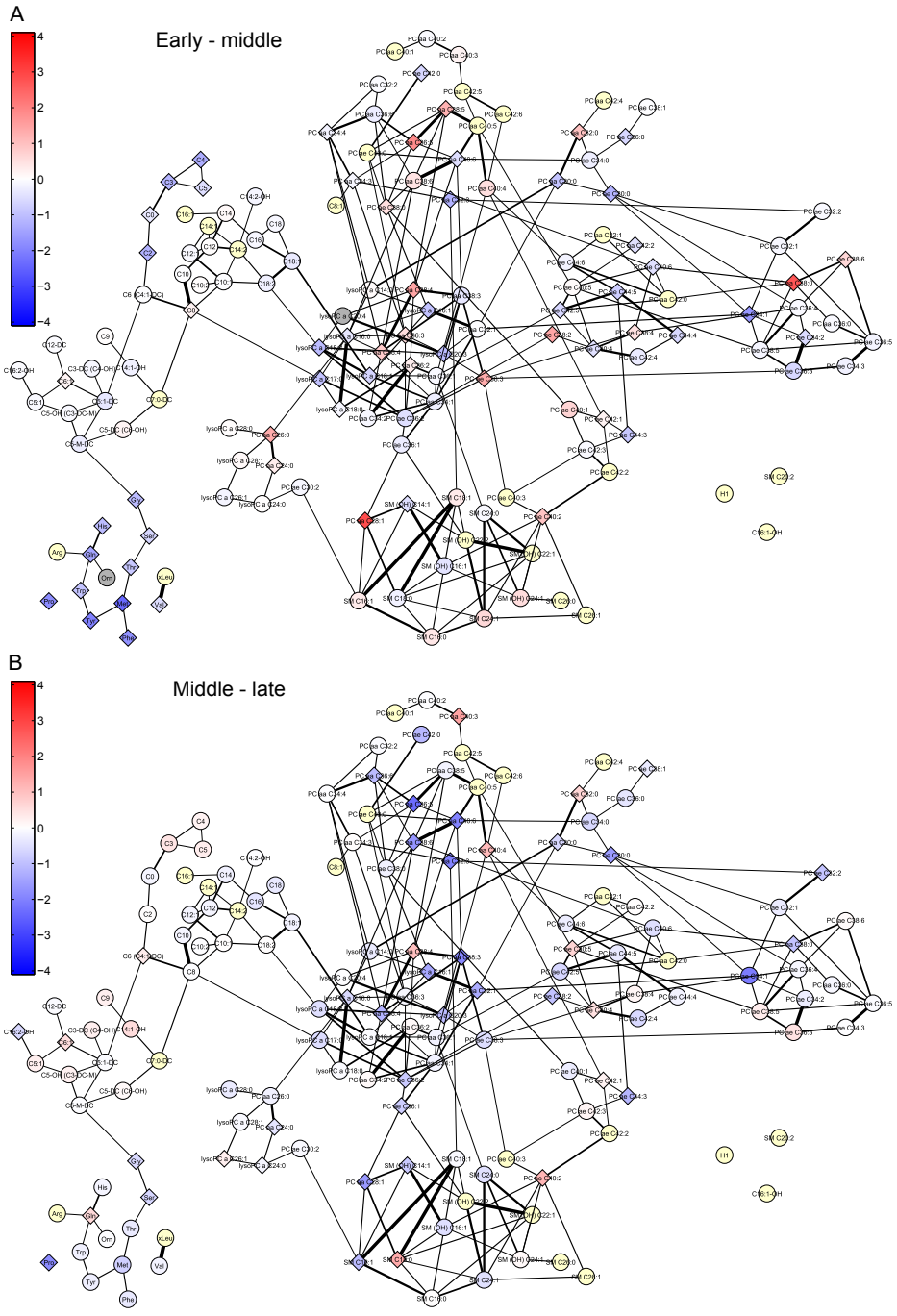

Figure 3.32: This is the same Gaussian graphical model (GGM) as in Figure 3.22. The nodes are colored with the fold change between the two differentiationphases which are denoted in the top of a panel. Grey nodes are metabolites with a missing value after the fold change calculation. Yellow nodes are metabolites that are not measured in this experiment. Additionally, a t-test between the measuring points of these two phases was performed for every metabolite. Metabolites with a significant difference between the two phases have a diamond as the shape of the node. (A) The fold change between the early (day 0 and 2) and middle phase (day 4, 6, 8, 10 and 14) are displayed. The adjusted significance level after FDR is 0.0197. (B) The fold change between the middle phase (day 4, 6, 8, 10 and 14) and late phase (day 18 and 28) are shown. The adjusted significance level after FDR is 0.0086.

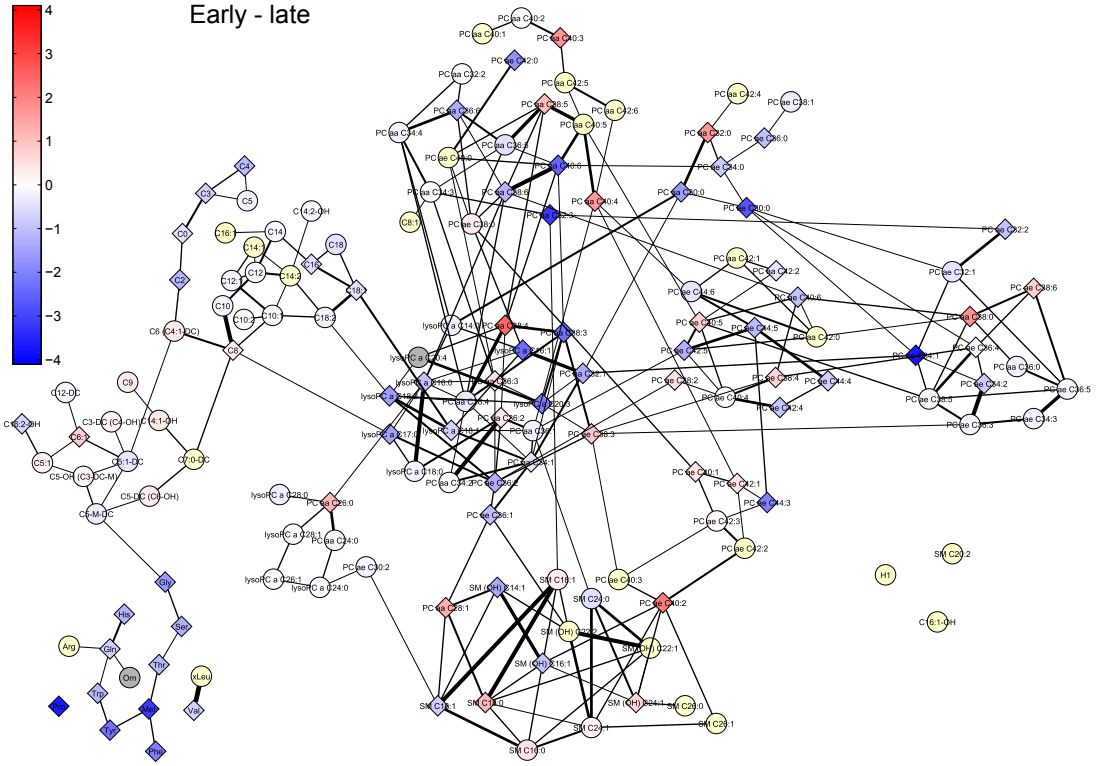

Figure 3.33: This is the same Gaussian graphical model (GGM) as in Figure 3.22. The nodes are colored with the fold change between the early phase (day 4, 6, 8, 10 and 14) and late phase (day 18 and 28) of differentiation. Grey nodes are metabolites with a missing value after the fold change calculation. Yellow nodes are metabolites that are not measured in this experiment. Additionally, a t-test between the measuring points of these two phases was performed for every metabolite. Metabolites with a p-value $<$ 0.0174 (adjusted significance level after FDR) have a diamond as the shape of the node.

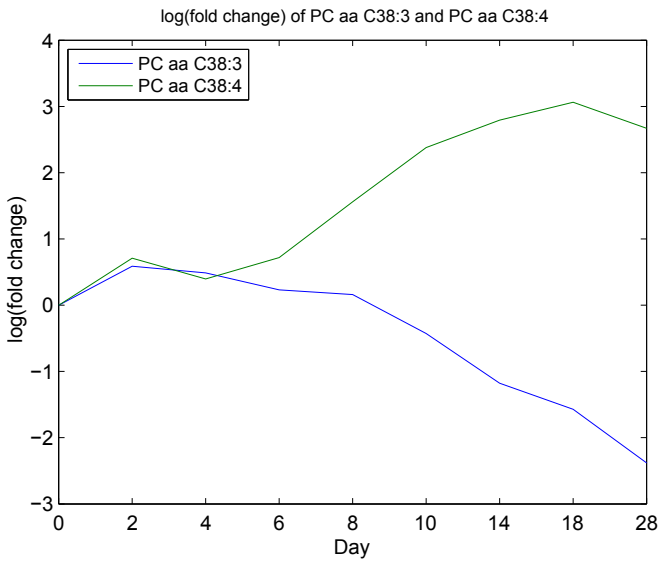

Figure 3.34: The logFCs of PC aa C38:3 (blue) and PC aa C38:4 are displayed. Starting at day 4, the logFCs of these two metabolites show an anti-correlation, which could be explained with a desaturation of one fatty acid chain of PC aa C38:3.

3.3 Analysis of the Intra- and Extracellular Metabolite Dependency

The clustering analysis and the enrichment analysis were methods that seperately dealt with the two environments, i.e. intra- and extracellular. In this section, we want to investigate the exchange of metabolites between the intra- and extracellular environment. For that, three different methods were performed to detect metabolites, for which a connection can be observed between the intra- and extracellular concentrations. Two of these methods are based on the Spearman's rank correlation coefficient (SCC) and the last one uses the t-values of Student's t-test. All methods were performed with the log(concentrations) of the 151 metabolites that passed the quality control.

3.3.1 Global Correlation

The SCC describes the linear dependence between the intra- and extracellular concentrations for a metabolite. The calculation was performed over 8 measuring days (from day 8 to day 28) as described in 2.7.1. Since it is always somehow arbitrary to determine a cutoff, we decided to use the p-value to evaluate whether the SCC is significantly different from 0. At a significance level α of 0.05, there are 16 metabolites with a significant SCC. A list of the metabolites is given in Table 3.7. After multiple testing correction with FDR, α was adjusted to 6.19×10^{-5}, which was only passed by one metabolite. This metabolite is glutamine, which has a SCC of 0.74. The intra- and extracellular concentrations are shown in Figure 3.35. Increase or decrease of the concentrations are highly correlated between these two environments, e.g. the increase from day 10 to day 28.

3.3.2 Window-based Correlation

Since it could be the case that an intra-/extracellular exchange only occurs during a certain timeframe of the differentiation process (e.g. from day 4 to day 8), the calculation of the SCC over all eight measuring days might not be the right approach. Thus, we applied the computation of the SCC over a certain time window of measuring days as described in 2.7.2. We performed this method with a range of the windowsize ws between 2 and 5. Due to the window-based approach, one metabolite has n $(= 8 - ws + 1)$ SCCs. Table 3.6 list the number of metabolites that had a SCC with a significant difference from 0 for at least one window. In this case, we used the Bonferroni correction to account for multiple testing [3], because FDR always adjusted the significance level α to 0.

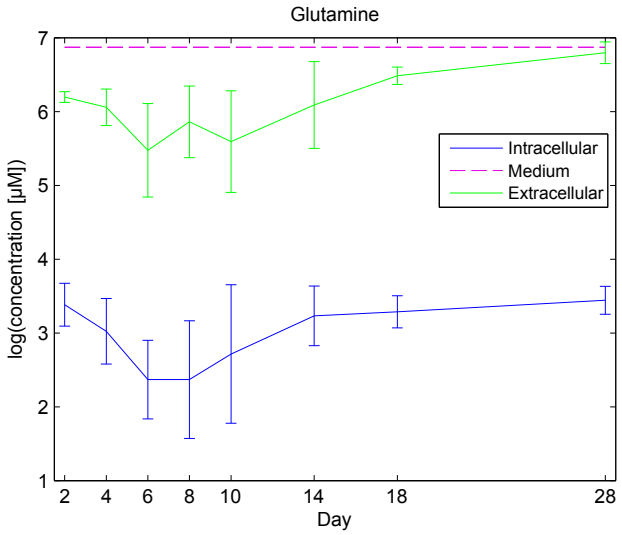

Figure 3.35: The intra- and extracellular log(concentrations) of glutamine are shown. The mean of the three experiments was calculated. The errorbars are plus-minus the standard deviation of the experiments. Glutamine has a Spearman's rank correlation coefficient of 0.74. The concentration of glutamine is also plotted for the medium to assess the effect of the medium on the extracellular concentrations.

Window size	2	3	4	5
Number of significant metabolites (uncorrected)	32	37	32	33
Number of significant metabolites (corrected)	0	0	0	0

Table 3.6: The number of metabolites that have a Spearman's rank correlation coefficient (SCC) which is significantly different from 0. The significance level α for uncorrected was 0.05. It was corrected for multiple testing with the Bonferroni correction [3]. It was used instead of FDR, because FDR always adjusted α to 0. The adjusted α were as follows: $\alpha = 4.73 \times 10^{-5}$ for a window size of 2 ($ws = 2$), 5.53×10^{-5} for $ws = 3$, 6.62×10^{-5} for $ws = 4$, 8.28×10^{-5} for $ws = 5$.

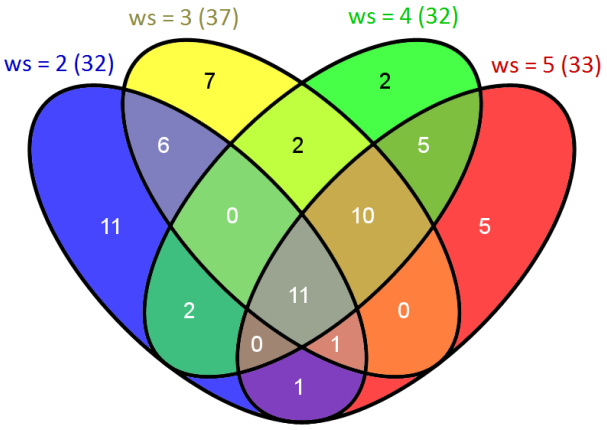

Figure 3.36: Venn diagram of the window-based SCC calculation with a varying windowsize ws (from 2 to 5). The numbers denote the quantity of metabolites with a SCC that is significant different from 0 (not corrected for multiple testing). The numbers in the brackets denote the number of metabolites detected by this method. An overview of the detected metabolites with the varying windowsize is given in Table E.1. The venn diagram was created with VENNY [23].

To visualize the differences of the results between the windowsizes, a Venn diagram was created (see Figure 3.36). Almost the half of the metabolites which are detected by $ws = 2$ are not found by $ws = 3$. The major of these are also not detected by one of the other two windowsizes. So, the result of $ws = 2$ is the most different one in comparison to the others. 21 metabolites are shared by the other three windowsizes, which is around two-thirds of their whole result. Table E.1 lists the detected metabolites of the different window sizes. A windowsize of 3 is going to be used for further analysis. This windowsize seemed to be the most reasonable considering the biological meaning of it which is the temporarily exchange of metabolites between the environments.

So far, this method was based on a window that moved along the measuring days. Similar to Section 3.2.2, we include information about the differentiation phases in the analysis. Instead of a moving window with a rigid size, we defined three windows according to the differentiation phases:

- Window for the early phase: day 2 and 4
- Window for the middle phase: day 6, 8, 10 and 14

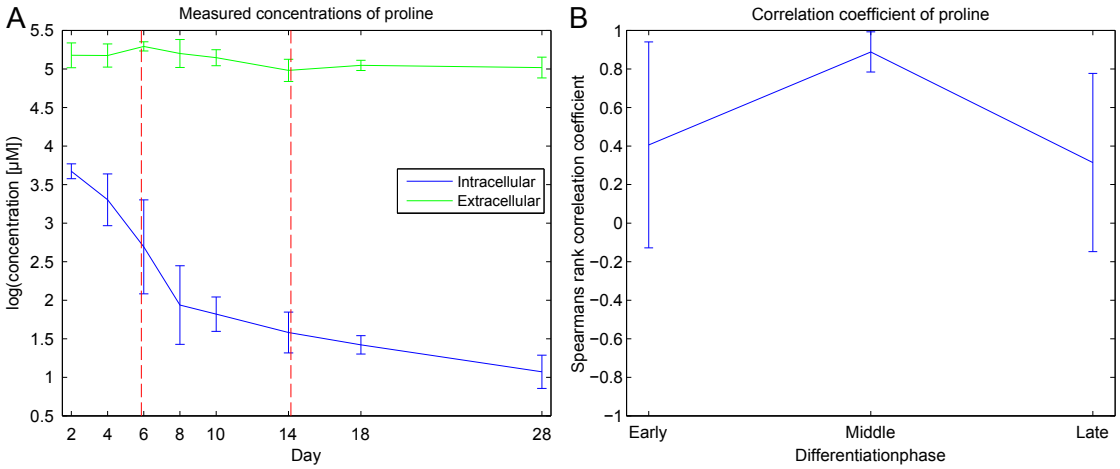

Figure 3.37: The log(concentrations) and Spearman's rank correlation coefficient (SCC) of proline are shown. (A) The intra- and extracellular log(concentrations) of proline are displayed. The two red dashed lines indicate the window with the significant SCC. (B) The SCC of the three windows for the differentiation phases. The errorbars were calculated with bootstrapping with a repetition of 10,000 times. This approach might not be accurate, since there are only two measuring days in the early and late phase. The SCC of proline over all 8 measuring days is 0.60.

- Window for the late phase: day 18 and 28

In contrast to the definition of differentiation phases for the enrichment analysis, day 4 is part of the early phase. One can justify the assignment of day 4 to both differentiation phases with the data of the differentiation markers. In this case, the assignment of day 4 to the early phase was necessary. Since day 0 is left out for the analysis of the intra-/extracellular exchange, day 2 would have been the only measuring day in the early phase and a SCC calculation of this window would have not been possible.

The predefined windows detected 20 metabolites that have at least one SCC which is signifcantly different from 0 (at a significance level of 0.05). The detected metabolites are listed in Table 3.7. There are 10 metabolites in the early phase, 7 in the middle phase and 4 in the late phase. But only one of the 21 SCCs is still significant after the Bonferroni correction to account for multiple testing. This metabolite is the amino acid proline. The log(concentrations) and the SCCs for the three phases are shown in Figure 3.37. The middle phase has the significant SCC with a value of 0.88. The SCC of the other two phases are with a value around 0.4 relatively high, but their p-values is not below 0.05 due to the low number of used data points for the calculation of the SCC.

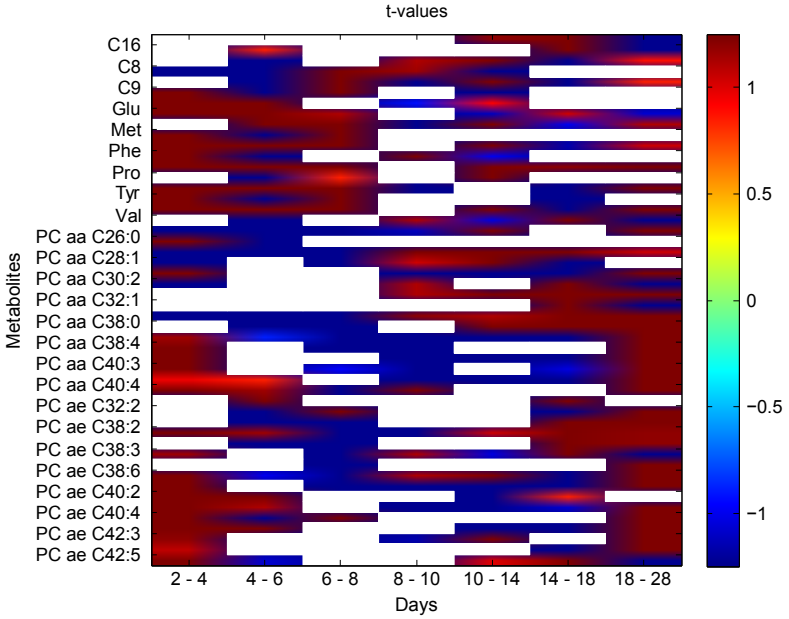

Figure 3.38: Representation of the t-values for 25 metabolites. A t-test between the two denoted measuring days on the x-axis was performed for every metabolite. t-values above the threshold 1.25 and below -1.25 are set to these two values. All t-values between -0.83 and 0.83 (this is two-thirds of 1.25) are colored white. This was done for readability. A positive t-value means a decrease of concentration and a negative value an increase. There are two lines per metabolite. The first one consists of t-values for the intracellular environment and the second line the t-values of the extracellular one. Each of these metabolites has at least two measuring day pairs at which the absolute t-value was equal or above the threshold of 1.25 for both environments, e.g. C16 at the days 14 - 18 and 18 - 28.

3.3.3 t-Value Based

The t-value based method uses t-tests of log(concentrations) between adjacent measuring days as described in Section 2.7.3. It seeks for metabolites that have at least two so called qualified measuring day pairs. A measuring day pair consists of the two days which are used to perform the t-test. It is called qualified, when the absolute t-values of both environment are equal or above the threshold t of 1.25. In this test-setting, this t-value corresponds to a p-value around 0.22 to 0.28.

This method detected 25 metabolites, which are listed in Table 3.7. A representation of the t-values of these metabolites is given in Figure 3.38. The majority of these metabolites has at least one of the qualified measuring day pairs at the first (day 2 to 4) or last (day 18 to 28). The most prevalent biochemical classes among these metabolites are amino acids and phosphatidylcholines. This could

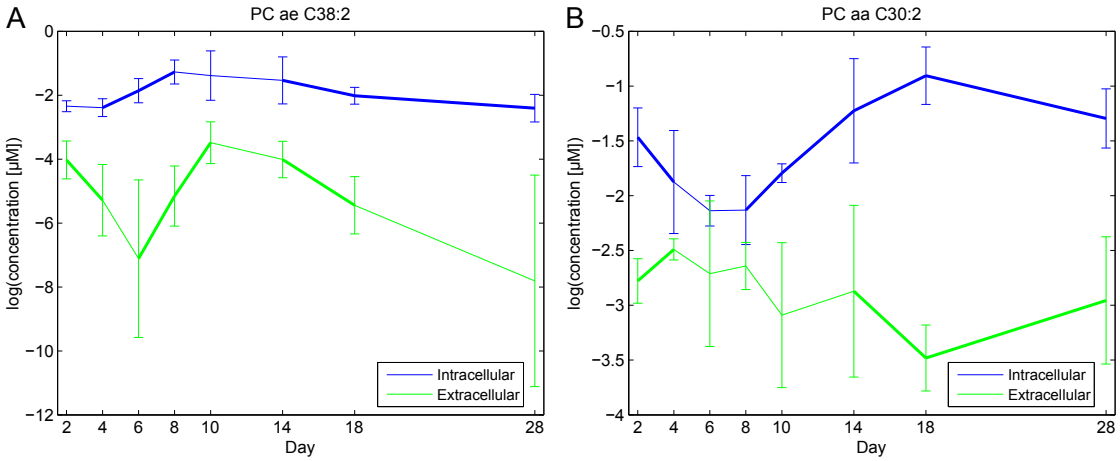

Figure 3.39: The intra- and extracellular log(concentrations) of PC ae C38:2 (A) and PC aa C30:2 (B). The line between two measuring days has a bigger width if the t-value of the corresponding measuring day pair was above the threshold.

have been expected, since this analysis is based on changes of concentrations and the enrichment analysis already showed that the metabolites of these biochemical classes undergo these changes. The log(concentrations) of two metabolites are displayed in Figure 3.39. PC ae C38:2 shows a tendency towards a positive correlation (see Figure 3.39A), whereas PC aa C30:2 seems to be negative correlated (see Figure 3.39). Interestingly, both of these metabolites were not detected by one of the earlier applied methods.

There are several modifications possible of this analysis to increase its sensitivity or specificity or to make it more suitable for a certain biological scenario. So far, a threshold of 1.25 is used and at least two qualified measuring day pairs are necessary. Obviously, one can in- or decrease the threshold to obtain more or less metabolites that fulfill these criteria. It is also possible to add criteria for a qualified measuring day pair or to alter the criteria. For example, the qualified measuring day pairs have to be consecutive or that the change of both environments have to be the same, e.g. there is an increase of the intra- and extracellular environment.

3.3.4 Conclusion of the Intra- and Extracellular Exchange

We applied three different methods to asses the exchange of metabolites between the intra- and extracellular environment. At first, we want to compare the results of these methods and then take a look at the metabolites and their biochemical classes. For the methods which are based on the calculation of the SCC, we used the detected metabolites which had a SSC significantly different from 0 before

Global	Window	Preset window	t-value
C3-DC (C4-OH)	C12:1	C14:1-OH	C16
Ala	C14	C18:1-OH	C8
Gln	C16:2-OH	C3-OH	C9
Glu	C18:2	C4:1	Glu
Pro	C4:1	Gln	Met
PC aa C28:1	C5:1	Gly	Phe
PC aa C38:0	C6 (C4:1-DC)	Pro	Pro
PC aa C38:1	Ala	Tyr	Tyr
PC aa C38:4	Gln	PC aa C24:0	Val
PC aa C40:3	Glu	PC aa C34:1	PC aa C26:0
PC ae C34:1	Lys	PC aa C34:4	PC aa C28:1
PC ae C34:3	Phe	PC aa C36:1	PC aa C30:2
lysoPC a C18:1	Pro	PC ae C34:2	PC aa C32:1
lysoPC a C24:0	Ac-Orn	PC ae C34:3	PC aa C38:0
SM (OH) C16:1	Met-SO	PC ae C38:8	PC aa C38:4
H1	PC aa C26:0	PC ae C40:1	PC aa C40:3
	PC aa C28:1	PC ae C40:6	PC aa C40:4
	PC aa C30:0	lysoPC a C17:0	PC ae C32:2
	PC aa C32:2	lysoPC a C18:1	PC ae C38:2
	PC aa C36:2	SM C22:3	PC ae C38:3
	PC aa C38:0		PC ae C38:6
	PC aa C38:1		PC ae C40:2
	PC aa C40:2		PC ae C40:4
	PC aa C40:3		PC ae C42:3
	PC aa C40:6		PC ae C42:5
	PC ae C34:1		
	PC ae C34:2		
	PC ae C34:3		
	PC ae C36:3		
	PC ae C38:3		
	PC ae C38:6		
	PC ae C40:1		
	PC ae C40:6		
	PC ae C44:6		
	lysoPC a C20:3		
	lysoPC a C28:0		
	SM (OH) C16:1		

Table 3.7: This table lists the metabolites that were detected by one of the four methods. *Global* describes the calculation of the Spearman's rank sum correlation coefficient (SCC) over all measuring days. *Window* is the window-based computation of the SCC. A union of the results with a windowsize $ws = 3$ and $ws = 4$ is used. *Preset window* is the use of three predefined windows according to the differentiation phases. *t-value* is the usage of t-values of t-tests.

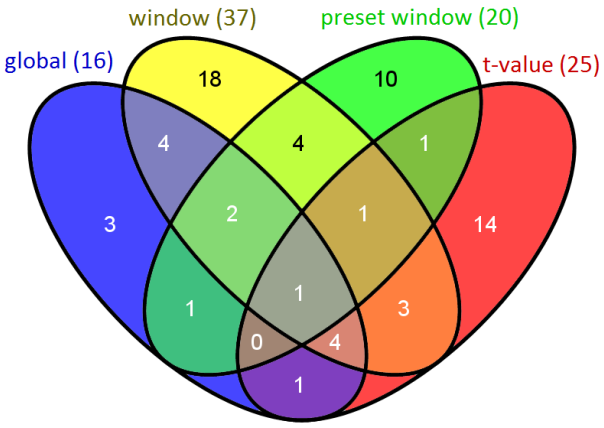

Figure 3.40: Four different analysis were applied to investigate the exchange between the intra-/extracellular environment. The number of detected metabolites are shown in this Venn diagram. *Global* describes the calculation of the Spearman's rank sum correlation coefficient (SCC) over all measuring days. *Window* is the window-based computation of the SCC. A union of the results with a windowsize $ws = 3$ and $ws = 4$ is used. *Preset window* is the use of three predefined windows according to the differentiation phases. *t-value* is the usage of t-values of t-tests. The numbers in the brackets denote the number of metabolites detected by this method. An overview of the detected metabolites is given in Table 3.7. The venn diagram was created with VENNY [23].

the correction for multiple testing. An overview of the results is given as a Venn diagram in Figure 3.40. In the following, we are going to refer to the results with the term which is denoted in the Venn diagram, i.e. *global*, *window*, *preset window* and *t-value*. The window-based correlation method is present with two results. One is based on the predefined windows according to the differentiation phases. The second one uses a rigid moving window with $ws = 3$.

First of all, around two-thirds of the metabolites which were detected by *global* are also found by *window*. This was expected, since both methods are based on the calculation of the SCC. *global* is the same as the window-based method with $ws = 8$. Overall, *preset window* has found 12 metabolites which are not part of the *window* result. Further investigation revealed that 8 of these are found in the window of the early or late phase. Both phases have a windowsize of 2 and the comparison of the different windowsizes in Section 3.3.2 already showed that $ws = 2$ finds a lot of metabolites that are not detected by $ws = 3$. Hence, the

preset window covers an aspect that *window* does not. In general, the overlap between *window* and *preset window* is very small considering that they are based on the same method.

There is one metabolite which is detected by every method. This metabolite is proline. As it was the only metabolite with a significant p-value after Bonferroni correction for the *preset window* method, its log(concentrations) were already displayed in Figure 3.37. Altogether, the overlap between the methods is very small. *window*, *preset window* and *t-value* each have around 50% of metabolites which are not detected by one of the other methods. *t-value* seems to be the most different one. This is probably due to the fact that it is the only method which is not based on the SCC.

There are several biological scenarios that could be explained with an exchange between the two environments. For example, an intracellular decrease of concentration and an extracellular increase could be due to a flux of the metabolite from the intracellular environment to the extracellular one. Vice versa, with an intracellular increase and an extracellular decrease, it could be other way. But there could be also scenarios, in which there is such a strong intracellular increase which leads to a flux out of the metabolite. So, the concentrations in both environments would increase. These effects could be observed on varying time spans. One method is unlikely to find all of these metabolites that undergo an remarkable exchange. For example, Figure 3.39 shows two metabolites which were found by *t-value*, but were not detected by the other methods. Their time courses are likely to indicate an exchange between the two environments. Thus, probably every method has its advantages and disadvantages. On the other hand, not every in- or decrease of concentration is due to an exchange between the two environments. Metabolite A can be catalyzed to another metabolite B, which would lead to a decrease of A and increase of B. These observations are not in context with an exchange between the two environments. So, not every metabolite, which is detected by one of these methods, might be target of an exchange. In conclusion, further investigation of the detected metabolites is necessary to assess whether an exchange of a certain metabolite occured. One possible approach could be the labeling of the corresponding metabolites to make a possible exchange traceable.

We implemented all methods in a way that they compare the intracellular measurements of a day with the corresponding extracellular measurements of the same day, e.g. intracellular measurements of day 4 to the extracellular measurements of day 4. One can argue that it takes some time until an exchange of metabolite manifests in the measured concentrations. We assume that this possibility is unlikely, because the exchange of metabolites occurs in such a small timeframe compared to the time distances between the measuring points, so that the two or more days are sufficient.

At the end, we performed an enrichment analysis as a hypergeometric test (see Section 2.6.2) with the results of the different methods to evaluate whether there is a group of metabolites enriched. Similar to the enrichment analysis in Section 3.2.2, we used the biochemical classes and GGM-based metabolite sets as the sets for the hypergeometric test. For both ways of the metabolite set definition, there was no set which was significantly enriched for any of the methods. The lowest p-values were obtained by the amino acids and sets which contain phosphatidylcholines. These are metabolites of biochemical classes which were also detected by the enrichment analysis in Section 3.2.2. Even though there was no significant enrichment of these metabolite sets for the exchange analysis, it seems like that the changes of concentrations of these metabolites rely to some extent on the exchange of metabolites between the two environments.

Chapter 4

Summary and Outlook

Obesity describes the excess of white adipose tissue (WAT) and is associated with cardiovascular diseases and diabetes type 2. Therefore, it is a major health risk factor in the Western world and developing countries [14, 39]. Besides preadipocytes, fibroblasts, nerves and diverse immune cells, WAT consists primarily of adipocytes [7]. The differentiation of preadipocytes to mature adipocytes is called adipogenesis. This process was observed in three independent experiments, in which preadipocytes differentiated to adipocytes. In this thesis, we analysed this metabolomics data to gain further insights of adipogenesis. Before we were able to carry out the analysis, we had to perform a quality control of the metabolites. Depending on the environment, this resulted in the exclusion of 21 to 37 metabolites. Still, there were metabolites with missing values, which were imputed with two methods. We also showed that the medium of day 0 has an influence on the extracellular measurements and discussed that the interpretation of the extracellular measurements differs from the intracellular one.

At first, a clustering analysis was performed to get an overview of the data. It was carried out with the Euclidean distance and Pearson product-moment correlation coefficient (PCC) as the distance measures. The results of the Euclidean distance helped us to find groups of metabolites with a similar regulation. Additionally, the clustering analysis already gave an indication that the amino acids are down-regulated and that there are diacyl phosphatidylcholines (PCs) that are strongly up-regulated.

An enrichment analysis was applied to perform a systematical analysis whether there are groups of metabolites with significant changes. For this, the enrichment analysis investigates the behaviour of biochemical classes or pathways. Regarding the latter, we used a Gaussian graphical model (GGM) based on KORA blood serum samples to include a pathway feature in the metabolite set definition [13, 18]. The enrichment analysis itself was then performed as a hypergeometric test with t-test results, two distribution-based tests on the log(fold changes) (logFCs) and

weighted enrichment analysis with t-values and logFCs. The weighted enrichment analysis appeared to be the most informative, because it does not have the issue with the hard cutoff like the t-test based hypergeometric test. In addition, the results of the distribution-based tests indicated an enrichment of metabolite sets which were not interesting for our purpose, since there was no notable change in their concentration during the whole experiment.

The results of the clustering analysis were confirmed by the enrichment analysis to some extent. The amino acids showed a significant enrichment, which is due to an down regulation of the concentrations, since they are involved in pathways of energy generation and oxidative stress [25]. Depending on the applied test for the enrichment analysis, several PCs were enriched. This is most likely due to their participation of lipid droplets, which are formed during adipogenesis [24]. The GGM was not only used to establish a metabolite set definition, but also to visualize the logFCs of the metabolites for the differentiationphases. This integrated analysis can be very useful for further research, since it contains a lot of information, e.g. the fold change, whether the change is significant and the connections to other metabolites.

We also examined whether there are metabolites that undergo an exchange between the intra- and extracellular environment. There are various biological scenarios that involve exchange. For instance, an increase of concentration in the extracellular environment could be due to an flux out of metabolites. Depending on the catalysed reactions within the cell, there could be an increase or decrease of the intracellular metabolite concenctrations. This results in distinct time-courses. The results of the applied methods indicate that one method is not able to detect all metabolites and thus, several methods are needed. From the applied methods, the t-value based approach and the local CC method with a windowsize of 3 seem to be the most promising one.

Outlook

We applied a wide spectrum of methods to analyze this dataset. There are two methods or variants that are similar to the already performed ones, which could be also applied and might provide interesting results. The first one deals with the metabolite set definition for the enrichment analysis. We used a modularity clustering algorithm to detect modules in the GGM. Depending on the outcome of the algorithm, every metabolite is assigned to one metabolite set. It is also possible to perform a so-called soft or fuzzy clustering, in which the assignment of metabolites to more than one cluster is possible, because many metabolites belong to several biological pathways or processes. Therfore, it would be interesting to see, how this changed metabolite set definition alters the outcome of the enrichment analysis. The second additional method deals with the intra- and extracellular

exchange of metabolites. This method analysis the trajectory of a metabolite which would exist when plotting the intracellular concentration on one axis against the extracellular concentrations on the other axis. Since this method is neither based on the Spearman's rank correlation coefficient nor on the t-values of a t-test, the results would have been interesting to compare to other already applied methods.

It was a demanding task to interprete the results of the clustering and enrichment analysis for the extracellular measurements. This might be due to different meaning of the extracellular measurements due to the change of medium. So far, we were not able to develope a method to recalculate either the extracellular or the intracellular measurements in a way, so that their meaning is comparable. A successful approach might also have an influence on the outcome of the analysis of the intra- and extracellular exchange.

There are several machine learning methods (e.g. support vector machines, ElasticNet, Lasso/Ridge Regression) which could be used to analyse the data. The methods would help to identify metabolite that are likely to have the most influence in the whole dataset. One could also include the data of the differentiation marker to make this analysis more powerful. Overall, this would further help to determine differentiation-specific pathways, which then could be used to perform a modeling of metabolic pathways during adipocyte differentiation.

Adipogensis and the role of a adipocyte as an endocrine cell are complex mechanisms which are not yet completely understood. The results of this thesis can be used as a starting point for further investigations to increase the knowlegdge about these mechanisms. This knowledge is essential to reduce the impact of obesity and the associated diseases as major health risk factor in the Western world and developing countries.

Appendix A

Methods and Materials

Table A.1: This table lists all 188 metabolites of the dataset. Metabolites that did not pass the quality control, either for intra- or extracellular, are denoted with an asterisk (*).

Acylcarnitines
C0
C10
C10:1
C10:2
C12
C12-DC
C12:1
C14
C14:1 *
C14:1-OH
C14:2 *
C14:2-OH
C16
C16-OH
C16:1 *
C16:1-OH *
C16:2
C16:2-OH
C18
C18:1
C18:1-OH
C18:2
C2

C3
C3-DC (C4-OH)
C3-OH
C3:1
C4
C4:1
C5
C5-DC (C6-OH)
C5-M-DC
C5-OH (C3-DC-M)
C5:1
C5:1-DC
C6 (C4:1-DC)
C6:1
C7-DC *
C8
C9

Amino acids
Ala
Arg *
Asn
Asp
Cit *
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Orn
Phe
Pro
Ser
Thr
Trp
Tyr
Val

Biogenic amines
ADMA *
Ac-Orn
Carnosine *
Creatine *
DOPA *
Dopamine *
Histamine *
Kynurenine
Met-SO
Nitro-Tyr *
OH-Pro *
PEA *
Putrescine *
SDMA *
Sarcosine *
Serotonin *
Spermidine
Spermine *
Taurine *
alpha-AAA
total DMA *

Diacyl phosphatidylcholines
PC aa C24:0
PC aa C26:0
PC aa C28:1
PC aa C30:0
PC aa C30:2
PC aa C32:0
PC aa C32:1
PC aa C32:2
PC aa C32:3
PC aa C34:1
PC aa C34:2
PC aa C34:3
PC aa C34:4
PC aa C36:0
PC aa C36:1

PC aa C36:2
PC aa C36:3
PC aa C36:4
PC aa C36:5
PC aa C36:6
PC aa C38:0
PC aa C38:1
PC aa C38:3
PC aa C38:4
PC aa C38:5
PC aa C38:6
PC aa C40:1 *
PC aa C40:2
PC aa C40:3
PC aa C40:4
PC aa C40:5 *
PC aa C40:6
PC aa C42:0 *
PC aa C42:1 *
PC aa C42:2
PC aa C42:4 *
PC aa C42:5 *
PC aa C42:6 *

Acyl-alkyl phosphatidylcholines

PC ae C30:0
PC ae C30:1
PC ae C30:2
PC ae C32:1
PC ae C32:2
PC ae C34:0
PC ae C34:1
PC ae C34:2
PC ae C34:3
PC ae C36:0
PC ae C36:1
PC ae C36:2
PC ae C36:3
PC ae C36:4
PC ae C36:5

PC ae C38:0
PC ae C38:1
PC ae C38:2
PC ae C38:3
PC ae C38:4
PC ae C38:5
PC ae C38:6
PC ae C40:1
PC ae C40:2
PC ae C40:3 *
PC ae C40:4
PC ae C40:5
PC ae C40:6
PC ae C42:0
PC ae C42:1
PC ae C42:2 *
PC ae C42:3
PC ae C42:4
PC ae C42:5
PC ae C44:3
PC ae C44:4
PC ae C44:5
PC ae C44:6

Lysophosphatidylcholines
lysoPC a C14:0
lysoPC a C16:0
lysoPC a C16:1
lysoPC a C17:0
lysoPC a C18:0
lysoPC a C18:1
lysoPC a C18:2
lysoPC a C20:3
lysoPC a C20:4
lysoPC a C24:0
lysoPC a C26:0
lysoPC a C26:1
lysoPC a C28:0
lysoPC a C28:1

Sphingomyelins
SM (OH) C14:1
SM (OH) C16:1
SM (OH) C22:1 *
SM (OH) C22:2 *
SM (OH) C24:1
SM C16:0
SM C16:1
SM C18:0
SM C18:1
SM C20:2 *
SM C22:3
SM C24:0
SM C24:1
SM C26:0 *
SM C26:1 *

Sugars
H1

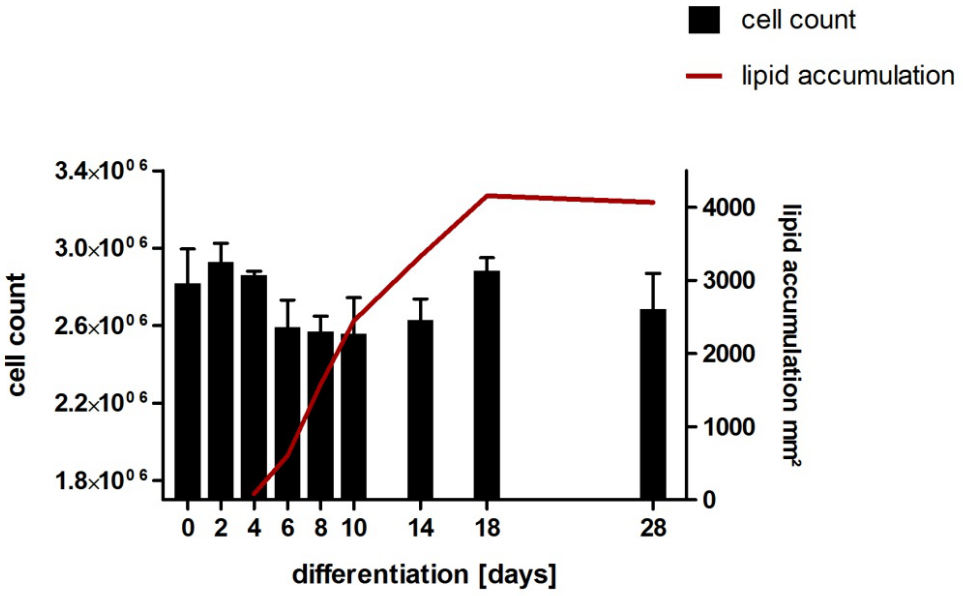

Figure A.1: The lipid accumulation and the cell count during the differentiation to adipocytes.

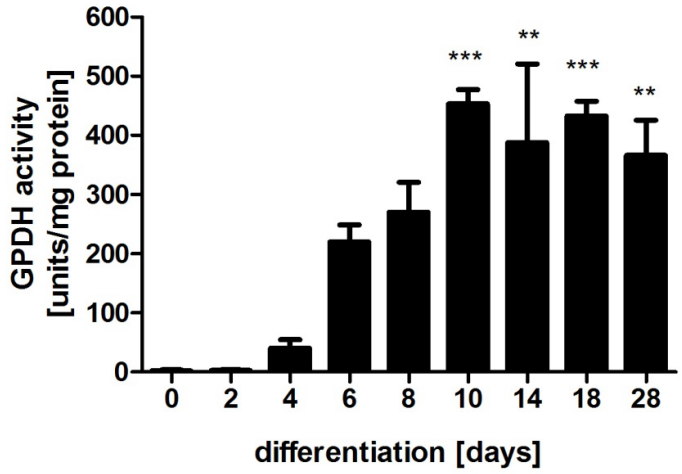

Figure A.2: The enzyme activity of Glycerol-3-phosphate dehydrogenase (GPDH) was measured during the differentiation to adipocytes.

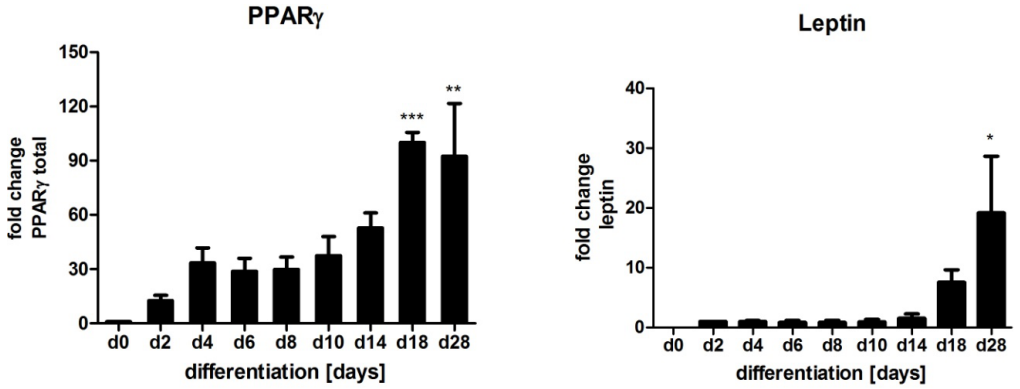

Figure A.3: The fold change in mRNA levels of peroxisome proliferator-activated receptor-γ (PPARγ) and leptin during the differentiation to adipocytes.

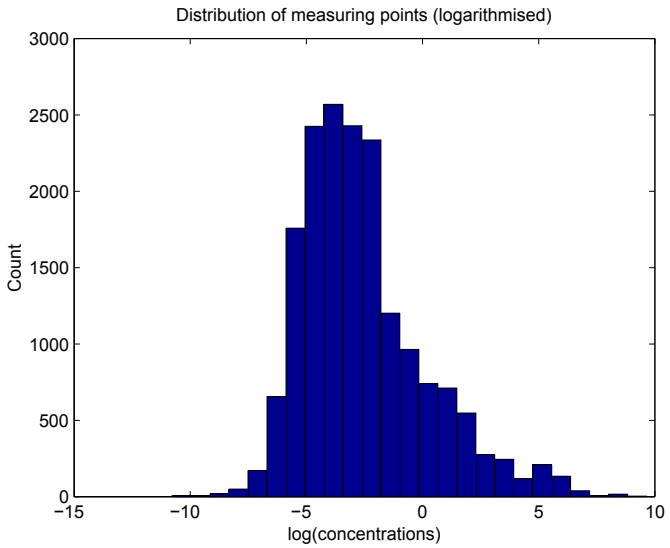

Figure A.4: This histogram shows the distribution of all measuring points of the data set. The concentrations were logarithmised.

Appendix B

Metabolomics Data Analysis and Preprocessing

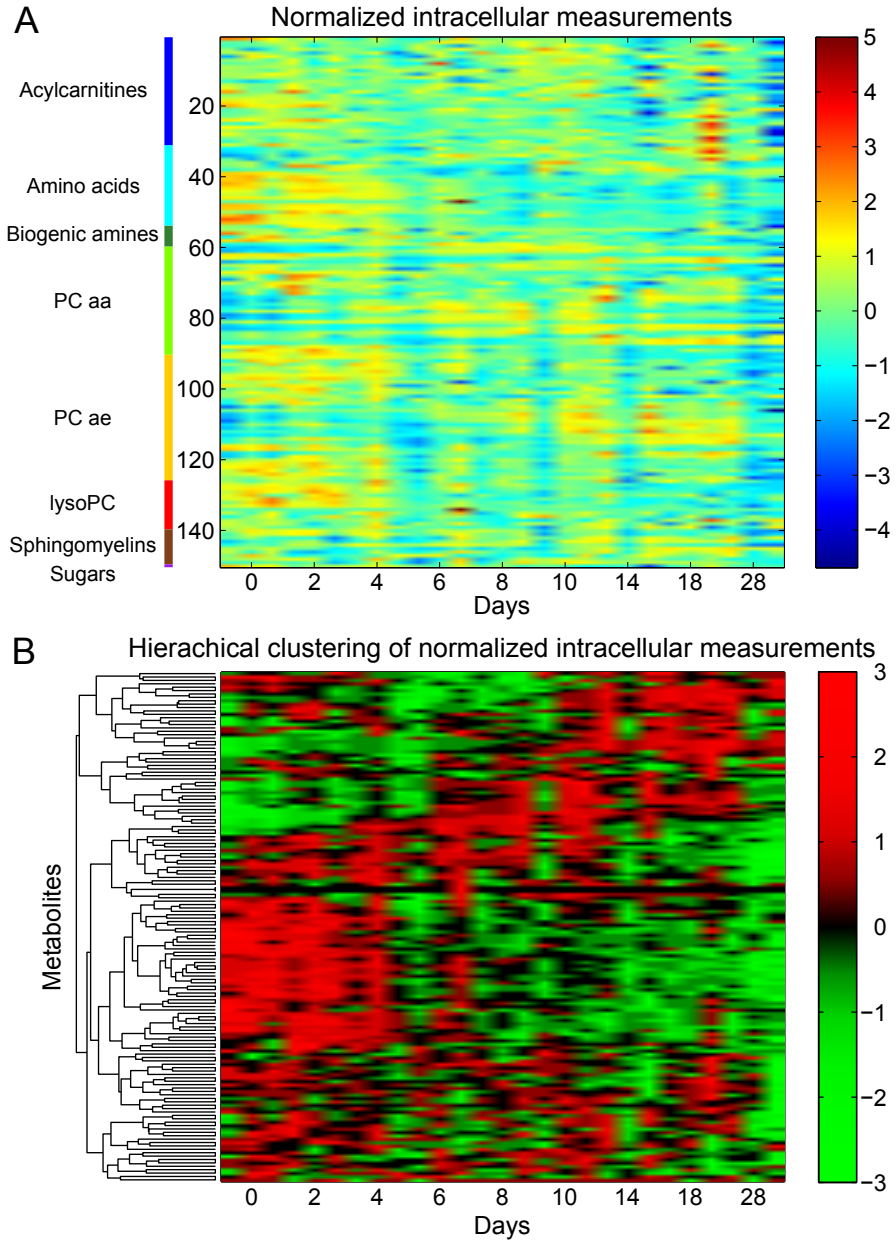

Figure B.1: Represenation of the normalized intracellular measurements. The average over the technical replicates was computed, so that there are three measuring points per day and every measuring day has three columns. (A) 152 metabolites are normalized with the z-score. The metabolites are ordered according to their biochemical classes. (B) A hierachical clustering was performed with Euclidean distance as metric and complete-linkage.

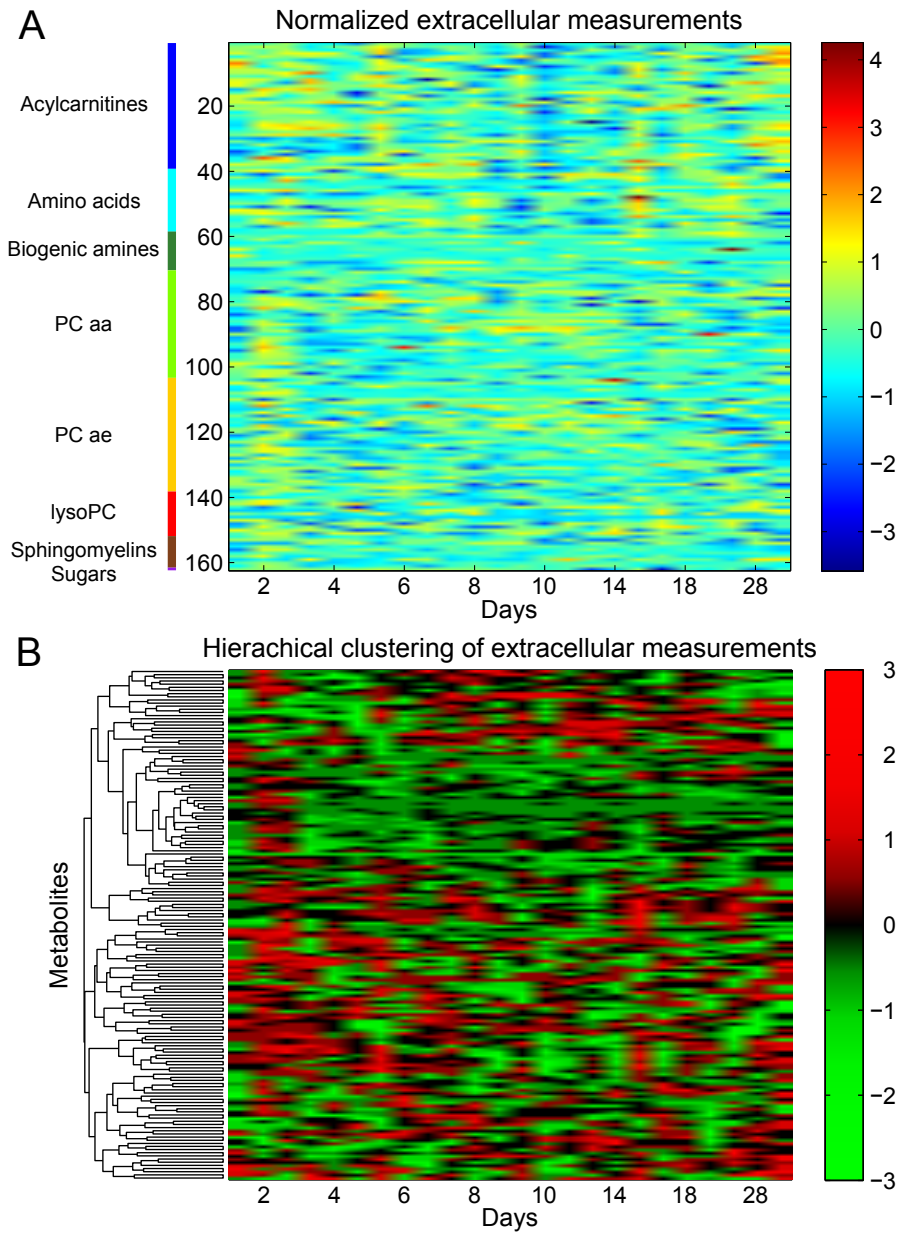

Figure B.2: Represenation of the normalized extracellular measurements. The average over the technical replicates was computed, so that there are three measuring points per day and every measuring day has three columns. (A) 167 metabolites are normalized with the z-score. The metabolites are ordered according to their biochemical classes. (B) A hierachical clustering was performed with Euclidean distance as metric and complete-linkage.

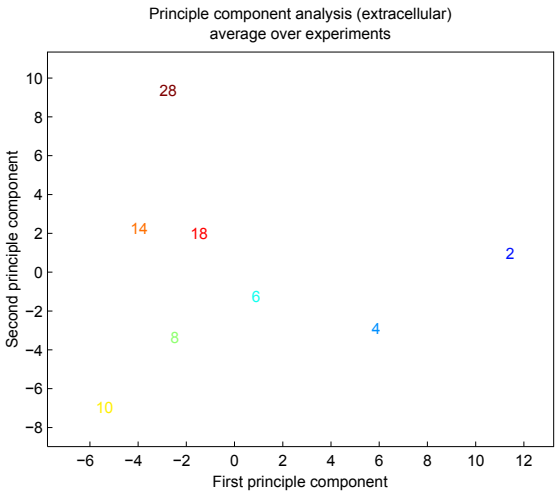

Figure B.3: A principle component analysis of the extracellular measuring days. The concrete day of the point is denoted as a number in the figure. The mean over the three experiments was calculated, so that there is only one value per measuring day.

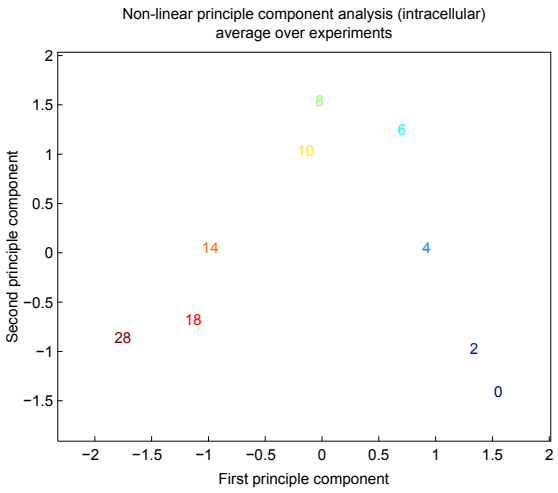

Figure B.4: A non-linear principle component analysis of the intracellular measuring days without imputed values. The PCA was computed with the GP-LVM toolbox [20]. The concrete day of the point is denoted as a number in the figure. The mean over the technical replicates and experiments was calculated, so that there is only one measuring point per day.

Appendix C

Clustering Analysis

Table C.2: Resulting clusters of the k-means clustering with extracellular logFCs and Euclidean distance as distance measure.

Cluster	Metabolites
Cluster #1	lysoPC a C28:0
Cluster #2	total DMA, PC ae C36:3, lysoPC a C17:0, SM C16:0
Cluster #3	C5, C5-OH (C3-DC-M), C5:1-DC, Ala, Gly, Lys, PC aa C24:0, PC aa C32:0, PC aa C32:1, PC aa C36:0, PC ae C34:0, PC ae C44:6, lysoPC a C16:1, lysoPC a C18:2
Cluster #4	C0, C10:1, C12, C12-DC, C14:2, C14:2-OH, C16, C16:1-OH, C16:2-OH, C18:1, C18:2, C3-DC (C4-OH), C3-OH, C5-DC (C6-OH), C5-M-DC, C5:1, C7-DC, C8, Asn, Cit, Gln, His, Ile, Leu, Met, Phe, Pro, Thr, Trp, Tyr, Val, ADMA, Ac-Orn, Creatinine, Histamine, Kynurenine, Serotonin, Spermidine, PC aa C28:1, PC aa C30:2, PC aa C32:2, PC aa C32:3, PC aa C34:4, PC aa C36:1, PC aa C36:6, PC aa C38:5, PC aa C40:4, PC aa C42:6, PC ae C34:3, PC ae C36:1, PC ae C36:2, PC ae C38:0, PC ae C38:4, PC ae C40:5, lysoPC a C16:0, lysoPC a C18:1, lysoPC a C20:3
Cluster #5	PC aa C36:2, PC aa C36:4, PC aa C36:5, PC ae C34:2

Cluster	Metabolites
Cluster #6	C10, C10:2, C12:1, C14, C14:1, C14:1-OH, C16-OH, C16:1, C16:2, C18, C18:1-OH, C2, C3, C3:1, C4, C4:1, C6 (C4:1-DC), C6:1, C9, Ser, DOPA, PC aa C26:0, PC aa C30:0, PC aa C34:1, PC aa C34:3, PC aa C38:1, PC ae C32:2, PC ae C34:1, PC ae C36:4, PC ae C38:1, PC ae C42:0, PC ae C42:1, PC ae C44:4, PC ae C44:5, lysoPC a C18:0, SM (OH) C16:1, SM C18:1
Cluster #7	Asp, Met-SO, alpha-AAA, PC aa C34:2, PC aa C38:0, PC aa C38:3, PC aa C40:1, PC aa C40:2, PC ae C30:0, PC ae C30:2, PC ae C38:2, PC ae C38:5, PC ae C40:1, PC ae C40:2, PC ae C42:3, PC ae C42:4, PC ae C42:5, lysoPC a C24:0, lysoPC a C26:0, SM C16:1, SM C18:0
Cluster #8	PC aa C36:3, PC ae C36:5, PC ae C38:3, PC ae C38:6, PC ae C40:6, lysoPC a C20:4, lysoPC a C26:1, lysoPC a C28:1, SM (OH) C14:1
Cluster #9	PC aa C38:4, PC ae C30:1, PC ae C44:3
Cluster #10	PC ae C36:0
Cluster #11	PC aa C38:6, PC aa C40:3, PC aa C40:6, PC aa C42:0, PC aa C42:2, PC ae C40:4, lysoPC a C14:0, SM (OH) C24:1
Cluster #12	Glu, Orn, PEA
Cluster #13	PC ae C32:1

Cluster	Metabolites
Cluster #1	Ac-Orn, PC aa C32:0, PC aa C32:2, PC aa C32:3
Cluster #2	C16, ADMA, Histamine, PC aa C30:0, PC aa C30:2, PC aa C34:2, PC aa C34:3, PC aa C36:2, PC aa C42:6, PC ae C30:1, PC ae C30:2, PC ae C32:1, PC ae C34:0, PC ae C34:2
Cluster #3	C10, C10:1, C10:2, C12, C12-DC, C12:1, C14, C14:1, C14:1-OH, C14:2, C14:2-OH, C16-OH, C16:1, C16:1-OH, C16:2, C16:2-OH, C18:1-OH, C18:2, C2, C3-DC (C4-OH), C3:1, C4, C4:1, C5, C5-M-DC, C5-OH (C3-DC-M), C5:1, C5:1-DC, C6 (C4:1-DC), C6:1, C7-DC, Ile, Orn, Phe, Pro, Ser, Thr, Val, Met-SO, Serotonin, Spermidine, alpha-AAA, total DMA, PC aa C24:0, PC aa C26:0, PC aa C28:1, PC aa C34:4, PC aa C36:0, PC aa C36:1, PC aa C36:4, PC aa C38:0, PC aa C38:1, PC aa C38:3, PC aa C40:1, PC aa C40:3, PC aa C42:0, PC aa C42:2, PC ae C30:0, PC ae C32:2, PC ae C34:1, PC ae C34:3, PC ae C36:0, PC ae C36:3, PC ae C36:4, PC ae C38:4, PC ae C38:5, PC ae C40:4, PC ae C42:1, PC ae C42:3, PC ae C42:4, PC ae C42:5, PC ae C44:3, PC ae C44:6, lysoPC a C14:0, lysoPC a C16:0, lysoPC a C17:0, lysoPC a C18:1, lysoPC a C18:2
Cluster #4	C0, C18:1, C3, C3-OH, C5-DC (C6-OH), Ala, Asn, Cit, Gln, Leu, Lys, Trp, Creatinine, DOPA, Kynurenine, PC aa C32:1, PC aa C34:1, PC aa C38:4, PC aa C38:6, PC aa C40:2, PC aa C40:4, PC aa C40:6, PC ae C36:1, PC ae C36:5, PC ae C38:0 PC ae C38:2, PC ae C38:3, PC ae C38:6, PC ae C40:2, PC ae C40:5, PC ae C40:6, PC ae C44:4, PC ae C44:5, lysoPC a C16:1
Cluster #5	C18, C8, C9, Asp, Gly, PEA, PC aa C36:3, PC aa C36:5, PC aa C36:6, PC ae C36:2, PC ae C38:1, PC ae C40:1, PC ae C42:0, lysoPC a C18:0, lysoPC a C20:3
Cluster #6	Glu, His, Met, Tyr, PC aa C38:5

Table C.1: Resulting clusters of the k-means clustering with intracellular logFCs and Euclidean distance as distance measure.

Cluster	Metabolites
Cluster #1	C14, Tyr, ADMA, PC aa C30:0, PC aa C30:2, PC aa C32:2, lysoPC a C18:1
Cluster #2	C10:2, C5:1, Val, PC aa C28:1, PC aa C32:1, PC aa C34:2, PC aa C42:2, PC ae C30:0, PC ae C30:1, PC ae C30:2, PC ae C32:2, PC ae C34:0, PC ae C44:5, PC ae C44:6
Cluster #3	C16, C3:1, C5-OH (C3-DC-M), C7-DC, DOPA, PC aa C26:0, PC aa C34:1, PC aa C36:0, PC aa C36:1, PC ae C34:1, PC ae C34:3, PC ae C36:2
Cluster #4	C4:1, C6:1, Serotonin, total DMA, PC aa C24:0, PC aa C42:0, PC ae C32:1, PC ae C34:2, PC ae C36:3
Cluster #5	C10:1, C14:1, C18:1-OH, C18:2, C2, C3-DC (C4-OH), C5, C5-M-DC, C5:1-DC, Pro
Cluster #6	C12, C12-DC, C12:1, C16:2, C6 (C4:1-DC), Phe
Cluster #7	C8, Creatinine, PC aa C34:4, PC aa C40:4, PC ae C42:3, lysoPC a C16:0
Cluster #8	PC ae C42:0, PC ae C42:4, PC ae C42:5, lysoPC a C16:1, lysoPC a C17:0
Cluster #9	C0, C16:1-OH, C18, C18:1, C3, C9, Ala, Asn, Asp, Gln, Glu, Gly, Leu, Lys, Met, Ser, Trp, Ac-Orn, Spermidine, alpha-AAA, PC aa C36:3, PC aa C36:6, PC aa C38:5, PC aa C38:6, PC aa C40:1, PC ae C38:1, PC ae C38:2, PC ae C40:1, PC ae C40:2, PC ae C40:4, PC ae C40:5
Cluster #10	C14:2, C16-OH, Histamine, PC aa C32:0, PC aa C32:3, PC aa C34:3, PC aa C38:1, PC aa C40:2, PC aa C40:3, PC aa C40:6, PC ae C42:1, lysoPC a C14:0, lysoPC a C18:0
Cluster #11	C10, C14:1-OH, C14:2-OH, C16:1, C16:2-OH, C3-OH, C4, C5-DC (C6-OH), Kynurenine, PC aa C42:6
Cluster #12	Cit, His, Ile, Orn, Thr, Met-SO, PEA, PC aa C36:2, PC aa C36:4, PC aa C36:5, PC aa C38:0, PC aa C38:3, PC aa C38:4, PC ae C36:0, PC ae C36:1, PC ae C36:4, PC ae C36:5, PC ae C38:0, PC ae C38:3, PC ae C38:4, PC ae C38:5, PC ae C38:6, PC ae C40:6, PC ae C44:3, PC ae C44:4, lysoPC a C18:2

Table C.3: Resulting clusters of the k-means clustering with intracellular logFCs and Pearson product-moment correlation coefficient as distance measure.

Cluster	Metabolites
Cluster #1	C5, C5-OH (C3-DC-M), C5:1, C5:1-DC, Ala, Gly, His, Lys, Val, PC aa C24:0, PC aa C32:1, PC aa C32:2, PC aa C36:0, PC aa C36:1, PC ae C30:0, PC ae C34:1, PC ae C36:4, lysoPC a C14:0, lysoPC a C17:0, lysoPC a C18:0, lysoPC a C20:3, SM (OH) C14:1, SM C18:0
Cluster #2	C10:1, C12-DC, C16:2, C5-DC (C6-OH), C5-M-DC, Cit, Gln, Phe, Trp, Tyr, Creatinine, DOPA, Kynurenine, alpha-AAA, PC aa C30:0, PC aa C32:3, PC aa C38:0, PC aa C38:6, PC aa C40:6, PC ae C32:1, PC ae C36:0, PC ae C38:1, PC ae C38:5, PC ae C42:3, lysoPC a C18:2, lysoPC a C28:0
Cluster #3	C10, C14, C14:1, C14:2, C16-OH, C16:2-OH, C18:2, C3, C4, C6:1, C8, C9, Ile, Leu, Met, PC aa C26:0, PC aa C34:2, PC ae C30:2, PC ae C40:6, PC ae C42:4, PC ae C44:4, PC ae C44:6, lysoPC a C24:0
Cluster #4	C16, C4:1, Orn, Thr, ADMA, PC aa C28:1, PC aa C30:2, PC aa C34:4, PC ae C34:2, PC ae C36:5, PC ae C44:5, SM C24:0
Cluster #5	C12:1, C16:1-OH, Asn, Asp, Glu, Ac-Orn, Met-SO, total DMA, PC aa C38:1, PC aa C38:4, PC aa C40:1, PC aa C40:2, PC aa C42:0, PC aa C42:2, PC ae C30:1, PC ae C32:2, PC ae C34:0, PC ae C36:2, PC ae C38:0, PC ae C38:3, PC ae C38:6, PC ae C40:2, PC ae C40:4, PC ae C42:5, lysoPC a C16:0, lysoPC a C26:1, SM C16:1, SM C18:1, SM C22:3
Cluster #6	C10:2, C12, C14:1-OH, C14:2-OH, C16:1, C18, C2, C3-DC (C4-OH), C3-OH, C3:1, Ser, PC aa C32:0, PC aa C34:3, PC aa C40:3, PC aa C40:4, PC aa C42:6, PC ae C36:1, PC ae C38:2, PC ae C40:5, PC ae C44:3, lysoPC a C16:1, lysoPC a C28:1, SM C16:0
Cluster #7	C0, C18:1, C18:1-OH, C6 (C4:1-DC), C7-DC, Pro, Spermidine, PC aa C34:1, PC aa C36:2, PC aa C36:3, PC aa C36:4, PC aa C36:5, PC aa C36:6, PC aa C38:3, PC aa C38:5, PC ae C34:3, PC ae C36:3, PC ae C38:4, PC ae C40:1, PC ae C42:0, PC ae C42:1, lysoPC a C18:1, lysoPC a C26:0, SM (OH) C16:1, SM (OH) C24:1

Table C.4: Resulting clusters of the k-means clustering with extracellular logFCs and Pearson product-moment correlation coefficient as distance measure.

Appendix D

Enrichment Analysis

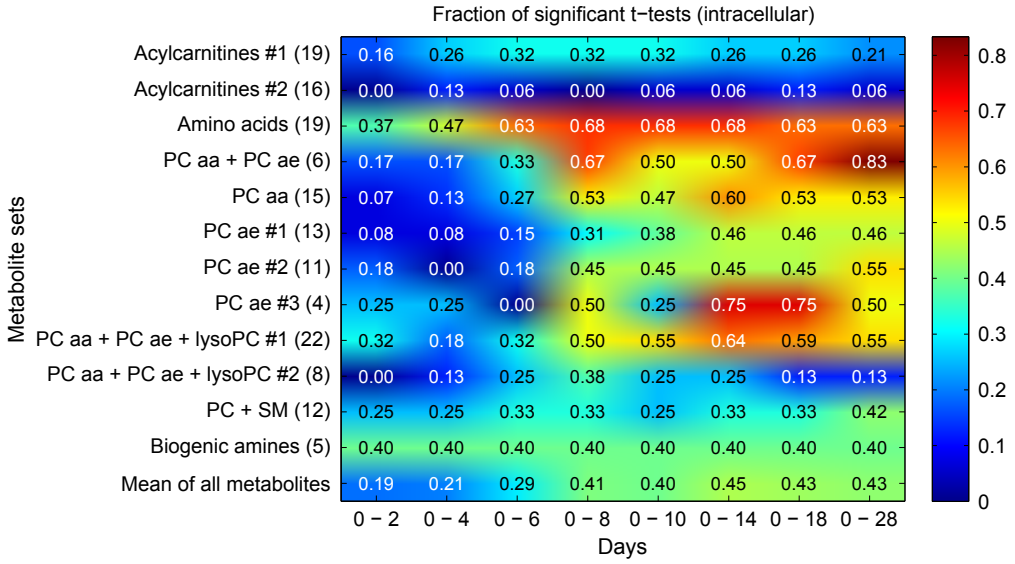

Figure D.1: A t-test for every metabolite was performed between a pair of measuring days which is declared on the x-axis. The fraction of significant metabolites for each metabolite set is shown. The metabolite sets are based on the GGM. Additionally, the fraction of all metabolites is shown (last row).

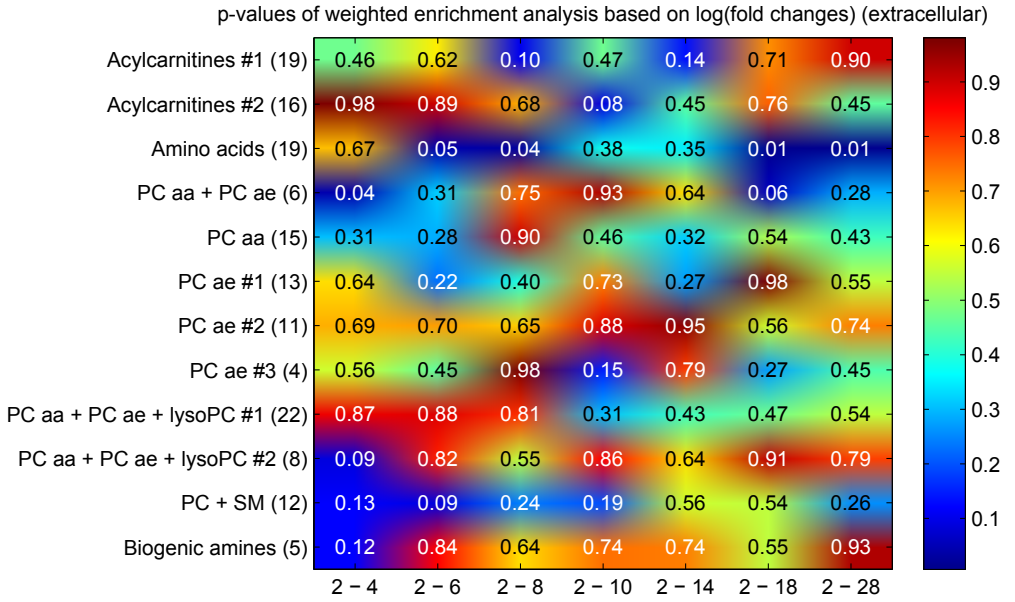

Figure D.2: The p-values of the hypergeometric test based on the extracellular log(fold changes) and GGM-based metabolite set definition (y-axis). A t-test was applied for every metabolite between the measuring points of the two days that are denoted on the x-axis. The outcome of the t-test was then used for the hypergeometric test.

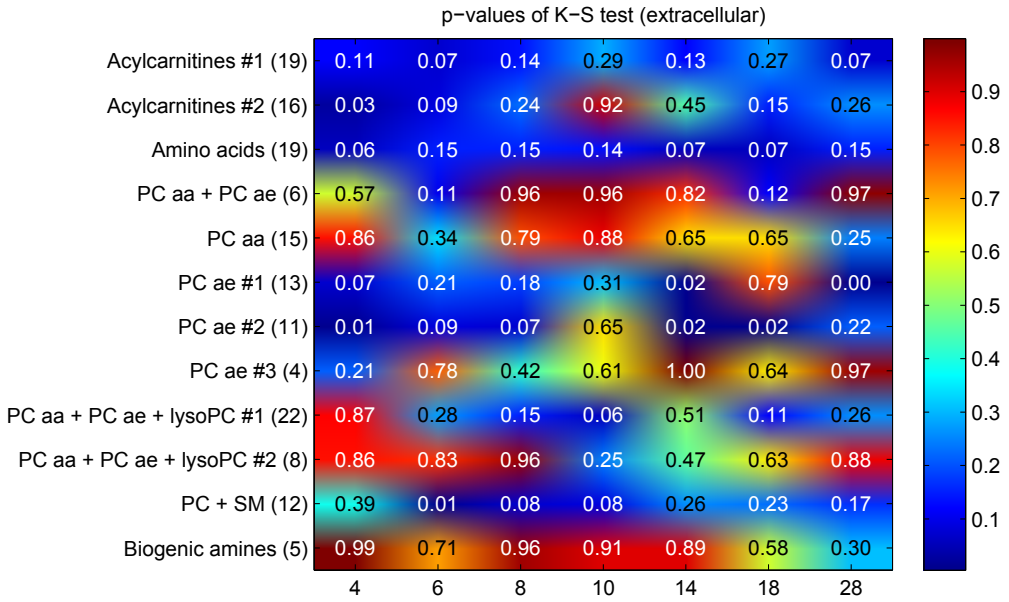

Figure D.3: The p-values of the Kolmogornov-Smirnov test based on the extracellular log(fold changes) and GGM-based metabolite set definition (y-axis).

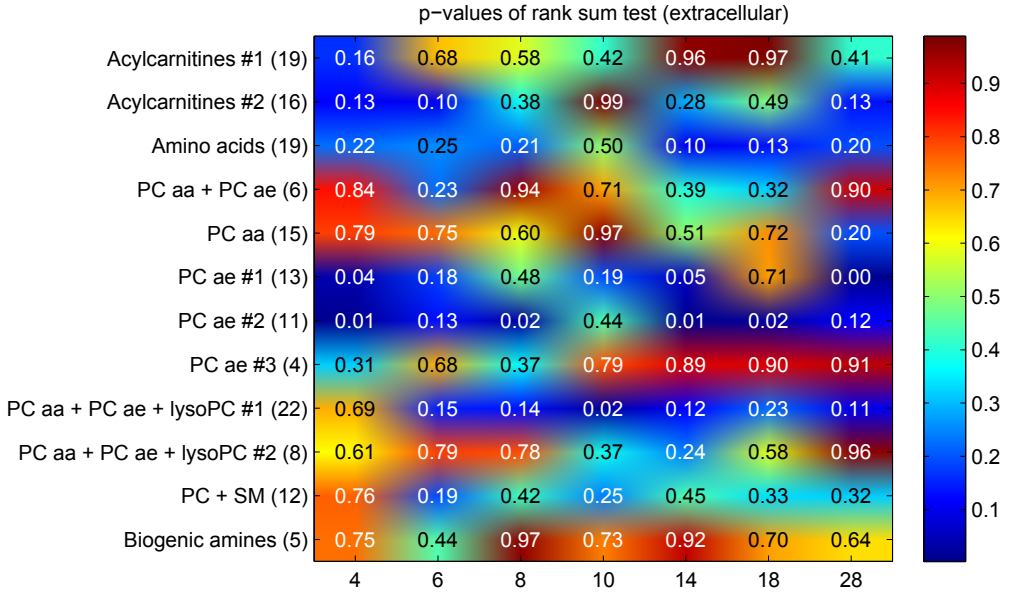

Figure D.4: The p-values of the wilcoxon rank sum test based on the extracellular log(fold changes) and GGM-based metabolite set definition (y-axis).

Appendix E

Analysis of the Intra- and Extracellular Metabolite Dependency

$ws = 2$	$ws = 3$	$ws = 4$	$ws = 5$
C10:1	C12:1	C12:1	C12:1
C12-DC	C14	C14	C2
C14:1-OH	C16:2-OH	C14:1-OH	C3-DC (C4-OH)
C18:1-OH	C18:2	C18:1-OH	C3-OH
C18:2	C4:1	C3-OH	C4:1
C4:1	C5:1	C4:1	C5:1
C5:1	C6 (C4:1-DC)	C5:1	C6 (C4:1-DC)
Ala	Ala	C6 (C4:1-DC)	Ala
Gln	Gln	Gln	Gln
Gly	Glu	Glu	Glu
Ile	Lys	Lys	Lys
Leu	Phe	Pro	Pro
Lys	Pro	Ac-Orn	Ac-Orn
Phe	Ac-Orn	Met-SO	PC aa C26:0
Pro	Met-SO	PC aa C26:0	PC aa C28:1
Tyr	PC aa C26:0	PC aa C28:1	PC aa C32:0
Ac-Orn	PC aa C28:1	PC aa C32:2	PC aa C32:2
PC aa C24:0	PC aa C30:0	PC aa C34:1	PC aa C34:2
PC aa C32:2	PC aa C32:2	PC aa C34:2	PC aa C38:0
PC aa C36:1	PC aa C36:2	PC aa C34:4	PC aa C38:1
PC aa C40:2	PC aa C38:0	PC aa C38:0	PC aa C38:4
PC ae C34:2	PC aa C38:1	PC aa C38:1	PC aa C40:3
PC ae C34:3	PC aa C40:2	PC aa C38:4	PC ae C34:1
PC ae C36:3	PC aa C40:3	PC aa C40:3	PC ae C34:2
PC ae C36:4	PC aa C40:6	PC ae C34:1	PC ae C34:3
PC ae C38:3	PC ae C34:1	PC ae C34:2	PC ae C36:3
PC ae C40:6	PC ae C34:2	PC ae C34:3	PC ae C38:6
PC ae C44:6	PC ae C34:3	PC ae C36:3	PC ae C40:2
lysoPC a C17:0	PC ae C36:3	PC ae C38:6	lysoPC a C17:0
lysoPC a C18:1	PC ae C38:3	PC ae C40:2	lysoPC a C24:0
SM (OH) C16:1	PC ae C38:6	SM (OH) C16:1	SM (OH) C16:1
SM C22:3	PC ae C40:1	SM C16:1	SM C16:1
	PC ae C40:6		H1
	PC ae C44:6		
	lysoPC a C20:3		
	lysoPC a C28:0		
	SM (OH) C16:1		

Table E.1: This table lists the results of the windows-based correlation method to evaluate the exchange between the intra- and extracellular environment. All listed metabolites have atleast on Spearmans's rank correlation coefficient that is significantly different from 0 according to their p-value. The p-values were not corrected for multiple testing. The windowsize (ws) ranges from 2 to 5.

Bibliography

[1] Abdi, H. and Williams, L.J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. ISSN 19395108.

[2] Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. 1995.

[3] Bland, J.M. and Altman, D.G. Statistics Notes Multiple significance tests : the Bonferroni method. *British Medical Journey*, 310(January):1995, 1995.

[4] Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. ISSN 1742-5468.

[5] Blow, N. Biochemistry's new look. *Nature*, 455(October), 2008.

[6] Chace, D.H., Sherwin, J.E., Hillman, S.L., Lorey, F., and Cunningham, G.C. Use of phenylalanine-to-tyrosine ratio determined by tandem mass spectrometry to improve newborn screening for phenylketonuria of early discharge specimens collected in the first 24 hours. *Clinical Chemistry*, 2409:2405–2409, 1998.

[7] Cinti, S. The adipose organ. *Prostaglandins, Leukotrienes and Essential Fatty Acids*, 73(1):9–15, 2005.

[8] Cristancho, A.G. and Lazar, M.a. Forming functional fat: a growing understanding of adipocyte differentiation. *Nature reviews. Molecular cell biology*, 12(11):722–34, 2011. ISSN 1471-0080.

[9] Fischer-Posovszky, P., Newell, F.S., Wabitsch, M., and Tornqvist, H.E. Human SGBS cells - a unique tool for studies of human fat cell biology. *Obesity facts*, 1(4):184–9, 2008. ISSN 1662-4025.

[10] Galic, S., Oakhill, J.S., and Steinberg, G.R. Adipose tissue as an endocrine organ. *Molecular and Cellular Enocrinology*, 316(2):129–139, 2010.

[11] Gregoire, F.M. Adipocyte Differentiation : From Fibroblast to. *Experimental Biology and Medicine*, 226:997–1002, 2001.

[12] Hashimoto, T., Igarashi, J., and Kosaka, H. Sphingosine kinase is induced in mouse 3T3-L1 cells and promotes adipogenesis. *Journal of lipid research*, 50(4):602–10, 2009. ISSN 0022-2275.

[13] Holle, R., Happich, M., Löwel, H., and Wichmann, H.E. KORA–a research platform for population based health research. *Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany))*, 67 Suppl 1:S19–25, 2005. ISSN 0941-3790.

[14] Hossain, P., Kawar, B., and El Nahas, M. Obesity and Diabetes in the Developing World âĂŤ A Growing Challenge. *The New England Journal of Medicine*, 356:213–215, 2007.

[15] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B.S., Mewes, H.W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D., Adamski, J., and Suhre, K. A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*, 42(2):137–41, 2010. ISSN 1546-1718.

[16] Kim, S. and Moustaid-moussa, N. Symposium : Adipocyte Function , Differentiation and Metabolism Secretory , Endocrine and Autocrine / Paracrine Function of the Adipocyte 1. *The Journal of Nutrition*, 130:3110–3115, 2000.

[17] Krug, S., Kastenmüller, G., Stückler, F., Rist, M.J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C., Frank, T., Engel, K.H., Hofmann, T., Luy, B., Zimmermann, R., Moritz, F., Schmitt-Kopplin, P., Krumsiek, J., Kremer, W., Huber, F., Oeh, U., Theis, F.J., Szymczak, W., Hauner, H., Suhre, K., and Daniel, H. The dynamic range of the human metabolome revealed by challenges. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 26(6):2607–19, 2012. ISSN 1530-6860.

[18] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011. ISSN 1752-0509.

[19] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *Journal of proteome research*, 11(8):4120–31, 2012. ISSN 1535-3907.

[20] Lawrence, N. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[21] Morrison, R.F. and Farmer, S.R. Symposium : Adipocyte Function , Differentiation and Metabolism Hormonal Signaling and Transcriptional Control of Adipocyte. *The Journal of Nutrition*, 130:3116–3121, 2000.

[22] Newman, M.E.J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–82, 2006. ISSN 0027-8424.

[23] Oliveros, J. VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html. 2007.

[24] Penno, A., Hackenbroich, G., and Thiele, C. Phospholipids and Lipid Droplets. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, pages 1–6, 2012. ISSN 13881981. doi:10.1016/j.bbalip.2012.12.001.

[25] Portius, D. *Metabolic profiling during adipocyte differentiation and analysis of metabolic pathways in hepatoma cells.* Diploma thesis, Martin-Luther-Universität Halle-Wittenberg, 2012.

[26] Rivals, I., Personnaz, L., Taing, L., and Potier, M.C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics (Oxford, England)*, 23(4):401–7, 2007. ISSN 1367-4811.

[27] Roberts, L.D., Virtue, S., Vidal-Puig, A., Nicholls, A.W., and Griffin, J.L. Metabolic phenotyping of a model of adipocyte differentiation. *Physiological genomics*, 39(2):109–19, 2009. ISSN 1531-2267.

[28] Ross, S.E. Inhibition of Adipogenesis by Wnt Signaling. *Science*, 289(5481):950–953, 2000. ISSN 00368075. doi:10.1126/science.289.5481.950.

[29] Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 03770427.

[30] Rubinov, M. and Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–69, 2010. ISSN 1095-9572.

[31] Schäfer, J. and Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)*, 21(6):754–64, 2005. ISSN 1367-4803.

[32] Shaham, O., Wei, R., Wang, T.J., Ricciardi, C., Lewis, G.D., Vasan, R.S., Carr, S.a., Thadhani, R., Gerszten, R.E., and Mootha, V.K. Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Molecular systems biology*, 4(214):214, 2008. ISSN 1744-4292.

[33] Spalding, K.L., Arner, E., Westermark, P.l.O., Bernard, S., Buchholz, B.a., Bergmann, O., Blomqvist, L., Hoffstedt, J., Näslund, E., Britton, T., Concha, H., Hassan, M., Rydén, M., Frisén, J., and Arner, P. Dynamics of fat cell turnover in humans. *Nature*, 453(7196):783–7, 2008. ISSN 1476-4687.

[34] Spector, A.A. and Yorek, M.A. Membrane lipid composition and cellular function. *Journal of lipid research*, 26:1015–1035, 1985.

[35] Theis, F.J., Latif, N., Wong, P., and Frishman, D. Complex principal component and correlation structure of 16 yeast genomic variables. *Molecular biology and evolution*, 28(9):2501–12, 2011. ISSN 1537-1719.

[36] Wabitsch, M., RE, B., Melzner, I., Braun, M., Möller, P., Heinze, E., Debatin, K., and Hauner, H. Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *International Journal of Obesity and Related Metabolic Disorders*, 25:8–15, 2001.

[37] Wilk, M.B. and Gnanadesikan, R. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968. ISSN 0006-3444.

[38] Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J.a., Lim, E., Sobsey, C.a., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., and Forsythe, I. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, 37(Database issue):D603–10, 2009. ISSN 1362-4962.

[39] Wu, Y. Overweight and obesity in China. *BMJ (Clinical research ed.)*, 333(7564):362–3, 2006. ISSN 1756-1833.

[40] Xuan, J.Y., Hughes-benzie, R.M., Mackenzie, A.E., and Alberta, S. A small interstitial deletion in the GPC3 gene causes Simpson-Golabi-Behmel syndrome in a Dutch-Canadian family. *Med Genet*, 36:57–58, 1999.

[41] Young, T. Proof Without for the Prejudice: Use of the Analysis of Histograms and Other Kolmogorov-Smirnov from Flow Systems and Other Sources. *The Journal of Histochemistry and Cytochemistry*, 1977.