Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Jessica Legradi
Oral-examination: 21.12.2011

Microarray based transcriptomics and the search for biomarker genes in zebrafish

Referees: Prof. Dr. Uwe Strähle
Prof. Dr. Thomas Braunbeck

Abstract

In the past, zebrafish genes were mapped to human or mouse orthologs in order to perform Gene Ontology or pathway analyses. Therefore, genes without orthologs were removed and zebrafish-specific pathways were not taken into account. After the zebrafish genome has been sequenced almost completely, a growing number of biological databases for zebrafish have been made available. The increasing availability of gene function descriptions and specific pathways improves the applicability of zebrafish for transcriptomics studies. To make full use of the enhanced capabilities, however, new methods need to be developed.

In this thesis, I describe results of two different transcriptional studies. In the first one, I analyzed gene expression data of zebrafish embryos treated with 10 different compounds at 24-48 hpf. I employed multivariate statistical methods to identify compounds that lead to similar expression pattern changes. Furthermore, I tried to identify similarities by comparing co-regulated genes. A gene function analysis of the significantly differentially expressed genes was performed in order to gain a better understanding of the modes of action of the compounds. The findings were validated using literature data. In order to identify biomarker genes, I grouped the compounds based on the identified modes of action and searched for genes that were only de-regulated after treatment with compounds with the same mode of action. I defined sets of biomarker genes for the following modes of action: disruption of mitochondrial potential, Acetylcholinesterase inhibition, Glutathione metabolism, and induction of apoptosis.

During the studies of the 10 compounds, it became obvious that commercially available zebrafish microarrays lack several important genes. To overcome this problem, I designed a new array that covers almost the whole zebrafish genome. I could show that the newly designed whole genome array clearly improves microarray experiments.

Additionally, we aimed at gaining deeper insights into the transcriptional regulation during zebrafish development. For this reason, I designed a new microarray consisting only of transcription factors. This array was employed to study six different developmental stages, covering the complete development from egg till larva. We were also interested in variations of transcription factor expression in certain tissues like muscle and brain. The microarray data was analyzed with a newly developed approach using two color arrays to detect expressed transcription factors. Using the new method, I could detect groups of transcription factors that exhibited a similar expression pattern over time. With the help

of Gene Ontology, I was able to identify different gene function mechanisms associated with specific developmental stages. Transcription factors with highest expression before gastrulation were mostly involved in protein metabolism, and factors expressed at similar levels during the whole development period were likely to be involved in organ development. Transcription factors with expression peaking at the end of the development seemed to be mostly involved in development of the nervous system and biosynthesis. Additionally, I defined biomarker genes specific for the 6 developmental stages and the tissue samples used in this study.

Zusammenfassung

Um Analysen der Annotation mit Genfunktionen oder Stoffwechselwegen durchzuführen, wurden Zebrafischgene in der Vergangenheit mit Orthologen im Menschen oder der Maus ersetzt. Gene bei denen das nicht möglich war, gingen in diesem Prozess verloren. Außerdem, wurden Stoffwechselwege, die nur im Zebrafisch vorkommen, ebenfalls nicht berücksichtigt. Mittlerweile, ist das Zebrafischgenom fast vollständig sequenziert. Darüber hinaus, stehen auch auch immer mehr biologische Datenbanken auch für Zebrafisch zur Verfügung. Diese steigende Verfügbarkeit von Annotationen mit Genfunktion und speziellen Stoffwechselwegen verbessert die Anwendbarkeit von Zebrafisch für transkriptomische Untersuchungen. Um die neu gewonnen Möglichkeiten möglichst gut auszuschöpfen, müssen allerdings auch neue Analysemethoden entwickelt werden.

In meiner Arbeit habe ich zwei verschiedene transcriptomische Analysen durchgeführt. In der ersten, wurden Zebrafischembryonen (24-48 hpf) mit einer von zehn Chemikalien behandelt und danach die Genexpressions analysiert. Mithilfe multivariater statistischer Verfahren, habe ich untersucht, welche Chemikalien ähnliche Expressionsmustern hervorrufen. Des Weiteren, habe ich versucht die Ähnlichkeiten zwischen Chemikalien mittels Genen zu definieren, deren Expression gleich reguliert wurde. Um toxikologische Mechanismen, die durch die verschiedenen Substanzen induziert wurden, zu identifizieren, wurde eine Funktionsanalyse der differientiel expremierten Gene durchgeführt und die Ergebnisse mit Literaturdaten verglichen. Danach, habe ich die Chemikalien aufgrund ihrer identifizierten toxischen Mechanismen gruppiert um so die Entwicklung neuer Biomarker zu ermöglichen. Auf Basis der Gene, deren Expression nur durch Substanzen mit dem gleichen toxischen Mechanismus dereguliert wurde, konnte ich Biomarker für verschiedene die Mechanismen definieren: Störung des Mitochondrialmembranpotentials, Acetylcholinesterase Hemmung, Glutathione Metabolismus und Induktion der Apoptose.

Während dieser Analyse wurde deutlich, dass viele interessant Gene nicht mithilfe kommerziell erhältlicher Zebrafischmicroarrays gemessen werden können. Um dieses Problem zu lösen, habe ich ein neues Array entwickelt, welches fast das ganze Zebrafisch Genom abdeckt. Ich konnte zeigen, dass dieses Array die Ergebnisse von durchgeführten Experimente deutlich verbesserte.

Des Weiteren wollte ich einen tieferen Einblick in die transkriptionelle Regulation während der verschiedenen Entwicklungsphasen des Zebrafisches bekommen. Deswegen habe ich auch ein Transktiptionsfaktorarray entworfen. Mit diesem Arrays wurden

sechs verschiedene Entwicklungsstadien, vom Ei bis zur Larve, untersucht. Wir waren auch an den Unterschieden zwischen den Geweben Hirn und Muskel interessiert. Die Microarrays wurden mit einer neu entwickelten Methode analysiert, die 2-Farbarrays verwendet, um exprimierte Transktipionsfaktoren zu ermitteln. Dadurch konnte ich Gruppen von Transktiptionsfaktoren ermitteln, die ein ähnliches Expressionsmuster über die verschiedenen Entwicklungsphasen zeigten. Durch Gene Ontology-Analysen wurden Mechanismen deutlich, die spezifisch für einzelne Entwicklungsstadien sind. Transcriptionsfaktoren, die vor Beginn der Gastrulation am stärksten exprimiert waren, waren meistens im Proteinmetabolismus involviert. Transktiptionsfaktoren, deren Expression sich in den verschiedenen Entwicklungsphasen nicht stark änderte, waren meistens an der Organentwicklung beteiligt. Die Transktiptionsfaktoren, die eher am Ende der Entwicklungsphase exprimiert waren, wiesen meist eine Beteiligung an der Entwicklung des Nervensystems und der Biosynthese auf. Zusätzlich habe ich noch Biomarker speziell für die sechs verwendeten Entwicklungsstadien und die Gewebearten definiert.

Acknowledgements

First, I would like to thank my supervisor Uwe Strähle for giving me the opportunity to do this PhD and for his support and advice during my PhD studies. Furthermore, I want to thank Urban Liebel for his guidance and the offered possibilities.

I have been extremely fortunate to have learned from and benefited from collaborations with many wonderful and brilliant people. I would like to acknowledge in particular Jens and Olivier for the joy of working. I would also like to thank the ITG secretaries for their help with all the bureaucracy. Thanks also to the people from building 341 for the great time there. Furthermore, I want to acknowledge my collaboration partners at the UFZ in Leipzig, the ECT in Frankfurt and at the IAI in Karlsruhe. Unfortunately, there is not enough space to mention all of them personally. My special thanks go to Ralf Mikut for introducing me to Gait-CAD. I also want to say thank you to all the people from the ITG for the time we spend together. In particular, I want to thank Anita, Sebastian, Nadine and Babs for their friendship during my time there.

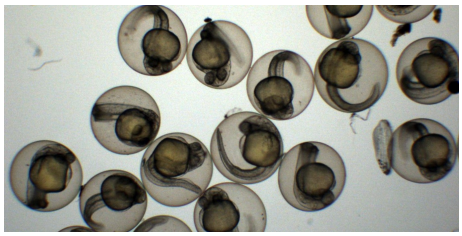
I also have to thank all my friends for being part of my life. Especially I have to thank my boyfriend for all his support and motivation. Finally, I would like to thank my family for their endless support and trust.

Table of Contents

Chapter 1

Introduction

1.1 Zebrafish as Model Organism



(a) zebrafish embryos

(b) adult zebrafish
(source: www.en.wikipedia.org)

Figure 1.1: Images of a zebrafish embryos and an adult zebrafish.

In the recent years, the zebrafish has become one of the most important vertebrate model organisms. It is used in developmental biology, disease modeling, chemical toxicology, regulatory physiology, behavioral studies, and many more disciplines. Zebrafish have distinct advantages compared to other model organisms such as mice and rats. They are inexpensive to maintain and easy to bread, especially compared to mammals. As all oviparious species, they fertilize and develop outside of the mother animal. Together with their transparency, this makes them an ideal organism for studying embryo development. Furthermore, a single female fish can lay up to 300 eggs every week in one clutch (Hill *et al.* July 2005). The high number of eggs and the small size of the embryos makes the zebrafish an ideal organism to perform any kind of high-throughput screen (Spitsbergen and Kent 2003). Since the late 1960's when the first zebrafish entered the lab, a large variety of different molecular biological methods have been established (Grunwald and Eisen 2002). Transient gene expression, in situ hybridization, and morpholino gene knockdowns are only a few examples (Hill *et al.* July 2005). The genome has been

almost completely sequenced and several thousand mutants and transgenic lines are available. Cell culture methods were developed to create cell lines from from adult tissues as well as from embryos (Spitsbergen and Kent 2003).

1.2 Transcriptomics and Biomarker Genes

The transcriptome is the total set of RNAs in an organism. The messenger RNA (mRNA) reflects the genes that are expressed at a specific time. Developmental or external environmental conditions can influence the level of expression. Transcriptomics is the genome-wide measurement of mRNA expression levels. Microarrays are one of the most prominent methods to study the transcriptome. Besides that, next-generation sequencing became quite popular in the recent years. Transcriptomics helps to understand molecular mechanisms, gene networks, and signaling pathways. Comparative transcriptomics, compares the expression levels of genes between different developmental stages, tissues, treatments, and species. Special attention is paid to investigations of transcription patterns during embryonic development and to the impact of environmental or nutritional factors on the transcriptome.

Transcriptomics can also help to identify biomarker genes. In general, biomarkers can be genes, proteins, or enzymes. Biomarker genes are genes whose changes in expression is associated with a specific biological effect. For example, a disease biomarker gene is used as an indicator of a disease or to predict the clinical outcome. A toxicity biomarker gene monitors a specific toxicological effect of a compound (Jain 2010).

Transcriptomics is also an often applied technique in zebrafish research. Many transcriptomics studies have been published, mainly in investigating chemical toxicity. Fan *et al.* 2010 studied the gene expression changes in developing zebrafish in order to find biomarker genes specific for developmental neurotoxicity. Alexeyenko *et al.* 2010 studied the gene expression changes in zebrafish embryos exposed to dioxin. The authors generated a dynamic gene expression network (interactome) based on orthologs and interaction data from other species.

Toxicogenomics is a sub-field of transcriptomics that deals with the interpretation of gene and protein activity in an organism in response to toxic substances. In the last years, zebrafish became a very prominent model organism in this field. Especially the embryos are often used to study teratogenic effects of xenobiotics. It was shown that the gene expression pattern of treated zebrafish embryos significantly changes, already at concentration far bellow any visible effect concentration (Voelker *et al.* 2007). The changes in the expression pattern are highly specific (barcode-like) for the used treatment (Yang *et al.* 2007). Gene expression profiling, for example with DNA microarrays, can help to characterize toxicological mechanism. Furthermore, modes of action of uncharacterized compounds can be identified (Neumann and Galvez 2002). Additionally, biomarker genes can be defined to predict the effects of a toxicant.

1.3 Overview

This work focuses on the development of new microarray based transcriptomics approaches and the detection of new biomarker genes in zebrafish. In total, I performed two different transcriptional analyses. First, I analyzed the modes of action of ten different compounds (Figure 1.1). For most of these compounds, no information regarding their modes of action in zebrafish or any other fish species were known. To identify the modes of action, I established a new analysis method based on gene function analysis. Additionally, I determined biomarker genes specific for the detected modes of action.

During the toxciogenomics analysis, I realized that a certain amount of interesting genes were missing on commercially available microarrays. Therefore, I decided to design my own whole genome zebrafish microarray. Due to the size of the genome, I had to split the design over two separate microarrays. I investigated the error introduced by the unavoidable splitting of RNA samples. Furthermore, I compared the commercially available arrays with the new design.

In the second transcriptional analysis, I studied the expression pattern of transcription factors during development and in adult muscle and brain. Determining the changes of the interactome during development is a major aim of developmental biology. Several studies were published investigating the early stages of embryogenesis (Mathavan *et al.* 2005; Vesterlund *et al.* 2011). However, no study has been performed covering the complete phase from egg till larva so far. Therefore, I designed a microarray covering all transcription factors of zebrafish. We performed experiments for 5 different stages and 4 different tissue samples. The microarray data were analyzed with a newly developed approach using two color arrays to detect expressed transcription factors. I carried out a time-series analysis for detecting functional patterns in the dataset. Additionally, I identified stage and tissue specific biomarker genes.

This thesis is structured into five chapters followed by the bibliography and an appendix.

Chapter 2 describes the microarray platforms used in this work. Furthermore the experimental set up of the different experiments is explained. The lab protocols used to perform the microarrays are also described.

Chapter 3 gives an overview of the studied toxicants including their chemical structure and the general application.

In Chapter 4, the bioinformatic and statistical methods applied in this work are explained. Used programs and databases are also named.

Chapter 5 presents the results of the bioinformatic and statistical analysis. The first part of this chapter deals with the results of the 10 compound study. This is followed by the results of the whole genome array. Last, the findings of the transcription factor screen are presented.

Chapter 6 summarizes the results and presents the conclusion drawn from the different

Compound name	Mode of action	Reference	Organism
Esfenvalerate	Affecting sodium and calcium ion channels in cell membranes	Viant *et al.* 2006a	chinook salmon
Methoxychlor	Methoxychlor-metabolites act like ER agonist	Ortiz-Zarragoitia and Cajaraville 2005	adult male zebrafish
Di-n-buthyl phthalate	AchE inhibitor; induction of liver peroxisome proliferation; ER agonist	Jee *et al.* 2009; Ortiz-Zarragoitia *et al.* 2006	bagrid catfish; zebrafish
Flucythrinate	Strong hPXR agonist, hERalpha agonist, hAR antagonist	Kojima *et al.* 2010	chinese hamster ovary cells
2,4-Dimethylphenol	Decrease in ATPase activity	Duchnowicz *et al.* 2005	human erythrocytes
Chlorpyrifos	AchE inhibitor; hPXR agonist, hERalpha agonist	Wheelock *et al.* 2005; Kojima *et al.* 2010	Juvenile chinook salmon, chinese hamster ovary cells
4-Chlorophenol	Estrogenic activity; Chlorophenols act as oxidative uncoupler	Ogawa *et al.* 2006; Comparative *et al.* 2001	Yeast; Aquatic organism
Chlorthalonil	Thiol-reactive	Baier-Anderson and Anderson 2000	striped bass macrophages
Propoxur	AchE inhibitor	Smulders *et al.* 2003	rat brain
1,2-Dibromoethane	Oxphos disruptor	Thomas *et al.* 2001	rat liver mitochondria

Table 1.1: Table of known modes of action found in the literature for the 10 compounds.

transcriptional analysis.

The Appendix consists of result tables and figures of the transcriptional analysis.

Chapter 2

Microarray Material and Methods

In this thesis several microarray based studies are analyzed and compared. This Chapter describes the methods that were used to perform the microarray experiments. Three different microarray platforms are used. Besides the common two-color control design, a special two color approach without controls was developed to study transcription factor time series data. Wilde type zebrafish were utilized for all experiments. Depending on the RNA sample, two different RNA extraction methods were applied. The amplification, labeling, hybridization and scanning steps were done according to standard procedures (Agilent 2006).

2.1 Microarray Platform

To find the most suitable microarray system for our work, we compared the most appropriate microarray platforms, which where available. In previous projects performed in our group, self-printed Compugen (Compugen, Tel Aviv, Israel) zebrafish cDNA arrays were utilized. The Compugen Zebrafish Oligo Library (Cat # XEBLIB384) was designed employing the gene information available in 2001. Although good results have been achieved using these arrays, we decided to look for an updated system. We focused our search on oligonucleotide arrays, which covered the largest part of the genome. Commercially available arrays have the advantage that they are printed with more than 12 times more probes on a slide as compared to our established in-house system. Additionally, they are also printed with a much higher spot quality. Since the commercially available slides are printed in a clean room, they also provide a much clearer background with less dust and scratches. For zebrafish, only Agilent (Agilent Technologies, Inc., Santa Clara CA, USA) offers an updated whole genome microarray.
In 2007, Agilent released the 4x44k two color cDNA array platform, consisting of 4 separate arrays on one slide. The zebrafish 22k array was already successfully used within the institute. Agilent also updates its array platforms on a regular basis, typically once a year. Additionally, they offer the possibility to design custom arrays with their eArray

system (https://earray.chem.agilent.com/earray/). This gave us the possibility to create own arrays, which fit perfectly to the requirements of our specific applications. The good experiences we already made with the system, its high quality, and the regularly updated system led to the decision to use the 4x44k Agilent array system for this project.
The Agilent Gene Expression 4x44k Microarrays consists of 4 identical blocks (arrays) each with 45220 spots. The single spots are approximately 65 μm in size. Each block can be used for hybridization of a different sample. In th efollowing, I will refer to a single block as array and to the whole array as slide. The 60 nucleotides long oligonucleotides on the array are called probes.

2.1.1 Agilent 4x44k Zebrafish v1 and v2

The zebrafish v1 array (id 015064) is basically a duplication of the old zerbrafish 22k array (id 015064) and was published in 2005. The zebrafish v2 array (id 019161), on the contrary, represents a completely new design. The probes on this array were based on

- RefSeq, Jan 2008
- Unigene (Release 54), Dec 2007
- TIGR (Release 17), Jun 2006
- (Release 48), Dec 2007
- UCSC (danRer5) Zv7, Jul 2007.

Agilent included several control spots on their 4x44k platform for enabling users to easily check the quality of the experiments. In total, 1470 spots of the array are used as positive and negative controls. The positive controls consist of different amounts of ten in vitro synthesized polyadenylated transcripts derived from the Adenovirus E1A transcriptome and are spiked into the samples to control the amplification, labeling, and hybridization processes. The negative controls should help to control for background noise. They have a special secondary structure or are derived from *Arabidopsis thaliana* or *Escherichia coli* genes. Ideally, zebrafish cDNA should not hybridize to them. 50 zebrafish oligos were replicated 5 times and are distributed equally over the entire array, to control for spacial problems. All control spots are spread randomly over the array (www.chem.agilent.com).

2.1.2 ITG Whole Genome Array

The Agilent v2 array seems not to include all the genes we were interested in. Therefore, I analyzed the usability of the Agilent zebrafish v2 array in prospect to our needs.
For our toxicity studies, I made a comparison of genes, which were published to be regulated by compounds, and the genes on the Agilent v2 array. To this end, I downloaded

all genes from *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *D. rerio* (zebrafish) from the Comparative Toxicity Database (CTD) in March 2009 (Davis *et al.* 2009). Afterwards, the human, mouse, and rat genes were mapped via their orthologs to zebrafish genes. The ortologous relationships between genes of the different organisms were downloaded from the Ensembl Zv7 database (Flicek *et al.* 2010). Finally, I compared the list of possible tox-genes with the genes on the used Agilent v2 array. 1302 genes of putative tox-genes were missing from the array. Importantly, some of these genes had toxicity information published specifically for zebrafish.
In future projects, it was planed to study toxicant-induced transcriptional changes in the early stages of zebrafish development. To check the usability of the Agilent v2 array, I looked for genes known to be expressed in the early stages of development. I downloaded via Biomart (Smedley *et al.* 2009) all genes from the ZFIN database (Sprague *et al.* 2008) that showed expression in the blastula high, blastula dome, 50% epibolie, or bud epibolie and compared them to the genes on the array. In total, 207 genes known to be expressed in the early stages, were not present on the Agilent v2 array.
Since the Agilent arrays do not cover all of our genes of interest, we decided to design our own zebrafish whole genome array, called ITG_WG_Danio. The array is based on 28717 cDNAs from Ensembl zebrafish Zv8, which I downloaded via Biomart (Smedley *et al.* 2009). To improve the quality of the arrays, I decided to use three different oligos per transcript. For 28159 transcripts, I was able to design 3 different probes using the Agilent eArray system (https://earray.chem.agilent.com/earray/). As the total amount of oligos exceeds the available space of a single array, I divided the oligos randomly over two arrays. As controls, I used the Agilent controls from the commercially available arrays.
To further improve the system, I included 3129 spots of self designed *Arabidopsis thaliana* controls on the two arrays. The *A. thaliana* controls were oligos designed for different *A. thaliana* genes and show no match with the zebrafish genome larger than 21 base pairs. The two newly designed arrays have the Agilent ids 024077 and 024078.

2.1.3 Transcription Factor Array

Combined with another screening project, we also wanted to study the transcriptional regulation during the development of zebrashish embryos. We manually curated a list of 2,370 transcription factor genes, which contained at least one Interpro (Hunter *et al.* 2009) or one Pfam domain (Finn *et al.* 2010) related to transcription or with an entry in the transcription factor database DBD (Kummerfeld and Teichmann 2006). I compared the resulting list with the genes on the Agilent zebrafish v2 array. Since 439 genes were missing on the Agilent v2 array, we decided to design a special array covering only transcription factors. I used the Ensembl (Flicek *et al.* 2010) cDNA sequences corresponding to our list of transcription factor genes. We also developed the idea to use other databases like Refseq (Pruitt *et al.* 2007) but only 1399 of the selected transcription factor genes could be mapped to the Refseq database. I also tested the usability of the 3‘UTR se-

quences for the oligo design. In most cases the sequence was too short or not specific enough to find unique regions that could be used for the oligo design. To improve the quality of the planned experiments and due to available space on the array, I decided to use 8 different oligos for each transcript. This was possible for 3,957 of the 4,009 transcripts of the selected transcription factor genes. Additionally, 529 unknown sequences from a zebrafish sequencing project were included. The oligo design was made using the Agilent eArray system. As controls, I used 30 known zebrafish housekeeping genes, the *A. thaliana* controls (Chapter 2.1.2) used for the ITG_WG_Danio array, and the 1,417 standard Agilent controls. The transcription factor array has the Agilent id 022326 and the name ITG_TF_rerio.

2.2 Experimental Microarray Design

A variety of microarray design strategies has been published previously. Depending on the underlaying questions, the array system, and available samples, we decided to use two different approaches. One is the common control design usually used for two color arrays. For the time series data of the transcription factor study, we decided not to use reference samples. Instead we developed a control free two color design strategy.

2.2.1 Two Color Control Design

We decided to pool several embryos into one sample. On the one hand, this was the only way to obtain enough RNA for performing microarray experiments. On the other hand, the pooling reduces the effect of biological variation. In the beginning, we compared the advantages and capabilities of different design strategies. For the ten compound study, it was important to be able to compare different treatments in order to find toxicant specific genes. Furthermore, the individual expression patterns induced through the toxicants were of high interest as they offer the possibility to study the toxicant's modes of action via pathway or Gene Ontology (Ashburner *et al.* 2000) analysis. Therefore, we decided to utilize a treatment-control design in which each treatment is hybridized together with a corresponding solvent-control (Figure 2.1). A common reference design, which is normally used in such experiments, would clearly improve the quality of the treatment comparisons. However, it would also make it almost impossible to distinguish between the signal changes that result from the different treatments and changes that are induced from using different breeds or fishtanks. To counteract the problems induced through the pooling and the individual treatment controls, we performed three biological repeats for each treatment. To avoid differences caused by the different labeling and hybridization efficiencies of the two dyes, we performed a dye-swap for each sample. Because of the good quality of the arrays and the high costs, we decided not to do technical repeats.

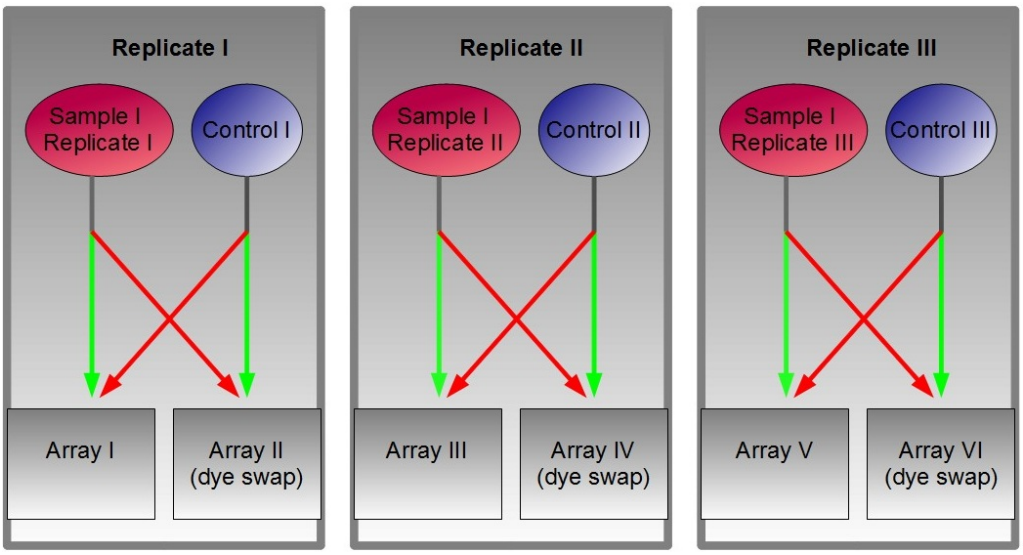


Figure 2.1: Two color control design. The green and red arrows represents the labeling color. For each replicate the labeling was also performed in reverse direction (dye swap), to correct for color induced dye bias.

2.2.2 Transcription Factor Design

For our transcription factor study, we wanted to compare expression patterns from six different developmental time points and 4 tissues. Since the design of a reference control for all our samples is not possible, we decided to use the two color system without any sample-control. We used the transcription factor array described in Chapter 2.1.3. To identify expressed genes, we used the *A. thaliana* controls (Chapter 2.1.2), and for improving the quality, we used four replicates. In Figure 2.2, the experiment design for this study is shown. One Array was loaded with two different RNA samples from the same stage. One sample was labeled with cy5 (red) and the other with cy3 (green). To obtain enough RNA, we pooled 100-300 embryos to get the samples for the early stages and 3-4 larvae for the later stages.

2.3 Zebrafish Lines

For the different studies, zebrafish wild type strains were employed. For the microarray toxicology studies, the *AB2O2* strain was chosen, and the transcription factor screen was performed with fish from the *ABO* strain. They were kept and bred as previously described (Westerfield 1993) in the fish facility of the Institute of Toxicology and Genetics at KIT. The crossing was performed by single matings. Male and female fish were separated the evening before spawning. In the morning, the female and the male were transfered together to a new spawning tank. This way, the eggs all have the same age.

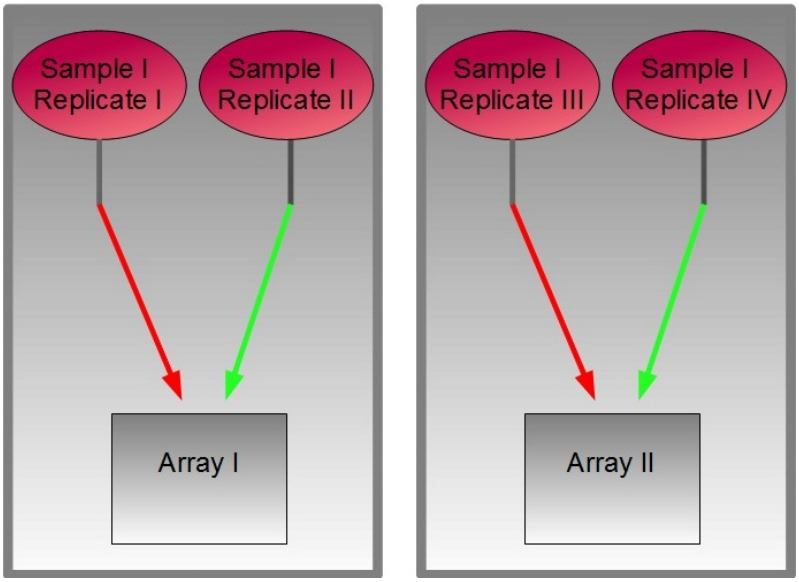


Figure 2.2: Two color no control design. The green and red arrows represents the labeling color. Each array consists of two replicates from the same sample.

2.4 Sample Preparation

2.4.1 Extraction of Total RNA via RNeasy Mini kit

The total RNA was extracted using the QIAGEN RNeasy Mini kit (QIAGEN, Venlo, Netherlands). First, the frozen samples were lysed in 1.2 ml RLT-Buffer and homogenized via pipetting. RNases were deactivated through addition of 12 μl ß-mercaptoethanol (ß-ME) to the lysis buffer. The lysate was then centrifuged, and the supernatant extracted. 1.2 ml of 70% ethanol was added to the lysate to provide ideal binding conditions. The sample was then loaded onto the RNeasy silica membrane. The spin columns were washed with 700 μl RW1 buffer, 2x 500 μl RPE buffer and centrifuged after each step to remove the washing solution. The total RNA was eluted in 50 μl RNase-free water. The quality of the total RNA was examined by denaturing agarose gel electrophoresis, and the quantity was checked by using the NanoDrop spectrometer (NanoDrop Technologies, Wilmingto, USA). The exact procedure is descriped in more detail in the RNeasy Handbook.

2.4.2 Extraction of RNA via Trizol

Trizol works by maintaining RNA integrity during tissue homogenization, while at the same time disrupting and breaking down cells and cell components. For the RNA extraction via Trizol, the samples were transfered into a 2.0 ml Eppendorf tube with a minimum quantity of PBS. After adding 1 ml Trizol, the sample was homogenized by pipetting and vortexing. Then, the sample was incubated for 5 minutes at room temperature. The

sample can now be stored at -80 °C or the extraction can be continued. 200 µl chloroform was added, the sample was homogenized again and incubated for 2 minutes at room temperature. After centrifuging for 15 minutes at 4 °C, the upper phase was transfered to a clean Rnase free 1.5 ml tube. 0.5 µl of glycogen solution (10 mg/ml) and 500 µl isopropylalcohol were added. The sample was shortly mixed, incubated for 30 minutes at -80°C and spun for 30 minutes at 4°C. The supernatant was removed and 500 µl 75% ethanol added to precipitate the RNA. After 5 minutes centrifuging, the supernatant was completely removed. The pellet was resuspended in 100 µl DEPC water and kept at -80 °C. To purify the RNA, the volume was adjusted to 100 µl with DEPC water. After adding 100 µl chloroform, the sample was vortexed for 1 minute and centrifuged for 30 minutes at 4 °C. The upper aqueous phase was transfered into a clean Rnase free 1.5 ml tube and mixed with 10 µl 3 M sodium acetate DEPC pH5.2 and 250 µl 97% EtOH. After incubation over night at -20 °C, the sample was spun for 30 minutes at 4 °C. The supernatant was removed, and 500 µl 75% ethanol was added. The RNA was stored in this stage at -80 °C. To utilize the RNA, the sample was centrifuged for 5 minutes, the supernatend was removed, and the pellet was resuspendend in 12 µl DEPC water. The quality of the total RNA was examined by denaturing agarose gel electrophoresis, and the quantity was checked by using the NanoDrop spectrometer (NanoDrop technologies, Wilmingto, USA).

2.4.3 Amplification, Labeling, and Purification

To obtain fluorescently labeled cRNA, we used Agilent's Low RNA Input Linear Amplification Kit PLUS (Agilent 2006). First, the dilutions of the two spike-mixes were prepared. 1.5-2.5 ng of total RNA were mixed with 2 µl of the corresponding spike control and 1.2 µl of the T7 promoter primer. Nuclease-free water was added to obtain a total reaction volume of 11.5 µl. The sample was then incubated for 10 minutes at 65°C. Afterwards, the samples were cooled down for 5 minutes on ice. In the next step, 8.5 µl of cDNA Master Mix (4 µl 5X Strand Buffer, 2 µl 0.1 M DTT, 1 µl 10 mM dNTP mix, 1 µl MMLV-RT, 0.5 µl RNaseOut) were added to each sample. Samples were then first incubated for 2 hours at 40 °C, then for 15 minutes at 65 °C, and lastly cooled on ice for 5 minutes. With the help of T7 RNA polymerase, the RNA was simultaneously amplified and labeled via incorporation of cyanine 3 or cyanine 5 cytidine-tri-phosphates (CTPs). To this end, 30 µl of the Transcription Master Mix (15.3 µl Nuclease-free water, 20 µl 4X Transcription Buffer, 6 µl 0.1 M DTT, 8 µl NTP mix, 6.4 µl 50% PEG, 0.5 µl RNaseOut, 0.6 µl Inorganic pyrophosphotase, 0.8 µl T7 RNA Polymerase, 2.4 µl Cyanine 3-CTP or Cyanine 5-CTP) were added to the samples, followed by an incubation step for 2 hours at 40 °C.
The amplified cRNA was then purified using RNeasy mini spin columns from Quiagen (QIAGEN, Venlo, Netherlands). Nuclease-free water was used to reach a total volume of 100 µl. 350 µl of buffer RLT and 250 µl of ethanol were added and mixed via pipetting. The sample was then transfered to the RNeasy column and washed twice with 500 µl of

RPE buffer. The cleaned sample was then eluted in 60 μl of RNase-free water.
For quantification of the cRNA, 1.5 μl of the samples were analyzed with a NanoDrop spectrometer (NanoDrop technologies, Wilmingto, USA).

2.4.4 Hybridization Procedure

The required volume of 825 ng of labeled cRNA was brought to a total volume of 52.8 μl by adding nuclease-free water. Afterwards, the samples were mixed with 11 μl 10X blocking agent and 2.2 μl 25X Fragmentation Buffer. In order to fragment the RNA, the samples were incubated at 60 °C for 30 minutes. The fragmentation process was stopped utilizing 55 μl of 2x GEx Hybridization Buffer HI-RPM. The samples were then immediately loaded onto the arrays. 100 μl sample solution were put on the arrays and cover slips were careful placed on top to avoid bubbles. The chips were placed in hybridization chambers and incubated at 65°C for 17 hours (Agilent 2006).

2.4.5 Washing Procedure

The arrays were removed from the hybridization chambers and washed twice in GE Wash Buffer 1 at room temperature for 1 minute and once in GE Wash Buffer 2 for 1 minute at 37°C. Afterwards, the arrays were dipped into drying solution to avoid droplets on the arrays. Slides were scanned immediately after finishing the washing procedure, to minimize the impact of environmental influences on the signal intensities (Agilent 2006).

2.5 Scanning and Image Acquisition

2.5.1 Scanner Settings

The arrays were scanned with the Axon 4000B from Molecular Devices (Molecular Devices, Inc., Sunnyvale, CA,United States). The software used for image acquisition and image analysis was GenePix Pro 6.1 (Molecular Devices, Inc., Sunnyvale, CA,United States). Both channels (532 nm for green and 635 nm for red) were scanned simultaneously with 100% laserpower. The scans were performed with a resolution of 5 μm without line averageing or adjusting of the focal plane. The PMT was adjusted to reach a signal ratio between the two color channels of approximately 1. The images were stored as 16 bit multiple TIFF files.

2.5.2 PMT Setting

In previous projects, the arrays were scanned with three different PMT-settings (low, medium, and high) in order to increase the signal detection limit. I tested this approach

for the new Agilent 4x44k arrays. To this end, an array was scanned with three different PMT settings, and the signal-to-noise ratios of the different scans have been compared. The signal-to-noise-ratio is a quantitative measure of the ability to distinguish true signal from background noise. For microarrays, it is calculated as:

$$SNR = \frac{Signal - Background}{Standard\ deviation\ of\ Background} \tag{2.1}$$

A SNR of three is commonly used as the lower limit for accurate detection. Signal can be detected below this value, but the accuracy of quantitative measurements decreases significantly. For only 1.6 % of the spots, I could see an improvement of the SNR (SNR > 3) using all 3 scanns compared to a single medium scan. One import aspect to consider is that the SNR dependens on the proper spot detection. A poorly aligned spot will have a larger standard deviation of the background and therefore a smaller SNR. This indicates that the true improvement of the low, median and high scans is below 1.6 %. Taking into account the dye bleaching effect of the scanning and the time needed for a scan, I decided not to use multiple scanes for this project.

2.5.3 Image Analysis

The spot acquisition was performed utilizing the GenePix Pro 6.0 software (Molecular Devices, Inc., Sunnyvale, CA,United States). An individual local background area around each spot was defined, which included 400 pixels of the spot and excluded neighboring spots. For each channel, the raw data was calculated as the median intensity of all foreground pixels with respect to all background pixels. The background is calculated using a circular region that is centered around the spot (Figure 2.3). The background area has a diameter that is three times the diameter of the corresponding spot. All of the pixels within this area are used to compute the background unless, they are part of a spot or a two pixel region around a spot. The signals and other statistical parameters calculated by the software were stored in GenePix Gene List format files (.gal) (Molecular Devices 2005).

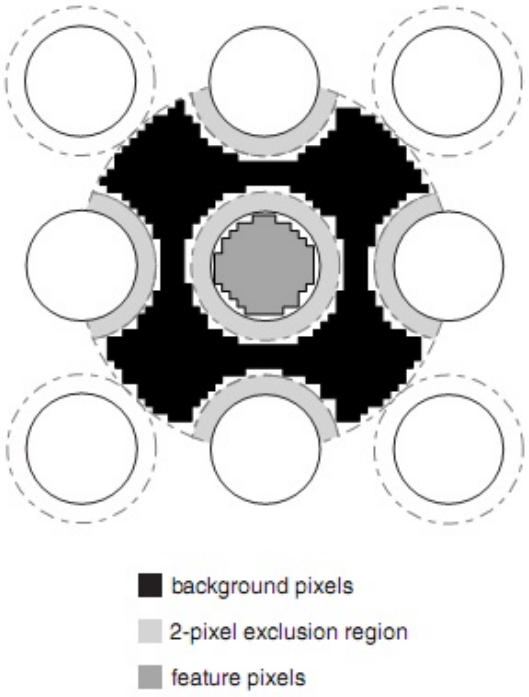


Figure 2.3: Spot detection in GenePix
(source: Molecular Devices 2005)

Chapter 3

Toxicants

For the toxicological studies, ten different compounds were selected which should cover a wide range of different toxicological mechanisms. Thereby, we hoped to be able to detect a larger variety of robust, sensitive toxicological biomarker genes. The experiments were performed from 24 hpf (hours post fertilization) to 48 hpf. The concentrations were selected to cause an acute phenotype in less then 10 % of the exposed animals.

Table 3.1: Used concentrations

Name	Solvent	Purity	Used Concentration
Propoxur	Water	Analytical standard Pestanal	150 mg/l
4-Chlorophenol	Water	Analytical standard Pestanal	50 mg/l
Chlorothalonil	DMSO	Analytical standard Pestanal	100 μg/l
Chlorpyrifos	Ethanol	Analytical standard Pestanal	7 mg/l
Di-n-butyl phthalate	DMSO	Supelco	1.5 mg/L
Esfenvalerate	Ethanol	Analytical standard Pestanal	80 μg/l
1,2-Dibromoethane	Water	Analytical standard Pestanal	400 mg/l
2,4-Dimethylphenol	Water	Analytical standard Pestanal	40 mg/l
Flucythrinate	Ethanol	Analytical standard Pestanal	125 μg/l
Methoxychlor	Ethanol	Analytical standard Pestanal	800 μg/l

Table 3.1 summarizes all selected compounds and their concentrations. The compounds were obtained from Sigma-Aldrich (Sigma-Aldrich GmbH, Seelze, Germany).

3.1 Toxicant Exposure of the Embryos

The toxicant exposure was performed in plastic Petri dishes with 20 ml exposure volume. To define the concentration, which was later used for the microarray experiments, the embryos were exposed from 24 to 48 hpf. At 48 hpf, the embryos are transferred to

control medium (ISO water) until 4 dpf. The concentration of toxicants was determined as the EC50 at 4dpf after treatment from 24 to 48 hpf. This design should help to discover a more robust genexpression response which is specific for the treatment. For the microarray experiments, the embryos were treated from 24 - 48 hpf. After exposure, the embryos were collected and immersed immediately in liquid nitrogen. The total RNA was extracted from three independently exposed batches of around 50 embryos each, and vehicle controls by using the QIAGEN RNeasy kit (QIAGEN, Venlo, Netherlands).

3.2 4-Chlorophenol

Figure 3.1: 4-Chlorophenol (source: www.en.wikipedia.org)

CAS: 106-48-9 4-Chlorophenol (C_6H_5CLO) belongs to the family of Chlorophenols. Chlorophenols can enter the environment throughout their production or life cycle. They are commonly used as a disinfectant in homes and hospitals, and as an antiseptic for root canal irrigant. Most of the Chlorophenols released into the environment dissolve in water, and only small amounts enter the air. They stick to soil and to sediments at the bottom of lakes, rivers, and streams. Low levels in water, soil, or sediment are broken down by microorganisms and are removed from the environment within a few days to weeks. Chlorophenols bioconcentrate in aquatic organisms such as fish. Exposure to high levels of chlorophenols have mainly effects on the skin, the liver and the immune system (*rats and mice*). Chlorophenols uncouple mitochondrial oxidative phosphorylation and produce convulsions (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.3 Pyrethroids

Pyrethroids are manufactured chemicals that are very similar in structure to the natural insecticides pyrethrins. But they are often more toxic to insects, as well as to mammals, and last longer in the environment. In air many of the pyrethroids are broken down or degraded rapidly by sunlight or other compounds found in the atmosphere. The compounds are extremely toxic to fish. They bind strongly to dirt. Therefore, they are normally not found in water. They have a toxic effect on the central nervous system and are likely to be cancerogenic (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.3.1 Esfenvalerate

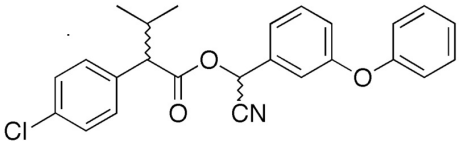


Figure 3.2: Esfenfalerate
(source: www.en.wikipedia.org)

CAS: 662-30-04-4 Esfenvalerate ($C_{25}H_{22}CLNO_3$) also known as Fenvalerate is a wildly used pesticide. It is used against a wide range of pests like flea, flies, and other insects. Most commonly it is used to control insects in food and cotton products, and for the control of stables. It can affect the endocrine, hematologic, neurologic, and reproductive system. It has been shown that Esfenvalerate has an influence on the levels of dopamine and muscarinic receptors from striatal membranes (*rat pubs*). It also influences the activity of acetylcholinesterase, monoamine oxidase and Na+- and K+-ATPase (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.3.2 Flucythrinate

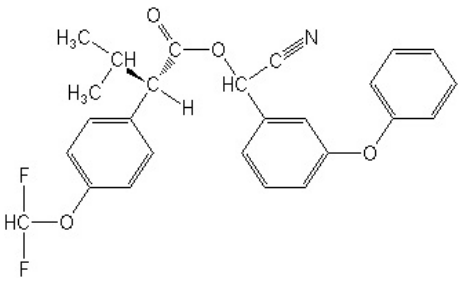


Figure 3.3: Flucythrinate
(source: www.en.wikipedia.org)

CAS: 662-30-04-4 Flucythrinate is considered to be toxic to humans. The use of Flucythrinate ($C_{26}H_{23}F_2NO_4$) has been restricted in the US and banned in the European Union since 2003 (Pesticide action network North America, www.panna.org). Flucythrinate is nearly insoluble in water and it has a strong tendency to bind to soil particles. It is therefore unlikely to contaminate groundwater. It affects the neurosystem (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.4 Methoxychlor

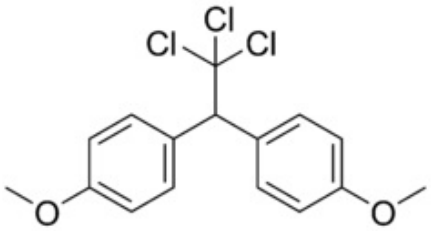


Figure 3.4: Methoxychlor
(source: www.en.wikipedia.org)

CAS: 72-43-5 Methoxychlor ($C_{16}H_{15}CL_3O_2$) is used as an insecticide against flies, mosquitoes, and a wide variety of other insects. The amount of Methoxychlor in the environment changes seasonally due to its use in farming and foresting. It does not dissolve readily in water and is mostly found in sediments. Its degradation may take many months. The use of Methoxychlor as a pesticide was banned in the United States in 2003 and in the European Union in 2002 (Pesticide action network North America, www.panna.org). Methoxychlor induces toxic effects in the endocrine,nervous and reproductive systems. Methoxychlor poses estrogen activity. It has been shown that Methoxychlor interacts with the members of the vascular endothelial growth factor (VEGF) and the angiopoietin families (Ang) and their receptors in a dose dependend manner (*female rat*). Furthermore it is known that Methoxychlor undergoes oxidative metabolism by cytochromes (P450) and produces substrates of the UDP-glucuronosyltransferases (UGTs) (*human liver*) (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.5 1,2-Dibromoethane

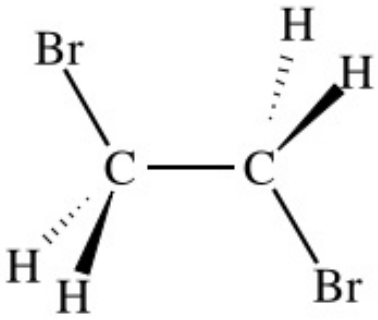


Figure 3.5: 1,2-Dibromoethane
(source: www.en.wikipedia.org)

CAS: 106-93-4 1,2-Dibromoethane ($BrCH_2CH_2Br$) has been used as a pesticide in soil, and on citrus, vegetable, and grain crops. Most of these uses have been stopped by the Environmental Protection Agency (EPA) since 1984. It can affect the skin, the liver, the urinary system, the kidneys, and the reproductive system. It is supposed to be cancerogenic for humans. 1,2-Dibromoethane is metabolized to active forms capable of inducing toxic effects by either of two systems, the microsomal monooxygenase system (cytochrome P-450 oxidation) or the cytosolic activation system (glutathione conjugation) (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.6 Chlorpyrifos

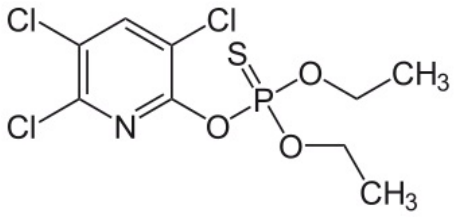


Figure 3.6: Chlorpyrifos
(source: www.en.wikipedia.org)

CAS: 2921-88-2 Chlorpyrifos ($C_9H_{11}C_{13}NO_3PS$) is an insecticide that inhibits acetylcholinesterase. It is widely used in homes and on farms to control insect pests. Chlorpyrifos is a neurotoxin and suspected endocrine disruptor. It sticks strictly to soil particles and does not mix well with water, so it is usually mixed with oily liquids before use. Toxicity induced by Chlorpyrifos results almost entirely from inhibition of neural acetylcholinesterase by itself and its bioactivation product chlorpyrifos oxon. Chlorpyrifos is bioactivated to chlorpyrifos oxon in the liver via cytochrome P450 (Ma and Chambers 1994; Sultatos and Murphy 1983). The majority of the neurological symptoms occur due to the subsequent cholinergic overstimulation. The cardiovasular effects are due to stimulation of muscarinic receptors in the heart (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/).

3.7 Propoxur

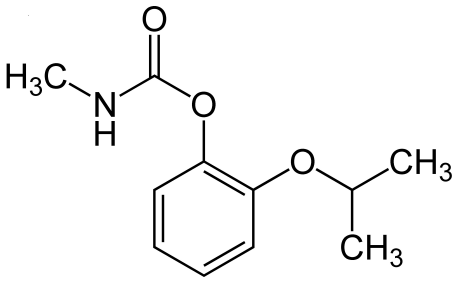


Figure 3.7: Propoxur
(source: www.en.wikipedia.org)

CAS: 114-26-1 Propoxur ($C_{11}H_{15}NO_3$) is a non-systemic insecticide with long residual effect used against turf, forest, and household pests and fleas. It is a synthetic analogue of the insect juvenile hormone. Unlike conventional insecticides that act as direct poisons, methoprene disrupts the morphologic development of insects. It is moderately to slightly toxic to fish and other aquatic species. It is thought to be a carcinogen, cardiovascular or blood toxicant, reproductive toxicant, and due to its cholinesterase inhibiting properties, neurotoxic (United States Environmental Protection Agency, www.epa.gov).

3.8 Chlorothalonil

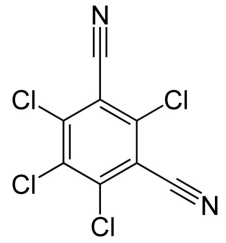


Figure 3.8: Chlorothalonil (source: www.en.wikipedia.org)

CAS: 1897-45-6 Chlorothalonil ($C_8Cl_4N_2$) is mainly used as a broad spectrum, non-systemic fungicide. It belongs to the top most used fungicides in the US. Chlorothalonil reduces fungal intracellular glutathione molecules to alternate forms which cannot participate in essential enzymatic reactions, ultimately leading to cell death. Chlorothalonil is highly toxic to fish and aquatic invertebrates. Available data on metabolism of chlorothalonil in rats and dogs indicate that the parent chemical is conjugated in liver to glutathione or cysteine-S-conjugates (United States Environmental Protection Agency, www.epa.gov).

3.9 2,4-Dimethylphenol

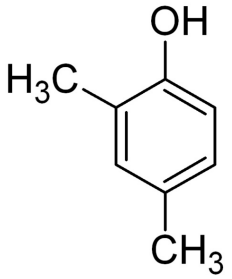


Figure 3.9: 2,4-Dimethylphenol (source: www.en.wikipedia.org)

CAS: 105-67-9 2,4-Dimethylphenol ($C_8H_{10}O$) belongs to the group of xylenols. They are very important for the chemical industry. Xylenols are used for the synthesis of pesticides, antioxidants, and pharmaceuticals. They are found in the wastewater of chemical and plastics producing companies. When released in water, they are biodegraded in a few days. 2,4-Dimethylphenol is used as microbiocide, fungicide and as solvent. Little is known about the underlying mode of action but due to its polar structure, it is classified as polar narcotic (United States Environmental Protection Agency, www.epa.gov).

3.10 Di-n-butyl phthalate

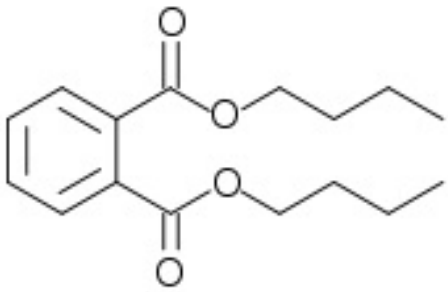


Figure 3.10: Di-n-butyl phthalate
(source: www.en.wikipedia.org)

CAS: 84-74-2 Di-n-butyl phthalate ($C_{16}H_{22}O_4$) is a commonly used plasticizer. The use has been restricted in the European Union for use in children's toys and cosmetics. Not much is known about the mode of action. Until now, it is classified as suspected teratogen and baseline narcotic substance. It is very toxic to aquatic organisms and badly soluble in water (United States Environmental Protection Agency, www.epa.gov).

Chapter 4

Bioinformatic Methods

This Chapter gives an overview of the different bioinformatic methods I used to analyze the microarray data. The primary analyses were performed using MATLAB (version R2010a, The MathWorks, Natick, Massachusetts, USA). For the gene function analysis, several freely available programs were selected.

4.1 Primary Microarray Analysis

The primary microarray analysis consists of five major parts.

- Quality control of the arrays. Exclude low quality arrays from further analysis.
- Spot filtering. Remove spots with bad quality.
- Data transformation. Transform the signal data in a more statistical more usable format - in general log-ratios (Equation 4.1).
- Data normalization. Normalize the data in order to remove bias, e.g. from dye effects.
- Detection of differentially expressed genes.

The primary analysis of the microarrays used in this thesis was performed completely in MATLAB. For the analysis, a special MATLAB application named Gait-CAD was further developed with an microarray section, which includes a method for analyzing two color microarray data. Since the design of the microarray experiments used for the transcription factor study is not the standard approach, the analysis could not be performed using standard methods and was therefore executed in MATLAB directly. The transcription factor analysis is described in Chapter 5.3.

4.1.1 Gait-CAD Microarray

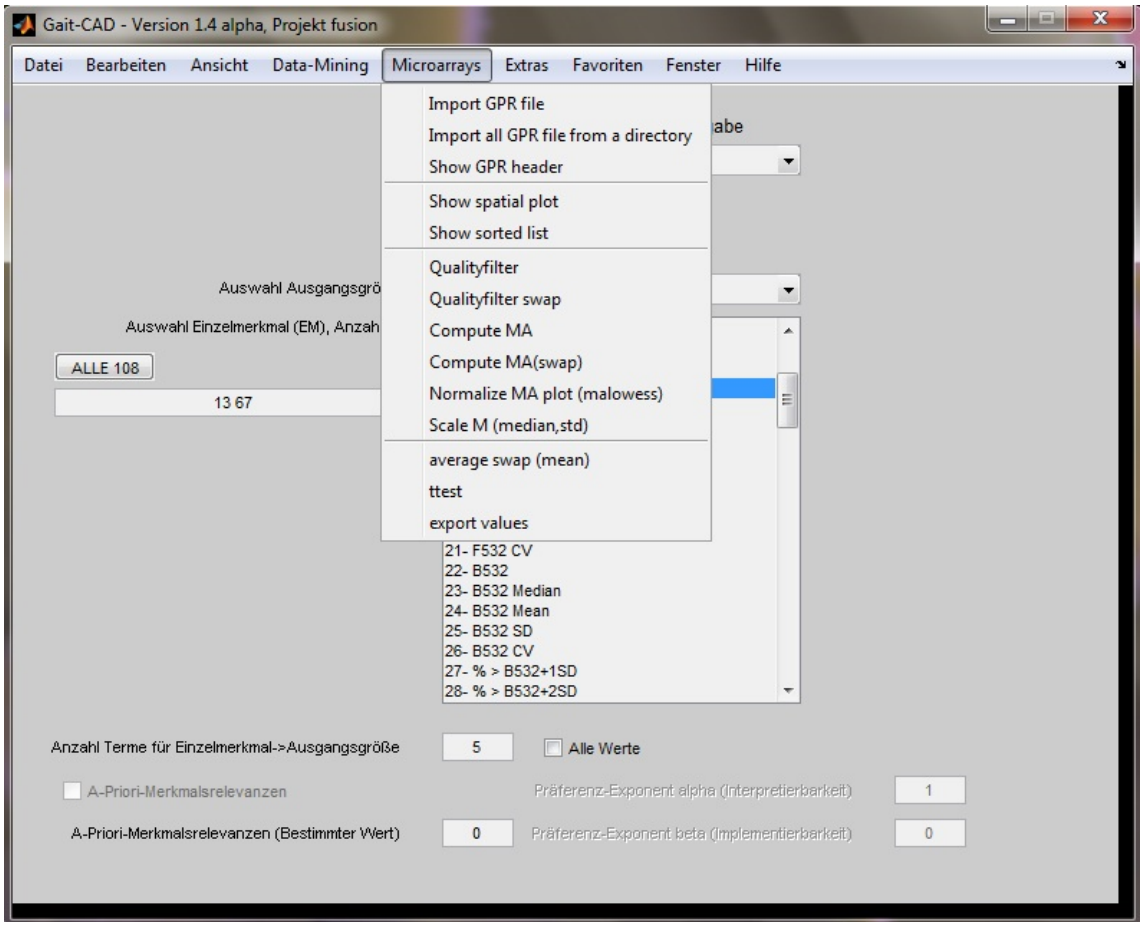


Figure 4.1: Gait-CAD screenshot

Gait-CAD (Mikut *et al.* 2008) is a graphical user interface, which allows to easily analyze different datasets via MATLAB without requiring any programing knowledge. Gait-CAD includes many statistical and data mining functions. As part of my work, I implemented a Gait-CAD add-on for microarray analysis, which provides different filtering functions, normalization methods, and statistical hypothesis tests. I used self-developed analysis functions and some functions from the MATLAB Bioinformatics Toolbox (http://www.mathworks.com/products/bioinfo/). In Figure 4.1, a screenshot of Gait-CAD and the microarray section is shown. In the following sections, I describe the different parts of the Gait-CAD microarray add-on in more detail.

Data Import

The data can be imported into Gait-CAT as gpr-files (GenePix Result-files). This type of files can be easily produced with many microarray image analysis programs includ-

ing Agilents Feature Extraction (Agilent Technologies Inc., Santa Clara, CA, USA) and GenePix (Molecular Devices Inc., Sunnyvale, CA, United States). The input file consists of a header with general scanning and image analysis information, for instance, color channels used, laser settings, date and time of scan, feature type, and used grid-file. The data part consists of the raw signal and background values, several statistical parameters such as mean, standard deviation, and the annotation provided by the microarray supplier.

Data Transformation

In order to improve the comparability of the data, the raw signal values are transformed to logarithmic scale. This can be done using uncorrected or background corrected values. One often used background correction method is to subtract the background values of single spots from the raw signal. In order to get a more symmetric distribution, I decided to use the logarithmic ratio transformation. If R denotes the signal value of the red color channel and G the signal value of the green one, then the log differential expression ratio for each spot is calculated as follows:

$$M = log_2 \frac{R}{G} \tag{4.1}$$

The log intensity of each spot is defined as:

$$A = log_2 \sqrt{RG} \tag{4.2}$$

On this scale, $M = 0$ represents equal expression, $M = 1$ represents a two fold change between the expression levels (Russell 2009).

Quality Filtering

The Gait-CAD add-on offers the possibility to filter the data based on different selectable classifiers, e.g. spot control types or spot names. It is also possible to perform a quality based spot filtering. Therefore a cut-off value for several quality parameters can be defined. The quality parameters are:

Flag A spot can be flagged during the image acquisition process. This is done either manually at the inspection of the array image if the spot is part of an artifact (dust, scratch), or by the spot detection algorithm if no spot could be found.

Diameter Minimum and Maximum of spot diameter. The size of a spot can be limited to exclude malformed spots that might be artifacts.

The following parameters can be set independently for each color channel of a spot.

SNR A minimum signal-to-noise-ratio (SNR) can be defined. In general, a SNR > three indicates that the spot signal can truly be distinguished from the background signal. This cut-off can be used to exclude spots with low foreground signals or unequally and high background signals.

$$SNR = \frac{mean\ (Foreground\ Pixel)\ -\ mean\ (Background\ Pixel)}{Standard\ deviation\ (Background\ Pixel)} \tag{4.3}$$

CV The coefficient of variation (CV) can be defined to filter spots with a non-uniform signal distribution, which also might indicate artifacts.

$$CV = \frac{Standard\ deviation\ (Foreground\ Pixel)}{mean\ (Foreground\ Pixel)} \tag{4.4}$$

Minimal signal A cut-off for the maximum percentage of Pixels per spot that are below the minimal signal can be defined. The minimal signal is defined as the intensity, which is two standard deviations above the background pixel intensity of the spot.

Maximal signal This defines the maximally accepted percentage of saturated pixel in a spot. Saturated signals can be used to calculate a signal ratio but this cannot be used in a comparisons with the other signals.

Afterwards, a general quality measure for each data point (spot) is calculated and used as a classifier for the data filtering. The qualifier indicates whether a given spot passes the defined quality check or not, and consequently, is considered a good or bad spot. In order to be considered a good spot, a spot has to pass all quality tests.

Plots and Sorted Lists

For visualizing the data, different displaying methods are available. Single or groups of parameters can be selected and the information is presented as sorted list (highest to lowest value) or plotted in a diagram. The following options are available for plotting the data:

Box plot A box plot (Figure 4.2) is used in statistics to represent descriptive parameters like mean, median, and variance of a dataset in a graph. With the help of box plots, it can easily be shown whether two datasets are significantly different or not. In microarray analysis, box plots are often used to compare the efficiencies of different normalization methods (Zhang 2006).

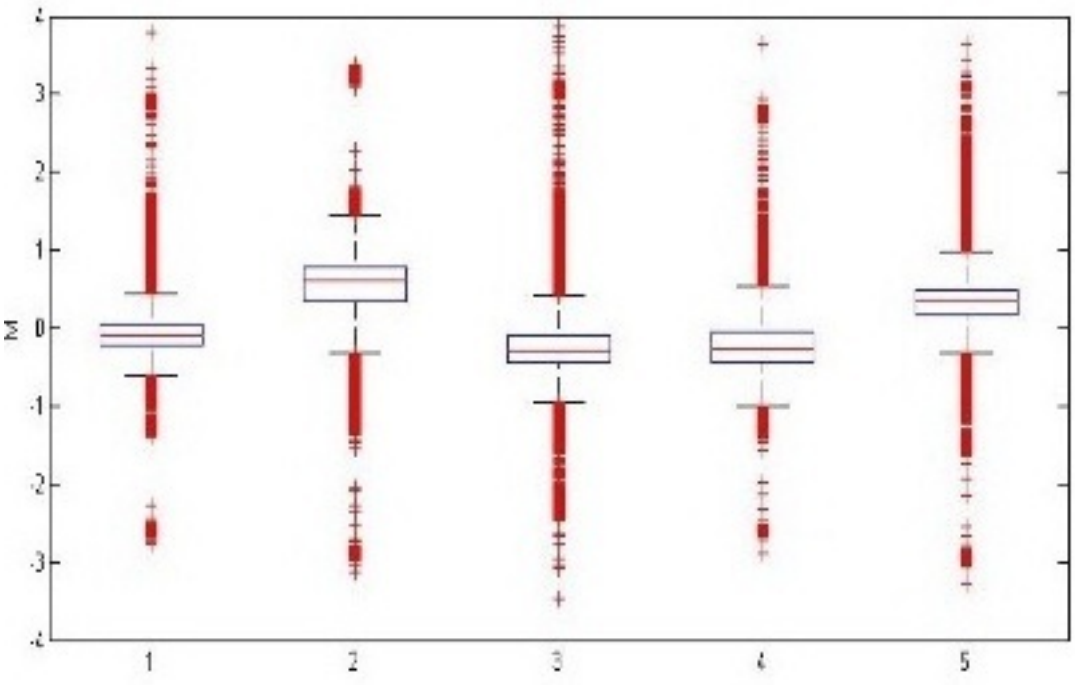


Figure 4.2: Box plots of the distribution of non-normalized M-values for five different microarray datasets. The central mark of the box is the median,and the edges are the 25th and 75th percentiles. The variability is indicated by the length of the whiskers. For microarray experiments the median should be ideally near 0. In the dataset 2 and 5 the distribution is clearly shifted towards 1. This could indicate dye bias or other labeling problems.

Spatial plot In a spatial plot, the values are presented according to their position on the array. Spatial effects can arise from hybridization problems or during the microarray production process. This bias cannot be corrected by most normalization methods. Therefore, it is very important to investigate the arrays for possible spatial effects (Zhang 2006). In Figure 4.3 the spatial plot of the red foreground signal from an microarray is shown.

Histogram An histogram is a representation of the distribution of a parameter (Figure 4.4). The histogram subdivides the data points into equal intervals called bins. For each bin, the number of points in that interval is presented. Histograms are used to study the distribution of parameters and to define cut offs (Zhang 2006).

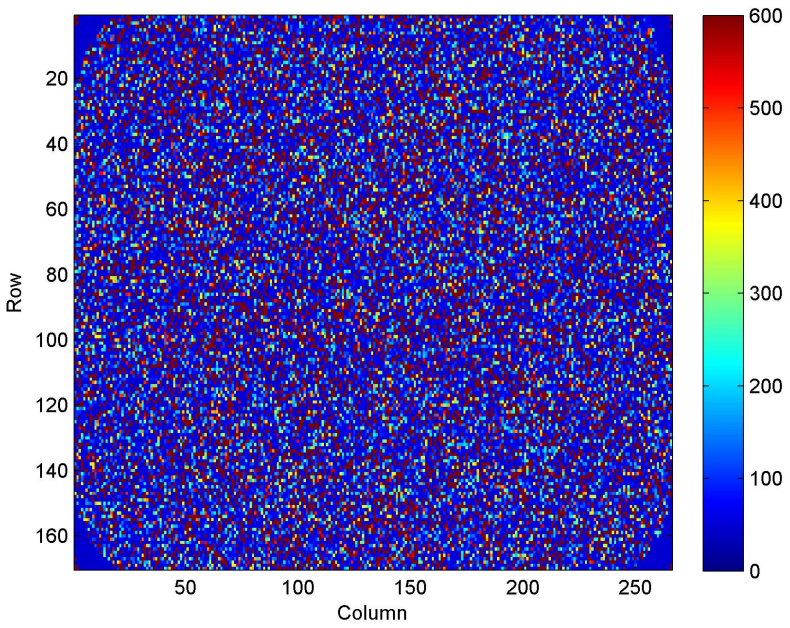


Figure 4.3: Spatial plot of the raw red foreground signal of an microarray. High (red), medium (yellow) and low (blue) signals are equally distributed over the array. No artifacts, empty regions, signal gradients or accumulations are detectable.

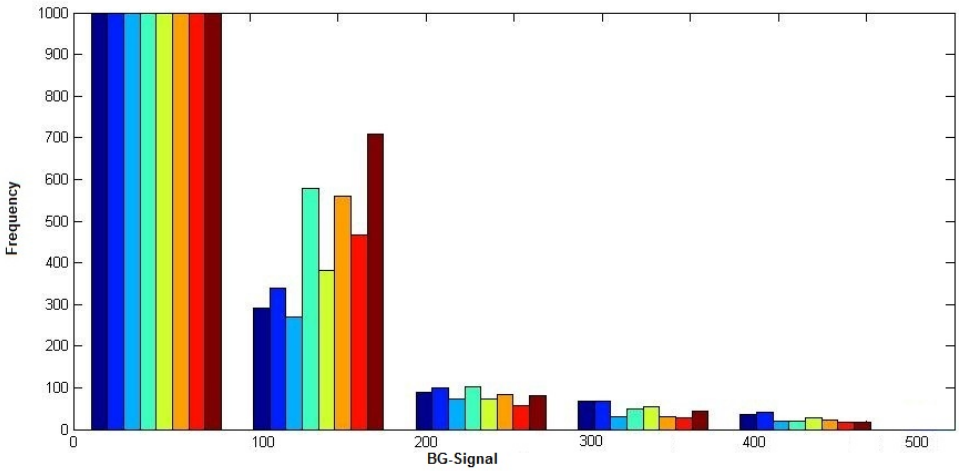


Figure 4.4: Histogram plot for the spot background signals of eight different datasets. The bars represent the number of spots with a background signal in the corresponding interval (0-100; 101-200; 201-300; 301-400; 401-500). The background signals for all datasets show a similar distribution. Most of the spots have a background signal bellow 100.

Scatter plot A scatter plot can be employed for visualizing the relationship or associations between two parameters in the same dataset. To display the relation of the two color channels, the ratio versus intensity plot (M-A plot) is most commonly used. In these plots, the y-axis displays the ration (M), and the X-axis the intensity (A) of the signals. Scatter plots of the M- and A-values can be used as quality indicator or gene selector (Zhang 2006). A scatter plot from M versus A is presented in Figure 4.5.

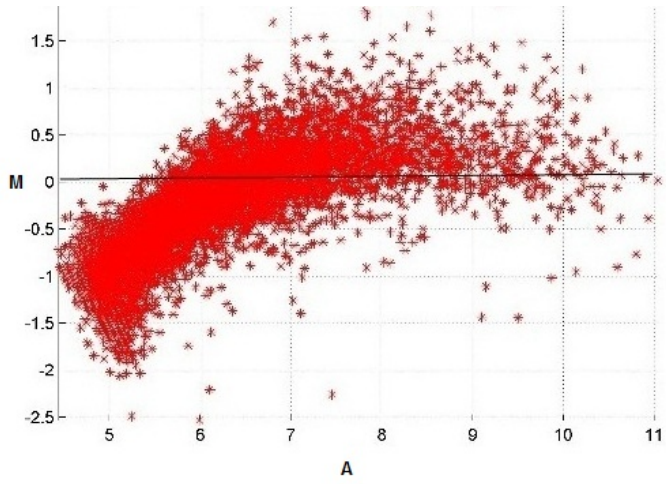


Figure 4.5: Scatter plot of M versus A before normalization. The plot shows the relationship between the total spot signals (A) and the ratio between the color channels (M). Each point represents a spot on the array. For low signals (A<6) the ratios are shifted towards the green color channel and for higher signals (A>8) towards the red channel. This is typical when dye bias occurs.

Normalization Methods

Microarray data is often influenced by non biological effects like differences in the labeling efficiencies of the used colors (dye bias). With the help of normalization methods the data should be corrected from this influences. Many different statistical methods have been developed, to address this problem. In order to select the best normalization method for the dataset it is useful to first have a close look at the data using the different available display functions. This helps to get an idea of the data distribution and the problems that may influence the data. Then, several normalization methods should be tested and compared regarding their effect on correcting the possible problems. The best one, sometimes more than one, is than chosen to perform the data normalization during analysis. In the following the normalization methods implemented in Gait-CAD are described.

LOWESS normalization The LOWESS (locally weighted scatter plot smoothing) normalization is used to perform an intensity-based normalization. Especially in two

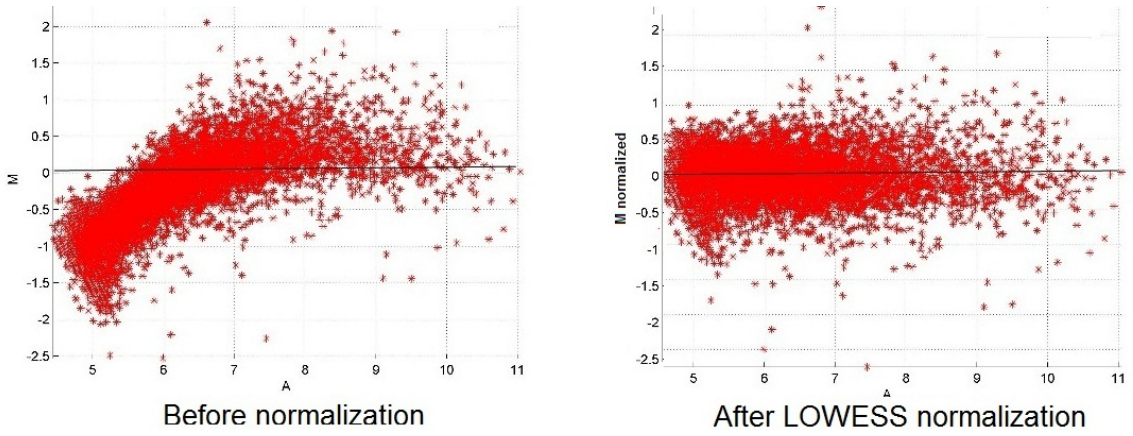


Figure 4.6: Intensity depended normalization (LOWESS). The M-A scatter plot before and after LOWESS normalization. Before the normalization a clear shift in the data is visible (dye bias). After normalization the data is centered around M equals 0.

color arrays, it is known that the used colors have different labeling efficiencies and stabilities over time. This color effect can be mainly seen in low signal data and leads to a shift of the signal ratio to the more stable and efficient color. Simplified, the LOWESS normalization corrects this shift. In an M-A-plot, this is visible as scattering the data equally around 0 (Simon *et al.* 2004). Figure 4.6 shows two M-A scatter plots, one of the raw data and the second one after LOWESS normalization.

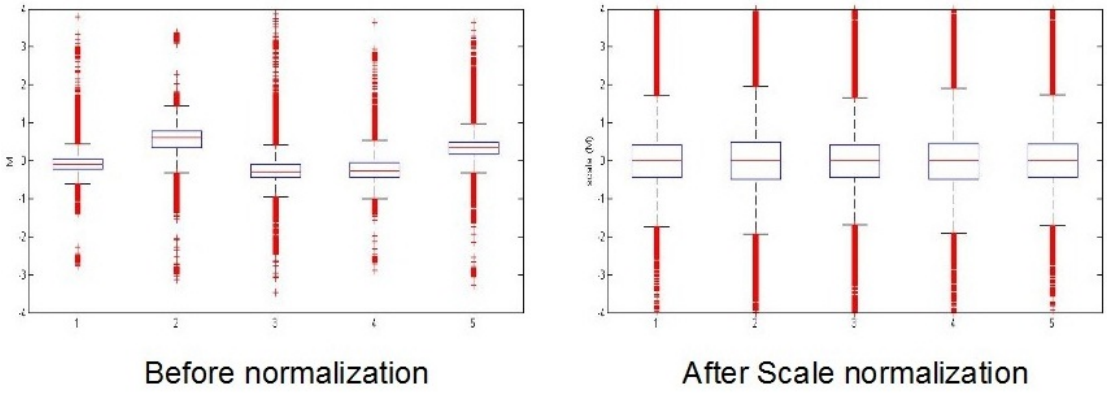


Figure 4.7: Scale normalization. Box plots for the M-values from five different microarrays. Before the normalization, the signal distributions are clearly different, regarding the mean and the variance.

Centering The centering normalization is a very conservative form of normalizing array data. It is mainly used to perform normalizations between different arrays. It

basically compacts all signal distributions in the same way and can therefore correct even for different scanner settings. For each value, centering subtracts the median of the distribution and divides through its standard deviation. This results in the median over all values being *1* and the standard deviation being *0* (Russell 2009).Figure 4.7 shows the box plots of a five microarray dataset, before and after the scaling normalization.

Spike Control Analysis

Since we are using mainly Agilent two color arrays, the analysis of the Agilent spike in controls is also implemented in the add-on. The log_2 signal of the spike controls is calculated and compared to the expected ratios provided by Agilent. The outcome is displayed in a scatter plot. If there is no problem, the five different spike groups are nicely separated and show a linear regression like shown in Figure 4.8.

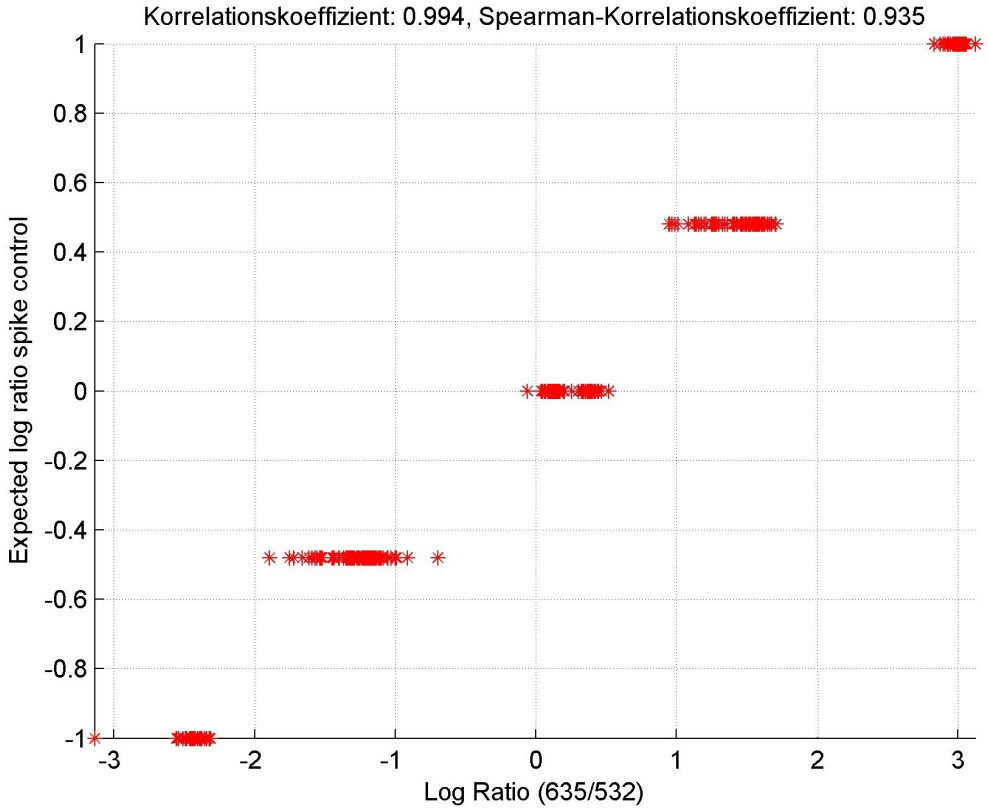


Figure 4.8: Spike controls. The scatter plot shows the expected log ratios against the calculated spot ratios. The five different spike control groups are clearly separated and show a linear relationship.

Dye Swap Handling

It is very common to use reverse labeling (dye swaps) to correct for the different labeling efficiencies of the different colors. Gene specific dye bias cannot be completely removed by normalization. To perform a dye swap, sample A is first labeled with red and sample B with green. In the reverse experiment, sample A is then labeled with green and sample B with red. The two experiments are averaged in the end to receive one gene specific dye-bias corrected dataset.

The Gait-CAD add-on is able to handle dye swap data by calculating a dye swap specific M-value. To calculate this value, the sign of the M-value of the reverse labeled experiment is flipped. The averaging (mean) is done considering the previously calculated quality measure (4.1.1). The average is then computed using:

- both values (when both values pass the quality filtering step)
- one value (if only one value passes the filtering step)
- set as Na N (when no valuable data point was found)

T-test

In order to investigate the differently expressed genes, an 'one-sample' t-test is used. First, the array replicates (minimum two) have to be fused together into one dataset. The implemented function performs a t-test of the null hypothesis, that the data is from a normal distribution with mean 0 and unknown variance. This null hypothesis was tested against the alternative hypothesis that the mean is not 0. Therefore, the t-test should only be applied to the normalized and centered M-values. The used t-test statistics was calculated as follows:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}, \tag{4.5}$$

where $\overline{x}$ is the sample mean, μ (=0) is the hypothesized population mean, s is the sample standard deviation and n the number of replicates. The implemented function calculates the p-value, which is the probability to obtain a value like the sample mean under the null hypothesis. P-values below 0.05 are generally considered to be statistically significant. Furthermore, the function determines the confidence interval, the median, the mean, and the standard deviation over all replicates (Russell 2009).

Batch Analysis

Since some microarray experiments consists of a larger number of arrays, I performed the whole analysis as batch analysis. The whole analysis was implemented as macro and then automatically executed for the complete batch of arrays. This function is part of the Gait-CAD program and is also applicable to functions that do not belong to the microarray add-on.

Export Function

The software includes several data export functions. It is possible to save the whole dataset or a selected part in text or Excel files, which then can be used for further analysis. Another function allows for saving the different plots and lists using various file formats.

4.1.2 Two Color Control Design Analysis

The goal of using two color microarray experiments is to detect differentially expressed genes between two samples or a control and a sample. In our case, we wanted to find genes indicating treatment with a specific toxicant. For this reason, a method was developed that is focused to detect very robust differentially expressed genes. In the following, the methods used to analyze the two color control microarray data are described.

1. Data upload. The data is uploaded into Gait-CAD

2. Data transformation. The M- and A-values were calculated based on non background corrected signals (Equation 4.2 and 4.1). If the background value is larger than the spot signal, subtracting the background from the signal results in a negative corrected signal value. These cases must be prevented, since a negative expression value makes biologically no sense. Furthermore, it is not possible to calculate a $\log_2$ ratio of negative values. Since genes with a small signal (smaller than background) in one color channel and a high signal in the other channel are of special interest as potential biomarker genes, we did not want to exclude them from the dataset.

3. Quality filtering. A quality measure for flagged and low-signal data is calculated. Spots were flagged during the image acquisition process as being artifacts or could not be found at all, are penalized. Spots with low signals in both color channels are also marked with a bad quality value. Low signals can lead to false, high signal ratios. For example, two low signals like 40 and 120 might produce a ratio of 3, but this might be caused by noise instead of a true biological signal. The selection of the cut-off value for the low signal filtering was based on the analysis of the background of all microarrays from one experiment. In our dataset, almost all background signals were below 200. Therefore, I decided to use 200 as low signal cut-off value (Chapter 4.1.1).

4. Normalization. LOWESS normalization and centering normalization was performed to correct for intensity dependent dye bias and array differences (Chapter 4.1.1).

5. Quality control. Several quality plots representing the data distribution of the single arrays were inspected. The spike controls were also analyzed, and the outcome saved.

6. Data fusion. The datasets of all arrays that belong to one treatment were fused together in one large dataset (all Dye Swaps of all replicates).

7. Dye swap averaging. The dye swaps were averaged taking into account the calculated quality measure (Chapter 4.1.1).

8. T-test. The p-values and other parameters were calculated for the averaged dye swap values (Chapter 4.1.1).

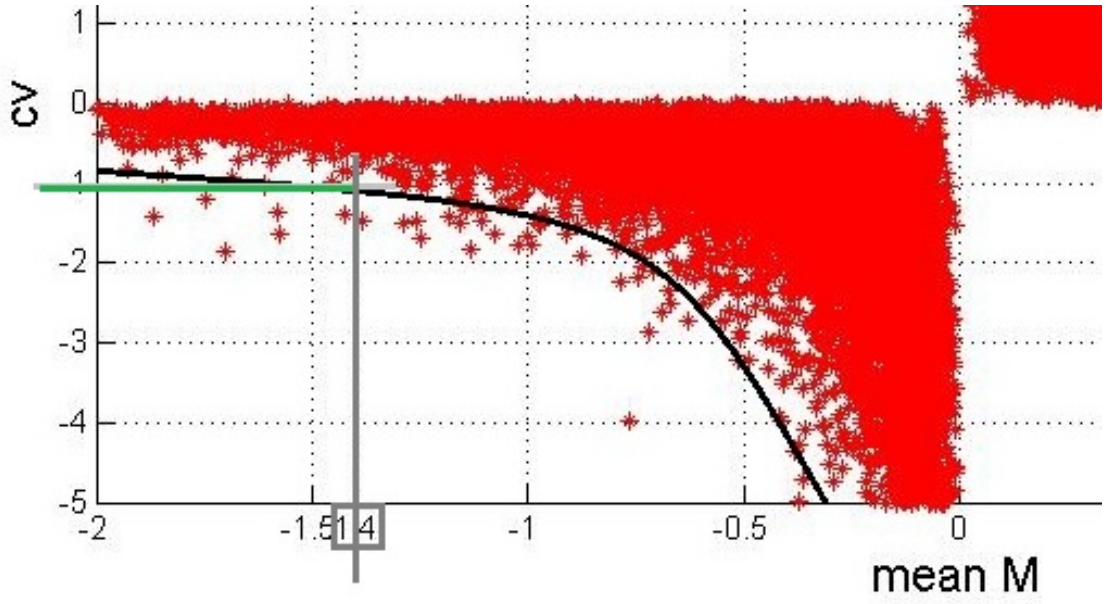


Figure 4.9: M-value cut-off

9. False discovery rate detection. The number of spots on the array is rather high (∼44000). If a t-test for this number of values is performed, 2200 false positives are expected at a p-value cut-off of 0.05. The use of adjusted p-values like Benjamini-Hochberg or Bonferroni cannot be performed directly on such a large dataset. For this reason, I decided to define a M-value cut-off based on the variation of the M-values for reducing the number of false positives. The false positives will have rather small M-values. The variance of M-values over the replicate is expected to be smaller when arising from a true signal as compared to being caused by noise. To test this, the mean M-values are plotted against the coefficient of variation (CV) over the replicates. The CV is defined as follows:

$$CV = \frac{Standard\ deviation\ (M\ values)}{mean\ (M\ values)} \tag{4.6}$$

A CV value of one is generally used as cut-off between small (CV < 1) and high variance (CV > 1). The M value cut-off is found when looking for the point where

the M-values scatter over the CV of one. In Figure 4.9 the black line describes the maximal variance for an particular M value. The gray horizontal lines represent the CV of 1. The green vertical line the M value cut-off for which the variance becomes greater than 1. Differentially expressed spots (p-value < 0.05) with M-values greater than 1.4 are assumed to have a small variance and therefore coming from true biological signal. Differentially expressed spots (p-value < 0.05) with M-values smaller than 1.4 might be false positives, and must be treated with caution.

10. Data export. The data is exported in a tab delimited text file for further analysis steps.

4.2 Multivariate Analysis Methods

In this work, multivariate statistics is mainly applied for analyzing the relationships of different gene expression patterns. Two different types of multivariate analysis were used. All calculations were performed with MATLAB (version R2010a, The MathWorks, Natick, Massachusetts, USA).

Principal components analysis (PCA) PCA calculates a set of variables that represent a summary of the dataset. With the help of these variables, similarities in the data can be detected.

Clustering Clustering assigns the data objects into groups. Objects that belong to the same cluster have a higher similarity than objects from different clusters.

4.2.1 Principal Component Analysis

Principal component analysis (PCA) is a mathematical technique that tries to minimize the number of variables in a dataset. For a set of objects, it calculates uncorrelated variables called principal components that describe the variability in the data source. This is done in such a way that the first principal component covers the highest variance, the second the next highest and so on. If we assume that the highest source of variance in our gene expression experiment is the treatment, the first two principal components should give us an indication which treatments induce a similar expression pattern and which are more dissimilar. To represent the results, the first two principal components for all treatments were plotted against each other (Figure 4.10). If two treatments are located close to each other in this plot, it is highly likely that they are similar.

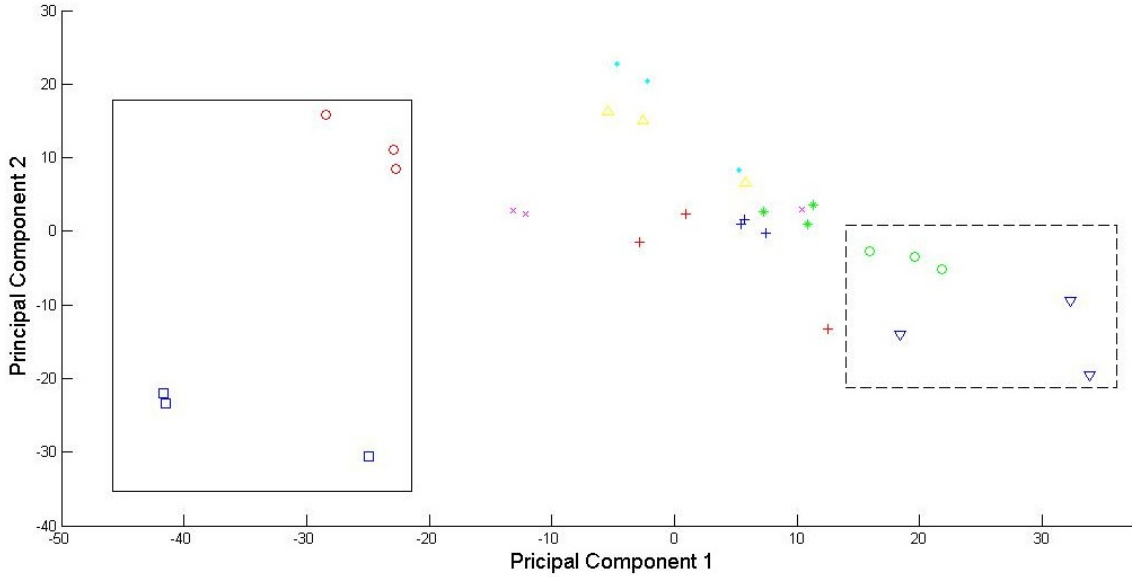


Figure 4.10: PCA analysis plot of a group of microarrays. Replicates labeled with the same symbol. A box marks microarrays, that have a similarity based on the first principal component.

4.2.2 Hierarchical Clustering

For the hierarchical clustering analysis, I used an agglomerative clustering approach. This means that in the beginning represents a cluster. A distance metric is utilized for calculating the similarity between two clusters. The linkage method defines which elements of a cluster are employed for determining the distance between two clusters. During the analysis, the clusters with the highest similarities are fused together until no similar groups can be found. Several linkage methods and distance metrics are available. The ones which performed best on our data are described in the following section (Simon *et al.* 2004).

Complete Linkage

Complete linkage, also called furthest neighbor, uses the largest distance between objects in the clusters. The different expression values for a single dataset were standardized, so that the mean was 0 and the standard deviation was 1. The distance metric is then defined as follows:

$$d(r,s) = max(dist(x_{ri}, x_{sj}), i \in (1,...,n_r), j \in (1,...,n_s) \tag{4.7}$$

r and s are two clusters
n_r and n_s denote the number of objects in the clusters
x_{ri} is the ith object in cluster r
x_{si} is the ith object in cluster s

Correlation distance

The correlation distance is calculated as one minus the sample correlation between the data points. It is defined as follows:

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}}, \tag{4.8}$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj} \qquad \bar{x}_t = \frac{1}{n}\sum_j x_{tj}. \tag{4.9}$$

x_s and x_t are the vectors of the cluster-representatives calculated with the linkage method. The distance assumes value between 0 (when correlation coefficient is +1, i.e. the two samples are most similar) and 2 (when correlation coefficient is -1).

The results of the cluster analysis are usually shown as dendrogram sometimes together with a heat map like in Figure 4.11.

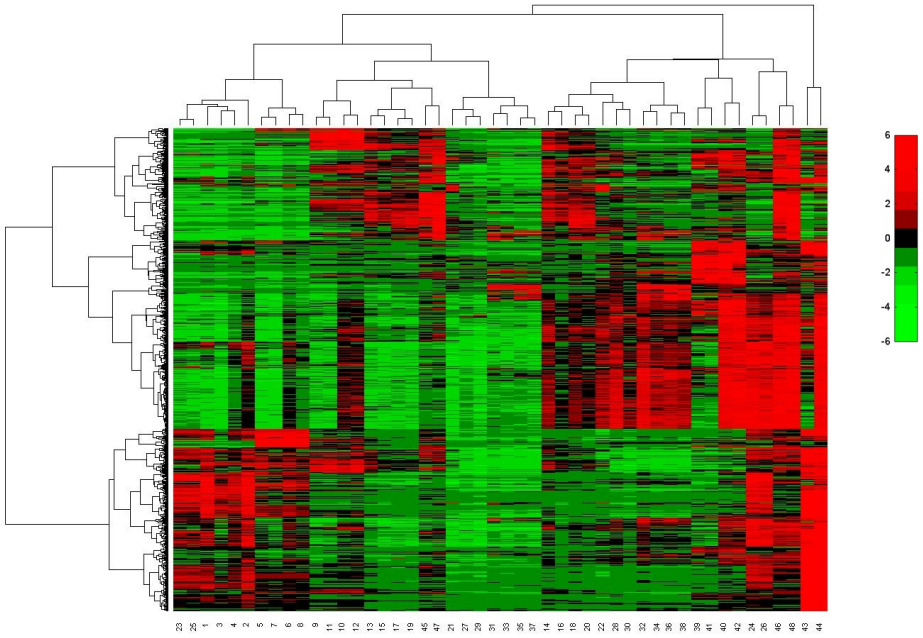


Figure 4.11: Two dimensional cluster plot of 42 microarray experiment. Columns represents microarrays and rows genes. The two dendrograms and the heat map is shown. Up-regulated genes are red labeled and down regulated genes are green labeled, black means very low or no signal.

4.2.3 K-means Clustering

K-means clustering is a method of cluster analysis which aims to partition the objects into a predefined number of clusters.

1. It starts with a random set of cluster centers (of the predefined number of clusters).

2. Each object is than fused with the cluster with the nearest center. To calculate the distance between object and cluster center, I utilized the previously described correlation distance metric (Equation 4.8, 4.9).

3. The new cluster centers are calculated and used as a starting point for the next cycle.

4. The objects are again fused with the nearest cluster and the new cluster centers calculated.

5. This procedure was repeated until stable clusters were obtained.

The result is the object-cluster assignment after the last calculation cycle. Since the results can be very different depending on the initial cluster centers, it is useful to repeat the analysis several times and to calculate an average result cluster.

4.3 Enrichment Analysis Methods

When investigating a subset of genes with respect to their presence in a set of interesting genes, it might be sometimes very difficult to interpret the results using the raw numbers. Simply by chance a certain number of genes will be part of the set of interesting genes. A statistical analysis is needed to evaluate the enrichments. Therefore, an enrichment ratio can be calculated (Equation 4.11=. The significance (p-value) of the enrichment is then computed using the hypergeometric test (Zhang *et al.* 2005). A small p-value indicates a high probability that the enrichment is not produced simply by chance.

$$k_{exp} = (n/m) * j \tag{4.10}$$

$$r = k/k_{exp} \tag{4.11}$$

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}} \tag{4.12}$$

n= number of genes in our interesting gene set
m= number of genes in our reference gene set
k= number of genes of the subset in our interesting gene set
j= number of genes of the subset in our reference gene set
k_{exp}= number of genes which are expected to be in our interesting gene set
r= ratio of enrichment
P= significance of the ratio of enrichment

4.4 Gene Function Analysis

Gene function analysis helps to get a better understanding of the underlaying mechanisms of a microarray experiment. First, the data are linked to gene function information. Then, an enrichment analysis is performed to find gene function categories that are overrepresented in the dataset. For these categories a regulation in the dataset is assumed. In order to gain a better understanding of the mechanisms of the microarray expression patterns, I decided to use Gene Ontology (Ashburner *et al.* 2000), KEGG pathways (Kyoto Encyclopedia of Genes and Genomes; Kanehisa and Goto 2000), and WikiPathways (Pico *et al.* 2008).

4.4.1 Gene Ontology

The Gene Ontology (GO) categories consists of defined terms representing gene product properties. The ontology covers three domains:

- cellular component, the parts of a cell or its extracellular environment.
- molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis.
- biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

The database entries can be accessed and downloaded from the Gene Ontology web page (www.geneontology.org). The enrichment results of the gene ontology terms can be displayed in Tables and as directed acyclic graph (DAG).

4.4.2 KEGG

KEGG is a manually curated database. For my analysis, I used the pathway section which consists of manually drawn pathway maps for several organisms. The pathways are categorized into:

- Global Map
- Metabolism
- Genetic Information Processing
- Environmental Information Processing
- Cellular Processes
- Organismal Systems
- Human Diseases
- Drug Development

This data is also accessible via Internet (www.genome.jp/kegg/).

4.4.3 WikiPathways

WikiPathways is a Wikipedia-like internet portal. Each pathway is represented by a wiki entry. The pathways are all manually curated and can be searched via the web-portal www.wikipathways.org.

4.4.4 Gene Set Analysis Toolkit V2

To perform the gene function analysis of our datasets, the Gene Set Analysis Toolkit V2 was used (Zhang *et al.* 2005; http://bioinfo.vanderbilt.edu/webgestalt/option.php). As gene ids Ensembl Gene ids were used. The reference gene set was comprised of all genes on the Agilent zebrafish v2 array. As statistical method, I applied the hypergeometric test with a Benjamini-Hochberg multiple testing correction. For each category, a minimum number of 2 genes was selected. If a category consists of only a few genes, it can be more easily significantly enriched. Nevertheless, I did not exclude these categories. This fact should be taken into account when interpreting the results. Since KEGG and Wikipathways generally produce very high adjusted p-values, the result list consist of the top10 results. Therefore, the KEGG and Wikipathways results must be handled carefully and manually judged whether the enriched categories are really enriched. Figure 4.12 shows an example output from the KEGG enrichment analysis performed with the Gene Set Analysis Toolkit V2.

This table lists the enriched KEGG pathways, number of Entrez IDs in your user data set for the pathway, the corresponding Entrez IDs, and the statistics for the enrichment of the pathway. The statistice column lists the number of reference genes in the category (C), number of genes in the gene set and also in the category (O), expected number in the category (E), Ratio of enrichment (R), p value from hypergeometric test (rawP), and p value adjusted by the multiple test adjustment (adjP). Finally, the pathway name is linked to KEGG where the user ids are highlighted, the number of user gene ids is linked to a table with information about the user ids, and the Entrez IDs are linked to Entrez Gene			
Steroid biosynthesis	3	494054 768185 550369	C=12;O=3;E=0.32;R=9.42;rawP=0.0034; adjP=0.0476
SNARE interactions in vesicular transport	3	30711 571872 30712	C=32;O=3;E=0.85;R=3.53;rawP=0.0521; adjP=0.2611
Melanogenesis	4	436815 353151 393801 30080	C=60;O=4;E=1.59;R=2.51;rawP=0.0746; adjP=0.2611

Figure 4.12: Output Table from an KEGG enrichment analysis performed with the Gene Set Analysis Toolkit V2. A description of the output parameter is shown in the first row of the Table.

4.5 GO similarity methods

If many genes are differentially expressed, it is normal that also many Gene Ontology terms (Ashburner *et al.* 2000) are enriched. If several microarray studies are to be compared, it is very difficult to do that with large lists of GO terms. Additionally, the interpretation of results is made difficult due to the high redundancy between individual GO categories. In order to simplify the large lists of GO terms, semantic similarity measures are used (Schlicker and Albrecht 2008). This helps to remove redundant GO terms

and to summarize the GO results. For this approach, I used REViGO (Supek *et al.* 2010; http://revigo.irb.hr). REViGO is a web service that reduces the lists of uploaded GO terms and also helps visualizing them in scatter plots, tag clouds, and interactive graphs. Figure 4.13 depicts an example output of REViGO.

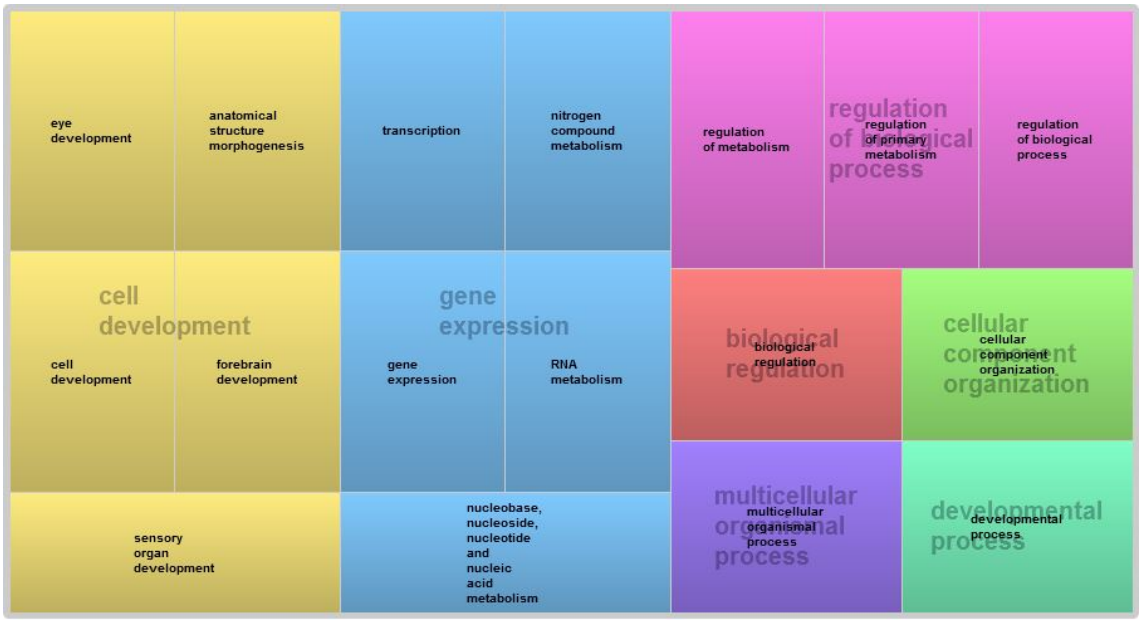


Figure 4.13: Output from an REViGO GO similarity analysis. The major GO categories are presented in light gray. The size of the category boxes represents the calculated adjusted p-value of the category enrichment.

4.6 Microarray Annotation

The annotation file provided for microarrays can be from low quality. Sometimes, the annotation is very old and the data cannot be linked to the updated information in databases like Ensembl. In other cases, the annotation is very limited and consist simply of company ids. When different microarray experiments from different array platforms are to be compared, the provided annotations are mostly not helpful. To solve this problem, I developed my own microarray annotation system. All microarray systems are simply mapped to the same genome information and thereby can be now linked and compared.

1. First the oligo sequences are blasted using a locally installed NCBI blast function (Altschul *et al.* 1990). As reference genome information, the Ensemble zebrafish cDNA Zv8.54 library was used.

2. A self implemented bioperl script extracts the information from the blast output and filters them (Stajich *et al.* 2002). An oligo is annotated only if all blast hits with a

length larger than half of the total length of the oligo belong to the same gene. The result is than transformed to a Table like format and saved as new annotation file.

3. Afterwards, FileMaker (FileMaker GmbH, Santa Clara, CA, USA) is used to link the new annotation file with the gene expression data.

Chapter 5

Results

Depending on the project and the research questions involved, different analysis methods are needed. In the following Chapter, the results of the different microarray data sets and the underlaying analysis is described. The interpretation and evaluation of the results are presented in Chapter 6.

5.1 10 Compound Study

The aim of the 10 compound study is to get a better understanding of the underlaying toxicity mechanisms and to find possible biomarker genes for that mechanism. Therefore, the microarray data were first analyzed as specified in Chapter 4.1.2. To identify similarities between the compounds and possible shared mechanisms, multivariate statistical methods and co-expression analysis was used. In the next step, the signal distributions of the expression data was examined. Furthermore, data sets of published microarray experiments were linked to our data and compared. Enrichment analysis of important gene sets were performed. Finally, a gene function analysis was made for discovering affected pathways.

5.1.1 Comparative Analysis

In order to find similarities between the expression patterns of the compounds, three different multivariate analyses methods were applied. Agglomerative hierarchical clustering (Chapter 4.2.2), Principal Component Analysis (PCA, Chapter 4.2.1), and the partitioning clustering algorithm K-means (Chapter 4.2.3). A critical questions is which genes represent the toxicity specific response in the data set. In the complete gene expression dataset, the toxicity mechanisms might be such a small part that the clustering might rather reflect the level of the induced damage and the ongoing repair and immune responses than the underlaying shared mechanisms of toxicity. That is why I created three different data sets. Missing values were set to *0* indicating no change in the expression

value ($M = \log_2 FC; M = 0 => FoldChange = 1$).

- *'all':* The complete gene expression data set.
- *'p-value 0.05':* The data from 14394 transcripts that showed a significant differential expression with a p-value < 0.05 in at least one treatment.
- *'194':* The 6 most up-regulated and 5 most down-regulated transcripts from each array (total 194 different transcripts). The selection was based only on transcripts with a p-value < 0.05. This list showed the best clustering performance regarding the replicates.

Hierarchical Clustering

In the following section, the results of the hierarchical cluster analysis are described. The first data analyzed was the complete gene expression data set (*all*). Although most of the transcripts show no differential expression, clear clusters are detectable (Figure 5.1). In most cases the replicates for the different treatments cluster together. The main clusters found in the dendrogram are:

- chlorophenol and propoxur
- dibromoethane and dimethylphenol
- methoxychlor and esfenvalerate
- dibutylphthalate and flucythrinate

The dendrogram of the *p-value 0.05* data set looks very similar to the complete data set (Figure 5.2). Also here for 6 compounds, the replicates cluster nicely together. The expression patterns of methoxychlor and esfenvalerate are very similar and are not dividable. For propoxur and flucythrinate, only 2 replicates clustered together. The main clusters are:

- chlorophenol and propoxur
- dibromoethane and dimethylphenol
- methoxychlor and esfenvalerate
- dibutylphthalate and flucythrinate

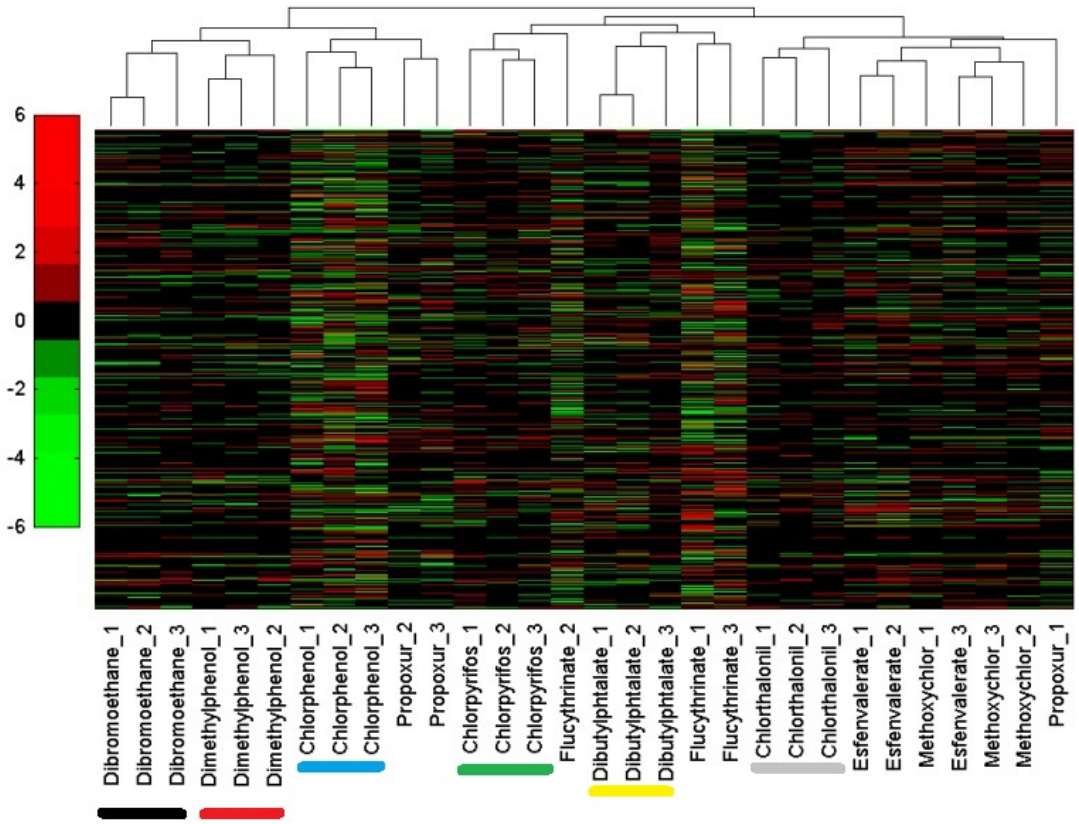


Figure 5.1: Cluster analysis from the complete gene expression data set (*all*). The columns indicate the 3 replicates for the 10 treatments. For 6 compounds, the replicates are clustered together. Esfenvalerate and methoxychlor seem to overlay. For flucythrinate and propoxur, one replicate clusters not with the other two. Similarities between dibromoethane and dimethylphenol were detectable. Chlorophenol and propoxur also cluster together, as well as flucythrinate and dibutylphthalate.

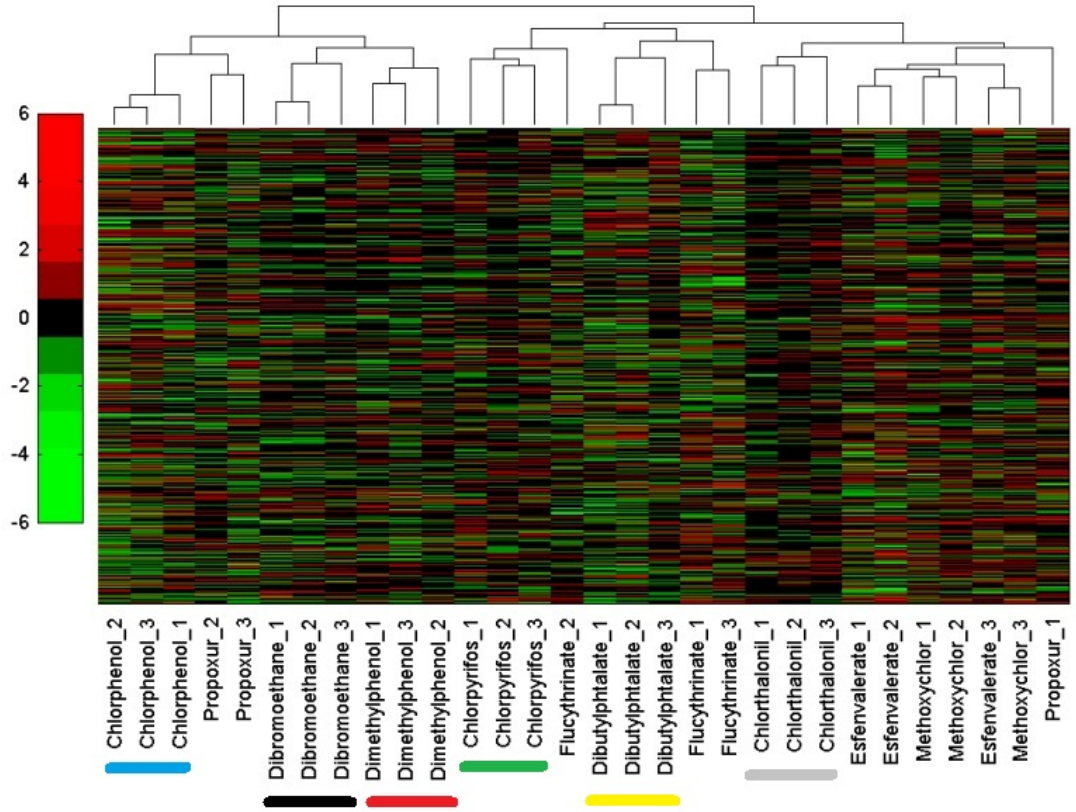


Figure 5.2: Cluster analysis from data set *p-value 0.05*. The replicates of 6 compounds cluster together. The expression patterns of methoxychlor and esfenvalerate seem to be very similar and not dividable. For flucythrinate and propoxur, one replicate clusters not with the other two. Similarities between dibromoethane and dimethylphenol were detectable. Chlorophenol and propoxur also cluster together, as well as flucythrinate and dibutylphthalate.

In the dendrogram of the *194* data set, all replicates cluster perfectly together (Figure 5.3). This shows that the highly differentially expressed transcripts are very specific for the used compounds. This could either be caused by a compound-specific mechanism or by difference in the toxicity response (immune system and repair mechanism). A high-resolution version of Figure 5.3 is provided on the supplementary CD. The main clusters for *194* are:

- chlorophenol and propoxur
- dibromoethane and dimethylphenol
- methoxychlor and esfenvalerate

In contrast to the other two data sets, no clustering of dibutylphthalate and flucythrinate can be observed in the *194* data set.

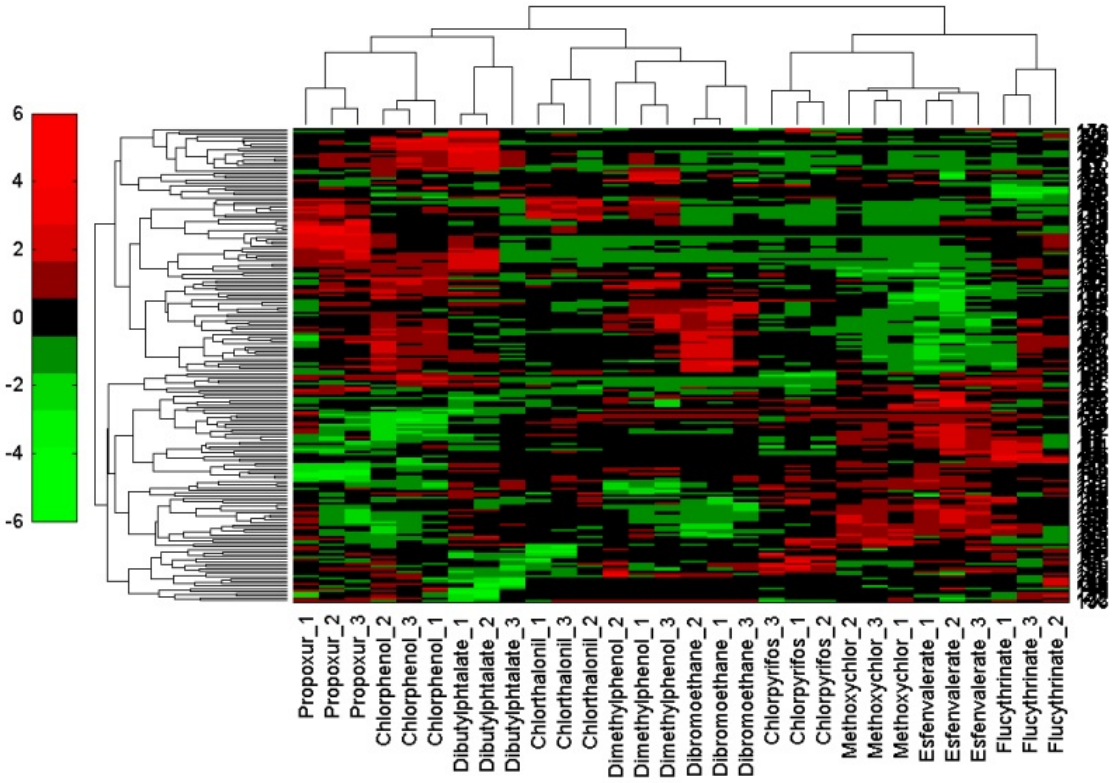


Figure 5.3: Cluster analysis from data set *194*. All replicates cluster together. Similarities between dibromoethane and dimethylphenol were detectable as well es between chlorophenol and propoxur. Esfenvalerate and methoxychlor are also clustered together.

Principal Component Analysis

Due to the high computing power required, no principal component analysis could be performed for the whole data set. For the *p-value 0.05* data set, three clusters can be observed (Figure 5.4) in the 1. and 2. principal component. The main clusters for *p-value 0.05* are:

- Chlorophenol and propoxur (solid line)
- dibromoethane and dimethylphenol (dotted line)
- methoxychlor and esfenvalerate (dashed line)

The principal component analysis of the *194* data set shows two clear clusters (Figure 5.5). A clustering of the replicates could not be observed. The main clusters for *194* are:

- Chlorophenol and propoxur (solid line)
- methoxychlor and esfenvalerate (dashed line)

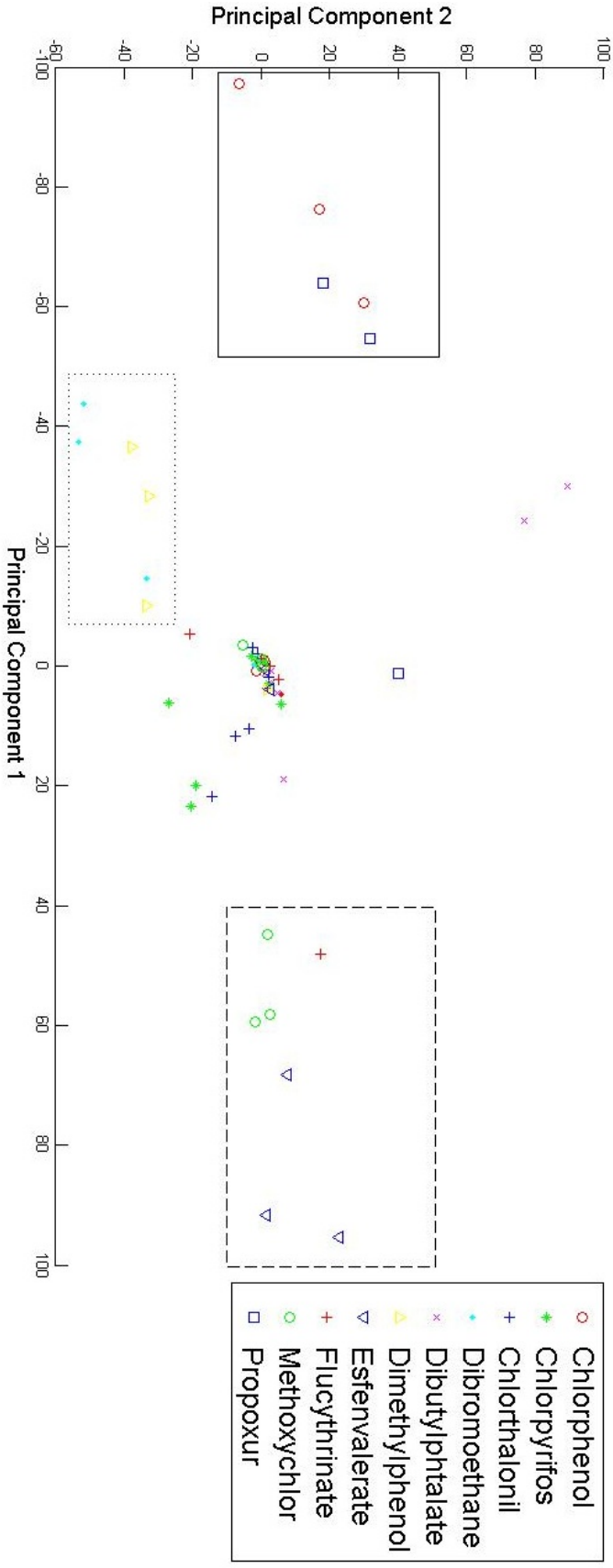


Figure 5.4: PCA from data set *p-value 0.05*. The boxes indicate groups of compounds that showed similarity based on the first two principal components. The x-axis describes the first principal component and the y-axis the second one.

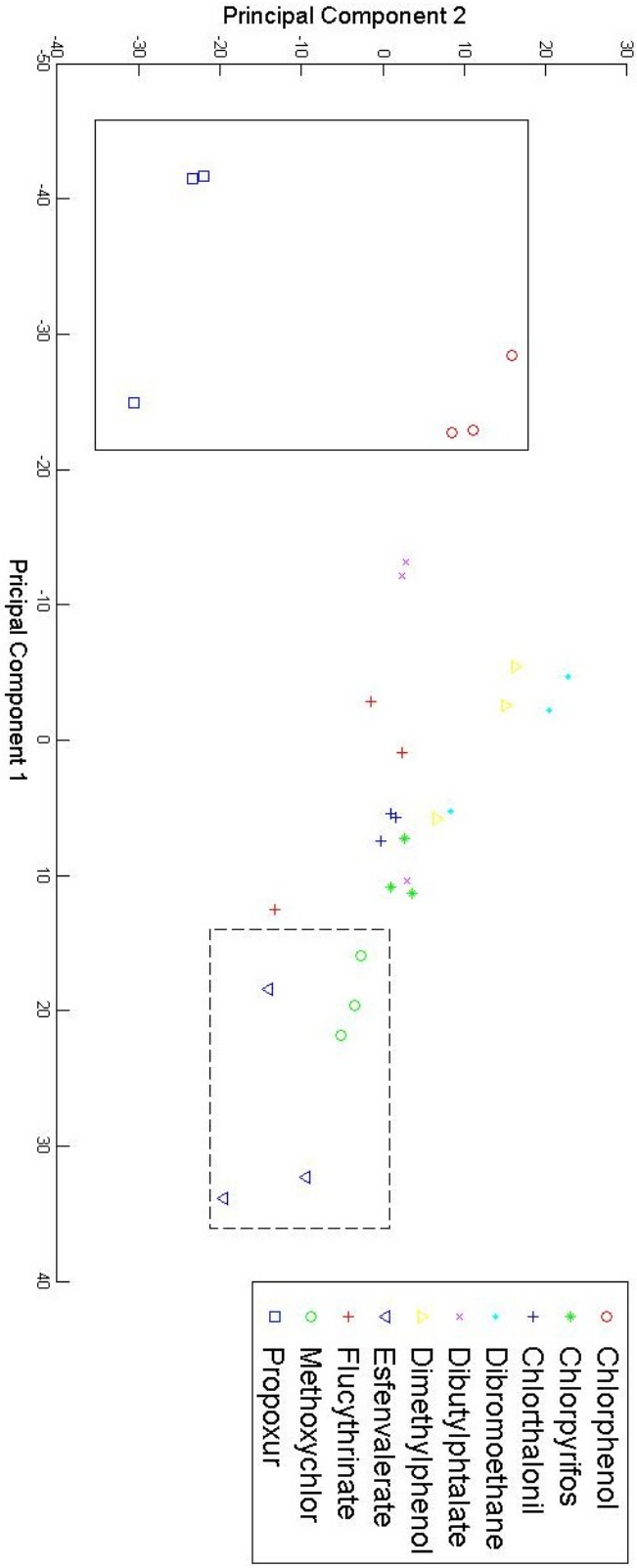


Figure 5.5: PCA from data set *194*. The boxes indicate groups of compounds that showed similarity based on the first two principal components. The x-axis describes the first principal component and the y-axis the second one.

K-means Clustering

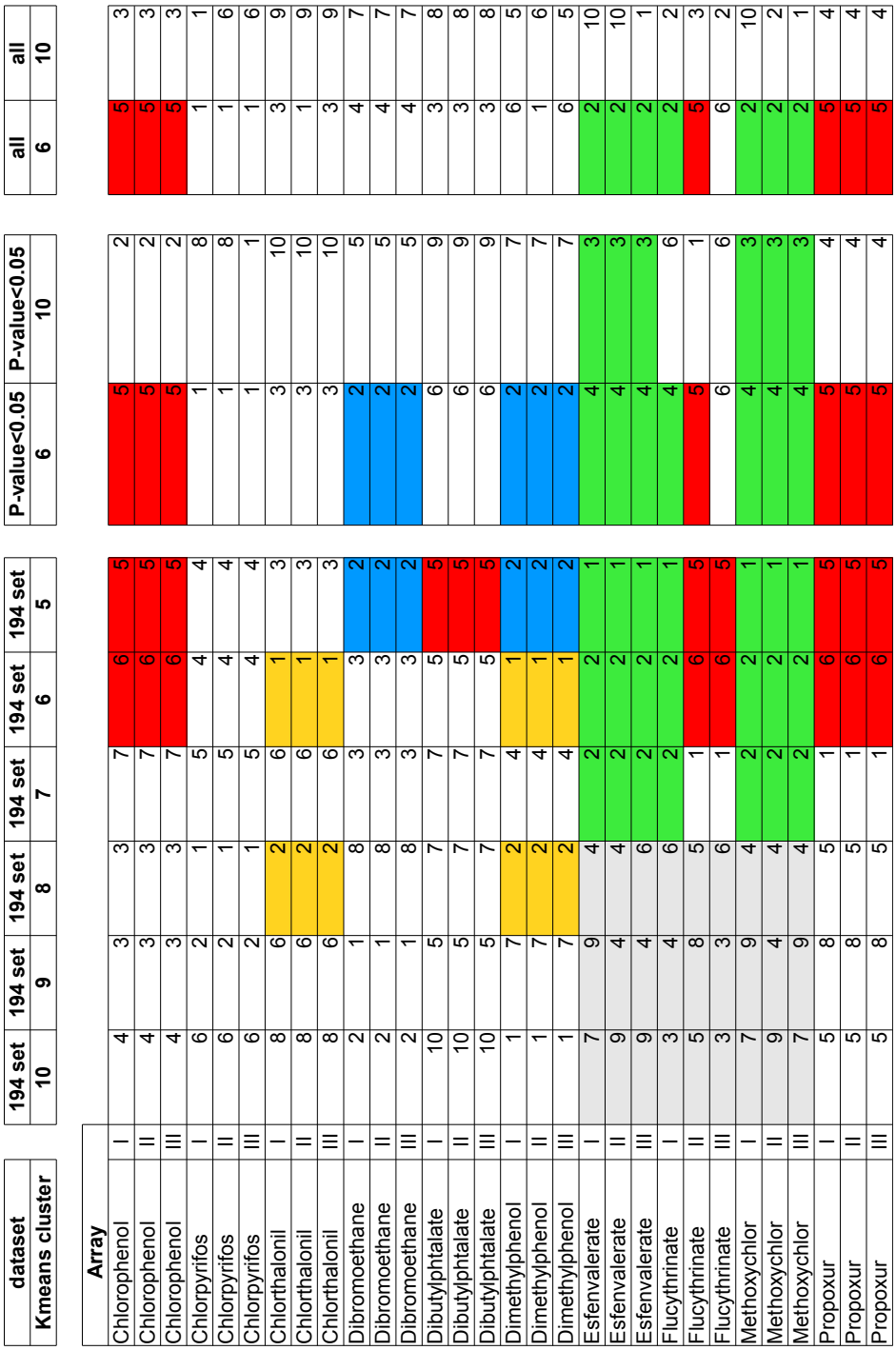


Figure 5.6: Results of the K-means cluster analysis for the three data sets, *194*, *p-value<0.05* and the *all*. The row K-means cluster indicates the pre-specified number of clusters. Since with every calculation the assignment of the cluster number changes, reoccurring compound clusters were color labeled.

K-means cluster analysis was performed for all three data sets. The results are shown in Figure 5.6. In the *all* data set, 5 compounds show a perfect clustering of the replicates if the number of clusters is set to 10. If the number of clusters is defined as 6, all replicates of two compounds cluster together. The main clusters for *all* are:

- chlorophenol and propoxur (red)
- methoxychlor and esfenvalerate (green)

In the *p-value 0.05* data set, 6 compounds show a perfect clustering of all replicates when 10 clusters are used. In the case of 6 cluster, three compound clusters can be detected. The main clusters for *p-value 0.05* are:

- Chlorophenol and propoxur (red)
- dibromoethane and dimethylphenol (blue)
- methoxychlor and esfenvalerate (green)

If 10 clusters were chosen, the *194* data set shows clustering of all replicates for 7 compounds. For 6 clusters, 3 compound-specific clusters are found. The main clusters for *194* are:

- chlorophenol and propoxur (red)
- chlorthalonil and dimethylphenol (yellow)
- methoxychlor and esfenvalerate (green)

Summary

To study the similarity of the expression patterns, a variety of statistical analysis methods was used. This was done to improve the quality of this analysis step. Each algorithm has its own characteristic of clustering the data. Moreover, the use of different statistical parameters (e.g distance measures) for a method can result in a completely different clustering result. Therefore, the occurrence of the same clusters in the results of different analysis methods clearly underlines the value of these clusters. In Figure 5.7 an overview of the results from the different methods is shown.

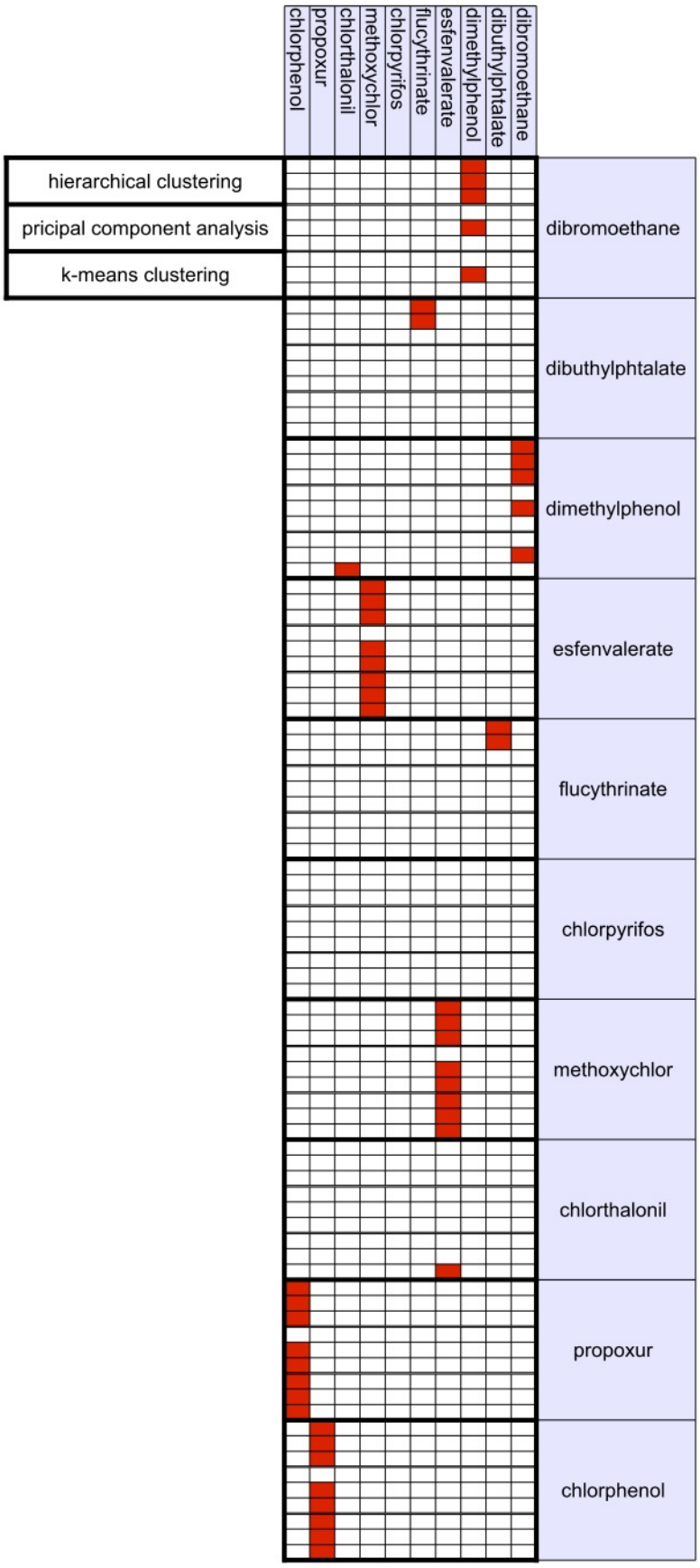


Figure 5.7: Overview over the results of the comparative analysis. Red boxes indicate that the two compounds clustered together. Each cluster method is represented with three boxes per compound, representing the three data sets used. Whereas *all* is the most left one, *p-value 0.05* the middle one and *194* the most right box.

As conclusion, a similarity between the following groups of compounds can be assumed:

- chlorophenol and propoxur
- dibromoethane and dimethylphenol
- methoxychlor and esfenvalerate

5.1.2 Co-regulated Genes

When compounds share a similar toxicity mechanism, inducing e.g. a certain pathway, they should express the same genes. Here I will perform a co-regulation analysis. In Chapter 5.1.1, I studied the similarity of the expression patterns using multivariate statistical analysis. This method uses all expression values of a defined data set. Therefore, the similarity is based on the similarity of the expression patterns. Whether compounds really share a similar toxicity mechanism or display the same levels of general toxicity response (immune system reaction, repair mechanisms) is unclear.

In the following, I will use the term co-regulation to describe genes that show an expression in response to exposure by several compounds. The direction of regulation (up or down regulation) is not taken into account.

In Table 5.1, a summary of all co-regulated transcripts is shown. No transcripts were found to be co-regulated by all compounds. Therefore no general toxicity response gene could be detected. The most expressed transcripts were predominantly compound-specific and not co-regulated by other compounds. The number of co-regulated transcripts is clearly decreasing when the number of compounds increases.

# compounds	total	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
194 set	194	111	49	24	3	5	2	0	0
P-value < 0.05	14394	5603	5060	965	306	60	15	3	0
M-value >1.4	2763	2118	452	146	44	3	0	0	0

Table 5.1: Number of co-regulated transcripts. A transcript is categorized as differentially expressed if the calculated p-value is smaller than 0.05. In the columns the numbers of transcripts, which are regulated by one compound or co-regulated by 2 to 8 different compounds are shown.

In Table 5.2, the numbers of regulated transcripts that the compounds share with other compounds are displayed. For example, dibromoethane induced 1316 differentially expressed transcripts. Alone 164 transcripts, it has in common with dibutylphthalate, but not exclusively. Some of the 164 transcripts could also be included by dibromoethane and dibutylphthalate, or by another compound.

	dibromoethane	dibutylphhalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol
dibromoethane	**1316**	164	263	284	67	104	256	69	146	418
dibutylphthalate	164	**1718**	159	208	81	92	199	65	142	505
dimethylphenol	263	159	**129**3	258	88	105	240	70	206	566
esfenvalerate	284	208	258	**146**8	64	102	360	61	119	425
flucythrinate	67	81	88	64	**913**	34	48	26	66	316
chlorpyrifos	104	92	105	102	34	**683**	107	41	83	213
methoxychlor	256	199	240	360	48	107	**1482**	61	149	366
chlorthalonil	69	65	70	61	26	41	61	**471**	87	141
propoxur	146	142	206	119	66	83	149	87	**1339**	525
chlorophenol	418	505	566	425	316	213	366	141	525	**4510**

Table 5.2: Co-regulated transcripts. The table shows the number of differentially expressed transcripts which the ten compounds share with each other. The total number of differentially expressed transcripts of a compound is shown in bold.

The number of co-regulated transcripts is dependent on the total number of regulated transcripts of a compound. The more transcripts a compound has differentially expressed, the higher the probability is that it shares transcripts with other compounds. To be able to identify groups of compounds with an enriched number of co-regulated transcripts, I calculated the percentage of all differentially expressed transcripts the compounds share with each other.

In Table 5.3, the percentage of co-regulation is shown. Dibromoethane shares 76.75 % of its regulated transcripts with other compounds and 12.46% with dibutylphthalate. Whereas dibutylphthalate shares 9.55 % of its regulated transcripts with dibromoethane. A group of compounds shows an enrichment of co-regulated transcripts only when all compounds of that group show a higher number (> mean + 1*std) of co-regulated transcripts.

Two groups of compounds with an enriched co-regulation were detected:

- esfenvalerate, methoxychlor (green numbers)
- chlorophenol, dimethylphenol (blue numbers)

	dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol
dibromoethane	**100**	9.55	20.34	19.35	7.34	15.23	17.27	14.65	10.9	9.27
dibutylphthalate	12.46	**100**	12.3	14.17	8.87	13.47	13.43	13.8	10.6	11.2
dimethylphenol	19.98	9.25	**100**	17.57	9.64	15.37	16.19	14.86	15.38	12.55
esfenvalerate	21.58	12.11	19.95	**100**	7.01	14.93	24.29	12.95	8.89	9.42
flucythrinate	5.09	4.71	6.81	4.36	**100**	4.98	3.24	5.52	4.93	7.01
chlorpyrifos	7.9	5.36	8.12	6.95	3.72	**100**	7.22	8.7	6.2	4.72
methoxychlor	19.45	11.58	18.56	24.52	5.26	15.67	**100**	12.95	11.13	8.12
chlorthalonil	5.24	3.78	5.41	4.16	2.85	6	4.12	**100**	6.5	3.13
propoxur	11.09	8.27	15.93	8.11	7.23	12.15	10.05	18.47	**100**	11.64
chlorophenol	31.76	29.39	43.77	28.95	34.61	31.19	24.7	29.94	39.21	**100**
mean + 1*std	23.86	18.14	28.41	23.25	19.25	21.84	21.36	21.48	23.1	11.73
total # co-regulation	76.75	63.39	82.52	75.95	56.19	74.52	73.75	73.46	69.75	52.04

Table 5.3: Percentage of co-regulated transcripts. The columns show the percentage of differentially expressed transcripts a compound shares with other compounds. The colored numbers indicate groups of compounds where all compounds have a high (> mean + 1*std) number of co-regulated transcripts.

	dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol
dibromoethane	**30.93**	2.85	2.55	3.68	1.2	3.07	3.24	3.4	2.54	2.33
dibutylphthalate	3.72	**35.45**	1.86	4.02	3.18	4.25	4.39	4.25	3.14	5.43
dimethylphenol	2.51	1.4	**31.71**	3.07	1.64	2.05	2.63	3.4	2.84	3.44
esfenvalerate	4.1	3.43	3.48	**35.22**	2.19	3.66	7.35	4.03	1.57	2.86
flucythrinate	0.84	1.69	1.16	1.36	**31.33**	0.59	1.08	1.7	1.19	3.19
chlorpyrifos	1.6	1.69	1.08	1.7	0.44	**31.04**	1.42	1.91	1.42	1.22
methoxychlor	3.65	3.78	3.02	7.43	1.75	3.07	**34.41**	4.46	3.44	2.35
chlorthalonil	1.22	1.16	1.24	1.29	0.88	1.32	1.42	**36.73**	1.19	1.06
propoxur	2.58	2.44	2.94	1.43	1.75	2.78	3.1	3.4	**34.58**	4.63
chlorophenol	7.98	14.26	11.99	8.79	15.77	8.05	7.15	10.19	15.61	**28.69**
mean + 1*std	5.28	7.72	6.65	6.39	7.98	5.35	5.89	6.56	8.22	4.39

Table 5.4: Percentage of co-regulated transcripts between two compounds. The columns show the percentage of differentially expressed transcripts a compound shares particularly only with one other compound. The colored numbers indicate compound groups with a high (> mean + 1*std) number of co-regulated transcripts.

In order to get a better overview whether two compounds have an enriched number of co-regulated transcripts, I restricted the list of transcripts on only the ones that were shared between two compounds. The percentage of co-regulation, specific for only two compounds can be seen in Table 5.4.

Regarding co-expression that is specific for two compounds, three groups are above average:

- esfenvalerate, methoxychlor (green numbers)
- propoxur, chlorophenol (red numbers)
- chlorophenol, dibutylphthalate (blue numbers)

Chlorophenol and dimethylphenol have only an enriched co-regulation when all differentially expressed transcripts were taken into account. This means that they have a co-regulation but parts of that transcripts were also regulated through other compounds. Propoxur and chlorophenol as well as chlorophenol and dibutylphthalate show an enriched co-regulation based only on transcripts regulated in these compounds. Therefore, it can be assumed that the underlaying mechanisms are specific for these compounds.

Gene Function Analysis

To get a better understanding of the mechanisms, gene function analysis was performed as described in Chapter 4.4. Since the numbers of co-regulated transcripts are in some cases not that high, significant results (p-value < 0.05) could not be found for all co-regulated compound groups.

In Table 5.5 the significant enriched categories of the gene function analysis from the co-regulated transcripts of methoxychlor and esfenvalerate is shown. In Table 5.6 the results for the co-regulated transcripts, only regulated in this compounds is presented.

Gene ontology molecular function	**KEGG**	**WikiPathways**
GTPase activity GO:0003924	Proteasome	Proteasome Degradation
isomerase activity GO:0016853	Gap junction	
	Fatty acid metabolism	

Table 5.5: Co-regulated in methoxychlor and esfenvalerate. Result of the gene function analysis (p-value < 0.05) for the co-regulated transcripts of methoxychlor and esfenvalerate.

The results of the gene function analysis for the co-regulated transcripts of chlorophenol and dimethylphenol are shown in Table 5.7.

The interpretation of the results and the link to the cluster analysis is done in the discussion part of my thesis (Chapter 6.1.2).

WikiPathways
IL2 Signaling Pathway
IL6 Signaling Pathway
SIDS Susceptibility Pathways
Proteasome Degradation

Table 5.6: Co-regulated only in methoxychlor and esfenvalerate. Result of the gene function analysis (p-value < 0.05) for the co-regulated transcripts specific for methoxychlor and esfenvalerate.

Gene Ontology biological process	**KEGG**
cell cycle GO:0007049	Cell cycle
cellular response to DNA damage stimulus GO:0034984	
cellular response to stress GO:0033554	
nitrogen compound metabolic process GO:0006807	
response to DNA damage stimulus GO:0006974	
DNA metabolic process GO:0006259	
nucleobase/nucleoside/nucleotide/nucleic acid metabolic process GO:0006139	
Gene ontology molecular function	
nucleotide binding GO:0000166	
nucleic acid binding GO:0003676	
purine nucleotide binding GO:0017076	
ligase activity, forming carbon-nitrogen bonds GO:0016879	
ribonucleotide binding GO:0032553	
purine ribonucleotide binding GO:0032555	
ligase activity GO:0016874	
adenyl nucleotide binding GO:0030554	
purine nucleoside binding GO:0001883	
DNA binding GO:0003677	
nucleoside binding GO:0001882	
ATP binding GO:0005524	
adenyl ribonucleotide binding GO:0032559	
binding GO:0005488	
protein serine/threonine kinase activity GO:0004674	
polo kinase kinase activity GO:0042801	
endonuclease activity GO:0004519	
acid-amino acid ligase activity GO:0016881	

Table 5.7: Chlorophenol and dimethylphenol. Significantly enriched categories (p-value < 0.05) for the co-regulated transcripts of chlorophenol and dimethylphenole.

5.1.3 Intensity Distribution Analysis

To further investigate the gene expression changes in response to the different treatments, I decided to have a closer look on the overall intensity distribution. Therefore, I compared for all treatments, the number of differentially expressed transcripts and their intensity levels. High numbers of differentially expressed transcripts indicate also a higher number of disturbed pathways. This could be a sign of a more non specific toxicity response (immune system ,apoptosis). Whereas a small number of regulated transcripts might be the result of a more specific response to the compound. In Figure 5.8, the number of differentially expressed transcripts for each compound is shown. The numbers were calculated based only on the transcripts that could be perfectly mapped to Zv8 (Chapter 4.6).

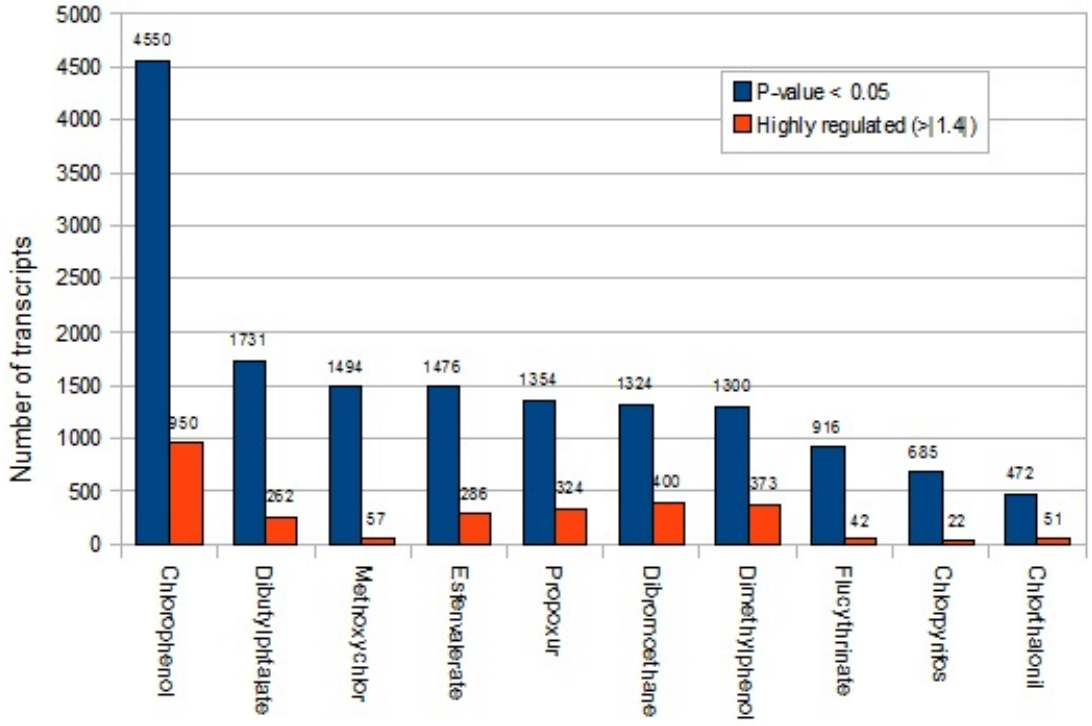


Figure 5.8: Number of differentially expressed transcripts. Only transcripts were counted which could be properly annotated (See Chapter 4.6).

I also compared the maxima of the M-values and the distribution between gene up- and down-regulation of the different toxicants. In Figure 5.9 the maximum and minimum M-values for each compound are shown. The distributions of the differently regulated transcripts (P-value < 0.05), for each compound is presented in Figure 5.10.

The chlorophenol data set show by far the highest number of differentially regulated transcripts (4550). In comparison, after treatment with chlorpyrifos and chlorthalonil fewer than 1000 transcripts were detected as being differentially expressed. Accordingly,

their number of highly regulated transcripts is also comparatively small. It is striking that methoxychlor has the third largest number of regulated transcripts, but one of the smallest numbers of highly regulated transcripts. It also shows very small maximum and minimum M-values. The signal distribution can be described as very broad and flat with an higher number of up-regulated transcripts than down-regulated ones. This cannot be explained by the small number of highly regulated transcripts. Chlorthalonil also shows high maximum values, although only a few transcripts were regulated. The treatment with propoxur leads to the highest maximum and minimum M-values and an average number of regulated genes. For most compounds, the signal distribution of the highly regulated transcripts was very symmetric between up and down regulation. On the contrary, chlorthalonil and dibutylphthalate show an increase in the up-regulated M-values. Chlorpyrifos induced more down-regulated transcripts, but the maxima were similar for up- and down-regulation.

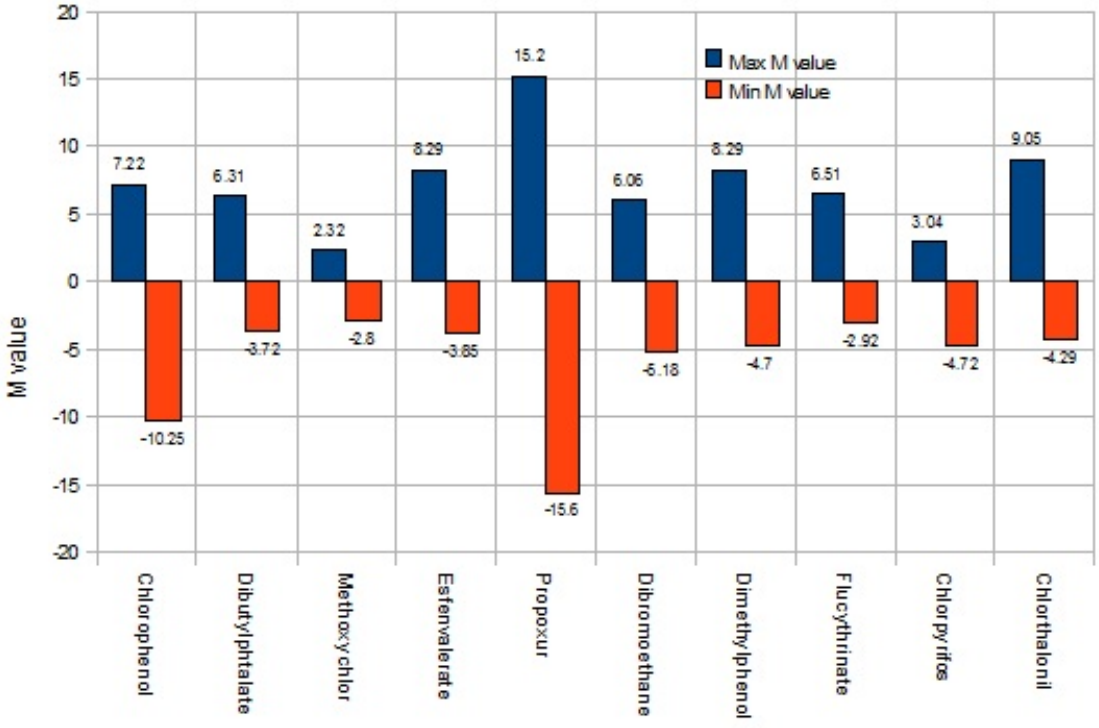
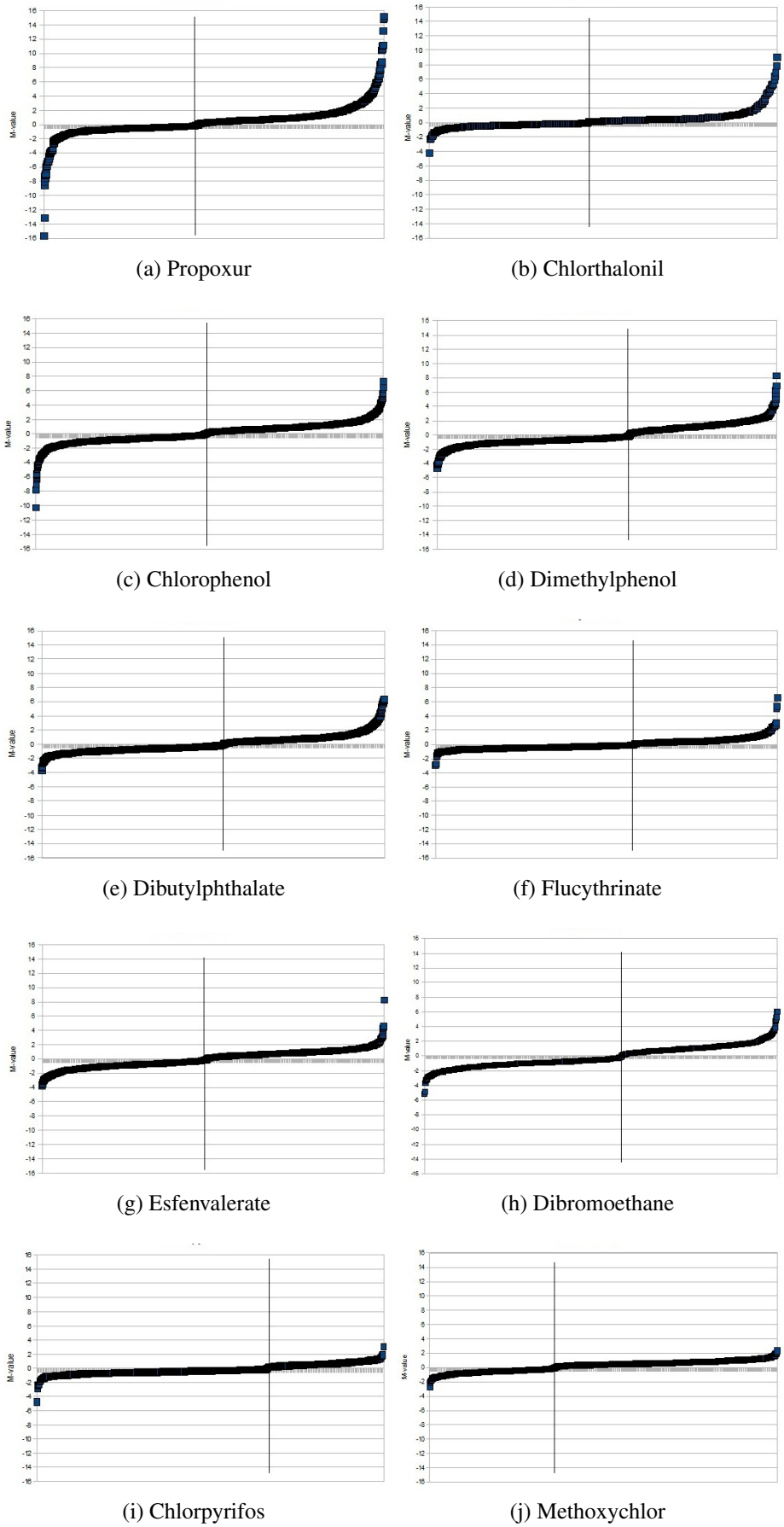


Figure 5.9: Maximun and Minimum M-values. Only transcripts were counted which could be properly annotated (See Chapter 4.6).



(a) Propoxur (b) Chlorthalonil

(c) Chlorophenol (d) Dimethylphenol

(e) Dibutylphthalate (f) Flucythrinate

(g) Esfenvalerate (h) Dibromoethane

(i) Chlorpyrifos (j) Methoxychlor

Figure 5.10: Intensity distribution of the differentially expressed transcripts (p-value < 0.05)

5.1.4 Linkage with other Microarray Studies

With the intention to further investigate the underlying mechanisms influencing the gene expression patterns, I linked other data sets to our data. Since these data sets were not produced using our microarray system, I mapped the transcripts from the second system to our array using FileMaker. Knowing that the signal values cannot be compared directly, I focused on the classification into up- or down-regulated as described in the corresponding publications (Yang *et al.* 2007,Stockhammer *et al.* 2009).

Biosensor Data

This data set was used in the past in our lab for studying different expression patterns of several compounds in dependency to developmental stages and compound concentrations (Yang *et al.* 2007). The data has been published in NCBI GEO as series GSE9357. For my comparison, I only used the data from treatments which were performed similar to the 10 compound study (24-48 hpf.). As was shown in the publication by Yang et al, the expression patterns are different depending on the developmental stage and the exposure scenario. The list of compounds and the treatment concentrations, of the data sets I used for the analysis is shown in Table 5.8.

Compound	Stage	Concentration
4-chloroaniline	48 hours	40 ppm
CdCL (cadmium chloride)	48 hours	5 ppm
DDT (dichlordiphenyltrichlorethan)	48 hours	15 ppm
TCDD (2,3,7,8-tetrachlordibenzo-p-Dioxin)	48 hours	500 ppt
Valproic Acid (2-propylpentanoic acid)	48 hours	50 ppm
MeHg (di-methyl mercury)	48 hours	60 ppb

Table 5.8: Biosensor compounds

As cut-off for the identification of regulated transcripts, a p-value < 0.025 and a logarithmic fold change > $|1.5|$ was used as described in the publication (Yang *et al.* 2007). A summary of the data is presented in Table 5.9.

Compound	DDT	Valproic Acid	TCCD	CdCl	MeHg	4-Chloroaniline
# regulated transcripts	280	98	992	16	556	30
Max ln(FC) value	2.95	6.33	6.45	2.83	3.86	4.75
Min ln(FC) value	-4.26	-2.29	-2.45	-1.99	-5.86	-4.75

Table 5.9: Compugen data. The number of differentially expressed transcripts (p-value < 0.025, ln(FC)>$|1.5|$) and the maxima of the ln(FC) values for each compound.

For the biosensor data the Compugen zebrafish microarray was used. This arrays consists of 16384 oligonucleotides of which 8125 could be mapped to Zv8 (see Chapter 4.6). 807 genes present on the Compugen array were not on the Agilent array employed in our study. 7256 oligonucleotides could be mapped to the Agilent 4x44k zebrafish v2 array. These oligonuceotides were used to link the Compugen data set with the 10 Compound data set. The linked data can be found in the comparison_data table on the supplementary CD.

To get a better understanding of the similarity of the gene expression patterns from the linked data sets, I performed a co-regulation analysis as described in Chapter 5.1.2. Table 5.10 delineates the percentage of transcripts that a compound from the biosensor study shares with the 10 compound study. For cadmium chloride this means, that 20.63 % of its differentially expressed transcripts were also differentially expressed in dibromoethane. In Table 5.11 the percentage of transcripts a compound of the 10 compound study shares with the biosensor compounds is shown. Based on this table, 0.99 % of the differentially expressed transcripts form dibromoethane were also differentially expressed in cadmium chloride.

The higher values for chlorophenol, MeHg and TCDD are based on their higher numbers of differentially expressed transcripts. They share higher numbers of co-regulated transcripts with nearly all other compounds. This might be an indication that the mechanism is of a more general toxicity response (e.g. immune system or apoptosis).

	CdCL	DDT	4-Chloroaniline	MeHg	TCDD	Valproic Acide
dibromoethane	**20.6**3	12.26	8.33	18.29	14.27	20.86
dibutylphthalate	12.70	8.71	8.33	10.86	11.41	9.82
dimethylphenol	17.46	16.13	8.33	16.19	9.99	17.79
esfenvalerate	9.52	9.35	8.33	15.05	11.15	14.72
flucythrinate	3.17	5.48	4.17	4.76	4.41	4.29
chlorpyrifos	11.11	8.39	4.17	5.71	6.36	8.59
methoxychlor	14.29	12.26	4.17	17.71	12.84	18.40
chlorthalonil	6.35	2.90	4.17	5.33	3.63	5.52
propoxur	12.70	9.03	12.50	8.57	8.30	5.52
chlorophenol	**33.33**	**28.06**	**41.67**	**29.33**	**24.38**	**37.42**
mean + 1*std	22.54	18.21	21.74	20.89	16.64	24.37

Table 5.10: Percentage of co-regulated transcripts. The columns show the percentage of differentially expressed transcripts a compound from the biosensor data set shares with the 10 compound study. The bold numbers indicate compounds with a high (> mean + 1*std) number of co-regulated transcripts.

	Dibromoethane	Dibutylphthalate	Dimethylphenol	Esfenvalerate	Flucythrinate	Chlorpyrifos	Methoxychlor	Chlorthalonil	Propoxur	Chlorophenol
CdCL	0.99	0.47	0.85	0.41	0.22	1.02	0.61	0.85	0.60	0.47
DDT	2.89	1.57	3.87	1.98	1.86	3.81	2.56	1.91	2.09	1.93
4-Chloroaniline	0.15	0.12	0.15	0.14	0.11	0.15	0.07	0.21	0.22	0.22
MeHg	**7.29**	3.32	**6.57**	**5.38**	2.74	4.39	**6.28**	**5.94**	3.36	3.41
TCDD	**8.36**	**5.12**	**5.96**	**5.86**	**3.72**	**7.17**	**6.68**	**5.94**	**4.78**	**4.17**
Valproic Acide	2.58	0.93	2.24	1.63	0.77	2.05	2.02	1.91	0.67	1.35
mean + 1*std	7.07	3.85	5.92	5.04	3.03	5.66	5.86	5.32	3.77	3.51

Table 5.11: Percentage of co-regulated transcripts. The columns show the percentage of differentially expressed transcripts a compound from the 10 compound study shares with the compounds from the biosensor data set. The bold numbers indicate compounds with a high (> mean + 1*std) number of co-regulated transcripts.

TCDD and chlorophenol are the only compounds which showed an enriched co-regulation for each other. They have 188 transcripts co-regulated. Based on this list gene function analysis was performed. The results are shown in Table 5.12.

WikiPathways	Gene Ontology biological process
FGF signaling pathway	negative regulation of cellular process GO:0048523
canonical wnt - zebrafish	tube morphogenesis GO:0035239
	multicellular organismal development GO:0007275
	developmental process GO:0032502
	negative regulation of biological process GO:0048519

Table 5.12: Gene function analysis for TCDD and chlorophenol co-regulated transcripts. Enriched categories were significant with a p-value < 0.05.

Cadmium chloride shows an enriched co-regulation with dibromoethane and leads to the assumption that they might share a mechanism. In Table 5.13 the co-regulated genes from cadmium chloride and dibromoethane are shown. Since the number of genes is so small no further analysis could be performed.

Ensembl Gene ID	Gene Name	Ensembl Description
ENSDARG00000006900	impdh2	inosine-5'-monophosphate dehydrogenase 2
ENSDARG00000011989	crx	cone-rod homeobox
ENSDARG00000016301	zgc:65894	hypothetical protein LOC335798
ENSDARG00000032619	tob1a	transducer of ERBB2, 1a
ENSDARG00000036427	slc3a2	solute carrier family 3, member 2
ENSDARG00000036834	zgc:109868	cytokeratin-like
ENSDARG00000041394	dnajb1b	DnaJ (Hsp40) homolog, subfamily B, member 1
ENSDARG00000043561	psmc1b	proteasome(prosome/macropain) 26S subunit,ATPase,1b
ENSDARG00000058039	bhlhe22	class E basic helix-loop-helix protein 22
ENSDARG00000059053	zgc:162495	solute carrier family 13 member 4

Table 5.13: Co-regulated genes from cadmium chloride and dibromoethane.

Immune Response Data

It can be assumed that a part of the gene expression changes found after exposure with a specific compound are the result of reactions of the immune system of the organism. These reactions represent a more general response and no toxicity-specific mechanism. In order to get a better understanding of the compound-specific reactions in the organism, it would be an advantage to be able to filter the transcripts belonging to the immune system from the expression data. Therefore, I used a list of genes that was published in 2009 by Stockhammer *et al*. (Stockhammer *et al.* 2009). In this paper, they defined a transcriptional profile of the innate immune system in the zebrafish embryo after *Salmonella* infection. The infections were performed between 27 hpf and 48 hpf, which represents a similar time point as used in our toxicity experiments (24-48 hpf). They also used wild type zebrafish (AB-strain) for their experiments. As no sequence-information of the oligonucleotides used by Stockhammer *et al*. was available, I linked the published list of genes expressed after infection (Supplementary Table II) to our data via id-translation. To this end, I extracted the Unigene identifiers and mapped them to Ensembl gene identifiers. This procedure resulted in 1649 up-regulated genes and 1848 down-regulated genes. 2841 genes could be mapped to the Agilent zebrafish v2 array. It cannot be excluded that some of the genes might not be exclusively part of the reaction of the immune system, but nevertheless, this data set can give an overview of the general immune response. In Figure 5.11, the percentage distribution of the immune system related genes in the regulated (p-value < 0.05) and highly regulated (p-value < 0.05 and M > $|1.4|$) data sets are presented. The percentage of immune response genes in the highly regulated data set is for all compounds always higher than in the regulated data set. This indicates that the compounds induced a strong (M > $|1.4|$) reaction of the immune system. To get an better understanding of this effect an enrichment analysis (Chapter 4.3) for the immune response

genes was performed.

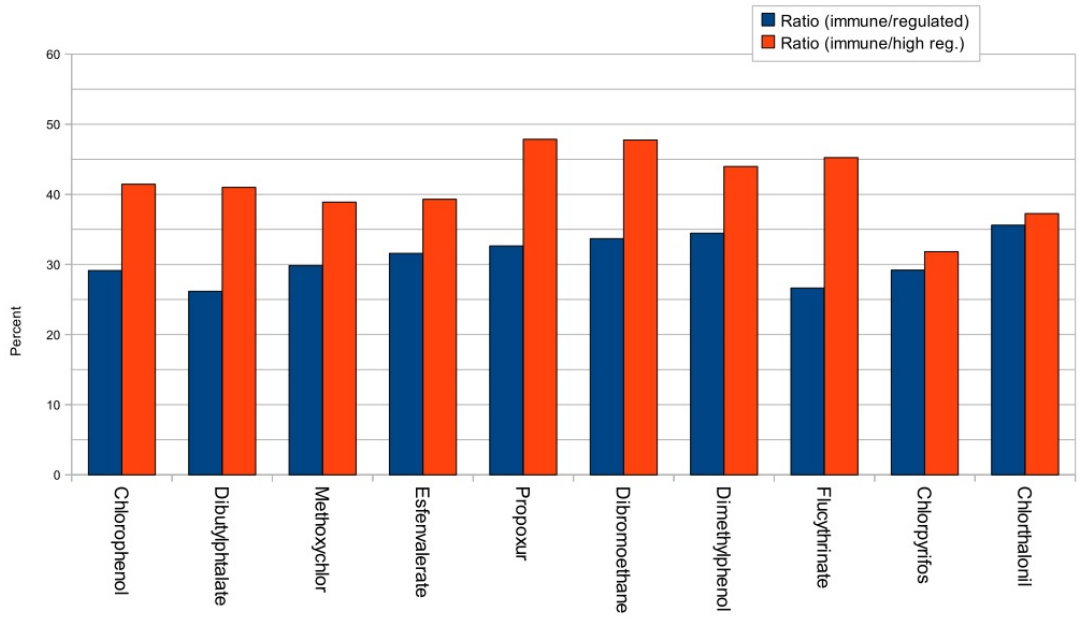


Figure 5.11: Overview over the induced immune response genes. The bars represent the percentage of genes, which could be linked to the immune system, of the regulated (p-value < 0.05) and highly regulated data sets (p-value < 0.05; M > $|1.4|$).

	Regulated data set				
	chlorophenol	dibutylphthalate	methoxychlor	esfenvalerate	propoxur
Ratio of enrichment	1.43	1.28	1.46	1.55	1.6
P-value	2.51E-57	8.40E-10	1.39E-19	5.75E-26	5.04E-28
	dibromoethane	dimethylphenol	flucythrinate	chlorpyrifos	chlorthalonil
Ratio of enrichment	1.65	1.69	1.31	1.43	1.75
P-value	1.07E-31	1.89E-34	1.97E-06	1.60E-08	4.53E-101
	Highly regulated data set				
	chlorophenol	dibutylphtalate	methoxychlor	esfenvalerate	propoxur
Ratio of enrichment	2.03	2.01	1.91	1.93	2.35
P-value	1.44E-51	1.94E-14	1.40E-03	1.48E-13	8.09E-29
	dibromoethane	dimethylphenol	flucythrinate	chlorpyrifos	chlorthalonil
Ratio of enrichment	2.34	2.16	2.22	1.56	1.83
P-value	6.45E-35	2.77E-25	2.45E-04	1.43E-01	4.00E-03

Table 5.14: Enrichment statistics of the immune response genes for the 10 compounds. An p-value < 0.05 shows that the enrichment of the immune response genes in a data set is statistically significant. Ratio of enrichment values > 1 indicate an over representation of immune response genes in the data set, compared to what would be expected by chance.

In Table 5.14, the ratios of the enrichment analysis of the immune response genes is shown. The corresponding p-value proves statistically that the number of immune response genes is truly enriched in the regulated and highly regulated data set. Only for the highly regulated data set of chlorpyrifos, the p-value is above 0.05. This, however, might be due to the fact that it shows the smallest number of highly regulated transcripts (22 transcripts). In the highly regulated data sets, the enrichment ratio was higher than in the regulated one. This indicates that the general immune system reaction represents a main effect in the highest regulated genes. This shows that it is very important to investigate the immune response if a specific mechanism is searched. The immune response list is also included in the comparison_data Table on the supplementary CD.

I also performed a hierarchical cluster analysis for the immune response genes. The result is shown in Figure 5.12. The dendrogram looks very similar to the results of the cluster analysis performed in Chapter 5.1.1.

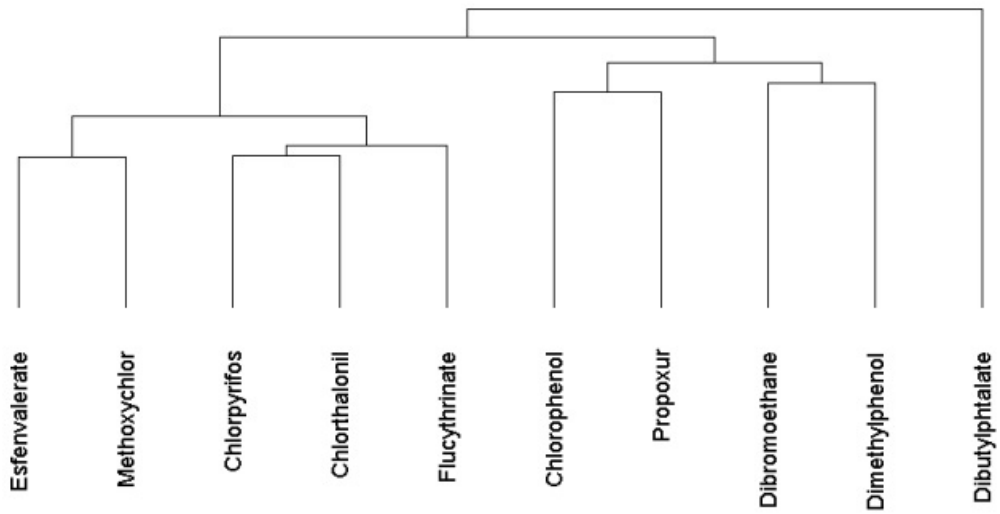


Figure 5.12: Result of the hierarchical cluster analysis. Performed only with the genes linked to the immune response.

5.1.5 Gene Set Analysis

To further investigate what happened in the treated organism, I decided to have a closer look on the 'death pathways' (apotosis, necrosis ans autophagy) and transcriptional processes. Apoptosis, necrosis and autophagy are of course very common in toxicity induced expression patterns. With this analysis, I wanted to get an idea how prominent this pathways are in the treatments. To study the transcriptional processes, I focused on the differentially regulation of the transcription factors.

Apoptosis, Necrosis and Autophagy Genes

Unfortunately, no data set was available containing a list of that genes. On that account, I checked Gene Ontology terms and the gene descriptions provided by Ensembl for the occurrence of the terms death and apotosis, necrosis or autophagy. The resulting list of genes was mapped to the Agilent zebrafish v2 array. For necrosis no genes could be identified and only 11 transcripts were linked to autophagy. Therefore, this two pathways were not further investigated. But 271 transcripts on the array could be linked to apoptosis. The percentage of 'apoptotic' transcripts for the significantly regulated gene sets (p-value < 0.05) for each compound is shown in Figure 5.13. Even if not all genes involved in apotic processes have been identified, this list should give a good overview of the general degree of apoptotic damage in the treated organism.

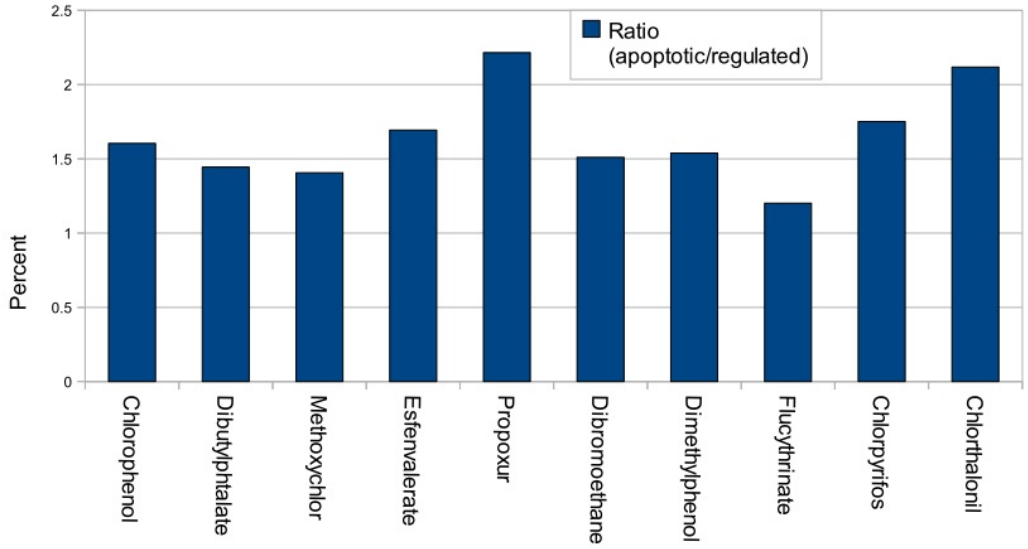


Figure 5.13: Overview over the induced apoptotic Genes. The bars represent the percentage of genes, which could be linked to apoptosis in the regulated (p-value < 0.05) data set.

The percentage of apoptotic transcripts was very low for all compounds. Although chlorophenol showed a large number of differentially expressed transcripts, there was no enrichment of apoptotic genes detectable as compared to the other compounds. In Table 5.15, the results from the enrichment analysis (Chapter 4.3) of the apotosis genes are shown.

		Regulated data set			
	chlorophenol	dibutylphthalate	methoxychlor	esfenvalerate	propoxur
Ratio of enrichment	1.29	1.16	1.13	1.36	1.78
P-value	1.06E-002	2.51E-001	3.17E-001	7.44E-002	1.70E-003
	dibromoethane	dimethylphenol	flucythrinate	chlorpyrifos	chlorthalonil
Ratio of enrichment	1.21	1.23	0.96	1.4	1.7
P-value	2.19E-001	1.96E-001	5.95E-001	1.51E-001	7.31E-002

Table 5.15: Enrichment analysis for the apotosis genes. A p-value < 0.05 shows that the enrichment of the immune response genes in a data set is statistically significant. Ratio of enrichment values > 1 indicate an over representation of apoptosis genes in the data set, compared to what would be expected by chance.

Only for chlorophenol and propoxur a significant enrichment (P-value < 0.05) was found. This means that there are more apoptotic genes differentially expressed than would be expected by chance. Therefore, one can assume that the exposure concentrations of this compounds are in a range were apoptosis is induced. Nevertheless, no high enrichment was found, so the influence of apoptosis on the whole expression data set is small and other processes seem to be more prominent.

Transcription Factors

In order to investigate the regulation of transcriptional process by the compounds, a gene set analysis for transcription factors was performed. Therefore, a list of possible transcription factors was used (Chapter 2.1.3). 2626 transcripts related to transcriptional processes could be found on the Agilent v2 Array in total. Figure 5.14 gives an overview of the percentage of transcription factors in the different compound data sets.

Generally, less than 16% of the regulated transcripts belong to transcription factor genes. For methoxychlor and chlorthalonil the occurence of genes involved in transcription in the very highly differentially expressed transcripts was lower than for the other compounds. Other processes might be more important in these data sets than transcription. For dibromoethane, dimethylphenol and flucythrinate even more transcripts annotated with transcription were differentially expressed in the high regulated data set than in the regulated data set.

In Table 5.16 the results of the enrichment analysis are presented. Based on all differentially expressed transcripts, chlorophenol, dibutylphthalate, dimethylphenol, and chlorpyrifos showed a significant enrichment (p-value < 0.05) of transcriptional genes. If only the highly expressed transcripts were taken into account, only dibromoethane and dimethylphenol were statistically significant enriched for transcriptional processes.

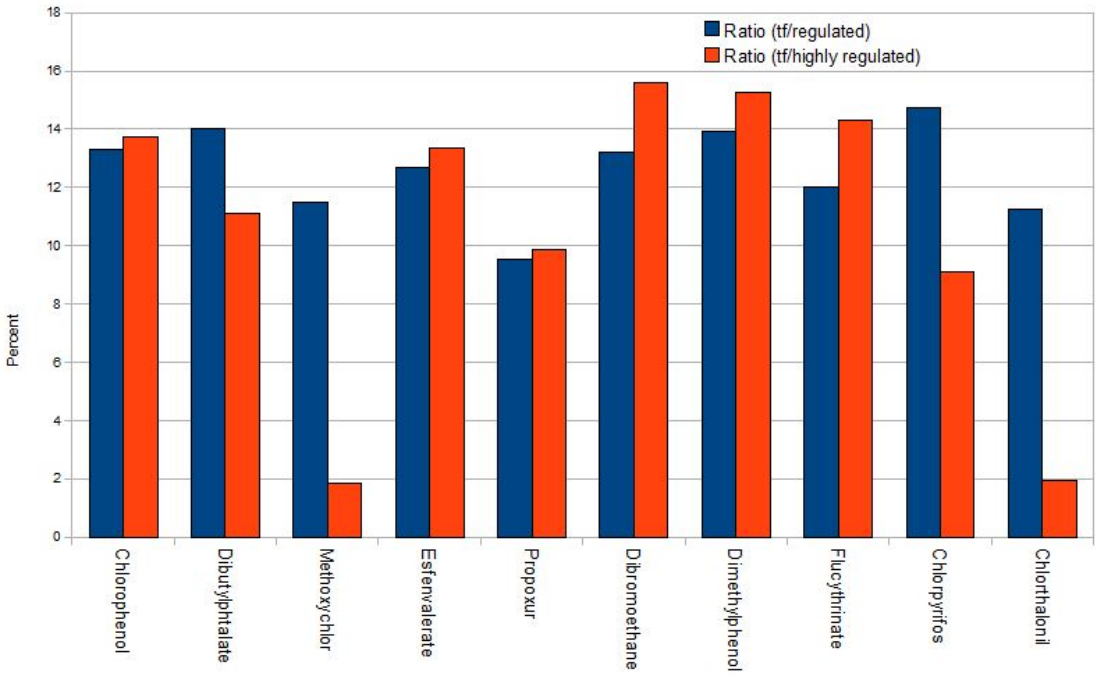


Figure 5.14: Overview over the induced transcription factors. The bars represent the percentage of genes involved in transcription, of the regulated (p-value < 0.05) and highly regulated data sets (p-value < 0.05; M > |1.4|).

		Regulated data set			
	chlorophenol	dibutylphthalate	methoxychlor	esfenvalerate	propoxur
Ratio of enrichment	1.10	1.16	0.95	1.05	0.79
P-value	2.50E-003	6.10E-003	7.73E-001	2.52E-001	9.99E-001
	dibromoethane	dimethylphenol	flucythrinate	chlorpyrifos	chlorthalonil
Ratio of enrichment	1.09	1.15	0.99	1.22	0.93
P-value	1.06E-001	2.19E-002	5.48E-001	1.97E-002	7.40E-001
		Highly regulated data set			
	chlorophenol	dibutylphthalate	methoxychlor	esfenvalerate	propoxur
Ratio of enrichment	1.14	0.92	0.15	1.10	0.82
P-value	6.36E-002	7.15E-001	9.99E-001	2.84E-001	9.09E-001
	dibromoethane	dimethylphenol	flucythrinate	chlorpyrifos	chlorthalonil
Ratio of enrichment	1.29	1.26	1.18	0.75	0.16
P-value	2.17E-002	3.71E-002	3.99E-001	7.64E-001	9.99E-001

Table 5.16: Transcription factor enrichment statistics. An p-value < 0.05 shows that the enrichment of the transcription factor genes in a data set is statistically significant. Ratio of enrichment values > 1 indicate an over representation of transcription factor genes in the data set, compared to what would be expected by chance.

5.1.6 Gene Function Analysis

For gaining a better understanding of the mechanisms in the gene expression patterns of the different compounds, a gene function analysis like described in Chapter 4.4 was performed. Since it is not clear where in the data set the information about the toxicity mechanism is located. A specific toxicity mechanism might be stronger induced than a general toxicity response. Therefore the gene set of the highly regulated transcripts might be better suited to find them. But it would also be possible that the complete set of differentially expressed transcripts is need to find the underlaying mechanisms. Pathways that show up- or down-regulation might be of higher interest than pathways that show a more mixed regulation. For this reason, I created several data sets and performed a gene function analysis of each of them. This should help to obtain more information and a better understanding of the regulation of specific pathways. The following data sets were used:

- *All:* All differentially expressed transcripts (p-value < 0.05).
- *All up:* All up-regulated transcripts.
- *All down:* All down-regulated transcripts.
- *Highly:* Highly regulated transcripts (p-value < 0.05, M > $|1.4|$).
- *Highly up:* Highly up-regulated transcripts.
- *Highly down:* Highly down-regulated transcripts.

The Gene Ontology and two pathway databases (KEGG and WikiPathwas) were used to find enriched functions or processes in the data sets. To improve the Gene Ontology analysis, the GO categories were summarized via similarity measures (Chapter 4.5). The results of the analysis for each compound can be found in the appendix Chapter A. The interpretation is done in the discussion of the individual compound results in Chapter 6.1.1.

5.2 Whole Genome Array

Here I want to address the question of the usability of the system and the problem of splitting the RNA samples.

5.2.1 Whole Genome Array versus Agilent Arrays

The whole genome design (Chapter 2.1.2) I created, consists of two 44k Agilent arrays. Since there is clearly a higher cost and time factor of using two 44xk arrays instead of one, I wanted to determine if there is really an improvement through the new whole genome design. To perform the comparison, the arrays were annotated as described in Chapter 4.6 and only the genes and transcripts were counted that gave a significant and specific hit in the blast search. First, I checked the arrays in total. I included also the newest Agilent zebrafish v3 array, which was published in the middle of 2010.

	Whole Genome Array	Agilent v2	Agilent v3
Genes	21690	14869	17719
Transcripts	23873	15390	18609

Table 5.17: Comparison of whole genome array and Agilent arrays. For each array type the numbers of genes and transcripts are shown which gave an significant hit in the blast search. The whole genome array contains the most genes and transcripts.

If the complete gene lists are compared, the whole genome array is obviously better as it contains the largest number of genes and transcripts (Table 5.17). To evaluate the improvements for real microarray experiments, I compared the list of significantly differentially expressed genes from existing whole genome array experiments with the content of the Agilent arrays (Table 5.18).

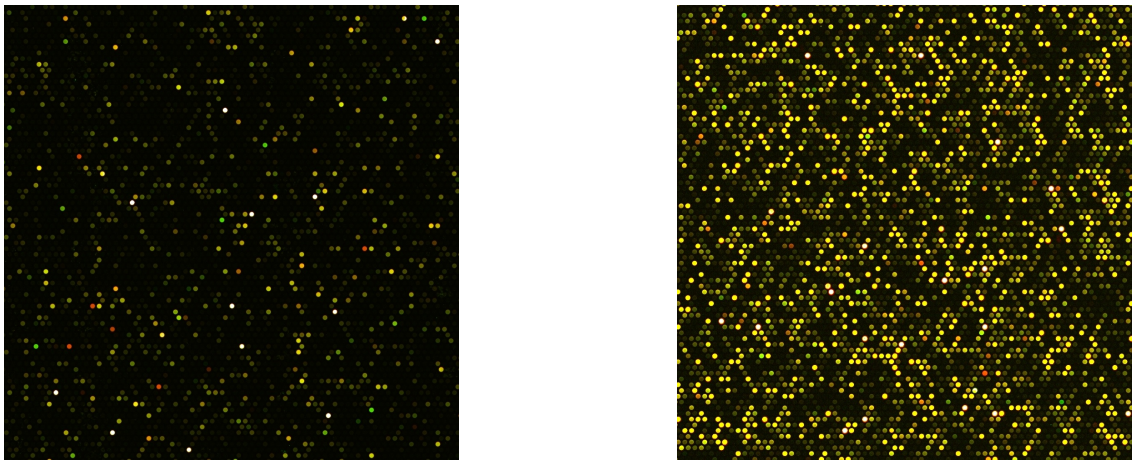
	Treatment A	Treatment B	Treatment C
Whole Genome Array			
Significantly regulated transcripts (p-value < 0.05)	679	467	897
Found on Agilent zebrafish v2	606	414	803
Improvement with Whole Genome Array in %	10.75	11.35	10.48
Found on Agilent zebrafish v3	644	441	852
Improvement with Whole Genome Array in %	5.15	5.57	5.02

Table 5.18: Comparison of whole genome array and the Agilent arrays. The significant regulated transcripts from an microarray experiment performed with the whole genome array were taken and compared based on there occurrence on the Agilent arrays. The whole genome arrays delivers 10 % more transcripts compared to the Agilent v2 array and around 5 % more than the Agilent v3 array.

The whole genome array leads to an improvement of around 10% compared to the Agilent v2 and still around 5% to the Agilent v3 array. An update of the whole genome array based on the new gene build Zv9 might further increase this factor.

5.2.2 Early Stages (10 hpf)

Not much miroarray data is published for such early stages as we used with the whole genome array (10 hpf). The first problem was to receive enough labeled RNA sample for the experiments. Several RNA-extraction methods were tested and the Trizol extraction (Chapter 2.4.2) worked best and was therefore used for the experiments. In the next step, we had to evaluate whether we receive enough signals to perform a microarray analysis. The normalization methods in general assume that most of the data comes from genes with no differentially expression between the treatment and the control. When only a few genes are expressed at this early stage, this might render the normalization of the data nearly impossible. The Figures 5.15a and 5.15b were made utilizing our Axon Scanner with comparable settings and a similar amount of sample RNA. In the early stage sample, clearly less spots are seen but still enough to carry out a microarray analysis. Importantly, the most spots are yellow, indicating similar gene expression in sample and control, so the normalization algorithms should work. The quality control plots of the data produced during the microarray analysis looked also normal. Based on the scanner images and the quality plots, the early stages seemed to be no problem for the microarray analysis.



(a) Early stage (10 hpf)

(b) Later stage (48 hpf)

Figure 5.15: Microarray scanner pictures of two different sample stages

5.2.3 Splitting RNA Samples

Since each microarray experiment with the whole genome array consists of two arrays, this is also problematic regarding the sample treatment. To resolve this situation, we simply used the two color control design (Chapter 2.2.1) and split our RNA samples in two equal parts and put similar amounts of RNA onto the two corresponding arrays. Consequently, the same RNA sample is used for the two arrays of one experiment. It has been postulated that splitting RNA samples over several arrays might introduce some

errors as the sample can never be completely homogeneous. To study the influence of these errors on our data, I performed several clustering analyses on the spike-in controls described in Chapter 2.1.1. Therefore, I used the spike-in control data from an whole genome array experiment with two replicates. The controls are used for all arrays and are added to the samples before the labeling process. Due to minimal pipetting differences, the amount of spike-in controls was always a little bit different for each microarray.

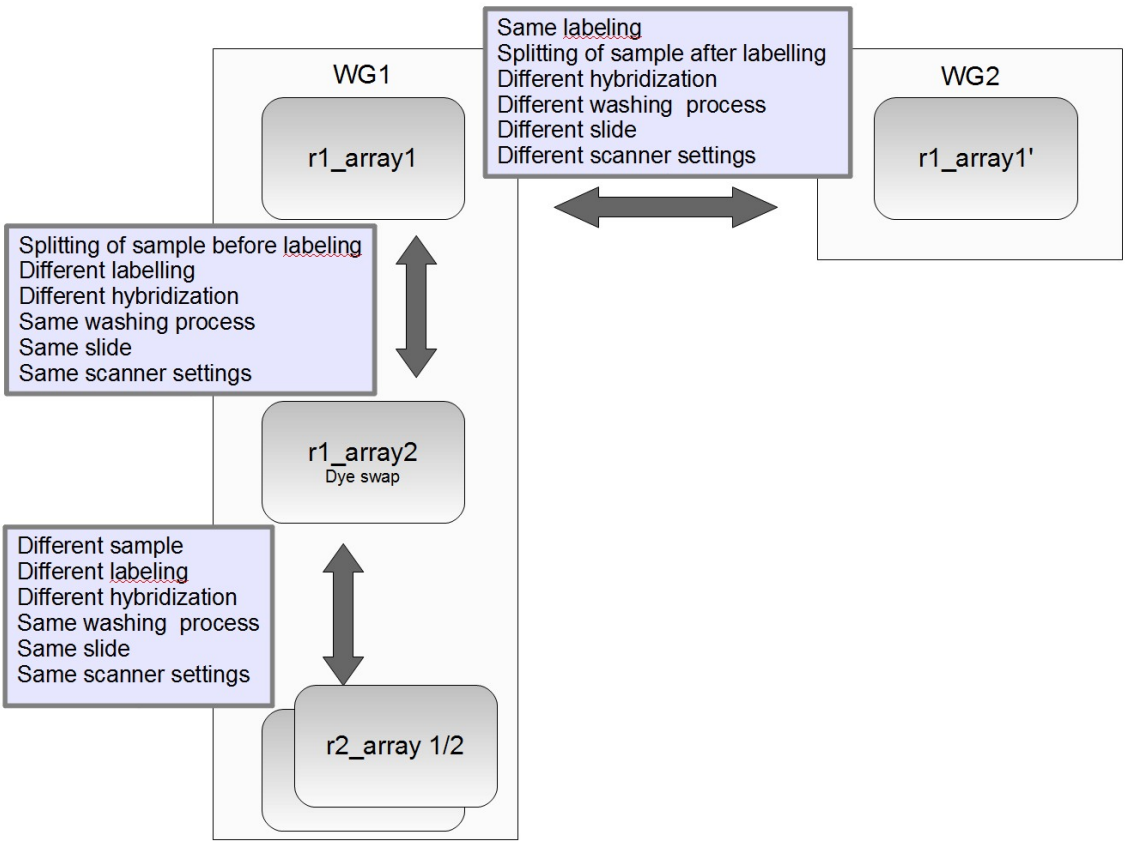


Figure 5.16: Overview on the similarities and differences between the arrays of an whole genome array experiment. WG1 and WG2 are the two slides belonging to the whole genome array. Each slide consists of 4 arrays. The replicates are labeled with r1 and r2.

In Figure 5.16 the similarities and differences of the arrays used in this microarray experiment are shown. Arrays belonging to a dye swap, have the same sample RNA but the sample RNA was split before the labeling process. On the contrary the sample RNA is split after the labeling process when used for the two whole genome arrays. The replicates consist of different sample RNAs but are hybridized on the same slide. To get a better understanding of the effect introduced through the RNA splitting, hierarchical cluster analysis was performed as described in Chapter 4.2.2. For the cluster analysis the signal data from the spike-in controls without any normalization or filtering was used. Only the M-values were calculated (Equation 4.1).

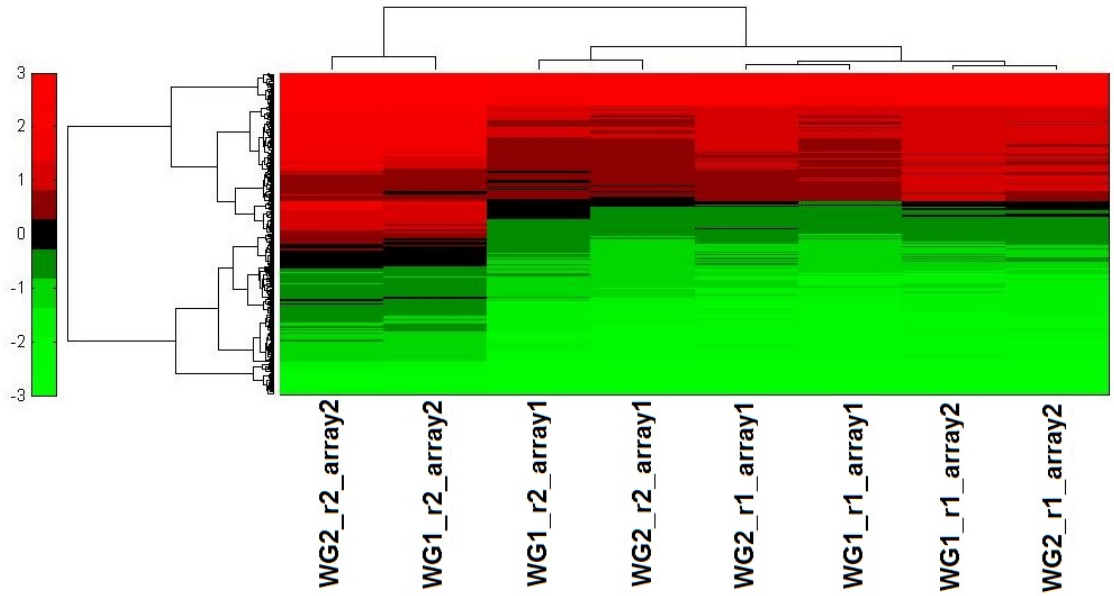


Figure 5.17: Cluster analysis of the spike-in control data of an experiment performed with the whole genome array. Each column represents one array. WG1 and WG2 are the two slides belonging to one whole genome array. The replicates are labeled with r1 and r2. The data was not normalized or filtered. The corresponding whole genome arrays cluster together.

In Dendrogram 5.17 the spike-in control data of the corresponding whole genome arrays cluster clearly together. Only for one replicate the dye swap arrays are clustered near by. This indicates that the error introduced through the splitting of the RNA sample over two different arrays is smaller than the error produced in a dye swap experiment. Dye swap experiments are very common in microarray analysis and the introduced errors known to be not problematic (Simon *et al.* 2004). Therefore, it can be assumed that the splitting of the RNA samples is no problem for the experiments performed with the two whole genome arrays.

5.3 Transcription Factor Study

With the help of the transcription factor study, we hoped to obtain deeper insights in transcriptional regulation during the different developmental phases of the zebrafish. We used a specially designed microarray, consisting of transcription factors (Chapter 2.1.3). Six different stages covering all embryonic stages (Table 5.19) and 4 different adult tissues (Table 5.20) were used for this study.

Period	Time ($\sim$)	Stage	Replicates
Cleavage	0.75 hpf	2-cell	8
Gastrula	4.5 hpf	early gastrula (30% epiboly)	4
Segmentation	10-12 hpf	1-6 somites	4
Pharyngula	24 hpf	24 hpf	6
Hatching	48 hpf	48 hpf	6
Larval	120 hpf	5 dpf	8

Table 5.19: Stages used for the transcription factor study

Tissue	Time	Replicates
diencephalon	> 90dpf	4
telencephalon	> 90dpf	4
head (brain)	> 90dpf	4
tail (muscle)	> 90dpf	4

Table 5.20: Tissues used for the transcription factor study

The RNA was extracted using Trizol (Chapter 2.4.2). The microarray experiments were performed without any control samples (Chapter 2.2.2) but using two colors (cy3 and cy5). Therefore, the analysis cannot be performed as for the 10 compound study.

5.3.1 Quality Control

At the beginning of the analysis, it is important to check the quality of the data. The quality of the arrays needs to be evaluated and possible bad arrays detected. In the next quality control step, problematic spots on the arrays itself must be removed from the data set.

Array Level

The following parameters were used to judge the array quality:

Raw image A manual inspection of the raw scanner images.

Signal histogram The scanner software GenePix provides the possibility to produce intensity histograms. The histograms indicate whether the array signals are well distributed over the detection range of the scanner. Labeling and hybridization problems or wrong scanner settings can so be detected.

Spike controls We used the Agilent provided spike-in controls (Chapter 2.1.1). I analyzed them as described in Chapter 4.1.1.

Diameter I compared the minimum and maximum diameters of the spots on the arrays. Variations in the diameters can occur due to spot detection problems based on too low signal or impurities on the array surface.

Saturated Spots Saturated spots disturb the analysis as the true signal cannot be calculated. Many saturated spots can be a sign that the scanner settings are not adjusted properly.

Coefficient of variation CV High CV values indicate spots with non-uniform signal distribution which might be due to artifacts.

Correlation Coefficient between replicates I used the Pearson correlation coefficient to calculate the similarity between arrays. Replicates should show a high correlation.

In general, the quality of the array was good. Only for the samples from the tail tissue, problems were detected. The tail data from two replicates showed problems in all quality categories. This data will still be included in the further analysis but should be handled with care in the interpretation of the results.

Spot Level

Artifacts on the array, low signals, or spot detection problems can lead to spot signals that are not representative of the biological experiment. In general, such spots can be identified manually or using quality control parameters. In my case, I used three different spot quality measurements, besides the manual inspection of the array scanner images.

Spot diameter The diameter should be between 35 and 75 μm.

Pixel signal variation of a spot The variation of the signal within a spot should be below 70%.

Number of saturated pixels in a spot A spot should have no saturated pixels.

These spots are excluded from the analysis.

5.3.2 Expressed Transcription Factors

As we did not use controls, I had to find a way to distinguish which transcription factors are expressed in our different samples. For high signal values, it is clear that the transcript is expressed, but for smaller ones it is not clear where the background noise ends and the true expression signal starts. To detect the background noise, I used the 7915 *A. thaliana* negative controls spots (Chapter 2.1.2). To use as cut-off, the highest value of that controls might not be useful since cross hybridizations or other impacts may result in a too high cut-off value. Therefore, I decided to test three distribution based parameters.

99th Percentile The 99th percentile is the value below which 99% of the negative control signals are.

Median + 2*std This cut-off is the median of all negative controls plus two times the standard deviation over all controls.

Median + 2*std/median This cut-off is the median calculated from all controls plus two times the standard deviation divided through the median.

To judge the quality of the different cut-offs, I used published data. Gene expression data from several development stages can be downloaded from Zfin. This information is also available via Ensembl Biomart. I used the list of transcription factors that are found on our array and downloaded all available gene expression descriptions. The Zfin descriptions covers also the whole embryonic development. The stages were categorized into 35 subgroups. To be able to compare this information with our array data, I fused the subgroups into 7 major groups. If one transcript is expressed in only one or a few subgroups, the whole major group will be counted as expressed. I used the foreground minus background signal for calculating the number of expressed transcripts based on the different cut-offs. Spots with bad quality were removed from the data set. Signals of all 8 oligos mapping to one transcript were averaged. I counted a transcript as expressed if it was expressed at least in one replicate. The literature data were compared with the list of expressed transcripts from our samples. The results are shown in Table 5.21.

All cut-offs showed a good detection rate of around 85% of the literature data. In the comparison of all cut-offs, the 99th quantile performed a little bit better than the other two measures.

Literature stages	Cleavage	Blastula	Gastrula	Segmentation	Pharyngula	Hatching	Larval	
Tf-study samples	2-cell	30% Epiboly	30% Epiboly	1-6 Somites	24 hpf	48 hpf	5 dpf	
# exp. trans. in literature	368	375	734	756	759	746	155	
99th quantile	336	334	552	623	654	642	132	Mean
In %	91.3	89.07	75.2	82.41	86.17	86.06	85.16	85.05
Median + 2*std	330	334	552	621	653	642	125	Mean
In %	89.67	89.07	75.2	82.14	86.03	86.06	80.65	84.12
Median + 2*relstd	335	334	552	623	654	642	132	Mean
In %	91.03	89.07	75.2	82.41	86.17	86.06	85.16	85.01

Table 5.21: Cutoff comparison

5.3.3 Normalization Methods

The normalization is a critical step, especially if signals of different experiments are compared. In Figure 5.18, the differences in the intensity distributions are depicted. The y-axis represents the signal (intensity), as measured by the scanner. The order is based on the date when the arrays were performed. The first sample belongs to the red color replicate, the following one to the green replicate. A clear dye-based effect can be observed. There are only weak differences detectable between the replicates. Interestingly the data when the arrays were performed showed an influence.

Some normalization methods can only be used on data, where the different data sets have a similar amount of data points. In our case, the different stages could express a different amount of transcripts. Early stages might have much less transcripts expressed than latter stages. To get an idea about the number of expressed transcripts in the different stages and tissues, I used the list calculated for the cut-off measure comparison (Table 5.22). A transcript is counted as expressed, if it is expressed in at least one replicate. The amount of expressed transcription factors were all in a similar range, so no special normalization method will be needed.

	2-cell	30% epiboly	1-6 somites	24 hpf	48 hpf	5 dpf	diencephalon	telencephalon	head	tail
# tf	3071	2778	2838	2968	2940	3195	2889	2627	2521	2797

Table 5.22: Number of expressed transcription factors for the different tissues and stages.

To find the best suitable normalization method, I tested several approaches. The normalizations are always performed on the whole raw signal data set.

Quantile normalization This normalization technique makes distributions identical in their statistical properties. All data sets are normalized together. The transcripts are sorted according to the expression values in the data set and the mean is calculated for each rank in the sorted lists. Then, the highest expression value is set to the highest average value and so on for all expression values. This is done for all data points in all data sets.

In Figure 5.19 the boxplot for the quantile normalized signal data is shown. In comparison to Table 5.18 all samples have now a similar signal distribution.

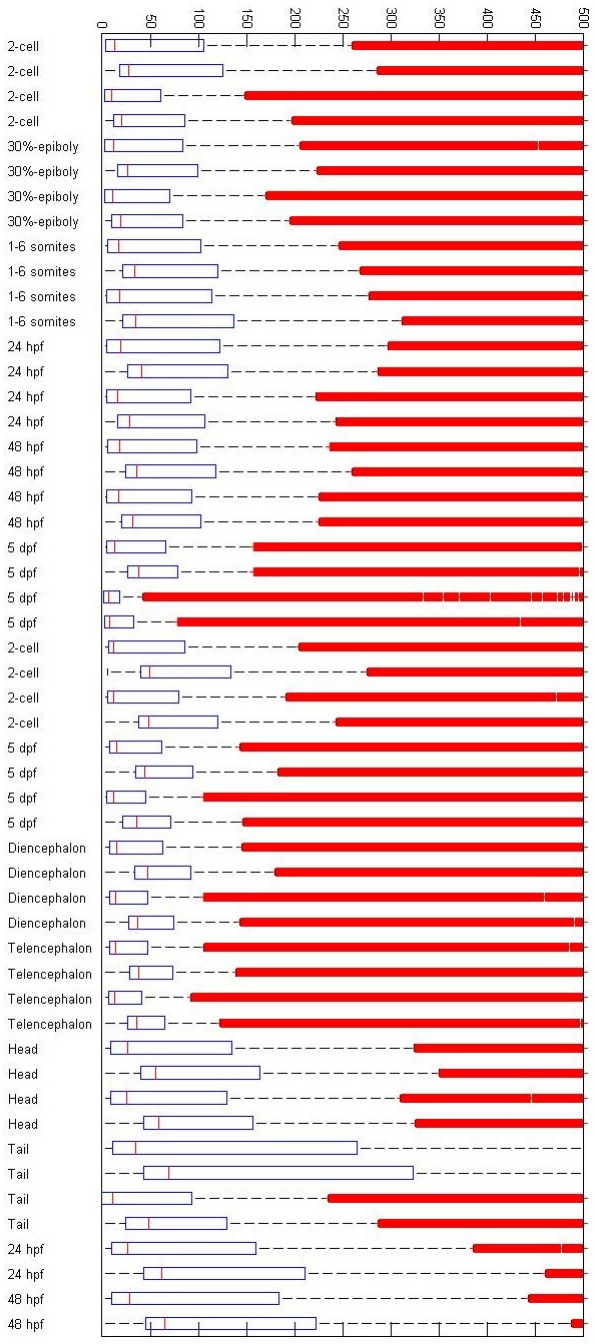


Figure 5.18: Box plot showing the distribution of the signals for all used microarrays. The red line in the box shows the median. The box represents the middle 50 % of the data. The red spots ('bars') are values above the 1.5 interquartile range (IQR). The differences in the signal distributions can be clearly seen. The median is shifted towards 0, indicating that most of the data points have very low signals.

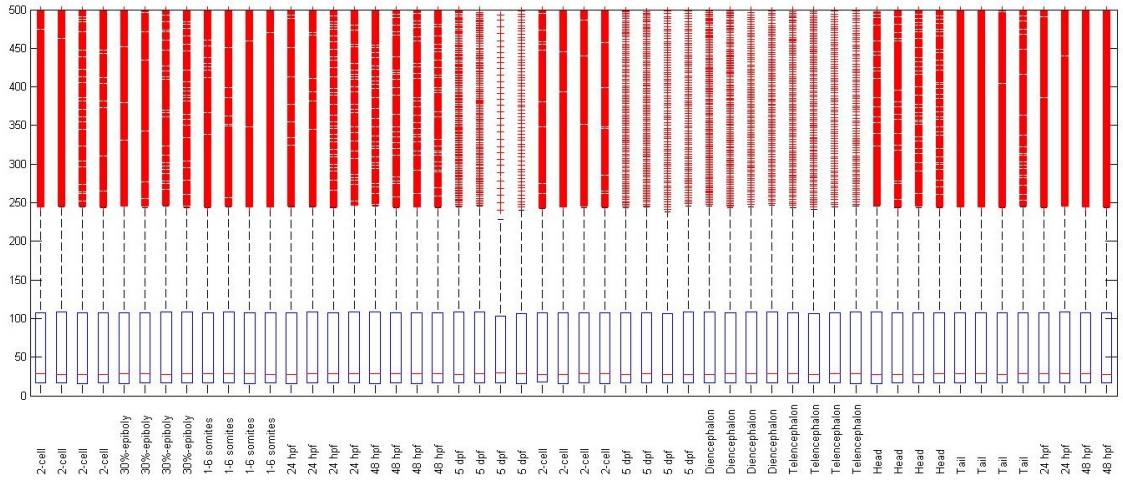


Figure 5.19: Boxplot of the signal distribution for the whole data set after quantile normalization. Compared to Figure 5.18 the signal distribution is here more equal. A description of the boxplot can be found in Figure 5.18.

Scaling This method scales all data sets to have a mean of 0 and a standard deviation of 1. Therefore, for each data point the median of the corresponding data set is subtracted and then it is divided through the standard deviation of the data set. This operation has the disadvantage of compressing the signal range, and consequently was not further investigated.

Rank invariant set normalization This method is based on a set of 'invariant transcripts' that do not change significantly between a data set and the reference set. To find them, all data points are ranked according to their intensity. Then, data points with similar ranks are identified. These items are then used to calculate the adjustment curve for the Lowess normalization, which corrects the data set based on the adjustment curve. As reference set, I used the median over all data sets. This method is highly depended on the invariant data points and was not able to normalize our data such that all data sets have the same distribution (Figure 5.20).

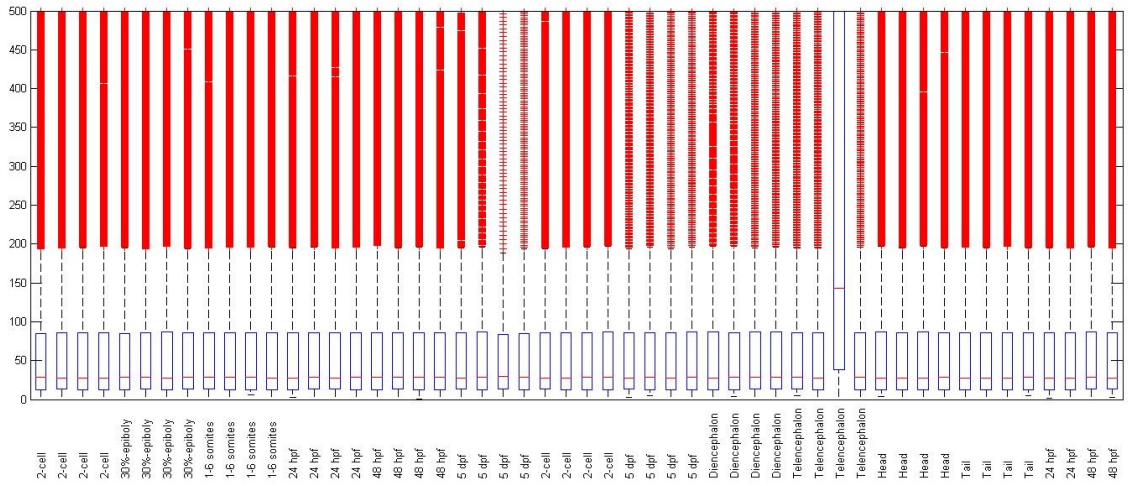


Figure 5.20: Boxplot of the Rank invariant set normalized signal data for all stages and tissues. A description of the boxplot can be found in Figure 5.18. The method was not able to normalize the data such that all data sets have the same distribution.

Subgroup normalization Instead of using the whole data set only a subgroup of data points can be used to calculate an adjustment curve, for instance, for housekeeping genes, which should be expressed at the same level in all samples. Even if a gene exists that is a true housekeeping gene in all developmental stages, it might show varying expression in all tissue samples. Therefore, this approach was not considered. The Agilent spike-in controls can also be chosen as subgroup (Chapter 2.1.1), but the intensity distribution of the controls looks different to the one of the sample data. This could be due to differences in the sample quality or spike control batch. Hence, this approach was also excluded.

Of all tested normalization methods, quantile normalization performed best.

5.3.4 Transcription Factor Array Analysis

Based on the detailed investigation of our data set and the comparisons of several analysis methods, I decided to use the following approach for analyzing the data:

1. The foreground minus background signal was used for the analysis (FG-BG)

2. The raw data were normalized using quantile normalization

3. The 99th quantile cut-off was calculated

4. Signals below the cut-off were removed from the data set

5. Spots with bad quality were excluded from the data set

6. The remaining signals of the 8 oligos from each transcript were averaged (mean)

7. The replicates were averaged using the median of the transcript signals

8. Transcription factors that were expressed in less than 50% of the replicates were removed.

In Figure 5.21, an overview of the number of expressed transcription factors in the different stages and tissues is shown.

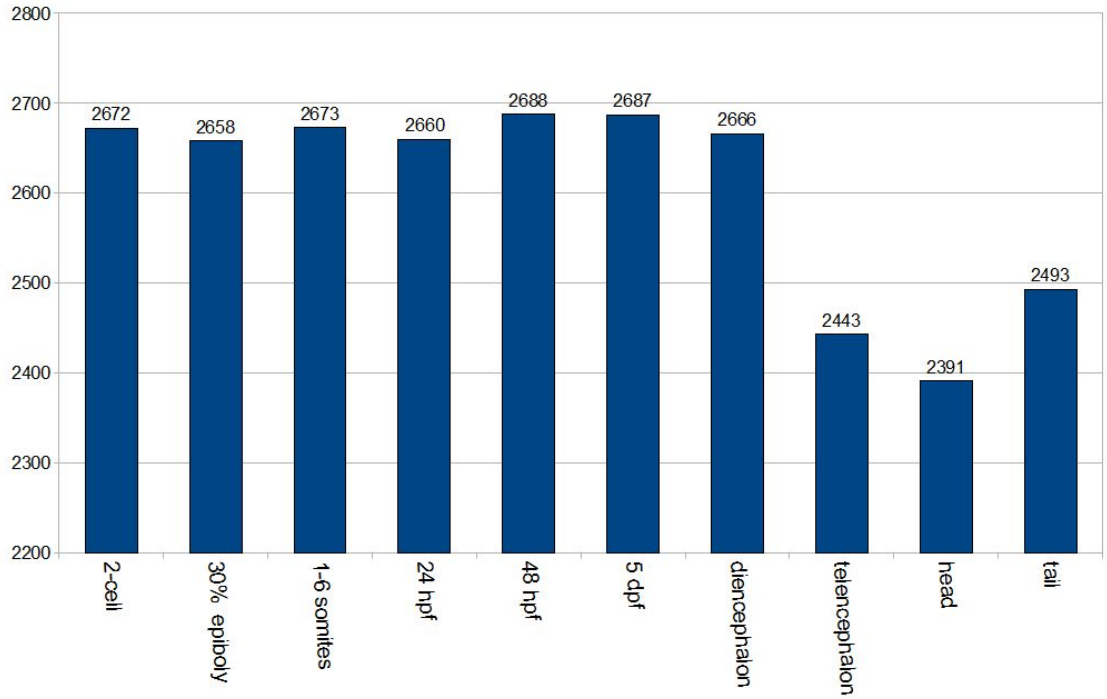


Figure 5.21: Number of expressed transcription factors after all analysis steps.

It is not clear whether the number of transcription factors that were expressed in the tail, head and telencephalon sample was really that low or whether this small number was caused by quality problems of the samples. The fact that the number of transcription factors in the whole head sample is smaller than in the two brain parts (diencephalon and telencephalon) might be caused by the RNA detection limitation of the microarrays. In the head sample the transcription factors from the telencephalon and the diencephalon could be expressed at such low levels compared that their signals are not detectable in the whole head sample.

5.3.5 Clustering Analysis

In order to study the similarity of the different gene expression patterns, I used hierarchical clustering (Chapter 4.2.2). With this, I want to identify similarities in the level of gene expression of the transcriptionfactors in the different samples. First, I clustered the raw signal data shown in Figure 5.22.

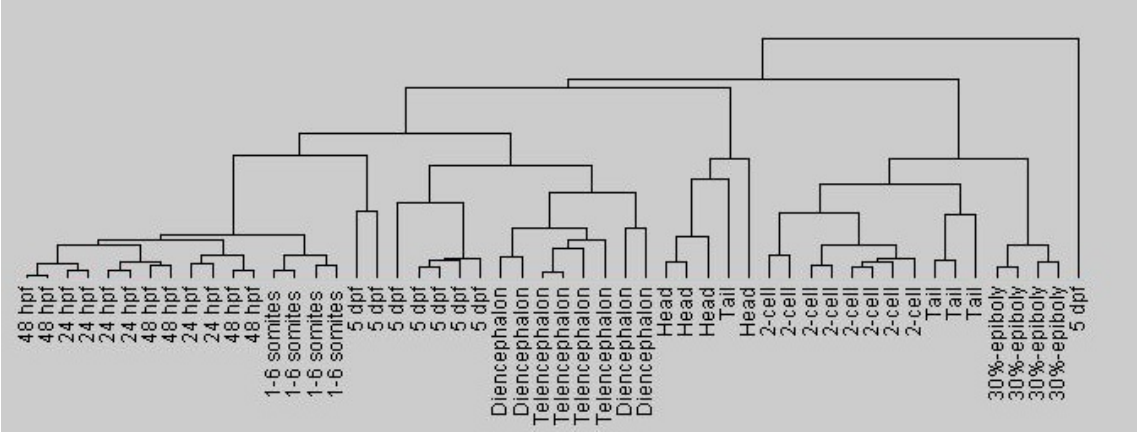


Figure 5.22: Result of the hierarchical cluster analysis. Performed on the raw signal data from all microarrays.

Only the replicates of the 2-cell, 30%-epiboly, and the 1-6 somites stage and the head tissue cluster nicely together. To improve this results, the cluster analysis was also done on the normalized data set (quantile normalization).

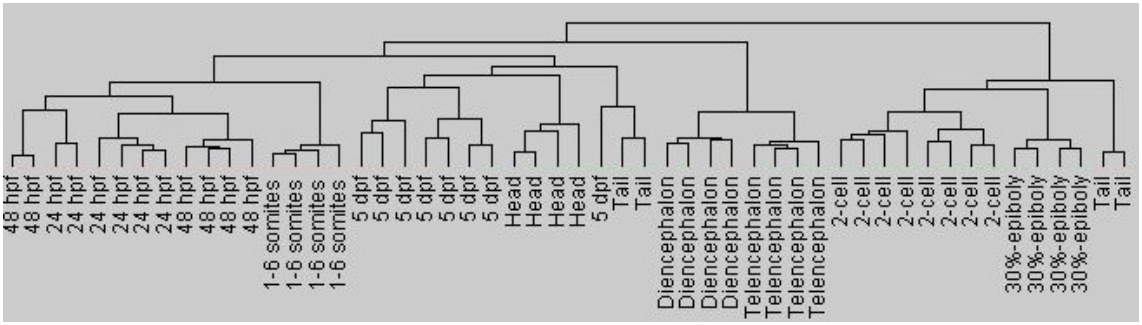


Figure 5.23: Cluster analysis of the normalized signal data set (quantile normalization, all microarrays)

In the dendrogram presented in Figure 5.23, the head and the brain tissue replicates give nice clusters. For the number of embryonic stages, the 24 hpf and 48 hpf replicates could not be separated. This might indicate that these expression patterns are very similar. The tail replicates are also not clustered together, maybe because of the array problems

identified in the quality analysis step.
After the normalization step, I identified expressed transcription factors via a cut-off based method and removed bad quality spots (Chapter 5.3.4). I also performed a cluster analysis on this data set. In Figure 5.24 the dendrogram of this analysis is shown. According to this analysis, the early stages, 2-cell, and 30%-epiboly show a similar expression pattern. The 1-6 somites, 24 hpf, and 48 hpf also share a similar transcription profile. The two brain samples diencephalon and telencephalon cluster also nicly together. The head shows more similarity with the 5 dpf stage than with the diencephalon and telencephalon sample.

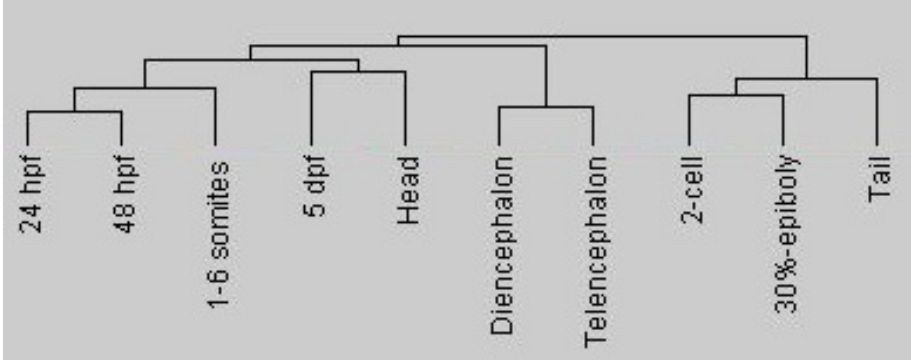


Figure 5.24: Cluster analysis of the analyzed data set

Interestingly, the tail sample clusters together with the very early stages in development. Myogenesis starts at the segmentation stage, and therefore, the tail sample would have been expected to exhibit a higher similarity to the 1-6 somites stage (Lo *et al.* 2003). It has been discovered that the first inducing mechanisms of myogenisis begin around the late blastula period (Ochi *et al.* 2008). However, it is unlikely that this explains the results of the cluster analysis.

Gene name	Ensembl ID	2-cell	30%-epiboly	1-6somites	24 hpf	48 hpf	5 dpf	Diencephalon	Telencephalon	Head	Tail
myod1	ENSDART00000027661	0	0	1	1	1	1	0	0	1	1
myf5	ENSDART00000014818	0	0	1	1	1	1	0	0	1	0
mef2cb	ENSDART00000044083	1	1	0	1	1	1	1	0	1	1
mef2ca	ENSDART00000097433	1	1	1	1	1	1	0	0	0	1
myogenin	ENSDART00000014062	0	0	1	1	1	1	0	0	1	1

Table 5.23: Expression pattern of known muscle specific transcriptionfactors in the data set. 1 indicates is expressed and 0 is not expressed in the particular sample.

Therefore, I decided to have a closer look at the expression patterns of some well

known muscle specific transcription factors. In Table 5.23, the expression patterns of five muscle specific transcription factors, in the data set are shown. No differences to other published studies are detectable (Lo *et al.* 2003). This indicates that the array quality and the sample integrity seem to be fine.

In order to further investigate the similarity of the very early stages and the tail sample, I performed a gene function analysis as described in Chapter 4.4. Unfortunately, no significantly enriched pathways could be detected. However, the GeneOntology analysis revealed an enrichment of GO-terms involved in:

- chondrocyte differentiation
- methylation
- regulation of apoptosis
- cell cycle
- biological processes
- chromatin modification

The occurrence of many transcription factors related to chondrocyte differentiation leads to the conclusion that the tail sample seems not be as representative for muscle tissue as expected. Since we only cut the complete tail and did not extract muscle tissue, the sample also contains other tissues, such as bone. This mixture of tissues might also be the reason for the clustering of the tail sample with the very early stages. In order to improve this study, a more specific muscle sample should be analyzed.

5.3.6 Gene Function Analysis Time Series Data

On the supplementary CD, an Excel file can be found, which contains the expression pattern of the transcription factor screen transformed to either 1 (expressed) or 0 (not expressed). With the help of this file, co-regulated transcripts can be found. For example, 1703 transcription factors were expressed continuously over all stages. It is also possible to search for transcription factors of interest and find similarly expressed ones.
However, besides the fact that a transcription factor is expressed, the changes of the expression over time (profile) might also be of interest, for example, if there is a very high expression at a certain stage (peak). To find groups of transcription factors that share the same profile I used the program STEM (Ernst and Bar-Joseph 2006). This software allows for detecting significant expression profiles in time series data and the genes that are associated with these profiles. STEM calculates all possible profiles for a certain amount of time points. It compares the uploaded time series data with the profiles and performs statistical tests to detect the profiles which are significantly enriched in the data set. At the

end it shows basically the most common profiles of the data set. The profiles and the corresponding genes can be downloaded. Furthermore a Gene Ontology analysis of the gene lists can be performed.In Figure 5.25, the most significant profiles of the transcription factor time series data set is shown.

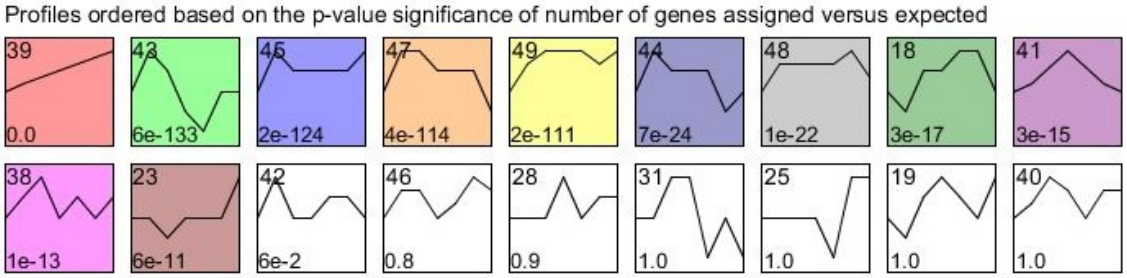


Figure 5.25: Significant profiles of the transcription factor data set. The boxes represent the different profiles. Significantly enriched profiles are colored. The profiles are ordered by p-value, which is shown in the lower left corner of the profiles. The profile number can be found in the upper left corner.

11 different significant profiles could be detected. The software provides the possibility to have a closer look at the expression profiles ("'zoom in"'). I had to define a time point 0, since we have no data from time point 0, all expression values for that time point are set to 0. This needs to be taken into account when interpreting the results of the analysis. In Figure 5.26, the "'zoom in"' for profile 43 is presented. The "'zoom in"' images of the 11 profiles can be found on Appendix B. The transcription factors of that profile show the highest expression at the 2-cell stage. At 30% epiboly the expression goes down and at the later stages is nearly gone. This transcription factor seems to be expressed till gastrulation starts.

I also performed a Gene Ontology analysis for the transcription factors of the 11 profiles. Therefore I downloaded the Gene Ontology annotation of the transcription factors from Ensembl Biomart. Then, STEM calculated the enriched GO categories for each profile. Since this list can be very long and difficult to interpret, I used the GO similarity analysis descried in Chapter 4.5 to simplify the data. In Figure 5.27, the simplified Gene Ontology results for profile 43 can be seen. As expected from the "'zoom in"', gastrulation is an enriched GO category. The results of the GO analysis and the corresponding profiles can be found in Appendix B. The gene lists are included in the supplementary CD. In the discussion part of my thesis I will describe the different profiles in more detail (Chapter 6.3)

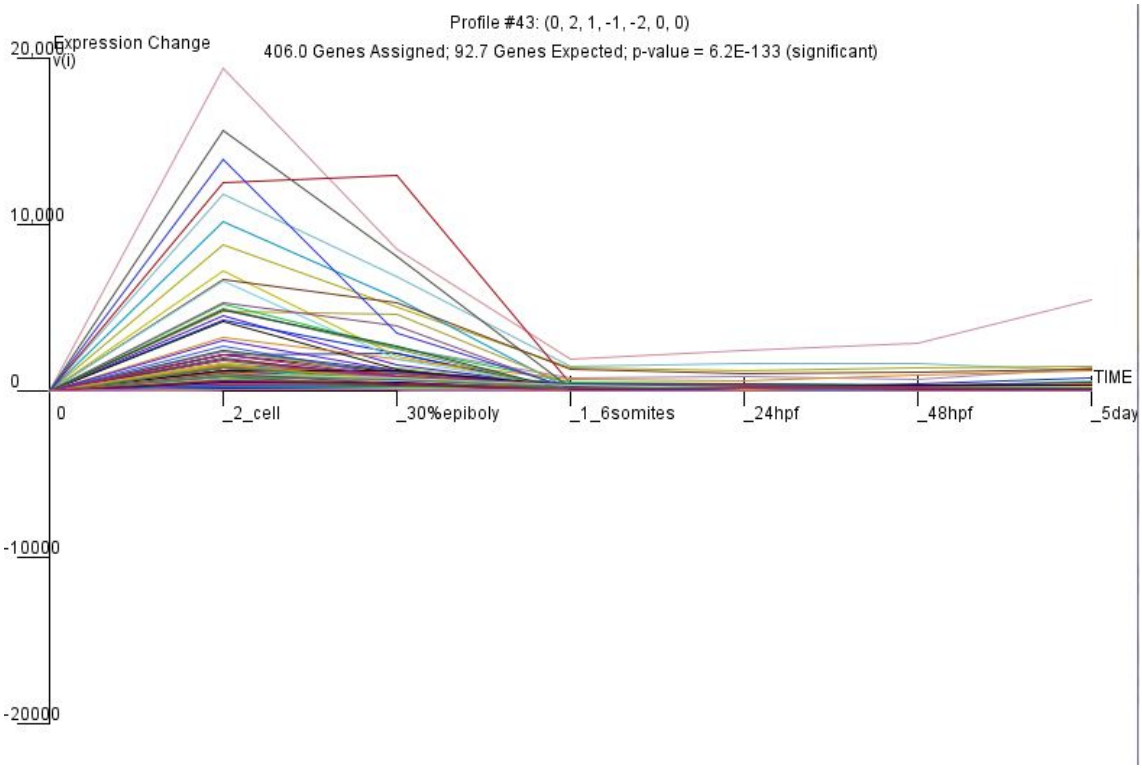


Figure 5.26: "'Zoom in"' on the expression signals of profile 43.

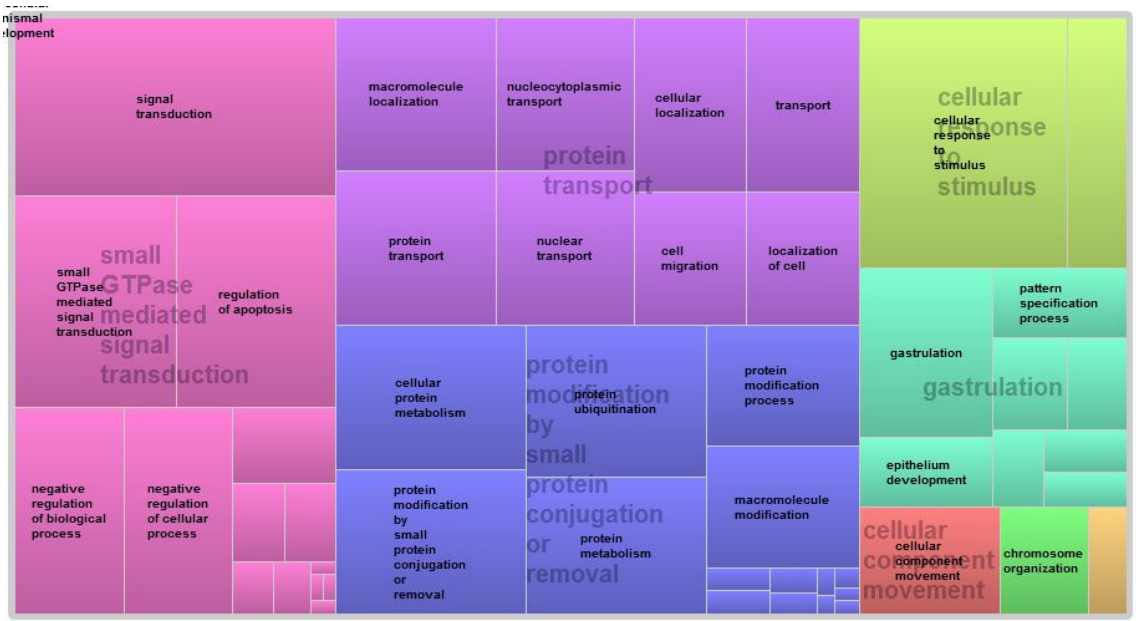


Figure 5.27: GO analysis of expression profile 43

Chapter 6

Discussion

6.1 10 Compound Study

For most of the compounds used in this study, little is known about their effect on zebrafish development. Therefore, our results will give the first insights into the xenobiotic metabolism of these chemicals in the zebrafish embryo. In the following chapter, I will summarize the results of the microarray analysis and link them with literature data in order to get a better understanding of the induced modes of action of the compounds. Based on these results, I define biomarker genes, which are specific for different modes of action of the compounds used in this study.

6.1.1 Results of the Microarray Analysis

The interpretation of microarray results can be very difficult and no standard method is available, especially not for identifying toxicity induced mechanism. Since there is no well established database for tox-pathways available, I used KEGG and WikiPathways to perform the gene function analysis of the data (Chapter 4.4). Therefore, it was necessary to link the obtained pathways to the suggested modes of action of the used compounds (Table 1.1). An extensive literature search was performed, in order to connect the identified pathways with possible mechanisms of toxicity. The results of the enrichment analysis and the intensity distribution analysis were used to further support the findings (Chapter 5.1.3, Chapter 5.1.5).

Esfenvalerate

Esfenvalerate is a pyrethroide and known to interfere with sodium and calcium channels in adult chinook salmon (Viant *et al.* 2006a). Chorionated fish embryos reacted less sensitive to the toxicity effects of esfenvalerate, which led to the assumption that the chorion may have a protective effect (Viant *et al.* 2006a). In our study, we exposed the embryos

to a very high concentration (80 μg/l) without detecting any mortality. In comparison, the LC50 for newly hatched rainbow trout is at around 2 μg/l (Barry *et al.* 1995). This of course questions the possibility to detect the mode of action of esfenvalerate in our study. However the microarray signal intensity distribution, the maximum and minimum signals of the esfenvalerate treated embryos, looked normal. The amount of differentially expressed transcripts was also in a normal range (Chapter 5.1.3). My gene function analysis clearly indicated the activation of the proteasomal degradation pathway and a repressed RNA degradation pathway (Appendix A). These findings agree very well with previously published studies. The treatment of catfish with fenvalerate decreased significantly the total RNA and protein content in brain, liver, and skeletal muscle. The authors suspected this might be due to reduced enzyme activity, changes in protein and RNA turnover (synthesis/degradation), and a general inhibitory effect on metabolism (Tripathi and Verma 2004). A decrease of ATP concentration and changed metabolism was also seen in chinook salmon treated with esfenvalerate (Viant *et al.* 2006b). The only direct hint for an effect on the calcium channel was given by an induced GTP binding function. GTP-binding proteins can be targets of xenobiotics and it is assumed that some pyrethroids bind to G-proteins and thereby alter their GTP-binding capabilities. G-proteins can interact with sodium and calcium channels (Dolphin 1998 and it was proposed that pyrethroids may influence calcium and sodium channels via interaction with G-proteins (Rossignol 1991). Therefore, a change in the GTP-binding functionality might indicate an effect on the calcium and sodium channels.

In our study the embryo showed less sensitivity to esfenvalerate induced toxicity, possible caused by protection by the chorion. Based on the microarray data no direct effect on sodium or calcium channels could be seen, but an increase in GTP-binding might indirectly lead to this effect. In agreement with other studies in fish, a clear effect on the protein and RNA metabolism was found (Tripathi and Verma 2004; Viant *et al.* 2006b). The cause of this effect could not be identified. Further investigations on protein and enzyme levels may help to reveal the cause of toxicity of esfenvalerate in zebrafish embryos.

Methoxychlor

Methoxychlor is a known endocrine disruptor. However, the exact mode of action of Methoxychlor toxicity is still unknown. In fish it is metabolized to mono- and bisdemythelated metabolites. Mono- and/or di-hydroxylated products are also produced sometimes (Berg 2003). It has been shown in rainbow trout that the metabolites have the potential to act as weak ER agonist (Thorpe 2000). Holdway and Dixon 1986 reported a protective mechanism in chorinated embryos, which prevents methoxychlor toxicity in flag fish embryos. We were also able to observe the same protective effect in zebrafish embryos during our study using a concentration of 800 μg/l for the microarray exposure. In another study also performed in zebrafish, all larvae died at concentrations higher than 10 μg/l 7 days after hatching (Versonnen *et al.* 2004). This shows a strongly reduced sensitivity of the zebrafish embryos compared to the larvae. Whether this is caused by a

protective effect of the chorion or a lack of metabolic capacity is unknown. The microarray analysis revealed a reduced number of highly expressed transcripts (Chapter 5.1.3). In the gene function analysis, the proteasome, spliceosome, and the RNA degradation pathways were induced (Appendix A). Methoxychlor is known to cause protein and DNA damage in mouse ovary by increasing superoxide production through impairment of mitochondrial respiration (Gupta *et al.* 2006). In our data, an induction of reactive oxygen species (ROS) or disturbed mitochondrial respiration could not be identified. Therefore, the cause of the changes in protein and RNA metabolism remain unclear.

In the chorionated zebrafish embryo, methoxychlor showed less toxicity compared to 7dpf larvae (Versonnen *et al.* 2004). In the microarray data, I could not identify an effect of methoxychlor on the estrogen receptor. However, a change in the protein and RNA metabolism could be observed. A study in mouse ovary also found a change in protein metabolism caused by increased ROS production by disruption of the mitochondrial respiration (Gupta *et al.* 2006). This could not be confirmed with our data. Therefore, further investigations are needed to identify the cause of the changes in protein and RNA metabolism.

Di-n-buthyl phthalate

Phthalate esters are suspected to act as endocrine disruptors by mimicking the effects of natural estrogens. Dibutylphthalat can alter the vitellogenin (VTG) protein and gene expression levels in treated zebrafish larvae but there was no clear induction (Ortiz-Zarragoitia *et al.* 2006). In our data, I found a strong induction of VTG gene expression with a 3.74 fold (M-value) up-regulation. Other evidence hinting at altered estrogen levels were not found. Besides the capability to influence estrogen levels, dibutylphthalat is also known to act as peroxisome proliferator. The effect of peroxisome proliferation is caused by interactions with nuclear hormone receptors like pregnane X receptor (PXR), constitutive androstane receptor (CAR), and the peroxisome proliferation-activated receptors (PPARs) Ortiz-Zarragoitia *et al.* 2006. Nuclear receptors are involved in the regulation of many metabolic pathways. Wyde and colleagues could show in fetal rat liver that dibutylphthalat can modulate nuclear receptors and thereby influence the metabolism of lipids, steroids, and other biological processes, including lipid homeostasis, cholesterol metabolism, and steroidogenesis (Wyde *et al.* 2005). The gene function analysis of the microarray data revealed an up-regulation in lipid biosythesis and metabolism, cholesterol biosynthesis, and other metabolic pathways. Besides that, I also identified the up-regulation of oxidative phorsphorylation and the electron transport chain pathways (Appendix A). Not much is known about the effects of dibutylphthalat on the respiratory chain in fish species. In mitochondria of male rats, dibutylphthalat seems to act as an energy transfer inhibitor, and at the same time, to influence ATPase activity. It was suggested that dibutylphthalat may act as uncoupler of the mitochondrial oxidative phosphorylation (Inouye *et al.* 1978).

The microarray data led us to the conclude that dibutylphthalat seemed to act via two

different modes of action in zebrafish embryos. On the one hand, dibutylphthalat seemed to interact with several nuclear hormone receptors. In zebrafish larvae, VTG gene expression is altered via dibutylphthalat treatment (Ortiz-Zarragoitia *et al.* 2006). First, VTG was highly induced. VTG might be induced via the estrogen receptor (Hill and Janz 2003). I could also detect an induction of the lipid and cholesterol metabolism. Like shown in fetal rat liver, this leads to the assumption that dibutylphthalat also interacts with the nuclear hormone receptor PXR, CAR and PPARs (Wyde *et al.* 2005). Second, the microarray data revealed an effect on the mitochondrial respiration of the zebrafish embryos. It was suggested that dibutylphthalat acts as uncoupler of the mitochondrial oxidative phosphorylation in male rats (Inouye *et al.* 1978). Based on our data dibutylphthalat seems to also act as uncoupler in treated zebrafish embryos.

Flucythrinate

Flucythrinate is a type II pyrethroide. In a reporter gene assay using COS-7 simian kidney cells flucythrinate was detected to have strong PXR agonist, weaker ER agonist, and AR agonist capabilities (Kojima *et al.* 2010). In our microarray data, the gene CYP3A65 was highly induced (M-value > 1.5). CYP3A65 can be activated through regulation of its upstream transcription factors, such as PXR (Tseng *et al.* 2005). The lack of highly regulated genes in the microarray data limited the gene function analysis. A detailed investigation of the regulated pathways and genes involved suggested an anti-apoptotic effect (Appendix A). Several genes with known anti-apototic capabilities were highly upregulated. In colon cancer cells, it has been shown that PXR can have an anti-apoptotic effect. Nevertheless, it is not known whether flucythinate has any anti-apoptotic capability. Based on our microarray results, flucythrinate seemed to have an antiapototic effect, possibly induced by PXR activation. This effect needs to be confirmed in future investigations.

Gene name	Reference	Organism
stat3	Lu *et al.* 2006	murine embryonic fibroblast
hsp90	Erdmann *et al.* 2007	neoblastoma cells
socs3	Jo *et al.* 2005	mice
bag3	Virador *et al.* 2009	HeLa human cancer cells

Table 6.1: Table of genes with known antiapototic properties which are highly upregulated in the flucythrinate microarrays.

2,4-Dimethylphenol

Dimethylphenol is categorized as polar narcotic (Tsai and Chen 2007). Little is known about the mode of action behind dimethylphenol toxicity. In human erythrocytes, a de-

crease in ATPase activity has been shown Duchnowicz *et al.* 2005. Our microarray data indicate the down-regulation of several subunits of the F-ATPase complex. Additionally, the mitochondrial glutathione reductase gene (zgc:110010) was highly induced. The mitochondrial glutathione reductase belongs to the mitochondrial antioxidant defense system. An up-regulation indicates production of reactive oxygen species (ROS) (Fleury *et al.* 2002). Other up-regulated genes were involved in apoptosis (tp53, caspase8). Several pathways linked with DNA damage (response to DNA damage, DNA replication, damaged DNA binding) were also activated (Appendix A). However, a possible DNA damaging effect of dimethylphenol is not yet known. Our data suggest ROS as a possible mechanism for this effect.

In the zebrafish embryos, dimethylphenol seemed to interact with the F-ATPase complex. Based on the microarray data, this seems to induce a change in the mitochondrial respiration leading to an induction of ROS. The higher levels of ROS then might have led to DNA damage in the treated embryos.

Chlorpyrifos

The microarray data for chlorpyrifos gave no clear results with respect to possible modes of action (Appendix A). The number of expressed transcripts was very small. It was also the only compound with more down- than up-regulated transcripts (Chapter 5.1.3). Chlorpyrifos is almost not soluble in water, therefore, ethanol was used as solvent. The LC50 in 8dpf old zebrafish was reported to be around 0.5 mg/l (Kienle *et al.* 2009). In our experiment, we used 7 mg/l without observing any mortality in the embryos. A protective effect of the chorion from chlorpyrifos toxicity has not been seen but is suggested by the high treatment concentration used in our experiments. Surprisingly, CYP1A was highly induced. It is known that CYP1A is specifically induced in fish by polycyclic aromatic hydrocarbons (PAH). Since chlorpyrifos is no PAH, CYP1A should not be induced (Levine and Oris 1999). However, it was also shown, especially in zebrafish, that CYP1A can be induced by activation of the aryl-hydrocarbon receptor (AhR) (Alderton *et al.* 2010). In a mouse hepatoma reporter cell line, chlorpyrifos showed AhR-mediated transcriptional activity (Takeuchi *et al.* 2008).

The high concentration used and the low number of expressed transcripts suggests that the uptake of the compound in the embryos is rather low. The gene function analysis revealed unfortunately nothing. Only the induction of CYP1A might give a hint, for an activation of the AhR via chlorpyrifos treatment.

4-Chlorophenol

Of all tested compounds, chlorophenol induced the biggest expression changes (Chapter 5.1.3). In the gene enrichment analysis, the gene sets for apoptosis and transcription were found to be enriched (Chapter 5.1.5). The gene function analysis revealed the activation

of many pathways involved in apoptosis (p53 signaling, apoptosis, death, and regulation of caspase activity) (Appendix A). Furthermore, genes known to be activated during apoptosis were induced (casp8, tp53), but the cause for the apoptotic activity remains unclear. Chlorophenol is known to disrupt oxphos in aquatic organisms (Comparative *et al.* 2001). The only hint for oxphos disruption was an increased level of mitochondrial glutathione reductase gene expression (zgc:110010). This gene is activated by increased ROS production in the mitochondria, which can be caused by oxphos disruption (Fleury *et al.* 2002). The mode of action, oxphos disruption is concentration dependent. High concentrations inhibit respiration, decrease ATPase activity and lead to break-down of electron-transport-processes. Low concentrations, on the other hand, lead to an increase in ATPase activity (Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/). In yeast-two-hybrid systems, chlorophenol also showed estrogen receptor activity (Ogawa *et al.* 2006). The gene expression of vitellogenin (VTG) was highly induced in our experiment leading to the hypothesis that 4-chlorophenol also has ER activity in zebrafish embryos.

The used concentration of chlorophenol seemed a little bit to high, as most of the genes and regulated pathways were linked to apoptosis. Nevertheless, I could find evidence that chlorpyrifos might act through two different modes of action in the embryos. First, VTG was highly expressed indicating an ER activation. Secondly, ROS was induced in the mitochondria, which leads to the assumption that chlorophenol has an effect on the mitochondrial respiration. Based on the literature data, an ATPase inhibition might be the cause for the disruption of respiration.

Chlorthalonil

Chlorthalonil is metabolized via glutathione (GSH) conjugation in the Phase II detoxification pathway in adult fish. The glutathione metabolites are then excreted through the bile and urinary systems (Davies and White 1985; Davies 1985a,b). The results of my gene function analysis suggested a high activation of the glutathione metabolism. Metabolism of xenobiotics by cytochrome P450 was also induced (Appendix A). The gene with the highest up-regulation (9 fold) was glutathione-S-transferas (gstp1). Glutathione-s-transferase (GST) mediates the metabolism of chlorthalonil in the liver and gill in channel catfish. GST induction has been suggested as biomarker gene for chlorthalonil toxicity in fish (Gallagher *et al.* 1991). The toxicity of chlorthalonil might be caused by the glutathione depletion followed by interactions with other thiol-rich proteins (Davies and White 1985).

In the microarray data the metabolism of chlorthalonil via the glutathione metabolism was clearly detectable. Since the microarray experiment showed the smallest number of differentially expressed transcripts, it can be assumed that glutathion is not yet completely depleted and that no toxic mechanism are activated.

Propoxur

The carbamate propoxur is well known as Ache inhibitor (Smulders *et al.* 2003). Besides that, it is assumed that propoxur induces oxidative stress through lipid peroxidation. The authors showed a decrease of glutathione reductase (GR), glutathione S-transferase (GST), and glutathione peroxidase (GPX) enzyme levels and a decrease of glutathione (GSH). GPX detoxifies lipid hydroperoxide and hydro peroxide by using GSH. GR is the enzyme that produces the reduced GSH needed for detoxification of GPX. GST uses GSH during xenobiotic metabolism (Seth *et al.* 2001). In the microarray results, the expression levels of mitochondrial glutathione reductase (zgc:110010), glutathione S-transferase (gstp2), and glutathione peroxidase (gpx1a) were highly up-regulated. The propoxur microarrays showed the highest expression signals and gstp2 was the second highest induced gene. In the gene function analysis, the glutathione pathway was also shown to be activated. Abd-Elraof *et al.* 1981 suggested that propoxur is metabolized via the Phase I cytochrome P-450 pathway. Our data did not confirm this hypothesis. Many pathways linked to apoptosis were induced (apoptosis, death, p53 signaling) (Appendix A); differentially expressed genes were also found to be enriched with apoptosis genes (Chapter 5.1.5), although apoptotic genes like tp53 and casp8 were only slightly induced.

The microarray data confirmed the finding that propoxur induces oxidative stress (Seth *et al.* 2001). This might be due to lipid peroxidation. The induced oxidative stress seems to be so high that it induces apoptosis. The metabolism of propoxur via Phase I detoxification could not be seen in the embryos (Abd-Elraof *et al.* 1981).

1,2-Dibromoethane

Dibromoethane is a well-known carcinogen in rats and mice. It is known to be metabolized in the liver by cytosolic glutathione-S-transferase into S-2-bromoethylglutamthione, a glutathione (GSH) conjugate. Microsomal oxidation produces bromoacetaldehyde, which also produces a conjugate with GSH. It is suggested that microsomal metabolites preferentially bind to proteins while the gluthatione conjugates prefer to bind to DNA (White *et al.* 1983; Botti *et al.* 1989). In the microarray data the microsomal glutathione-s-transferase was induced (mgst1). In the gene function analysis the proteasomal degradation complex was down regulated (Appendix A). This effect of dibromoethane has not been shown till now. A DNA damaging effect could not be detected (White *et al.* 1983; Botti *et al.* 1989). In rat liver mitochondria, dibromoethane disrupts oxidative phosphorylation via respiratory enzyme inhibition (Thomas *et al.* 2001). This effect could also not be seen in the microarray data.

In the zebrafish embryos, dibromoethane seems to be metabolized via microsomal oxidation. This was indicated by the up regulation of mgst1. Several genes of the proteasome complex were down regulated. Further effects could not be identified. The used concentration might be to small to induce further toxic effects in the zebrafish embryos.

6.1.2 Clustering and Gene Co-regulation

The comparative and gene co-regulation analysis tries to identify groups of compounds which share similar modes of action. This can help to gain new insights into the modes of action of the compounds. In the following, I summarize the findings from the cluster analysis (Chapter 5.1.1) and the co-regulation analysis (Chapter 5.1.2), and investigate the common modes of action.

Methoxychlor and Esfenvalerate

Methoxychlor and esfenvalerate were clustered together by almost all clustering methods applied. They also had an above average number of co-regulated transcripts (Chapter 5.1.2). The pathways detected in the gene function analysis gave also quite similar results (Appendix A, Appendix A. The proteasomal degradation pathway was activated for both compounds. The FGF signaling pathway and the BMP signaling pathway were down-regulated in both cases. In contrast, the RNA degradation pathway was down-regulated in the esfenvalerate data but upregulated in the methoxychlor data. The gene function analysis of the co-regulated genes led to the assumption that both compounds interact with G-proteins. GTPase activity, Gap-junctions, and fatty acid turnover can all be linked to G-protein signaling (Rossignol 1991; Rouach *et al.* 2006; Pashkov *et al.* 2011). It is known that esfenvalerate influences calcium and sodium channels, presumably through interactions with G-proteins (Rossignol 1991). The real mode of action of methoxychlor is still unclear, but there are first hints that it also interacts with G-proteins and thereby alters the calcium flux (Wu *et al.* 2006). The changes in the proteasome degradation pathway, the RNA degradation pathway, and the IL2 and IL6 pathways might be part of the secondary response since no direct effects of esfenvalerate or methoxychlor on these pathways are known.

Chlorophenol and Propoxur

Chlorophenol and Propoxur cluster together and also share an above-average number of co-regulated genes (Chapter 5.1.2). The two compounds express the highest amount of apoptotic genes, and they have been the only compounds with a significant enrichment of apoptotic genes (Chapter 5.1.5). The comparison of the gene function analysis results does not suggest further shared possible modes of action besides apoptosis. Therefore, it can be assumed that these two compounds share the same 'level' of toxicity rather than the same mode of action.

Dimethylphenol and Dibromoethane

In the cluster analysis, dimethylphenol and dibromoethane seem to share a similar gene expression pattern (Chapter 5.1.1), although they do not share a significant amount of

co-regulated genes. In the gene function analysis of the data, some similarities were also obvious. Both compounds down-regulated the proteasome and both showed a decreased activity in the microtubule-based movement and process. It is known that both compounds have an effect on the mitochondrial membrane potential. I could not identify a link between the two compounds and the down-regulation of the proteasome. However, it is known that rotenon, a electron-transport-chain complex I inhibitor, causes proteasome inhibition. Chou *et al.* 2010 showed that rotenon induces mitochondrial inhibition, reactive oxygen species, reactive nitrogen species, influences the microtubule assembly, and inhibits the proteasome. How these effects are linked and what causes the proteasome inhibition is unclear. Nevertheless, based on the gene function analysis results of dimethylphenol and dibromoethane, I assume that these compounds act through a similar toxicity mechanism as rotenon.

Chlorophenol and Dimethylphenol

Chlorophenol and dimethylphenol show no similarity in the cluster analysis but in the co-regulation analysis. The gene function analysis of the co-regulated genes resulted mostly in pathways involved in regulation of DNA damage (Chapter 5.1.2). In the gene function analysis results of the whole data set, both compounds also shared several pathways (Appendix A, Appendix A). P53 signaling, glycineserine and threonine metabolism, senescence and autophagy, and androgen receptor signaling pathways were induced. These pathways could also indicate a high degree of DNA damage. Therefore, this leads me to the conclusion that the mode of action shared by chlorophenol and dimethylphenol seems to be DNA damage. For both compounds, no direct effects on the DNA are known. Both compounds induced reactive oxygen species suggesting that ROS might be the cause of the induced DNA damage (COOKE *et al.* 2003). Since both compounds are substituted phenols, an effect based on the phenol group can also not be excluded.

Chlorophenol and Dibutylphthalate

Chlorophenol and dibutylphthalate have no similarity based on the clustering of the gene expression patterns, but they share a significantly large number of genes. Interestingly, the number of co-regulated genes was significant for genes which are expressed only after treatment with these two compounds (Chapter 5.1.2). The gene function analysis of these genes did not lead to significant results. A comparison of the regulated pathways also showed no similarities (Appendix A). Therefore, the shared mode of action remains unclear. Both compounds act as estrogen receptor agonists, but an analysis of the lists of co-regulated genes could not proof this. A secondary effect, such as induction of apoptosis or response of the immune system, could be the cause, but this could not be confirmed either with the data.

6.1.3 Biomarker genes

The goal of this study was to identify new biomarker genes that are specific for different modes of action. I used the results of the microarray analysis to identify compounds sharing toxic mechanisms (Chapter 6.1.1 and Chapter 6.1.2). Compounds with the same modes of action were than used to identify toxicity specific biomarker genes. This was done by searching for genes that were highly expressed only in these compounds. Since a biomarker gene should change significantly only for a specific mode of action, I decided to focus on highly expressed genes.

Disruption of Mitochondrial Respiration

An effect on the mitochondrial respiration could be seen in the microarray results of chlorophenol, dibutylphthalte, dibromoethane, and dimethylphenol. When all compounds were taken into account, only one gene came up as possible biomarker gene (Table 6.2). This gene encodes a membrane-bound protein which is a member of the ELO family. This proteins participate in the biosynthesis of fatty acids. Elovl4 plays an important role in photoreceptor cells. In NIH3T3 and HEK293 cells elovl4 is localized preferentially to the endoplasmic reticulum (ER) and was not found in the mitochondria (ßcitealt Karan2004). Nothing is known about the relationship between mitochondrial respiration and elovl4. Therefore, I decided to further specify my list of compounds. I excluded compounds which showed no clear effect on the mitochondrial respiration.

dibromoethane	dibutylphthalate	dimethylphenol	chlorthalonil	chlorophenol	Ensembl Description	Gene Name
-2.08	-1.58	-1.75	-0.44	-4.1	elongation of very long chain fatty acids-like 4 [Source:RefSeq peptide;Acc:NP_956266]	elovl4

Table 6.2: Expression values (M-value) of possible biomarker genes for disruption of mitochondrial respiration when chlorophenol, dimethylphenol, dibutylphthalate, and dibromoethane were taken into account.

Since the effect of dibromoethane on the mitochondrial respiration was only found indirectly by co-regulation of dimethylphenol, I decided to exclude the compound. The gene stathmin-2 (stmn2b) was highly repressed (Table 6.3). Stmn2b (previous name: SCG10) is neuron-specific, membrane-associated, and concentrated in growth cones. Its

expression is high in the developing nervous system (Riederer *et al.* 1997). In the literature, no direct link between stathmin-2 and the mitochondrial respiration could be found.

dibutylphthalate	dimethylphenol	chlorophenol	Ensembl Description	Gene Name
-1.55	-2.5	-1.87	stathmin-2 [Source:RefSeq peptide;Acc:NP_001019393]	stmn2b

Table 6.3: Expression values (M-value) of possible biomarker genes for disruption of the mitochondrial respiration when chlorophenol, dimethylphenol and dibutylphthalate were taken into account.

Dibutylphthalte is the only oxidative phorsohorylation uncoupler, all other compounds seem to rather act as inhibitors of the electron-transport-chain. When dibutylphthalate was excluded, more genes could be identified as potential biomarker genes (Table 6.4). Especially the mitochondrial uncoupling protein 4 (UCP4) and the ATPase atp1a1a.4, might be good candidates for biomarker genes. It is known that uncoupling proteins are regulated by ATP. Furthermore, an impairment of the mitochondrial respiration reduces the level of ATP and thereby induces uncoupling proteins (Criscuolo *et al.* 2006). The ATPase is part of the mitochondrial respiration system and a regulation in case of a disruption can be assumed. For the other genes, no hints of a regulation of mitochondrial respiration could be found in the literature.

The list of identified genes needs to be further analyzed and validated. Based on the literature data, however, ucp4 and atp1a1a.4 seem to be good candidates as possible biomarker genes for the disruption of the mitochondrial respiration.

Estrogen Receptor Activity

Only chlorophenol and dibutylphthalte effected estrogen receptor activity as shown by changes of the known biomarker gene vitellogenin (Table 6.5). It is known that vitellogenin is regulated in different fish species in an estrogen specifc manner and is therefore a good biomarker gene for estrogen receptor activity (Sumpter and Jobling 1995). No other evidence strengthened this hypothesis. However, one would assume that both compounds should display stronger similarities regarding other modes of action. The two compounds regulated over 200 genes in a similar way whose expression is not effected by any of one of the other compounds. 32 of them are highly expressed, but none of them could be linked to the estrogenic system (Table 6.6).

dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol	Ensembl Description	Gene Name
-4.81	0	-3.9	0	0	0	0	-1.07	0	-4.46	cone-rod homeobox [Source:RefSeq peptide;Acc:NP_694419]	crx
-5.18	0	-3.72	0	0	0	0	-0.95	0	-3.95	cone-rod homeobox [Source:RefSeq peptide;Acc:NP_694419]	crx
-3.72	0	-1.74	1.83	0	0	0	-0.55	0	-3.61	orthodenticle homolog 5 [Source:RefSeq peptide;Acc:NP_851848]	otx5
-2.37	0	-2.08	0	0	0	0.46	0.17	0	-1.96	coronin, actin binding protein, 1B [Source:RefSeq peptide;Acc:NP_001103177]	coro1b
-2.42	0	-3.65	0	0	-0.72	0	0	0	-4.11		si:dkey-1h6.1
-3.64	0	-2.82	0	0	0	0	0	0	-3.77	orthodenticle homolog 5 [Source:RefSeq peptide;Acc:NP_851848]	otx5
-3.22	0.21	-2.05	0	0	0	0	0	0	-3.44	Visual system homeobox 1 (Transcription factor VSX1) [Source:UniProtKB/Swiss-Prot;Acc:O42250]	vsx1
-2.72	-1	-1.54	0	0	0	0	0	0	-3.03	Neurogenic differentiation factor 1 (NeuroD1)(NeuroD) [Source:UniProtKB/Swiss-Prot;Acc:O42202]	neurod
-2.51	0	-2.64	1.68	0	0	0	0	0	-2.71	hypothetical protein LOC572084 [Source:RefSeq peptide;Acc:NP_001073527]	zgc:158291
-2.48	0	-1.89	1.37	0	-0.37	0.55	0	0	-2.61	hypothetical protein LOC767746 [Source:RefSeq peptide;Acc:NP_001070183]	zgc:153426
-2.72	0	-1.68	1.3	0	0	0	0	0	-2.35	barH-like 2 homeobox protein [Source:RefSeq peptide;Acc:NP_991303]	barhl2
-1.57	0	-2.19	1.24	-0.94	0	0	0	0	-2.27	RNA binding protein with multiple splicing 2 [Source:RefSeq peptide;Acc:NP_956553]	rbpms2
-2.23	0	-1.88	1.67	0	-0.68	0.66	0	0	-2.2	tubulin alpha-1D chain [Source:RefSeq peptide;Acc:NP_998195]	tuba2
-2.38	0	-2	1.39	0	0	0	0	0	-2.1	barH-like 2 homeobox protein [Source:RefSeq peptide;Acc:NP_991303]	barhl2
-1.8	0	-2.4	0	-0.72	0	0	0	0	-1.75	DNA-binding protein inhibitor ID-4 [Source:RefSeq peptide;Acc:NP_001035079]	zgc:123214
-1.85	0	-1.67	1.01	0	0	0	0	0	-1.69	hypothetical protein LOC335798 [Source:RefSeq peptide;Acc:NP_956269]	zgc:65894
1.76	0	1.63	-1.38	0	0	-0.72	0	0	1.94		cxcr4a
1.63	0	1.72	-1.18	0	0	0	0	0	2.33		cxcr4a
2.19	0	2.26	0	0	0	-0.85	0	0	3.15	mitochondrial uncoupling protein 3 [Source:RefSeq peptide;Acc:NP_955817]	ucp4
1.88	0.76	2.04	0	0	0.33	-0.8	0	0	3.71	ATPase, Na+/K+ transporting, alpha 1a.4 polypeptide [Source:RefSeq peptide;Acc:NP_571764]	atp1a1a.4
1.52	0	4.01	0	0	0	0	0	-2.4	4.24	hypothetical protein LOC447866 [Source:RefSeq peptide;Acc:NP_001004605]	zgc:92034

Table 6.4: Expression values (M-value) of possible biomarker genes for disruption of the mitochondrial respiration when chlorophenol, dimethylphenol and dibromoethane were taken into account.

Besides vitellogenin, no other good candidate emerged as possible biomarker genes for estrogen receptor activity. Consequently, an extension of the data set with compounds that act more specifically on the estrogen receptor would be required to obtain a better understanding of regulated pathways and might lead to a more promising set of biomarker genes.

dibutylphthalate	chlorophenol	Ensembl Description	Gene Name
3.74	4,33	hypothetical protein LOC678536 [Source:RefSeq peptide;Acc:NP_001038759]	vtg1

Table 6.5: The expression levels (M-value) of vitellogenin in the dibutylphthalate and chlorophenol microarrays.

Pregnan-X-Receptor Activity

In our data set, two compounds seem to have an effect on the pregnan-x-receptor (PXR) activity (Tabel 6.7). Dibutylphthalate and flucythrinate showed an up-regulation of the cyp3a65 gene. This gene is known to be regulated by PXR (Tseng *et al.* 2005). Unfortunately, no other evidence could be found to proof this. Additionally, no other gene was highly deregulated only by this two compounds. Therefore, it is not possible to suggest further biomarker genes specific for PXR activity.

Acetylcholinesterase Inhibition

Acetylcholinesterase (AChE) inhibition is a well studied toxicological mechanism. In our data, I was not able to detect any AChE inhibitory effect. It is possible that none of the 4 predicted AChE inhibitors were able to inhibited AChE in the zebrafish embryos, probably due to a low sensitivity of the embryos. On the other hand, very little is known about the effects of AChe inhibitors on gene expression levels. Although it is not possible to clearly determine whether specific compounds acted as AChE inhibitors, I still tried to detect possible biomarker genes. Propoxur (Smulders *et al.* 2003), chlorophenol(Liu and Liu 2011), dibutylphthalate (Jee *et al.* 2009), and chlorpyrifos (Sandahl *et al.* 2005) are the compounds predicted to act as AChE inhibitors. I excluded chlorpyrifos from that list because there was an obvious problem with the uptake of the compound in the embryo (Chapter 6.1.1). In Table 6.8 genes are shown with were highly regulated only in the three remaining compounds. Hspb11 is a promising candidate and is currently under investigation by a collaboration partner as possible biomarker gene for effects on AChE (data not published yet). This might indicate that the compounds had an effect on AChE activity, and that the other genes might also be good candidates as biomarker genes for this effect.

dibutylphthalate	chlorophenol	Ensembl Description	Gene Name
-2.1	-6.24	Neurogenic differentiation factor 2 (NeuroD2) [Source:UniProtKB/Swiss-Prot;Acc:Q9W6C8]	neurod2
-1.69	-5.55	retinoschisin [Source:RefSeq peptide;Acc:NP_001003438]	rs1
-2.13	-5.03	hypothetical protein LOC619266 [Source:RefSeq peptide;Acc:NP_001028919]	zgc:109965
-1.55	-4.67	recoverin [Source:RefSeq peptide;Acc:NP_001025419]	zgc:114180
-1.55	-3.84	Neurogenic differentiation factor 6-A (NeuroD6-A)(Protein atonal homolog 2-A) [Source:UniProtKB/Swiss-Prot;Acc:Q6NYU3]	atoh2a
-1.84	-3.27	recoverin [Source:RefSeq peptide;Acc:NP_956258]	rcv1
-1.6	-2.32		LOC568355
-1.7	-1.85	hyaluronidase PH-20 [Source:RefSeq peptide;Acc:NP_001074154]	spam1
-1.67	-1.67		A5PN32_DANRE
-1.66	-1.58	Zinc finger protein DPF3 [Source:UniProtKB/Swiss-Prot;Acc:A9LMC0]	dpf3
-2.08	-1.56	adenomatosis polyposis coli [Source:RefSeq peptide;Acc:NP_001137312]	apc
1.57	1.53	Charged multivesicular body protein 5 (Chromatin-modifying protein 5) [Source:UniProtKB/Swiss-Prot;Acc:Q7T339]	chmp5
1.98	1.56	Transmembrane protein 55B-B (EC 3.1.3.-)(Type I phosphatidylinositol 4,5-bisphosphate 4-phosphatase-B)(PtdIns-4,5-P2 4-Ptase I-B) [Source:UniProtKB/Swiss-Prot;Acc:Q66I51]	T55BB_DANRE
2.92	1.57	hypothetical protein LOC415223 [Source:RefSeq peptide;Acc:NP_001002133]	zgc:86757
3.05	1.68	Inversin [Source:UniProtKB/Swiss-Prot;Acc:Q8UVC1]	invs
2.54	1.69	hypothetical protein LOC415223 [Source:RefSeq peptide;Acc:NP_001002133]	zgc:86757
1.53	1.75	hypothetical protein LOC100149419 [Source:RefSeq peptide;Acc:NP_001137527]	NP_001137527.1
1.88	1.82		
3.34	1.89	asparagine synthetase [Source:RefSeq peptide;Acc:NP_957457]	asns
1.82	2.11	Ester hydrolase C11orf54 homolog (EC 3.1.-.-) [Source:UniProtKB/Swiss-Prot;Acc:Q6NWE0]	zgc:85789
3.01	2.12		wu:fk66d05
3.15	2.24	stanniocalcin-2 [Source:RefSeq peptide;Acc:NP_001014827]	stc2
4.31	2.32	neutral amino acid transporter A [Source:RefSeq peptide;Acc:NP_001002513]	slc1a4
5.19	2.44	rhomboid domain-containing protein 1 [Source:RefSeq peptide;Acc:NP_001017614]	zgc:110266
2.45	2.49	Protein spinster homolog 1 (Spinster-like protein)(Protein not really started) [Source:UniProtKB/Swiss-Prot;Acc:Q7ZU13]	spns1
3.2	2.56	prolactin [Source:RefSeq peptide;Acc:NP_852102]	prl
2.62	2.58	Novel protein similar to vertebrate carnitine palmitoyltransferase 1A (Liver) (CPT1A) Fragment [Source:UniProtKB/TrEMBL;Acc:B7ZCZ8]	si:ch211-236l14.3
2.63	2.67	Zgc:112399 protein [Source:UniProtKB/TrEMBL;Acc:A7MCA6]	zgc:112399
1.52	2.76	malonyl-CoA decarboxylase, mitochondrial [Source:RefSeq peptide;Acc:NP_001082972]	zgc:162977
2.06	3.16	hypothetical protein LOC768172 [Source:RefSeq peptide;Acc:NP_001070783]	zgc:153911
4.83	3.17	Hepatocellular carcinoma down-regulated mitochondrial carrier homolog A [Source:UniProtKB/Swiss-Prot;Acc:Q1ECW7]	zgc:136752
5.66	3.39	asparagine synthetase [Source:RefSeq peptide;Acc:NP_957457]	asns

Table 6.6: The expression levels (M-value) of genes expressed only in the dibutylphthalate and chlorophenol microarrays.

dibutylphthalate	flucythrinate	Ensembl Description	Gene Name
4.99	2.39	cytochrome P450, family 3, subfamily A, polypeptide 65 [Source:RefSeq peptide;Acc:NP_001032515]	cyp3a65

Table 6.7: The expression value (M-value) of cyp3a65 in dibutylphthalate and flucythrinate

Glutathione Metabolism

Some compounds do not have direct toxic effects on the organism since they are directly metabolized to less harmful substances. One of the ways to detoxify compounds is by the glutathione metabolism. In our study, propoxur (Chapter 6.1.1) and chlorthalonil (Chapter 6.1.1) showed an effect on this metabolic pathway. Although the glutathione metabolism should protect the organism, it can also be the cause for toxicity. The produced metabolites can be more toxic than the original compounds. Additionally, the glutathione, which is needed for the metabolic process, can be depleted. Glutathione is used in many metabolic and biochemical reactions, and a lack of the protein impairs normal cell functions (Di Giulio 2008). For this reason, I decided to also look for biomarker genes specific for effects on the glutathione metabolism. Table 6.9 summarizes the results. UDP glycosyltransferase 1 (ugt1ab), glutathione peroxidase 1 (gpx1a), glutathione S-transferase pi (gstp1) are all known to be a part of the glutathione metabolic pathway (Di Giulio 2008). Therefore, these genes are the most promising candidate biomarker genes. However, further investigations are necessary to test whether the expression levels of these genes are really glutathione dependent.

dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol	Ensembl Description	Gene Name
0	5.67	0	0	0	0	0	0.84	2.26	4.54	phosphoserine aminotransferase isoform 1 [Source:RefSeq peptide;Acc:NP_956113]	psat1
0	-2.52	0	0	0	0	0	0	-1.76	-2.09	hypothetical protein LOC553753 [Source:RefSeq peptide;Acc:NP_001018560]	zgc:113307
0	2.82	0	0	0	0	0	0	1.81	1.63	purine nucleoside phosphorylase 5b [Source:RefSeq peptide;Acc:NP_001004628]	zgc:101740
0	4.51	0	0	0	0	0	0	2.06	2.24	neutral amino acid transporter A [Source:RefSeq peptide;Acc:NP_001002513]	slc1a4
0	6.27	0	0	0	0	0	0	4.97	2.81	membrane-spanning 4-domains, subfamily A, member 17A.5 [Source:RefSeq peptide;Acc:NP_001004629]	ms4a17a.5
0	5.3	0	0	0	0	0	0	4.97	4.23	membrane-spanning 4-domains, subfamily A, member 17A.5 [Source:RefSeq peptide;Acc:NP_001004629]	ms4a17a.5
1.42	1.8	0	0	0	0	0	0	8.42	4.29	Heat shock protein beta-11 [Source:UniProtKB/Swiss-Prot;Acc:A5JV83]	hspb11

Table 6.8: Expression values (M-value) of possible biomarker genes for AChE inhibition.

dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol	Ensembl Description	Gene Name
0	0	0	0	0	-1.03	0	-1.91	-3.72	0	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 1 [Source:HGNC Symbol;Acc:16925]	si:ch211-135f11.1
0	0	0	0	0	0	0	1.77	2.45	0.64	retinol dehydrogenase 12, like [Source:RefSeq peptide;Acc:NP_001009912]	rdh12l
0	0.79	0	0	0	0	0	2.18	1.85	0	apoptosis-inducing factor, mitochondrion-associated, 2 [Source:HGNC Symbol;Acc:21411]	LOC557507
0	0	0	0	0	0	0	2.25	3.13	0	hypothetical protein LOC447803 [Source:RefSeq peptide;Acc:NP_001004542]	zgc:91887
0	0	0	0	0	0	0	2.43	4.27	0	UDP glycosyltransferase 1 family, polypeptide A1 precursor [Source:RefSeq peptide;Acc:NP_001032505]	ugt1ab
0	0	1.44	0	0	0	0	2.52	3	0	glutathione peroxidase 1 [Source:RefSeq peptide;Acc:NP_001007282]	gpx1a
0	0	0	0	0	0.87	0	3.91	4.27	0.99	transmembrane protease, serine 13a [Source:RefSeq peptide;Acc:NP_001152984]	TMPRSS13 (1 of 3)
0	0	0	0	0	0	0	5.14	4.66	0	SULT1 isoform 5 [Source:UniProtKB/TrEMBL;Acc:Q49IK6]	sult1st5
0	0	0	0	0	0	0	5.34	6.35	0	S100 calcium binding protein Z [Source:UniProtKB/TrEMBL;Acc:Q503K9]	s100z
0	0	0	0	0	0	0	9.05	6.85	-0.05	glutathione S-transferase pi [Source:RefSeq peptide;Acc:NP_571809]	gstp1

Table 6.9: Expression values (M-value) of possible biomarker genes for an activation of the glutathione metabolic pathway.

Induction of Apoptosis

Apoptosis can be induced by many compounds. In most cases, apoptosis occurs as a secondary toxicity effect in response to an impairment of another pathway. In the microarray analysis, I defined a set of genes specific for apoptosis based on Gene Ontology terms (Chapter 5.1.5). This list consisted of 271 transcripts. A comparison of differentially regulated genes with this list suggested that propoxur and chlorophenol induced apoptosis. During the analysis of the regulated pathways, dimethylphenol also appeared to induce apoptosis. This was confirmed by the strong induction of the known apoptosis biomarker genes, caspase 8 and tp53 (Chapter 6.1.1). In order to get a better set of genes, I searched for genes that were only highly regulated by chlorophenol, propoxur, and dimethylphenol. Table 6.10 contains the resulting gene list. For rasd1, mmp, thioredoxin, and tp53 a connection with apoptosis could be found in the literature (Vaidyanathan *et al.* 2004; Nordskog *et al.* 2003; Masutani *et al.* 2005). This suggests that the list of genes presented in Table 6.10 is a good indicator for induced apoptosis.

6.1.4 Linkage to other studies

In the present work, I tried to link our microarray data to other studies previously performed in zebrafish. Only few studies have been published using the early developmental stage and none which employed fish at the same stage and the same microarray system. Nevertheless, I could link the data from our study with two other studies. As described in Chapter 5.1.4, I mapped the genes of the different platforms to our Agilent system. This renders the datasets comparable on the basis whether a gene is de-regulated or not. The expression levels can not be compared and any multivariate statistics analysis, such as clustering, is also not possible.

Biosensor Data

When I compared the data of the biosensor study (Yang *et al.* 2007) with our 10 compound study, I could find only four compounds that showed similarity (Chapter 5.1.4). Chlorophenol and TCDD have 188 regulated transcripts in common. The gene function analysis of these genes revealed an effect on the canonical WNT and FGF signaling pathways. Biological processes in development were also affected. In the literature, no similarity of the effects of the two compounds could be found. TCDD is known to influence the canonical WNT pathway through the aryl hydrocarbon receptor in zebrafish (Mathew *et al.* 2008). However, such an effect has not been previously shown for chlorophenol. The only link between TCDD and chlorophenol I could find is that TCDD is known to be a trace by-product in the synthesis of chlorophenols (Beischlag *et al.* 2008). However, it is not to be expected that the amount of TCDD in our used chlorophenol sample (Pestanal analytical standard grade) is so high that it can alter gene expression. To answer the question why these two compounds regulate these genes, further detailed experiments are

dibromoethane	dibutylphthalate	dimethylphenol	esfenvalerate	flucythrinate	chlorpyrifos	methoxychlor	chlorthalonil	propoxur	chlorophenol	Ensembl Description	Gene Name
-0.5	0	2.89	0	0	0	0	-0.19	1.6	1.64	homologous-pairing protein 2 homolog [Source:RefSeq peptide;Acc:NP_001002124]	psmc3ip
0	0	1.56	-0.92	0	0	-0.46	0.28	2.08	1.95	gremlin-2 [Source:RefSeq peptide;Acc:NP_001017704]	grem2
0	0	-4.11	0	0	0	0	0	-6.28	-4.99	solute carrier family 22 member 2 [Source:RefSeq peptide;Acc:NP_998315]	zgc:64076
0	0	-1.7	0	0	0	0	0	-2.64	-2.46	serum/glucocorticoid regulated kinase 1-like [Source:RefSeq peptide;Acc:NP_001070770]	si:ch211-195b13.1
0	0	1.77	0	0	0	0	0	1.76	1.74	kinesin-like protein KIF2C [Source:RefSeq peptide;Acc:NP_001108065]	si:ch211-61f14.1
0	0	2.23	0	0	0	-0.45	0	1.58	1.76		LOC570063
0	0	1.82	0	0	0	0	0	2.34	1.99	thioredoxin [Source:RefSeq peptide;Acc:NP_001002461]	zgc:92903
0	0	2.15	0	1.2	0	0	0	2.56	3.52	hypothetical protein LOC561924 [Source:RefSeq peptide;Acc:NP_001076294]	zgc:162630
1.13	0	1.89	0	0	0	0	0	5.28	3.63	matrix metalloproteinase-9 [Source:RefSeq peptide;Acc:NP_998288]	mmp9
0	0	1.84	0	0	0	0	0	3.25	3.87	hypothetical protein LOC641321 [Source:RefSeq peptide;Acc:NP_001032194]	zgc:123218
0	0	2.35	-1.34	0	0	0	0	1.81	4.13	Cellular tumor antigen p53 (Tumor suppressor p53) [Source:UniProtKB/Swiss-Prot;Acc:P79734]	tp53
0	0	4.17	0	0	0	-1.34	0	2.28	4.82	dexamethasone-induced Ras-related protein 1 [Source:RefSeq peptide;Acc:NP_956826]	rasd1

Table 6.10: Expression values (M-value) of genes only highly regulated in chlorophenol, propoxur and dimethylphenol.

necessary. In case of the other two compounds, cadmiumchloride and dibromoethane, the number of co-regulated genes is so small (10 transcripts) that no gene function analysis could be performed. In the literature no evidence of shared modes of action could be found. Even if they co-regulated more genes than most other compounds in these studies, the small number might still be simply by chance. In general, it proved to be quite difficult to link data sets of different studies. Especially when the used microarray platforms are so different (Compugen 22k and Agilent 4x44k) with only around 7000 genes shared between both arrays.

Immune Response Data

Based on the list of genes involved in immune response published by Stockhammer *et al.* 2009, all compounds showed an effect on the immune system. This is not surprising as the immune system is the defense system that protects the organism from external induced damage. With this analysis, I hoped to identify compounds that have a specific immunotoxic effect. Whether none of the compounds were immunotoxic or the set of genes was too generic for this purpose remains unclear. Genes like tp53 and several caspases are contained in the list of immune response genes. Therefore, it can be assumed that this list describes a very broad gene response including apotosis and therefore is not specific for the basic immune reaction.

6.1.5 Conclusion

In this study I analyzed the gene expression data of zebrafish embryos treated from 24-48 hpf with 10 different compounds. I used multivariate statistical methods to identify compounds with similar expression patterns. Furthermore, I tried to identify similarities by counting the number of co-regulated genes. To understand the modes of action of the compounds, I performed a gene function analysis of the significantly differentially expressed genes. I validated my findings using literature data. In order to identify biomarker genes, I grouped the compounds based on the identified modes of action and searched for genes that were only de-regulated after treatment with compounds with the same mode of action. I defined sets of biomarker genes for the modes of action: disruption of mitochondrial potential, Acetylcholinesterase inhibition, Glutathione metabolism, and induction of apoptosis. These lists of biomarker genes are interesting hypotheses but require further validation through experiments. I also tried to link the data obtained from the ten compounds to other toxicity microarray studies performed in zebrafish embryos. Unfortunately, the comparability of the used microarray platforms was too small to obtain any usable results.

In this work, I described the modes of action of 10 different compounds. For some of the compounds, this was the first time they have been studied in a fish species. I could show that most compounds act through several modes of action at the same time. The

detected toxicity mechanisms were not always expected based on the available literature. This underlines how important it is to first identify the modes of action and search for biomarker genes based on these results. I defined lists of biomarker genes for four different modes of action. In this study, we showed that the zebrafish embryo is a very useful tool to study the toxicity of chemical compounds. Nevertheless, my results also show that it needs to be taken into account that the chorion might influence the uptake of a compound.

6.2 Whole Genome Array

During my studies, I realized that the commercially available zebrafish microarrays always lack several important genes. To overcome this problem, I designed an array that covers almost the whole zebrafish genome. This array design led to an improvement of around 10% compared to the Agilent v2 and around 5% to the Agilent v3 array (Chapter 5.2.1). An update of the whole genome array based on the new gene build Zv9 might further increase this factor. I could also show that our new array can be used for very early stages in the development like gastrulation (Chapter 5.2.2). One disadvantage of my design is that due to the number of oligos needed to cover the whole genome, it consists of two arrays. Therefore, the RNA samples need to be split. This might introduce errors as the sample can never be completely homogeneous. In order to investigate possible negative effects of this design, I used spike-in controls. I could show that splitting the RNA sample after the labeling process induces less errors than splitting the samples before the labeling process (Chapter 5.2.3). For dye swap experiments, the samples are usually split before the labeling process. Dye swap experiments are very common in microarray analysis and the introduced errors are known to be not problematic (Simon *et al.* 2004). The newly designed whole genome array can clearly improve microarray experiments. Splitting the RNA is not a major problem and data from the first studies performed with this array look very promising (data not published yet).

6.3 Transcription Factor Study

For this study, I analyzed the expression patterns of the transcription factors during zebrafish development, in the adult brain, and muscle tissue. In the following, I summarize the results and present a list of biomarker genes specific for 5 different developmental stages and the examined tissue samples.

6.3.1 Developmental Stages

The analysis of the different developmental stages showed that in all stages a similar amount of around 2670 transcription factors is expressed. In the cluster analysis, three

main clusters were detectable (Chapter 5.3.5. The very early stages (2-cell, 30%- epiboly) formed a cluster as well as the middle embryonic stages (1-6 somites, 24hpf and 48hpf), and the late embryonic stage (5 dpf) dataset represented the third group. This suggests that at least two major transcriptional regulation changes exist. The first at the beginning of the early gastrulation, and a second one when the embryos hatch.

Gene Ontology Analysis

To further investigate the transcriptional changes during development, I decided to perform a more detailed analysis of the changes of the expression over time. I aimed at detecting transcription factors that showed a similar pattern in their expression over time. Furthermore, I wanted to know which patterns (profiles) are the most common ones. With the help of the program STEM (Ernst and Bar-Joseph 2006), I could detect 11 significantly enriched profiles (Chapter 5.3.6). To further evaluate the profiles, I performed a Gene Ontology analysis with the genes associated with each profile. The results are presented in Table 6.11 and Appendix B. Profiles having the highest expression (peak) at the 2-cell and 30% epiboly stage were related with gastrulation and protein metabolism. Tay *et al.* 2006 showed also a peak in protein expression at around 6 hpf. Profiles describing a similar expression over the whole development (Profile 49 and 48) were linked with organ development. The profiles that peaked at the 5 dpf stage were enriched in nervous system development and biosynthesis according to the GO analysis.

Time Depended Biomarker Genes

Based on the results of the developmental stages in the transcription factor study, I defined a set of biomarker genes that are specific for each of the six developmental stages used. 289 transcription factors were expressed only in one stage. These biomarker genes can be used to identify, for example, developmental delays in compound exposure experiments. The number of specific transcripts for each stage can be found in Table 6.12.

The early 2-cell and the late 5dpf stage showed the highest amount of specifically expressed transcription factors. Due to the size of the list, it is only included on the supplementary CD.
In order to detect whether certain treatments caused a developmental delay, I used the 24 hpf biomarker gene set on the 10 compound data. However, none of the genes was differentially regulated. This might be because the concentrations were chosen not to cause any phenotypic effect.

6.3.2 Tissues

I analyzed four different tissue samples. The tail sample represented a muscle rich tissue; the other three samples were whole head, representing the brain, and two specific parts

Profile	2-cell	30%-epioly	1-6 somites	24 hpf	48 hpf	5 dpf	Gene Ontology terms
39						■	Nervous system development, biosynthesis, gene expression, regulation of biological quality
43	■						Gastrulation, protein transport, cellular response to stimulus, cellular component movement, chromosome organization, small GTPase mediated signal transduction, protein modification by small protein conjugation or removal
45	■					■	Protein modification by small protein conjugation or removal, nucleocytoplasmic transport, cellular metabolism, organelle organization, response to stress, small GTPase mediated signal transduction
47	■	■					Embryo development, cell differeantiation, organelle organization, cellular component organization, protein metabolism, signal transduction
49	■	■	■	■	■		Organ morphogenesis, chromatin organization, DNA metabolism, cellular component organization, cellular response to stimulus
44	■						Protein modification by small protein conjugation or removal
48	■	■	■	■	■	■	Organ development
18					■		Cellular developmental process, negative regulation of cellular process, biosynthesis
41			■				Positive regulation of cellular process, nucleic acid metabolism, pattern specification process, cellular response to stimulus
38		■					Gastrulation, signal transduction, protein modification by small protein conjugation or removal
23						■	Cell development, cytoskeleton organization , nervous system, cellular nitrogen compound metabolism, regulation of cellular process, cellular component organization

Table 6.11: Gene Ontology results from the 11 profiles. Red cells mark the peak of the profile.

	2-cell	30% epiboly	1-6 somites	24 hpf	48 hpf	5 dpf
unique	119	40	27	6	16	81

Table 6.12: Stage specific expressed transcription factors

of the brain, the diencephalon and the telencephalon. In the cluster analysis, the head sample clustered together with the 5dpf larva stage. Interestingly, the tail sample clustered with the pre-gastrula stages. Further analysis revealed that this seems to be caused due to bone and other tissue impurities in the tail sample. The two brain tissues shared no high similarity with any of the developmental stages. They also did not show a high similarity with the head, but as expected, they had more similarity with the head than with the tail sample (Chapter 5.3.5). Interestingly, the diencephalon showed the highest amount of expressed transcription factors (2666). The other tissues were slightly bellow (head 2391, telencephanoln 2443, tail 2493) (Chapter 5.3.4). Based on the results of the microarray analysis, I defined sets of biomarker genes specific for the four tissues. The lists are shown in Appendix C. Transcription factors expressed in the head sample were not excluded from being a possible biomarker gene specific for the telencephalon or the diencephalon and the other way around.

6.3.3 Conclusion

The transcription factor study should help to obtain deeper insights into the transcriptional regulation during zebrafish development. Additionally, we were also interested in the different transcription factors expressed in muscle and brain. For this reason, I designed a new microarray consisting only of transcription factors. We performed microarrays for 6 different developmental stages and four different tissue samples. In order to be able to compare all the different datasets, I developed a new analysis method. My approach is able to detect expressed transcripts without requiring a control dataset but still makes use of both color channels. In general, around 2670 transcription factors were expressed in the different developmental samples. I could detect two major changes in the transcriptional expression pattern during the development. One at the beginning of gastrulation and a second one at around 48 hpf when the embryos hatch. I could also detect groups of transcription factors that exhibited a similar expression pattern over time. The Gene Ontology analysis of the patterns revealed that transcription factors with highest expression before gastrulation were mostly involved in protein metabolism. Transcription factors expressed at similar levels during the whole development period were likely involved in organ development, and transcription factors peaking at the end of the development seemed to be mostly involved in the nervous system development and biosynthesis. Based on the results of the microarray analysis, I defined biomarker genes specific for the 6 developmental stages used in this study. The analysis of the tissue samples revealed that expression patterns of the adult tail shared high similarity with pre-gastrula stages whereas the adult head showed a similar expression like the 5 dpf larva. Further analysis revealed that this

seems to be caused due to bone and other tissue impurities in the tail sample. In all tissue samples, more then 2400 transcription factors were expressed. With the help of the microarray results, I designed biomarker genes specific for diencephalon, telencephalon, whole brain (head sample), and for tail tissue (tail sample). For most of the biomarker genes, I could find evidence that they are expressed in certain tissues or stages, but in all cases, it is known that they are also expressed in other stages or tissues. The detection limit of microarrays makes it quite difficult to use them for identification of specific biomarker genes. If genes are only expressed in a few cells, microarrays are not able to detect an expression signal. This means that genes need to be either highly expressed in a few cells or at moderate levels across the whole tissue or embryo. Furthermore, we used only four different tissues. Consequently, we cannot exclude the possibility that a transcription factor is expressed in any other tissue. The same applies for the biomarker genes specific for the developmental stages. The biomarker genes are not specific in the sense that they are expressed uniquely in one specific tissue or stage. They rather represent transcription factors exhibiting a striking expression pattern specific for only one of the samples in our study. Since transcription factors are key players in the regulation of gene transcription, the biomarker genes identified here may still play an important role in the transcriptional regulation in their associated stage or tissue.

Bibliography

Abd-Elraof, T.K., Dauterman, W.C., and Mailman, R.B. In vivo metabolism and excretion of propoxur and malathion in the rat: Effect of lead treatment. *Toxicology and Applied Pharmacology*, 1981. **59**(2):324 – 330. ISSN 0041-008X.

Agilent. *Agilent protocol*. Agilent, 2006.

Alderton, W., Berghmans, S., Butler, P., Chassaing, H., Fleming, A., Golder, Z., Richards, F., and Gardner, I. Accumulation and metabolism of drugs and CYP probe substrates in zebrafish larvae. *Xenobiotica*, 2010. **40**(8):547–557.

Alexeyenko, A., Wassenberg, D.M., Lobenhofer, E.K., Yen, J., Linney, E., Sonnhammer, E.L.L., and Meyer, J.N. Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity. *PLoS One*, 2010. **5**(5):e10465.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. Basic local alignment search tool. *J Mol Biol*, 1990. **215**(3):403–410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000. **25**(1):25–29.

Baier-Anderson, C. and Anderson, R.S. Suppression of superoxide production by chlorothalonil in striped bass (Morone saxatilus) macrophages: the role of cellular sulfhydryls and oxidative stress. *Aquatic Toxicology*, 2000. **50**(1-2):85 – 96. ISSN 0166-445X.

Barry, M.J., Logan, D.C., van Dam, R.A., Ahokas, J.T., and Holdway, D.A. Effect of age and weight-specific respiration rate on toxicity of esfenvalerate pulse-exposure to the Australian crimson-spotted rainbow fish (Melanotaenia fluviatilis). *Aquatic Toxicology*, 1995. **32**(2-3):115 – 126. ISSN 0166-445X.

Beischlag, T.V., Morales, J.L., Hollingshead, B.D., and Perdew, G.H. The aryl hydrocarbon receptor complex and the control of gene expression. *Crit Rev Eukaryot Gene Expr*, 2008. **18**(3):207–250.

Berg, M.V.D. Role of metabolism in the endocrine-disrupting effects of chemicals in aquatic and terrestrial systems. *Pure and Applied Chemistry*, 2003. **75**(11-12):1917–1932.

Botti, B., Ceccarelli, D., Tomasi, A., Vannini, V., Muscatello, U., and Masini, A. Biochemical mechanism of GSH depletion induced by 1,2-dibromoethane in isolated rat liver mitochondria. Evidence of a GSH conjugation process. *Biochim Biophys Acta*, 1989. **992**(3):327–332.

Chou, A.P., Li, S., Fitzmaurice, A.G., and Bronstein, J.M. Mechanisms of rotenone-induced proteasome inhibition. *NeuroToxicology*, 2010. **31**(4):367 – 372. ISSN 0161-813X.

Comparative, N.S., Park, K.J., Booth, F., and Hudson, P.J. Ann. Zool. Fennici 39: 21–28 ISSN 0003-455X Helsinki 6 March 2002 Finnish Zoological and Botanical Publishing Board 2002 Breeding losses of red grouse in Glen Esk, 2001.

COOKE, M.S., EVANS, M.D., DIZDAROGLU, M., and LUNEC, J. Oxidative DNA damage: mechanisms, mutation, and disease. *The FASEB Journal*, 2003. **17**(10):1195–1214.

Criscuolo, F., Mozo, J., Hurtaud, C., Nübel, T., and Bouillaud, F. UCP2, UCP3, avUCP, what do they do when proton transport is not stimulated? Possible relevance to pyruvate and glutamine metabolism. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2006. **1757**(9-10):1284 – 1291. ISSN 0005-2728. Mitochondria: from Molecular Insight to Physiology and Pathology.

Davies, P. The toxicology and metabolism of chlorothalonil in fish. II. Glutathione conjugates and protein binding. *Aquatic Toxicology*, 1985a. **7**(4):265 – 275. ISSN 0166-445X.

Davies, P. The toxicology and metabolism of chlorothalonil in fish. III. Metabolism, enzymatics and detoxication in Salmo spp. and Galaxias spp. *Aquatic Toxicology*, 1985b. **7**(4):277 – 299. ISSN 0166-445X.

Davies, P. and White, R. The toxicology and metabolism of chlorothalonil in fish. I. Lethal levels for Salmo gairdneri, Galaxias maculatus, G. truttaceus and G. auratus and the fate of 14C-TCIN in S. gairdneri. *Aquatic Toxicology*, 1985. **7**(1-2):93 – 105. ISSN 0166-445X.

Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., and Mattingly, C.J. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res*, 2009. **37**(Database issue):D786–D792.

Di Giulio, R. *The Toxicology of Fishes*. CRC Press, 2008. ISBN 041524868X.

Dolphin, A.C. Mechanisms of modulation of voltage-dependent calcium channels by G proteins. *J Physiol*, 1998. **506 (Pt 1)**:3–11.

Duchnowicz, P., Szczepaniak, P., and Koter, M. Erythrocyte membrane protein damage by phenoxyacetic herbicides and their metabolites. *Pesticide Biochemistry and Physiology*, 2005. **82**(1):59 – 65. ISSN 0048-3575.

Erdmann, F., Jarczowski, F., Weiwad, M., Fischer, G., and Edlich, F. Hsp90-mediated inhibition of FKBP38 regulates apoptosis in neuroblastoma cells. *FEBS Letters*, 2007. **581**(29):5709 – 5714. ISSN 0014-5793.

Ernst, J. and Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 2006. **7**:191.

Fan, C.Y., Cowden, J., Simmons, S.O., Padilla, S., and Ramabhadran, R. Gene expression changes in developing zebrafish as potential markers for rapid developmental neurotoxicity screening. *Neurotoxicol Teratol*, 2010. **32**(1):91–98.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. *et al.* The Pfam protein families database. *Nucleic Acids Res*, 2010. **38**(Database issue):D211–D222.

Fleury, C., Mignotte, B., and Vayssière, J.L. Mitochondrial reactive oxygen species in cell death signaling. *Biochimie*, 2002. **84**(2-3):131 – 141. ISSN 0300-9084.

Flicek, P., Aken, B.L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S. *et al.* Ensembl's 10th year. *Nucleic Acids Res*, 2010. **38**(Database issue):D557–D562.

Gallagher, E.P., Kedderis, G.L., and Giulio, R.T.D. Glutathione S-transferase-mediated chlorothalonil metabolism in liver and gill subcellular fractions of channel catfish. *Biochemical Pharmacology*, 1991. **42**(1):139 – 145. ISSN 0006-2952.

Grunwald, D.J. and Eisen, J.S. Headwaters of the zebrafish [mdash] emergence of a new model vertebrate. *Nat Rev Genet*, 2002. **3**(9):717–724. ISSN 1471-0056.

Gupta, R., Schuh, R., Fiskum, G., and Flaws, J. Methoxychlor causes mitochondrial dysfunction and oxidative damage in the mouse ovary. *Toxicology and Applied Pharmacology*, 2006. **216**(3):436 – 445. ISSN 0041-008X.

Hill, A.J., Teraoka, H., Heideman, W., and Peterson, R.E. Zebrafish as a Model Vertebrate for Investigating Chemical Toxicity. *Toxicological Sciences*, July 2005. **86**(1):6–19.

Hill, R.L. and Janz, D.M. Developmental estrogenic exposure in zebrafish (Danio rerio): I. Effects on sex ratio and breeding success. *Aquatic Toxicology*, 2003. **63**(4):417 – 429. ISSN 0166-445X.

Holdway, D.A. and Dixon, D. Effects of methoxychlor exposure of flagfish eggs (Jordanella floridae) on hatchability, juvenile methoxychlor tolerance and whole-body levels of tryptophan, serotonin and 5-hydroxyindoleacetic acid. *Water Research*, 1986. **20**(7):893 – 897. ISSN 0043-1354.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res*, 2009. **37**(Database issue):D211–D215.

Inouye, B., Ogino, Y., Ishida, T., Ogata, M., and Utsumi, K. Effects of phthalate esters on mitochondrial oxidative phosphorylation in the rat. *Toxicology and Applied Pharmacology*, 1978. **43**(1):189 – 198. ISSN 0041-008X.

Jain, K.K. *The Handbook of Biomarkers*. Humana Press, 2010. ISBN 160761684X.

Jee, J.H., Koo, J.G., Keum, Y.H., Park, K.H., Choi, S.H., and Kang, J.C. Effects of dibutyl phthalate and di-ethylhexyl phthalate on acetylcholinesterase activity in bagrid catfish, Pseudobagrus fulvidraco (Richardson). *Journal of Applied Ichthyology*, 2009. **25**(6):771–775. ISSN 1439-0426.

Jo, D., Liu, D., Yao, S., Collins, R.D., and Hawiger, J. Intracellular protein therapy with SOCS3 inhibits inflammation and apoptosis. *Nat Med*, 2005. **11**(8):892–898.

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

Karan, G., Yang, Z., and Zhang, K. Expression of wild type and mutant ELOVL4 in cell culture: subcellular localization and cell viability. *Mol Vis*, 2004. **10**:248–253.

Kienle, C., Köhler, H.R., and Gerhardt, A. Behavioural and developmental toxicity of chlorpyrifos and nickel chloride to zebrafish (Danio rerio) embryos and larvae. *Ecotoxicology and Environmental Safety*, 2009. **72**(6):1740 – 1747. ISSN 0147-6513.

Kojima, H., Takeuchi, S., and Nagai, T. Endocrine-disrupting Potential of Pesticides via Nuclear Receptors and Aryl Hydrocarbon Receptor. *Journal of Health Science*, 2010. **56**(4):374–386.

Kummerfeld, S.K. and Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res*, 2006. **34**(Database issue):D74–D81.

Levine, S.L. and Oris, J.T. CYP1A expression in liver and gill of rainbow trout following waterborne exposure: implications for biomarker determination. *Aquatic Toxicology*, 1999. **46**(3-4):279 – 287. ISSN 0166-445X.

Liu, Q.Y. and Liu, J.X. Kinetic Studies for the Inhibition Effect of 4-Chlorophenol on Acetylcholinesterase Activity. In *Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on*. ISSN 2151-7614, 2011 pages 1 –4.

Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z. *et al.* 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Res*, 2003. **13**(3):455–466.

Lu, Y., Fukuyama, S., Yoshida, R., Kobayashi, T., Saeki, K., Shiraishi, H., Yoshimura, A., and Takaesu, G. Loss of SOCS3 Gene Expression Converts STAT3 Function from Anti-apoptotic to Pro-apoptotic. *Journal of Biological Chemistry*, 2006. **281**(48):36683–36690.

Ma, T. and Chambers, J.E. Kinetic parameters of desulfuration and dearylation of parathion and chlorpyrifos by rat liver microsomes. *Food Chem Toxicol*, 1994. **32**(8):763–767.

Masutani, H., Ueda, S., and Yodoi, J. The thioredoxin system in retroviral infection and apoptosis. *Cell Death Differ*, 2005. **12**(S1):991–998. ISSN 1350-9047.

Mathavan, S., Lee, S.G.P., Mak, A., Miller, L.D., Murthy, K.R.K., Govindarajan, K.R., Tong, Y., Wu, Y.L., Lam, S.H., Yang, H. *et al.* Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet*, 2005. **1**(2):260–276.

Mathew, L.K., Sengupta, S.S., LaDu, J., Andreasen, E.A., and Tanguay, R.L. Crosstalk between AHR and Wnt signaling through R-Spondin1 impairs tissue regeneration in zebrafish. *The FASEB Journal*, 2008. **22**(8):3087–3096.

Mikut, R., Burmeister, O., Braun, S., and Reischl, M. The Open Source Matlab Toolbox Gait-CAD and its Application to Bioelectric Signal Processing. *Proc.,DGBMT-Workshop Biosignalverarbeitung, Potsdam*, 2008. pages 109–111.

Molecular Devices, C. *GenePix Pro 6.0 MICROARRAY ACQUISITION AND ANALYSIS SOFTWARE FOR GENEPIX MICROARRAY SCANNERS Users Guide and Tutorial*, 2005.

Neumann, N.F. and Galvez, F. DNA microarrays and toxicogenomics: applications for ecotoxicology? *Biotechnol Adv*, 2002. **20**(5-6):391–419.

Nordskog, B., Blixt, A., Morgan, W., Fields, W., and Hellmann, G. Matrix-degrading and pro-inflammatory changes in human vascular endothelial cells exposed to cigarette smoke condensate. *Cardiovascular Toxicology*, 2003. **3**:101–117. ISSN 1530-7905. 10.1385/CT:3:2:101.

Ochi, H., Hans, S., and Westerfield, M. Smarcd3 regulates the timing of zebrafish myogenesis onset. *J Biol Chem*, 2008. **283**(6):3529–3536.

Ogawa, Y., Kawamura, Y., Wakui, C., Mutsuga, M., Nishimura, T., and Tanamoto, K. Estrogenic activities of chemicals related to food contact plastics and rubbers tested by the yeast two-hybrid assay. *Food Addit Contam*, 2006. **23**(4):422–430.

Ortiz-Zarragoitia, M. and Cajaraville, M.P. Effects of selected xenoestrogens on liver peroxisomes, vitellogenin levels and spermatogenic cell proliferation in male zebrafish. *Comp Biochem Physiol C Toxicol Pharmacol*, 2005. **141**(2):133–144.

Ortiz-Zarragoitia, M., Trant, J.M., and Cajaravillet, M.P. Effects of dibutylphthalate and ethynylestradiol on liver peroxisomes, reproduction, and development of zebrafish (Danio rerio). *Environ Toxicol Chem*, 2006. **25**(9):2394–2404.

Pashkov, V., Huang, J., Parameswara, V.K., Kedzierski, W., Kurrasch, D.M., Tall, G.G., Esser, V., Gerard, R.D., Uyeda, K., Towle, H.C. *et al.* Regulator of G protein signaling (RGS16) inhibits hepatic fatty acid oxidation in a CHREBP-dependent manner. *Journal of Biological Chemistry*, 2011.

Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R., and Evelo, C. WikiPathways: pathway editing for the people. *PLoS Biol*, 2008. **6**(7):e184.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 2007. **35**(Database issue):D61–D65.

Riederer, B.M., Pellier, V., Antonsson, B., Paolo, G.D., Stimpson, S.A., Lütjens, R., Catsicas, S., and Grenningloh, G. Regulation of microtubule dynamics by the neuronal growth-associated protein SCG10. *Proc Natl Acad Sci U S A*, 1997. **94**(2):741–745.

Rossignol, D.P. Binding of a photoreactive pyrethroid to [beta] subunit of GTP-binding proteins. *Pesticide Biochemistry and Physiology*, 1991. **41**(2):121 – 131. ISSN 0048-3575.

Rouach, N., Pébay, A., Même, W., Cordier, J., Ezan, P., Etienne, E., Giaume, C., and Tencé, M. S1P inhibits gap junctions in astrocytes: involvement of G and Rho GTPase/ROCK. *Eur J Neurosci*, 2006. **23**(6):1453–1464.

Russell, S. *Microarray Technology in Practice*. Academic Press/Elsevier, Burlington, 2009. ISBN 9780123725165.

Sandahl, J.F., Baldwin, D.H., Jenkins, J.J., and Scholz, N.L. Comparative thresholds for acetylcholinesterase inhibition and behavioral impairment in coho salmon exposed to chlorpyrifos. *Environ Toxicol Chem*, 2005. **24**(1):136–145.

Schlicker, A. and Albrecht, M. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, 2008. **36**(Database issue):D434–D439.

Seth, V., Banerjee, B.D., and Chakravorty, A.K. Lipid Peroxidation, Free Radical Scavenging Enzymes, and Glutathione Redox System in Blood of Rats Exposed to Propoxur. *Pesticide Biochemistry and Physiology*, 2001. **71**(3):133 – 139. ISSN 0048-3575.

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., and Zhao, Y. *Design and Analysis of DNA Microarray Investigations*. Springer, 2004. ISBN 0387001352.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. BioMart–biological queries made easy. *BMC Genomics*, 2009. **10**:22.

Smulders, C.J.G.M., Bueters, T.J.H., Kleef, R.G.D.M.V., and Vijverberg, H.P.M. Selective effects of carbamate pesticides on rat neuronal nicotinic acetylcholine receptors and rat brain acetylcholinesterase. *Toxicology and Applied Pharmacology*, 2003. **193**(2):139 – 146. ISSN 0041-008X.

Spitsbergen, J.M. and Kent, M.L. The state of the art of the zebrafish model for toxicology and toxicologic pathology research–advantages and current limitations. *Toxicol Pathol*, 2003. **31 Suppl**:62–87.

Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Knight, J. *et al.* The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res*, 2008. **36**(Database issue):D768–D772.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 2002. **12**(10):1611–1618.

Stockhammer, O.W., Zakrzewska, A., Hegedûs, Z., Spaink, H.P., and Meijer, A.H. Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to Salmonella infection. *J Immunol*, 2009. **182**(9):5641–5653.

Sultatos, L.G. and Murphy, S.D. Kinetic analyses of the microsomal biotransformation of the phosphorothioate insecticides chlorpyrifos and parathion. *Fundam Appl Toxicol*, 1983. **3**(1):16–21.

Sumpter, J.P. and Jobling, S. Vitellogenesis as a biomarker for estrogenic contamination of the aquatic environment. *Environ Health Perspect*, 1995. **103 Suppl 7**:173–178.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., and Smuc, T. Translational selection is ubiquitous in prokaryotes. *PLoS Genet*, 2010. **6**(6):e1001004.

Takeuchi, S., Iida, M., Yabushita, H., Matsuda, T., and Kojima, H. In vitro screening for aryl hydrocarbon receptor agonistic activity in 200 pesticides using a highly sensitive reporter cell line, DR-EcoScreen cells, and in vivo mouse liver cytochrome P450-1A induction by propanil, diuron and linuron. *Chemosphere*, 2008. **74**(1):155 – 165. ISSN 0045-6535.

Tay, T.L., Lin, Q., Seow, T.K., Tan, K.H., Hew, C.L., and Gong, Z. Proteomic analysis of protein profiles during early development of the zebrafish, Danio rerio. *PROTEOMICS*, 2006. **6**(10):3176–3188. ISSN 1615-9861.

Thomas, C., Will, Y., Schoenberg, S.L., Sanderlin, D., and Reed, D.J. Conjugative metabolism of 1,2-dibromoethane in mitochondria: disruption of oxidative phosphorylation and alkylation of mitochondrial DNA. *Biochem Pharmacol*, 2001. **61**(5):595–603.

Thorpe, K.L. DEVELOPMENT OF AN IN VIVO SCREENING ASSAY FOR ESTROGENIC CHEMICALS USING JUVENILE RAINBOW TROUT (ONCORHYNCHUS MYKISS). *Environmental Toxicology and Chemistry*, 2000. **19**(11):2812–2820.

Tripathi, G. and Verma, P. Fenvalerate-induced changes in a catfish, Clarias batrachus: metabolic enzymes, RNA and protein. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 2004. **138**(1):75 – 79. ISSN 1532-0456.

Tsai, K.P. and Chen, C.Y. An algal toxicity database of organic toxicants derived by a closed-system technique. *Environmental Toxicology and Chemistry*, 2007. **26**(9):1931–1939. ISSN 1552-8618.

Tseng, H.P., Hseu, T.H., Buhler, D.R., Wang, W.D., and Hu, C.H. Constitutive and xenobiotics-induced expression of a novel CYP3A gene from zebrafish larva. *Toxicology and Applied Pharmacology*, 2005. **205**(3):247 – 258. ISSN 0041-008X.

Vaidyanathan, G., Cismowski, M.J., Wang, G., Vincent, T.S., Brown, K.D., and Lanier, S.M. The Ras-related protein AGS1//RASD1 suppresses cell growth. *Oncogene*, 2004. **23**(34):5858–5863. ISSN 0950-9232.

Versonnen, B.J., Roose, P., Monteyne, E.M., and Janssen, C.R. Estrogenic and toxic effects of methoxychlor on zebrafish (Danio rerio). *Environmental Toxicology and Chemistry*, 2004. **23**(9):2194–2201. ISSN 1552-8618.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., and Kere, J. The zebrafish transcriptome during early development. *BMC Developmental Biology*, 2011. **11**(1):30. ISSN 1471-213X.

Viant, M.R., Pincetich, C.A., and Tjeerdema, R.S. Metabolic effects of dinoseb, diazinon and esfenvalerate in eyed eggs and alevins of Chinook salmon (Oncorhynchus tshawytscha) determined by 1H NMR metabolomics. *Aquat Toxicol*, 2006a. **77**(4):359–371.

Viant, M.R., Pincetich, C.A., and Tjeerdema, R.S. Metabolic effects of dinoseb, diazinon and esfenvalerate in eyed eggs and alevins of Chinook salmon (Oncorhynchus

tshawytscha) determined by 1H NMR metabolomics. *Aquatic Toxicology*, 2006b. **77**(4):359 – 371. ISSN 0166-445X.

Virador, V.M., Davidson, B., Czechowicz, J., Mai, A., Kassis, J., and Kohn, E.C. The Anti-Apoptotic Activity of BAG3 Is Restricted by Caspases and the Proteasome. *PLoS ONE*, 2009. **4**(4):e5136.

Voelker, D., Vess, C., Tillmann, M., Nagel, R., Otto, G.W., Geisler, R., Schirmer, K., and Scholz, S. Differential gene expression as a toxicant-sensitive endpoint in zebrafish embryos and larvae. *Aquat Toxicol*, 2007. **81**(4):355–364. Fish test.

Westerfield, M. *The Zebrafish Book; A Guide for the Laboratory Use of Zebrafish (Brachydanio rerio)*, volume 2nd edition. University of Oregon Press, Eugene, 1993.

Wheelock, C.E., Eder, K.J., Werner, I., Huang, H., Jones, P.D., Brammell, B.F., Elskus, A.A., and Hammock, B.D. Individual variability in esterase activity and CYP1A levels in Chinook salmon (Oncorhynchus tshawytscha) exposed to esfenvalerate and chlorpyrifos. *Aquatic Toxicology*, 2005. **74**(2):172 – 192. ISSN 0166-445X.

White, R.D., Gandolfi, A.J., Bowden, G.T., and Sipes, I.G. Deuterium isotope effect on the metabolism and toxicity of 1,2-dibromoethane. *Toxicol Appl Pharmacol*, 1983. **69**(2):170–178.

Wu, Y., Foster, W.G., and Younglai, E.V. Rapid effects of pesticides on human granulosa-lutein cells. *Reproduction*, 2006. **131**(2):299–310.

Wyde, M.E., Kirwan, S.E., Zhang, F., Laughter, A., Hoffman, H.B., Bartolucci-Page, E., Gaido, K.W., Yan, B., and You, L. Di-n-Butyl Phthalate Activates Constitutive Androstane Receptor and Pregnane X Receptor and Enhances the Expression of Steroid-Metabolizing Enzymes in the Liver of Rat Fetuses. *Toxicological Sciences*, 2005. **86**(2):281–290.

Yang, L., Kemadjou, J.R., Zinsmeister, C., Bauer, M., Legradi, J., Müller, F., Pankratz, M., Jäkel, J., and Strähle, U. Transcriptional profiling reveals barcode-like toxicogenomic responses in the zebrafish embryo. *Genome Biol*, 2007. **8**(10):R227.

Zhang, A. *Advanced Analysis of Gene Expression Microarray Data (Science, Engineering, and Biology Informatics)*. World Scientific Publishing Company, 2006. ISBN 9812566457.

Zhang, B., Kirov, S., and Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 2005. **33**(Web Server issue):W741–W748.

Zhou, J., Liu, M., Zhai, Y., and Xie, W. The Antiapoptotic Role of Pregnane X Receptor in Human Colon Cancer Cells. *Molecular Endocrinology*, 2008. **22**(4):868–880.

List of Figures

List of Tables

Appendix A

Gene Function Analysis Tables

Table A.1: 4-Chlorophenol

	KEGG	WikiPathways	GO biological process	GO molecular function
all			cellular amino acid metabolism	nucleic acid binding
			nitrogen compound metabolism	transcription regulator activity
			digestive tract development	
all down	Spliceosome	mRNA processing	mRNA processing	nucleotide binding
	Basal transcription factors	G Protein Signaling Pathways	nitrogen compound metabolism	nucleic acid binding
		Calcium Regulation in the Cardiac Cell		binding
		Myometrial Relaxation and Contraction Pathways		RNA binding
all up	p53 signaling pathway	Apoptosis	cellular amino acid metabolism	
		Senescence and Autophagy	regulation of caspase activity	
		Toll-like receptor signaling pathway		
		Cell cycle		
		Adipogenesis		
filter all	p53 signaling pathway	Myometrial Relaxation and Contraction Pathways	cellular amino acid metabolsim	transcription regulator activity
	Glycine, serine and threonine metabolism		regulation of caspase activity	transcription factor activity
			nitrogen compound metabolism	DNA binding
filter down	Steroid biosynthesis	Myometrial Relaxation and Contraction Pathways	regulation of transcription	transcription regulator activity
		Calcium Regulation in the Cardiac Cell	nervous system development	transcription factor activity
		G Protein Signaling Pathways		sequence-specific DNA binding
filter up	p53 signaling pathway	Apoptosis	cellular amino acid metabolsim	insulin-like growth factor binding
	Glycine, serine and threonine metabolism	Androgen Receptor Signaling Pathway	regulation of caspase activity	growth factor binding
	Aminoacyl-tRNA biosynthesis	Senescence and Autophagy	death	
			response to biotic stimulus	
			Multi-organism process	

Table A.2: Chlorpyrifos

	KEGG	WikiPathways	GO biological process	GO molecular function
all		One Carbon Metabolism		
all down				
all up	Ribosome			structural constituent of ribosome
				metal ion binding
				cation binding
				ion binding
filter all				oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
				monooxygenase activity
				inorganic cation transmembrane transporter activity
filter down				
filter up				

Table A.3: Chlorothalonil

	KEGG	WikiPathways	GO biological process	GO molecular function
all	Glutathione metabolism		serine family amino acid biosynthetic process	
	Pentose phosphate pathway			
	Fructose and mannose metabolism			
	Drug metabolism - cytochrome P450			
	Metabolism of xenobiotics by cytochrome P450			
all down				
all up	Glutathione metabolism			catalytic activity
	Metabolism of xenobiotics by cytochrome P450			oxidoreductase activity
	Drug metabolism - cytochrome P450			oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
	Pentose phosphate pathway			NADP or NADPH binding
	Fructose and mannose metabolism			oxidoreductase activity, acting on CH-OH group of donors
filter all	Glutathione metabolism		oxidation reduction	disulfide oxireductase activity
	Fructose and mannose metabolism		response to chemical stimulus	catalytic activity
	Drug metabolism - cytochrome P450		cellular homeostasis	coenzyme binding
	Metabolism of xenobiotics by cytochrome P450		metabolism	glutathione transferase activity
	Pentose phosphate pathway			antioxidant activity
	Metabolic pathways			
filter down				
filter up	Glutathione metabolism		oxidation reduction	disulfide oxireductase activity
	Metabolism of xenobiotics by cytochrome P450		response to chemical stimulus	catalytic activity
	Drug metabolism - cytochrome P450		cellular homeostasis	coenzyme binding
	Fructose and mannose metabolism		metabolism	glutathione transferase activity
			coenzyme metabolism	antioxidant activity
			hexose metabolism	cofactor binding
			carbohydrate metabolism	
			cofactor metabolism	

Table A.4: 1,2-Dibromoethane

	KEGG	WikiPathways	GO biological process	GO molecular function
all	Proteasome	Proteasome Degradation		
all down	Proteasome	Proteasome Degradation	proteolysis involved in cellular protein catabolism	nucleoside-triphosphatase activity
	Gap junction		protein polymerization	nucleotide binding
			Microtubule-based process	binding
				nucleic acid binding
all up	Ribosome			structural constituent of ribosome
	Amino sugar and nucleotide sugar metabolism			structural molecule activity
	Pentose and glucuronate interconversions			
	Insulin signaling pathway			
	Metabolic pathways			
filter all	Proteasome		protein polymerization	
filter down	Proteasome	Proteasome Degradation	protein catabolism	Nucleoside-triphosphatase activity
	Gap junction		cellular protein complex assembly	transcription regulator activity
			macromolecule metabolsim	binding
			Microtubule-based movement	nucleotide binding
			Microtubule-based process	nucleic acid binding
			nervous system development	double stranded RNA binding
			cellular component organization	GTP binding
filter up	Insulin signaling pathway	Diurnally regulated genes with circadian orthologs		
		Circadian Exercise		

Table A.5: Di-n-butyl phtalate

	KEGG	WikiPathways	GO biological process	GO molecular function
all			central nervous system neuron differentiation	
			lipid biosynthesis	
			lipid metabolism	
all down	Focal adhesion	Wnt Signaling Pathway NetPath	sensory organ development	metal ion binding
	ECM-receptor interaction	noncanonical wnt pathway	gene expression	transcription regulator activity
		Delta-Notch Signaling Pathway	biological regulation	binding
		canonical wnt - zebrafish	cellular component organization	nucleic acid binding
		Androgen Receptor Signaling Pathway	developmental process	DNA binding
		Notch Signaling Pathway	multicellular organismal process	ion binding
		TGF-beta Receptor Signaling Pathway		protein binding
				extracellular matrix structural constituent
all up	SNARE interactions in vesicular transport	Cholesterol Biosynthesis	carboxylic acid metabolism	catalytic activity
	Metabolic pathways	Electron Transport Chain	lipid biosynthesis	transferase activity, transfering acyl groups
	Terpenoid backbone biosynthesis		protein transport	cofactor binding
	Fatty acid metabolism		response to salt stress	coenzyme binding
	Valine, leucine and isoleucine degradation		lipid metabolism	
	Porphyrin and chlorophyll metabolism			
	Lysosome			
	Oxidative phosphorylation			
	Aminoacyl-tRNA biosynthesis			
filter all	Terpenoid backbone biosynthesis	Cholesterol Biosynthesis	carboxylic acid metabolism	Aminoacyl-tRNA ligase activity
	Aminoacyl-tRNA biosynthesis	Fatty Acid Biosynthesis	lipid biosynthesis	hexosaminidase activity
	Lysosome		exocrine pancreas development	catalytic activity
	Metabolic pathways		lipid metabolism	extracellular matrix structural constituent
	Valine, leucine and isoleucine degradation		carboxylic acid transport	intramolecular oxidoreductase activity, transposing C=C bonds
	Glycosaminoglycan degradation		amine metabolism	
	Fatty acid metabolism			
	Glutathione metabolism			

	Glycosphingolipid biosynthesis - ganglio series			
	Fatty acid elongation in mitochondria			
filter down	ECM-receptor interaction		exocrine pancreas development	extracellular matrix structural constituent
			biological adhesion	
			cell adhesion	
filter up	Aminoacyl-tRNA biosynthesis	Cholesterol Biosynthesis	carboxylic acid metabolism	aminoacyl-tRNA ligase activity
	Metabolic pathways	Fatty Acid Biosynthesis	lipid biosynthesis	NADP or NADPH binding
	Terpenoid backbone biosynthesis	Endochondral Ossification	lipid metabolism	catalytic activity
	Lysosome		amine metabolism	intramolecular oxidoreductase activity, transposing C=C bonds
	Valine, leucine and isoleucine degradation		carboxylic acid metabolism	hexosaminidase activity
	Glycosaminoglycan degradation			
	Fatty acid metabolism			
	Glutathione metabolism			
	Fatty acid elongation in mitochondria			
	Glycosphingolipid biosynthesis - ganglio series			

Table A.6: 2,4-Dimethylphenol

	KEGG	WikiPathways	GO biological process	GO molecular function
all		mRNA processing	cell cycle	nucleic acid binding
			DNA metabolism	ligase activity, forming carbon-nitrogen bonds
			nitrogen compound metabolism	nucleotide binding
			Microtubule-based movement	RNA binding
				binding
				nucleoside binding
				purine nucleotide binding
all down	spliceosome	RNA processing	proton transport	GTP binding
	Ubiquitin mediated proteolysis	TCA Cycle	generation of precursor metabolites and energy	nucleic acid binding
	Proteasome	Proteasome Degradation	GPI anchor biosynthesis	nucleotide binding
	Oxidative phosphorylation		oxidative phosphorylation	
	Glycolysis / Gluconeogenesis		cellular macromolecular complex assembly	
	Synthesis and degradation of ketone bodies			
	Butanoate metabolism			
all up	Cell cycle	DNA Replication	glutamine metabolism	ATP binding
	Pyrimidine metabolism	One Carbon Metabolism	response to DNA damage stimulus	transferase activity
	p53 signaling pathway	Cell cycle	cell cycle checkpoint	catalyic activity
	One carbon pool by folate	ERK1 - ERK2 MAPK cascade	primary metabolism	DNA primase activity
	Glycine, serine and threonine metabolism	G1 to S cell cycle control	Camera-type eye development	nucleoside binding
		SIDS Susceptibility Pathways	cell cycle	pyridoxal phosphate binding
		Id Signaling Pathway	metabolism	cofactor binding
			response to stimulus	DNA binding
				binding
				nucleotide binding
				NADP or NADPH binding
filter all	Butanoate metabolism	DNA Replication	macromolecule metabolism	NADP or NADPH binding
	p53 signaling pathway	One Carbon Metabolism	DNA metabolism	
	One carbon pool by folate	G1 to S cell cycle control	response to stress	
	Glycine, serine and threonine metabolism	Cell cycle	protein complex assembly	
	Gap junction	Osteoclast	microtubule-based movement	

	Cell cycle		microtubule-based process	
			metabolism	
			cell cycle checkpoint	
filter down	Gap junction	Proteasome Degradation	cellular macromolecular complex assembly	nucleic acid binding
	Synthesis and degradation of ketone bodies	mRNA processing	microtubule-based movement	phospholipid binding
	Butanoate metabolism		microtubule-based process	nucleotide binding
	Propanoate metabolism			
	Pyruvate metabolism			
filter up	p53 signaling pathway	DNA Replication	cellular amino acid and derivative metabolism	DNA primase activity
	Cell cycle	Cell cycle	amine metabolism	NADP or NADPH binding
	One carbon pool by folate	G1 to S cell cycle control	response to stress	damaged DNA binding
	Glycine, serine and threonine metabolism	One Carbon Metabolism	response to stimulus	catalyic activity
	Pyrimidine metabolism	Delta-Notch Signaling Pathway	metabolism	hydrolase activity, hydrolyzing N-glycosyl compounds
	DNA replication	SIDS Susceptibility Pathways	nitrogen compound metabolism	nucleoside binding
	Glutathione metabolism	ERK1 - ERK2 MAPK cascade	embryonic cleavage	nucleotide binding
		Diurnally regulated genes with circadian orthologs	cell cycle checkpoint	nucleic acid binding
		Senescence and Autophagy	induction of apoptosis by intracellular signals	cofactor binding
		Androgen Receptor Signaling Pathway		

Table A.7: Esfenvalerate

	KEGG	WikiPathways	GO biological process	GO molecular function
all	Proteasome	Proteasome Degradation		nucleic acid binding
	RNA degradation			RNA binding
				GTPase activity
all down	RNA degradation	FGF signaling pathway	regulation of metabolsim	nucleic acid binding
		BMP signaling pathway	endoderm development	RNA binding
		Senescence and Autophagy	cell division	binding
		ERK1 - ERK2 MAPK cascade		
all up	Proteasome	Proteasome Degradation	protein folding	Nucleoside-triphosphatase activity
	SNARE interactions in vesicular transport			GTP binding
	Terpenoid backbone biosynthesis			nucleotide binding
	Fatty acid metabolism			unfolded protein binding
filter all				
filter down	Progesterone-mediated oocyte maturation			
	MAPK signaling pathway			
filter up	Terpenoid backbone biosynthesis			

Table A.8: Flucythrinate

	KEGG	WikiPathways	GO biological process	GO molecular function
all				
all down				
all up	Non-homologous end-joining		myofibril assembly	growth factor binding
			purine ribonucleoside triphosphate metabolism	binding
			cellular component organization	Nucleoside-triphosphatase activity
filter all		L-2 Signaling Pathway	regulation of cell growth	protein binding
		IL-6 Signaling Pathway	biological regulation	insulin-like growth factor binding
			response to stress	growth factor binding
			response to stimulus	
			growth	
			cellular process	
filter down				
filter up	NOD-like receptor signaling pathway	IL-2 Signaling Pathway	regulation of growth	protein binding
	Adipocytokine signaling pathway	IL-6 Signaling Pathway	response to stimulus	insulin-like growth factor binding
	Jak-STAT signaling pathway		response to stress	growth factor binding
	Progesterone-mediated oocyte maturation		growth	binding
			biological regulation	

Table A.9: Methoxychlor

	KEGG	WikiPathways	GO biological process	GO molecular function
all	Proteasome	Proteasome Degradation		
	Spliceosome	mRNA processing		
	RNA degradation			
all down	MAPK signaling pathway	canonical wnt - zebrafish	regionalization	peptidase inhibitor activity
	Hedgehog signaling pathway	FGF signaling pathway	developmental process	glycine hydroxymehtyltransferase activity
	Notch signaling pathway	Hedgehog Signaling Pathway	multicellular organismal process	enzyme regulator activity
		BMP signaling pathway	cell proliferation	Procollagen-proline 4-dioxygenase activity
		Integrin-mediated cell adhesion	hexose metabolism	
		neural crest development		
		noncanonical wnt pathway		
all up	Proteasome	Proteasome Degradation	macromolecule localization	GTPase activity
	Spliceosome	mRNA processing	cellular macromoleuclar complex assembly	GTP binding
	RNA degradation	Eukaryotic Transcription Initiation	proteolysis involved in cellular protein catabolism	nucleotide binding
		TNF-alpha NF-kB Signaling Pathway		methionine adenosyltransferase activity
				unfolded protein binding
filter all				unfolded protein binding
				L-ascorbic acid binding
				nucleoside binding
				purine nucleotide binding
				carboxylic acid binding
filter down			hexose metabolic process	L-ascorbic acid binding
			monosaccharide metabolic process	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
			alcohol metabolic process	carboxylic acid binding
				vitamin binding
filter up			protein folding	unfolded protein binding
				nucleoside binding
				purine nucleotide binding
				metallopeptidase activity

Table A.10: Propoxur

	KEGG	WikiPathways	GO biological process	GO molecular function
all	Aminoacyl-tRNA biosynthesis	Keap1-Nrf2	cellular amine metabolism	Aminoacyl-tRNA ligase activity
	Glutathione metabolism		response to methylmercury	catalytic activity
	p53 signaling pathway		regulation of apoptosis	antioxidant activity
	Drug metabolism - cytochrome P450		oxidation reduction	cofactorbinding
	Glycine, serine and threonine metabolism		death	Insuline-like growth factor binding
	Metabolism of xenobiotics by cytochrome P450			
	Apoptosis			
all down				
all up	Glutathione metabolism	Keap1-Nrf2	cellular amino acid metabolism	ligase activity
	Aminoacyl-tRNA biosynthesis	IL-6 Signaling Pathway	response to stress	protein dimerization activity
	p53 signaling pathway	Apoptosis	regulation of biological quality	catalytic activity
	Metabolism of xenobiotics by cytochrome P450	TNF-alpha NF-kB Signaling Pathway	oxidation reduction	iron ion binding
	Apoptosis	Toll-like receptor signaling pathway	death	adenyl nucleotide binding
	Drug metabolism - cytochrome P450	ERK1 - ERK2 MAPK cascade	apoptosis	nucleoside binding
	Arachidonic acid metabolism	IL-2 Signaling Pathway	response to stimulus	cofactor binding
	MAPK signaling pathway	Oxidative Stress		coenzyme binding
	Cell cycle	EBV LMP1 signaling		
	Toll-like receptor signaling pathway	IL-3 Signaling Pathway		
filter all	Glutathione metabolism	Toll-like receptor signaling pathway	response to other organism	peroxidase activity
	MAPK signaling pathway	IL-6 Signaling Pathway	regulation of biological quality	growth factor binding
	Phenylalanine metabolism	Keap1-Nrf2	oxidation reduction	iron ion binding
	Arachidonic acid metabolism	MAPK signaling pathway	Multi-organism process	MAP kinase tyrosine/serine/threonine phosphatase activity
	Glycine, serine and threonine metabolism	ERK1 - ERK2 MAPK cascade	cellular amino acid and derivative metabolism	tetrapyrrole binding
	Tyrosine metabolism	neural crest development	response to stimulus	catalytic activity
	Toll-like receptor signaling pathway	Oxidative Stress		antioxidant activity
	Drug metabolism - cytochrome P450	Myometrial Relaxation and Contraction Pathways		glutathione transferase activity
	p53 signaling pathway	Nodal signaling pathway		
	Metabolism of xenobiotics by cytochrome P450	Signaling of Hepatocyte Growth Factor Receptor		

filter down		neural crest development		
		Nodal signaling pathway		
filter up	Glutathione metabolism	Toll-like receptor signaling pathway	regulation of biological quality	mAP kinase phosphatase activity
	MAPK signaling pathway	MAPK signaling pathway	oxidation reduction	insuline-like growth factor binding
	Arachidonic acid metabolism	IL-6 Signaling Pathway	response to methylmercury	catalytic activity
	Glycine, serine and threonine metabolism	Keap1-Nrf2	response to stimulus	growth factor binding
	Toll-like receptor signaling pathway	ERK1 - ERK2 MAPK cascade	Multi-organism process	antioxidant activity
	p53 signaling pathway	Oxidative Stress	biological regulation	cofactor binding
	Metabolism of xenobiotics by cytochrome P450	Androgen Receptor Signaling Pathway		transcription factor activity
	Phenylalanine metabolism	Apoptosis		iron ion binding
	Drug metabolism - cytochrome P450	Signaling of Hepatocyte Growth Factor Receptor		transcription regulator activity
	Jak-STAT signaling pathway	TGF Beta Signaling Pathway		

Appendix B

GO Analysis Figures

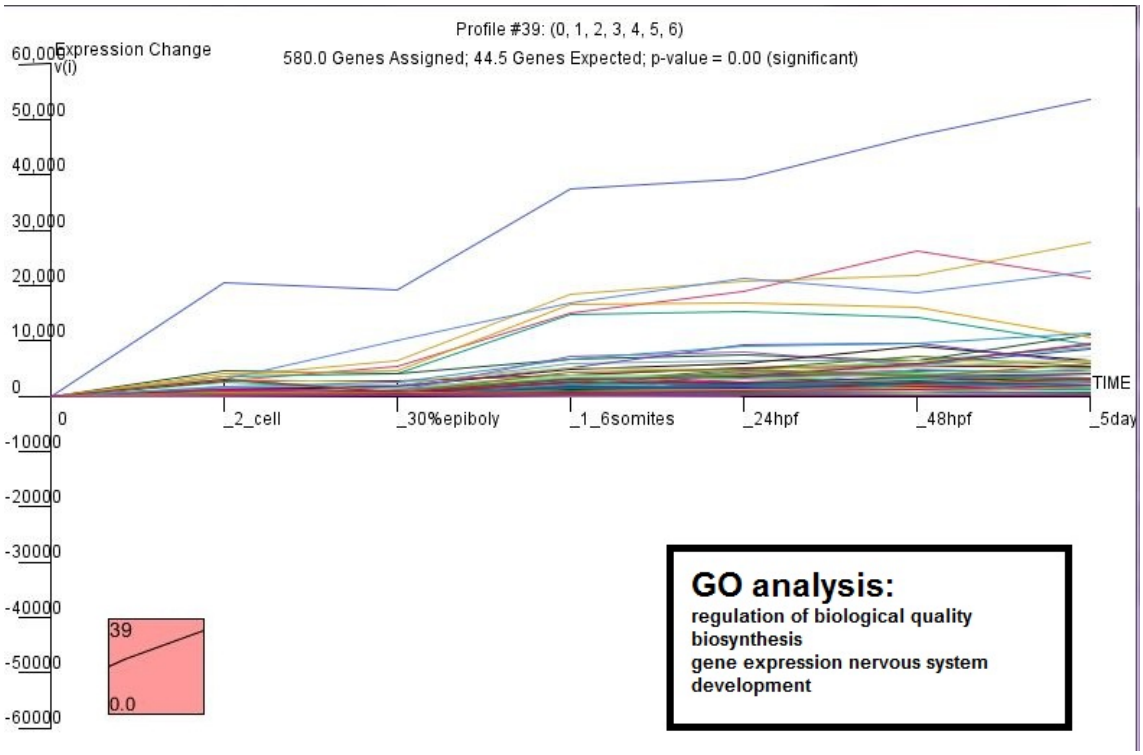


Figure B.1: Expression signals and GO analysis of profile 39.

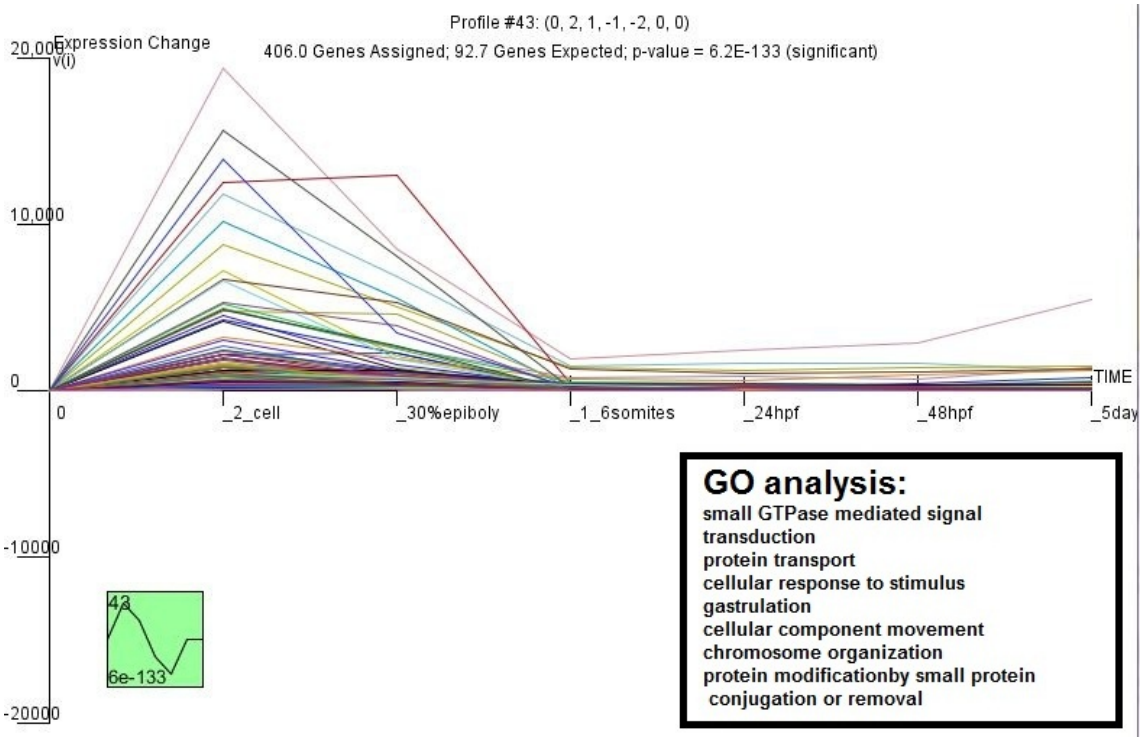


Figure B.2: Expression signals and GO analysis of profile 43.

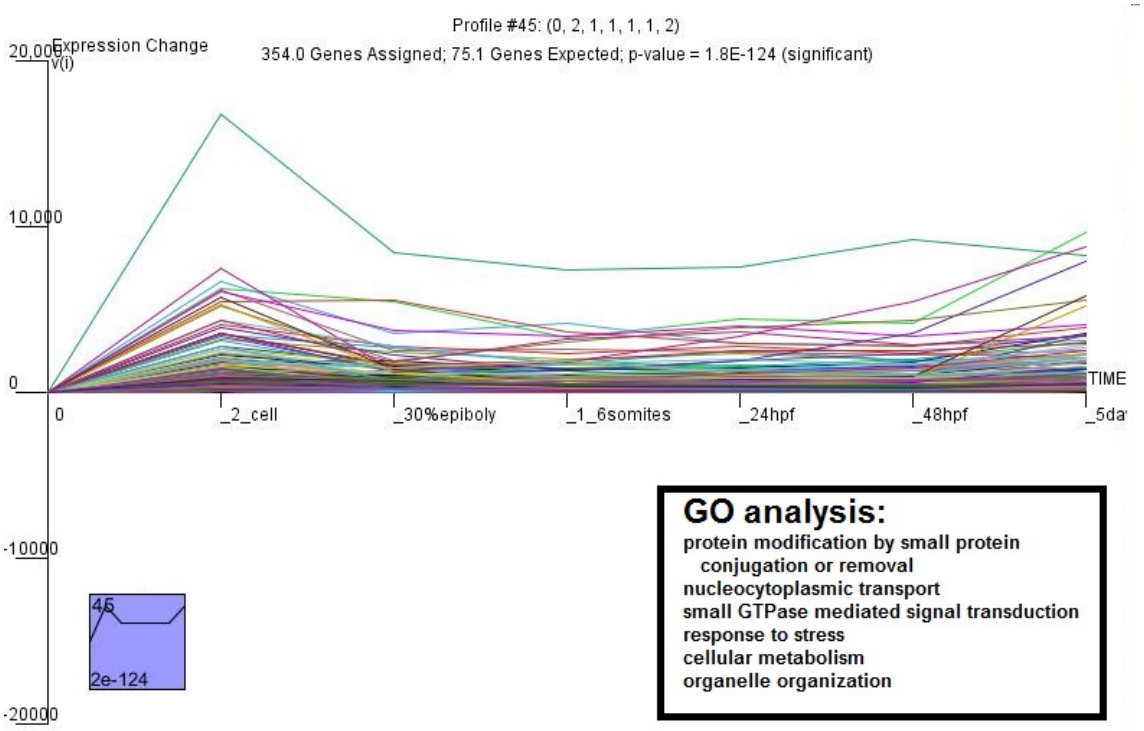


Figure B.3: Expression signals and GO analysis of profile 45.

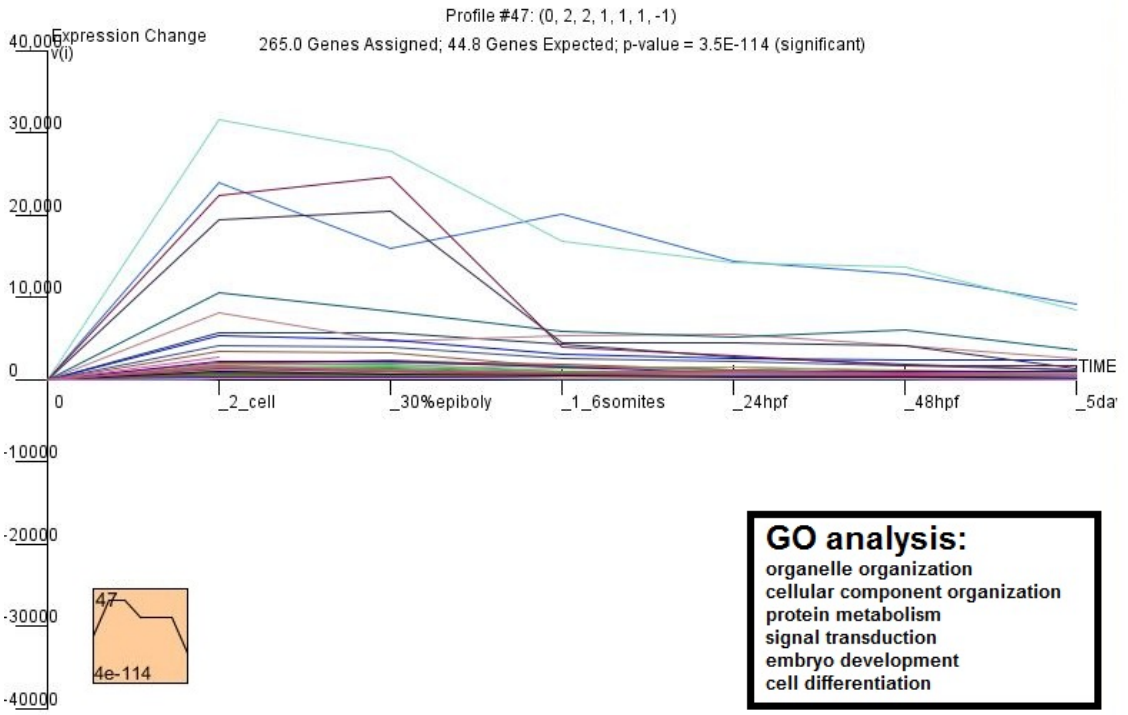


Figure B.4: Expression signals and GO analysis of profile 47.

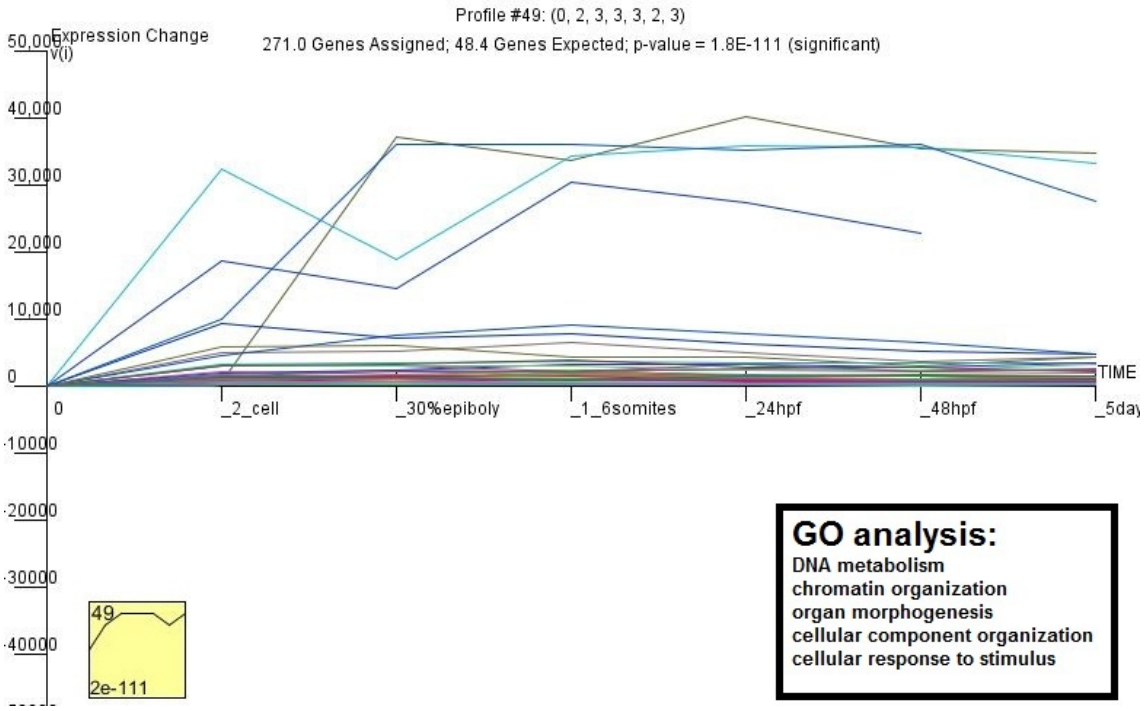


Figure B.5: Expression signals and GO analysis of profile 49.

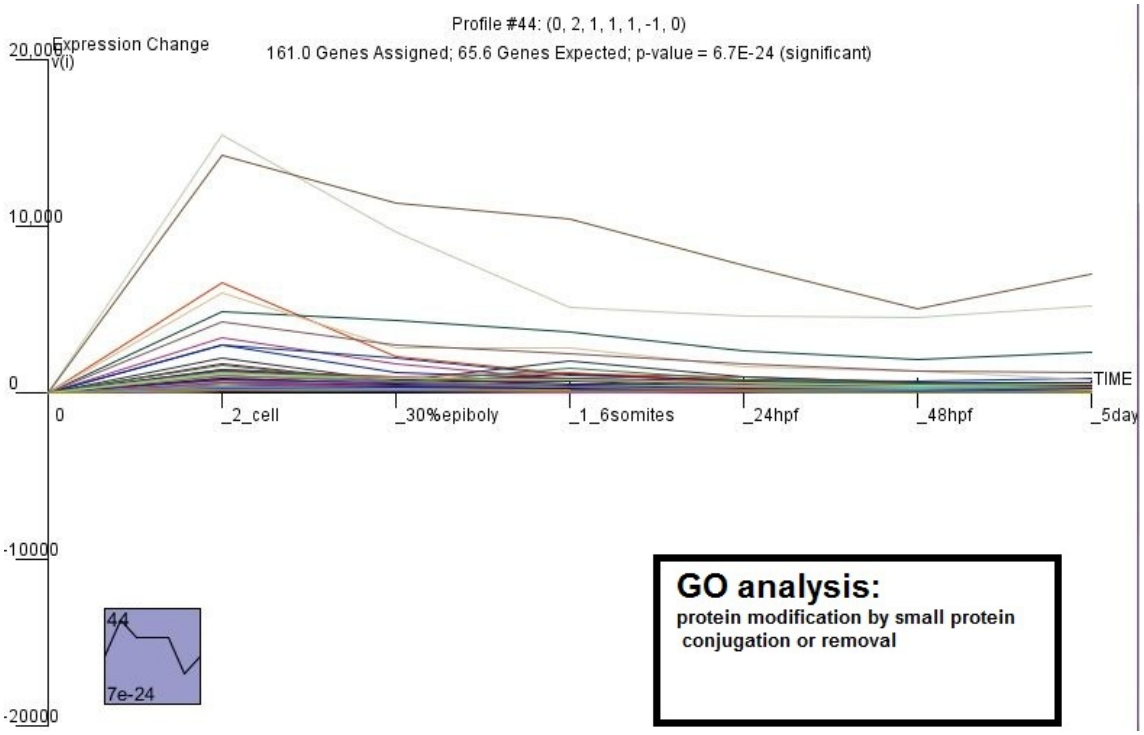


Figure B.6: Expression signals and GO analysis of profile 44.

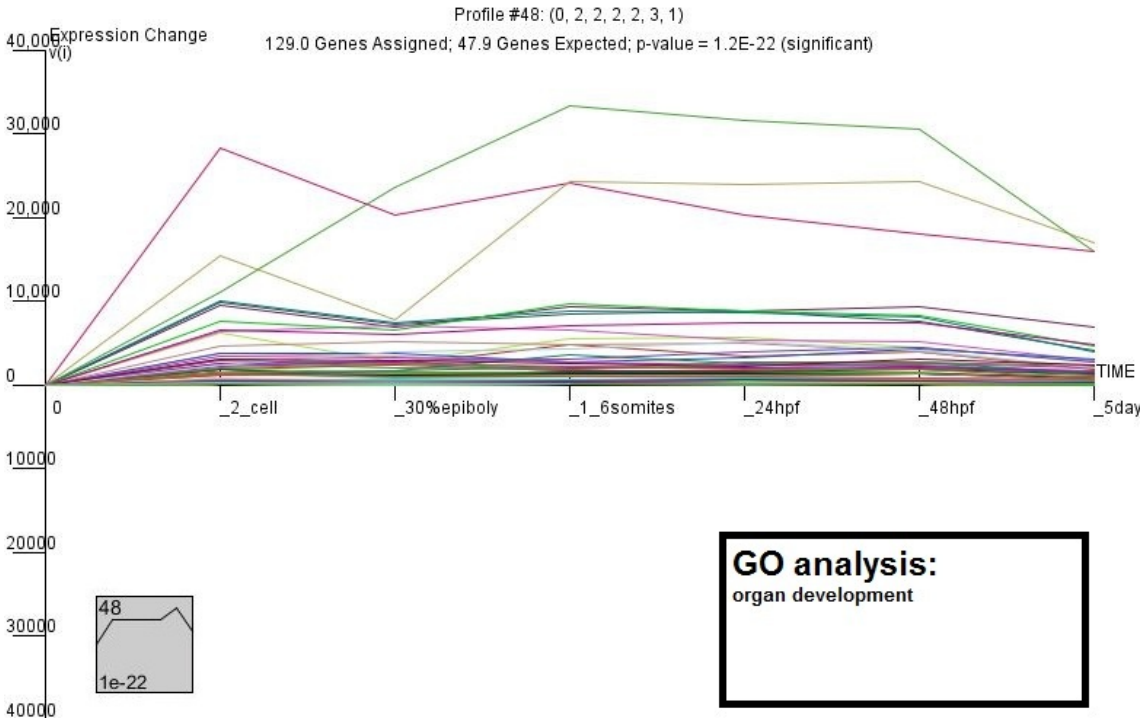


Figure B.7: Expression signals and GO analysis of profile 48.

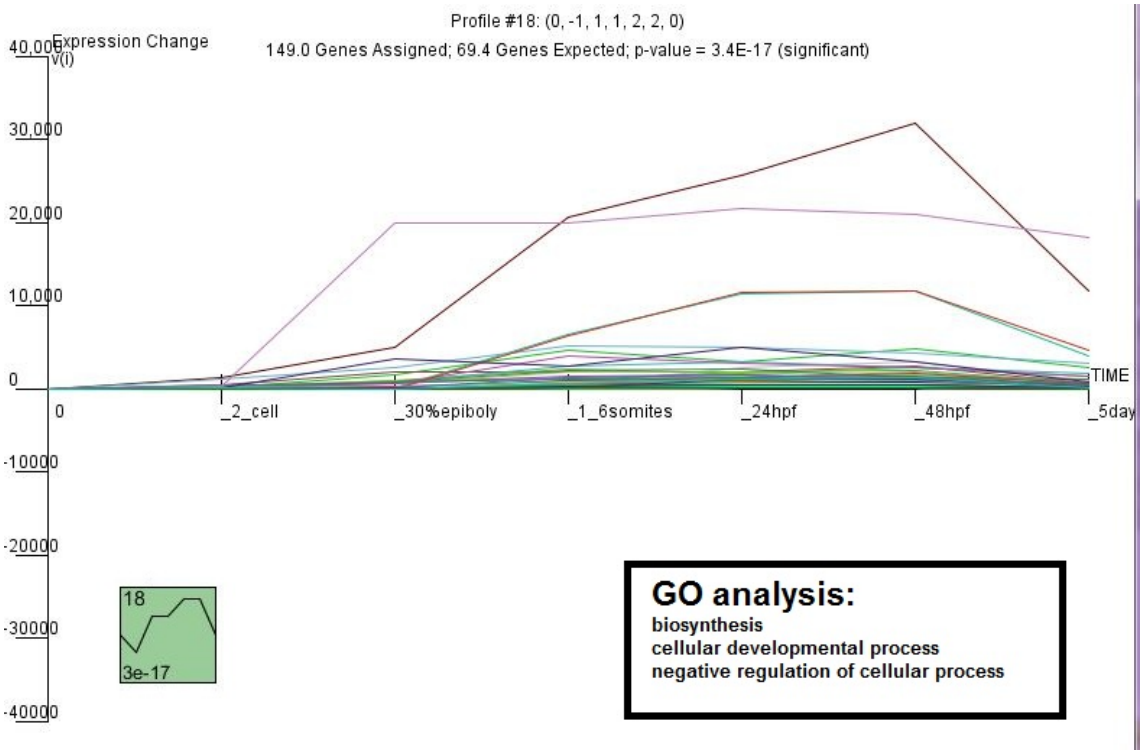


Figure B.8: Expression signals and GO analysis of profile 18.

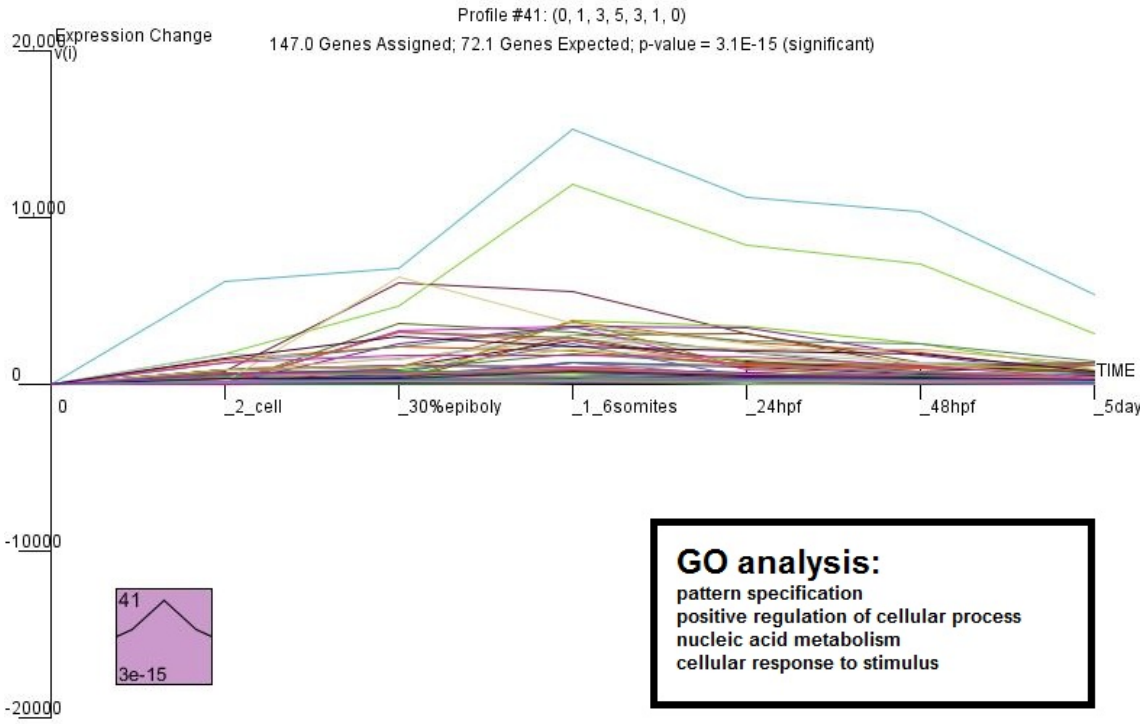


Figure B.9: Expression signals and GO analysis of profile 41.

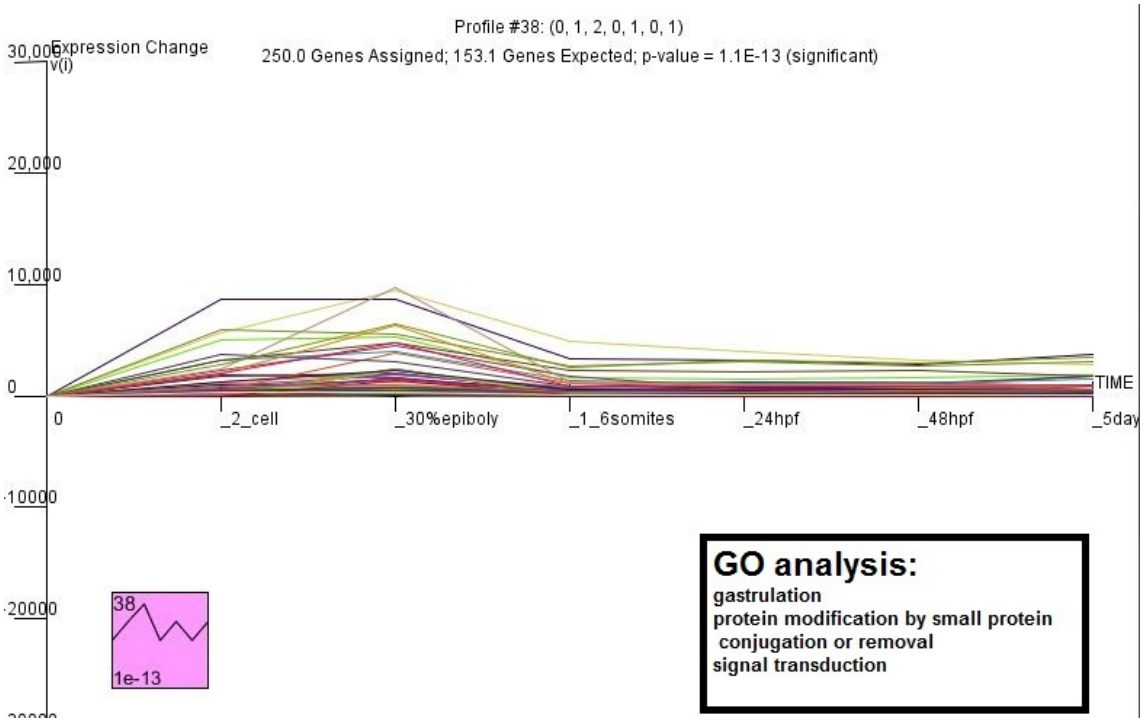


Figure B.10: Expression signals and GO analysis of profile 38.

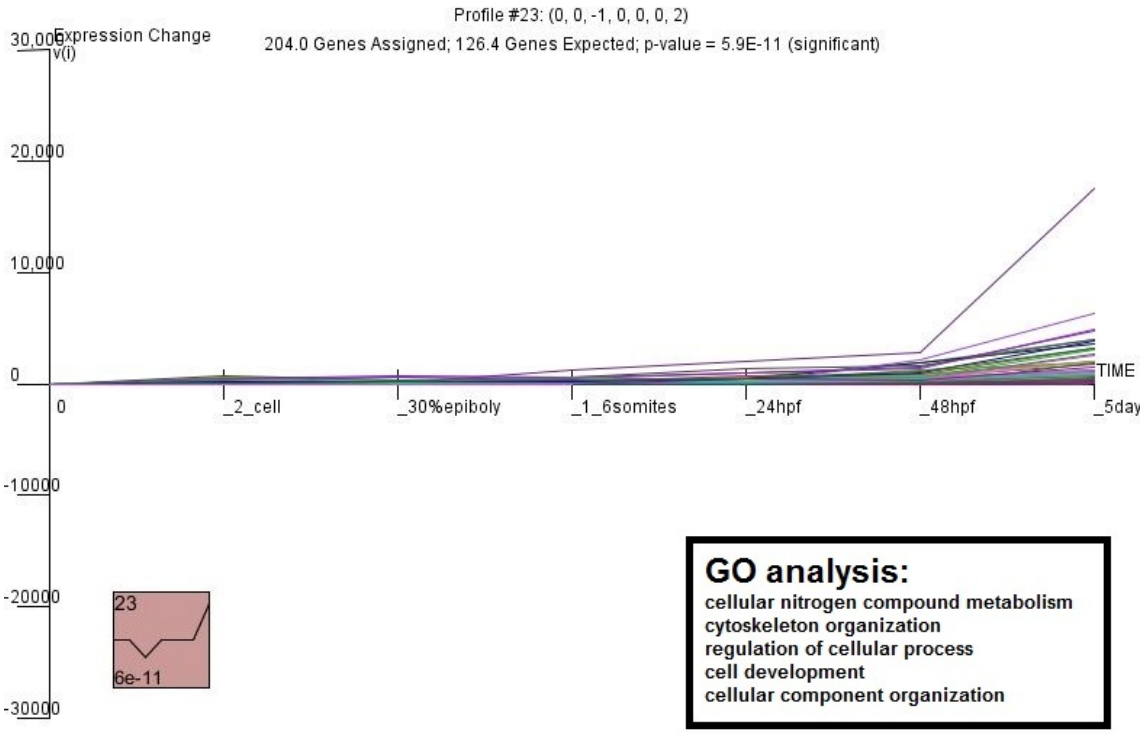


Figure B.11: Expression signals and GO analysis of profile 23.

Appendix C

Tissue Specific Transcription Factors

Ensembl Gene ID	Ensembl Transcript ID	Description	Associated Gene Name
ENSDARG00000019013	ENSDART00000004548	BarH-like 1.1 [Source:ZFIN;Acc:ZDB-GENE-060118-2]	barhl1.1
ENSDARG00000079182	ENSDART00000007777	neuronal PAS domain protein 3 [Source:HGNC Symbol;Acc:19311]	NPAS3
ENSDARG00000079012	ENSDART00000021009	core-binding factor, runt domain, alpha subunit 2; translocated to, 3 [Source:HGNC Symbol;Acc:1537]	CBFA2T3
ENSDARG00000021979	ENSDART00000028938	microtubule associated monoxygenase, calponin and LIM domain containing 3a [Source:ZFIN;Acc:ZDB-GENE-050126-2]	mical3a
ENSDARG00000013615	ENSDART00000040126	pre-B-cell leukemia transcription factor 3b [Source:ZFIN;Acc:ZDB-GENE-000405-3]	pbx3b
ENSDARG00000013539	ENSDART00000042377	IKAROS family zinc finger 1 (Ikaros) [Source:ZFIN;Acc:ZDB-GENE-980526-304]	ikzf1
ENSDARG00000013539	ENSDART00000046079	IKAROS family zinc finger 1 (Ikaros) [Source:ZFIN;Acc:ZDB-GENE-980526-304]	ikzf1
ENSDARG00000013539	ENSDART00000050481	IKAROS family zinc finger 1 (Ikaros) [Source:ZFIN;Acc:ZDB-GENE-980526-304]	ikzf1
ENSDARG00000036542	ENSDART00000053097	pbx/knotted 1 homeobox 1.2 [Source:ZFIN;Acc:ZDB-GENE-020123-1]	pknox1.2
ENSDARG00000055158	ENSDART00000061487	prospero-related homeobox gene 1 [Source:ZFIN;Acc:ZDB-GENE-980526-397]	prox1
ENSDARG00000044775	ENSDART00000065818	fucosyltransferase 7 (alpha (1,3) fucosyltransferase) [Source:ZFIN;Acc:ZDB-GENE-060929-997]	fut7
ENSDARG00000069512	ENSDART00000076285	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E7F4L9]	CABZ01034857.1
ENSDARG00000057395	ENSDART00000080033	single-stranded DNA binding protein 2 [Source:HGNC Symbol;Acc:15831]	SSBP2 (1 of 2)
ENSDARG00000060149	ENSDART00000084416	actin binding LIM protein 1a [Source:ZFIN;Acc:ZDB-GENE-080219-41]	ablim1a
ENSDARG00000061011	ENSDART00000086536	hepatic leukemia factor [Source:HGNC Symbol;Acc:4977]	HLF
ENSDARG00000067797	ENSDART00000097650	spleen focus forming virus (SFFV) proviral integration oncogene spi1 like [Source:ZFIN;Acc:ZDB-GENE-060825-351]	spi1l
ENSDARG00000068417	ENSDART00000098940	forkhead box L2 [Source:HGNC Symbol;Acc:1092]	FOXL2 (1 of 2)
ENSDARG00000001910	ENSDART00000099477	RAR-related orphan receptor A, paralog b [Source:ZFIN;Acc:ZDB-GENE-040426-855]	rorab
ENSDARG00000069335	ENSDART00000100808	B-cell CLL/lymphoma 6, member B [Source:HGNC Symbol;Acc:1002]	BCL6B
ENSDARG00000021979	ENSDART00000101035	microtubule associated monoxygenase, calponin and LIM domain containing 3a [Source:ZFIN;Acc:ZDB-GENE-050126-2]	mical3a
ENSDARG00000070929	ENSDART00000104532	SRY-box containing gene 14 [Source:ZFIN;Acc:ZDB-GENE-051113-268]	sox14

Table C.1: Transcription factors which were specifically expressed in the diencephalon sample.

Ensembl Gene ID	Ensembl Transcript ID	Description	Associated Gene Name
ENSDARG00000029766	ENSDART00000017326	nuclear receptor subfamily 1, group I, member 2 [Source:ZFIN;Acc:ZDB-GENE-030903-3]	nr1i2
ENSDARG00000087057	ENSDART00000041399	finTRIM family, member 34 [Source:ZFIN;Acc:ZDB-GENE-070912-110]	ftr34
ENSDARG00000032197	ENSDART00000045691	Kruppel-like factor 12b [Source:ZFIN;Acc:ZDB-GENE-071004-22]	klf12b
ENSDARG00000052094	ENSDART00000050856	notch homolog 1b [Source:ZFIN;Acc:ZDB-GENE-990415-183]	notch1b
ENSDARG00000043210	ENSDART00000088668	nuclear factor I/C [Source:ZFIN;Acc:ZDB-GENE-080305-2]	nfic
ENSDARG00000063031	ENSDART00000091716	RAD54-like 2 (S. cerevisiae) [Source:HGNC Symbol;Acc:29123]	RAD54L2
ENSDARG00000067850	ENSDART00000097755	jun D proto-oncogene [Source:ZFIN;Acc:ZDB-GENE-070725-2]	jund
ENSDARG00000068019	ENSDART00000098117	SRY (sex determining region Y)-box 9 [Source:HGNC Symbol;Acc:11204]	SOX9 (4 of 4)
ENSDARG00000031015	ENSDART00000105680	dystrobrevin, alpha [Source:ZFIN;Acc:ZDB-GENE-070117-2]	dtna

Table C.2: Transcription factors which were specifically expressed in the telencephalon sample.

Ensembl Gene ID	Ensembl Transcript ID	Description	Associated Gene Name
ENSDARG00000015506	ENSDART00000004361	Kruppel-like factor 5a [Source:ZFIN;Acc:ZDB-GENE-090312-167]	klf5a
ENSDARG00000037020	ENSDART00000005447	kinesin family member 1B [Source:ZFIN;Acc:ZDB-GENE-030820-1]	kif1b
ENSDARG00000029766	ENSDART00000017326	nuclear receptor subfamily 1, group I, member 2 [Source:ZFIN;Acc:ZDB-GENE-030903-3]	nr1i2
ENSDARG00000009094	ENSDART00000022731	GATA-binding protein 2b [Source:ZFIN;Acc:ZDB-GENE-040718-440]	gata2b
ENSDARG00000021200	ENSDART00000032502	zgc:113144 [Source:ZFIN;Acc:ZDB-GENE-050320-151]	zgc:113144
ENSDARG00000022251	ENSDART00000034068	zinc finger protein 536 [Source:ZFIN;Acc:ZDB-GENE-030616-624]	znf536
ENSDARG00000013539	ENSDART00000042377	IKAROS family zinc finger 1 (Ikaros) [Source:ZFIN;Acc:ZDB-GENE-980526-304]	ikzf1
ENSDARG00000013539	ENSDART00000046079	IKAROS family zinc finger 1 (Ikaros) [Source:ZFIN;Acc:ZDB-GENE-980526-304]	ikzf1
ENSDARG00000037421	ENSDART00000054460	early growth response 1 [Source:ZFIN;Acc:ZDB-GENE-980526-320]	egr1
ENSDARG00000074118	ENSDART00000061170	finTRIM family, member 15 [Source:ZFIN;Acc:ZDB-GENE-070912-394]	ftr15
ENSDARG00000052037	ENSDART00000073777	tripartite motif containing 35 [Source:HGNC Symbol;Acc:16285]	TRIM35 (41 of 43)
ENSDARG00000069512	ENSDART00000076285	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E7F4L9]	CABZ01034857.1
ENSDARG00000074884	ENSDART00000076291	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E7F4L8]	CABZ01034858.1
ENSDARG00000055359	ENSDART00000077702	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F1R6N2]	BX569789.2
ENSDARG00000055750	ENSDART00000078149	forkhead box P3a [Source:ZFIN;Acc:ZDB-GENE-061116-2]	foxp3a
ENSDARG00000058207	ENSDART00000081015	abl-interactor 2 [Source:HGNC Symbol;Acc:24011]	ABI2 (2 of 2)
ENSDARG00000059158	ENSDART00000082192	four and a half LIM domains 3 [Source:HGNC Symbol;Acc:3704]	FHL3 (2 of 2)
ENSDARG00000062420	ENSDART00000090242	nuclear factor I/A [Source:ZFIN;Acc:ZDB-GENE-050208-501]	nfia
ENSDARG00000067797	ENSDART00000097650	spleen focus forming virus (SFFV) proviral integration oncogene spi1 like [Source:ZFIN;Acc:ZDB-GENE-060825-351]	spi1l
ENSDARG00000052971	ENSDART00000098057	finTRIM family, member 19 [Source:ZFIN;Acc:ZDB-GENE-090506-5]	ftr19
ENSDARG00000003820	ENSDART00000099040	nuclear receptor subfamily 1, group D, member 2a [Source:ZFIN;Acc:ZDB-GENE-040504-1]	nr1d2a
ENSDARG00000069193	ENSDART00000100501	ets variant 7 [Source:HGNC Symbol;Acc:18160]	ETV7 (2 of 2)
ENSDARG00000069988	ENSDART00000102318	zgc:158706 [Source:ZFIN;Acc:ZDB-GENE-070112-1882]	zgc:158706
ENSDARG00000070852	ENSDART00000104372	zgc:195077 [Source:ZFIN;Acc:ZDB-GENE-080724-9]	zgc:195077
ENSDARG00000034429	ENSDART00000104593	tripartite motif containing 35 [Source:HGNC Symbol;Acc:16285]	TRIM35 (1 of 43)

Table C.3: Transcription factors which were specifically expressed in the head sample.

Ensembl Gene ID	Ensembl Transcript ID	Description	Associated Gene Name
ENSDARG00000007186	ENSDART00000003164	protein phosphatase 1, regulatory (inhibitor) subunit 8 [Source:ZFIN;Acc:ZDB-GENE-060503-681]	ppp1r8
ENSDARG00000009161	ENSDART00000019843	finTRIM family, member 55 [Source:ZFIN;Acc:ZDB-GENE-070424-161]	ftr55
ENSDARG00000017953	ENSDART00000079644	tumor protein p73 [Source:ZFIN;Acc:ZDB-GENE-030814-2]	tp73

Table C.4: Transcription factors which were specifically expressed in the tail sample.