HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER THESIS

Mr
Stephan Koch

Analyse von Fischexpressionsdaten zur Klassifizierung von molekularen Chemikalienwirkungen

2014

Department of **Mathematics, Natural and Computer sciences**

MASTER THESIS

Analyse von Fischexpressionsdaten zur Klassifizierung von molekularen Chemikalienwirkungen

Author:
Stephan Koch

Course of studies:
Molecular Biology/Bioinformatics

Seminar group:
MO12w1-M

First examiner:
Prof. Dr. rer. nat. Dirk Labudde

Second examiner:
Dr. Wibke Busch

Mittweida, Dezember 2014

Ein hartnäckiger Begleiter
der Erkenntnis ist die
Unwissenheit über die
eigene Unwissenheit.

Stanislaw Lem

Bibliographic description

Koch, Stephan: Analyse von Fischexpressionsdaten zur Klassifizierung von molekularen Chemikalienwirkungen, 63 pages, 32 figures, Hochschule Mittweida (FH), Department of Mathematics, Natural and Computer sciences

Master Thesis, 2014

Abstract

Gene expression experiments are used to understand the effects of environmental influences on the genome of an organism.
Chemicals are able to influence the expression of the genome. With experiments we can determine and observe the mechanisms of chemical actions on the gene expression level. In the past a lot of gene expression data was created. One favored model organism for such experiments is the zebrafish *danio rerio*. This fish has some special properties in his development, which favor such experiments. The embryonic development of the zebrafish is finished after five days. During that time the embryo is transparent. So it is easy to observe phenotypic and genomic influences of chemicals.
A lot of effects caused by chemicals on the zebrafish genome are already observered. This study aimed to compare expression profiles in order to investigate whether common genes and a general response to chemical exposure can be observed. To answer this question, gene expression data from cDNA- microarrays of different experiments was collected form the databases of NCBI and EBI. This microarray data was controlled for its quality. After this the microarray data was statistically analysed in a generall method. The data created this way, was compared for each chemical treatment.
This analysis of the microarray data shows a general answer to exposure with chemicals on the genomic level. Through the comparisons between the differential expression pattern of each chemical, we were able to determine 22 genes, that code for proteins whose function can be related to the binding and decomposition of chemicals.

Zusammenfassung

Durch Genexpressionsexperimente an Modellorganismen versucht man das Genom und die Reaktion des Genoms auf äußere Einflüsse zu erforschen.
Chemikalien sind in der Lage die Expression des Genoms zu beeinflussen. Durch die Experimente ist es möglich Rückschlüsse auf die Wirkweisen der Chemikalien auf genomischer Ebene zu ziehen. Durch derartige Chemikalienexperimente wurden in der Vergangenheit bereits große Mengen an Genexpressionsdaten erzeugt. Ein beliebter Modellorganismus für solche Versuche ist der Zebrabärbling *Danio rerio*, der in seiner embryonalen Entwicklung optimale Vorrausetzungen mit sich bringt. Bei ihm kann man phänotypische und genotypische Effekte der Chemikalien sehr gut beobachten, da er in diesem Zeitraum transparent ist und diese Phase nur fünf Tage anhält.
Anhand des Zebrabärblings soll erforscht werden, ob gleichartige Chemikalien auf genomischer Ebene ähnliche Wirkweisen besitzen und sich in Gruppen zusammenfassen lassen. Desweiteren stellt sich die Frage, ob es eine allgemeine (Gegen)Reaktion auf den Einfluss der Chemikalien gibt, welche auf genomischer Ebene sichtbar wird. Dazu werden Genexpressionsdaten aus cDNA-Microarrayexperimenten verwendet, die in den Datenbanken des NCBI und EBI abgespeichert sind. Nachdem geeigenete Experimentaldaten ausgewählt wurden und auf ihre Qualität geprüft, wurden diese einer vereinheitlichten statistischen Analyse unterzogen. Die Ergebnisse dieser statistischen Analyse wurden untereinander auf identisch differentiell exprimierte Gene verglichen.
Durch vereinheitlichte, allgemeine Analyse konnten Hinweise erbracht werden, dass es auf genomischer Ebene eine Antwort- und Gegenreaktion auf die Chemikalien gibt. Durch den Vergleich konnten 22 Gene ermittelt werden, die für das Binden und den Abbau der eingesetzten Chemikalie verantwortlich sind und experiementübergreifend auftraten.

Danksagung

Als erste möchte ich mich bei Prof. Labudde für die Betreuung meiner Masterarbeit bedanken. Desweiteren möchte ich mich bei meinen Betreuerinnen Dr. Wibke Busch und Dr. Kristin Reiche am UFZ bedanken, die für mich während meiner gesamten Zeit am UFZ da waren, mich fachlisch beraten und diese Chance auf eine Abschlussarbeit gegeben haben. Zusätzlich möchte ich mich bei Andreas Schüttler bedanken, der mir beim Programmieren mit `R` geholfen hat, Fehler beseitigte und meine Fragen beantwortete.
Weiterer Dank gilt meiner Familie, die mich in dieser Zeit unterstützte und Rückhalt bot.

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

1 Einleitung

In der Bio- und Ökotoxikolgie werden die schädigenden Einflüsse von Chemikalien und Umweltfaktoren auf Organismen untersucht. Ein wichtiges Thema ist dabei die Untersuchung von Substanzen, die durch den Menschen in die Umwelt gelangen. Sie werden auf ihre Giftigkeit und Einflüsse, die sie auf Organismen haben, untersucht. Um potentielle gefährliche Chemikalien frühzeitig zu erkennen, werden diese im Laborversuch an Modellorganismen getestet.
Aufgrund seiner einfachen Handhabung, schnellen Aufzucht und Generationswechsel ist der Zebrabärbling einer der beliebtesten Modellorganismen in der Ökotoxikologie und Genetik geworden. Anfang der 1990er Jahre begann man systematisch natürliche Gendefekte im Zebrabärbling zu untersuchen und zu beschreiben. So konnte 1996 eine erste Studie (Hafter et al. 1996 [1]) veröffentlicht werden, die 600 mögliche Defekte beschrieb. Ausgehend von diesen Ergebnissen, wurde das Genom des Zebrabärblings nach und nach entschlüsselt.
Parallel zu den genetischen Untersuchungen wurde der Fisch für toxikologische Versuche, speziell im embryonalen Alter verwendet. Durch seine besonderen physischen Gegebenheiten und die schnelle Aufzucht einer neuen Generationen an Versuchstieren konnten viele Erkenntnisse über den Einfluss von Chemikalien gewonnen werden [1].
In der Genexpressionsanalyse werden die genetischen und toxikologischen Methoden zusammengeführt. Über cDNA-Microarrays werden die Auswirkungen toxisch wirksamer Substanzen auf das Genom und dessen Expression des Zebrabärblings untersucht. Hierbei ist schon eine große Menge an Daten entstanden, die jedoch nur innerhalb des Rahmens in dem die Daten experimentell enstanden sind, analysiert wurden. Ziel dieser Arbeit ist es, subakute Effekte der Chemikalien auf die Expression des Zebrabärblinggenoms zu untersuchen und zwei Fragen zu beantworten:

- Gibt es Gene, die immer auf eine subakute Chemikalienbehandlung reagieren?
- Gibt es eine spezifische (Gegen)Reaktion auf ähnliche oder sogar alle Chemikalien?

Dazu müssen diese cDNA-Microarraydaten einheitlich analysiert werden, da bei den Experimenten, aus den sie stammen, verschiedene Methoden für die Datenanalyse gewählt wurden.
Diese einheitliche Methode soll eine allgemeine Analyse ermöglichen, mit der wir (ein) Stoffwechselsystem(e) auf Genomebene nachweisen können, welche als eine allgemeine Antwort auf zellulären Stress fungieren. Durch die allgemeine Analyse wollen wir diese(s) System(e) über differentiell exprimierte Gene nachweisen und diese ihrem Stoffwechselsystem oder Signalweg zuordnen. Dies sollte durch die statistische Analyse und Vergleich von cDNA-Microarraydatensätzen aus Zebrabärblingexperimenten mit Chemikalienbehandlung möglich sein.

2 Grundlagen

In diesem Abschnitt werden Ausgangspunkte und Grundlagen der Arbeit geklärt. Zunächst wird der Modellorganismus Zebrabärbling *Danio rerio* vorgestellt und unter welchem Rahmen Experimente mit diesem durchgeführt werden. Als nächstes erfolgt die Vorstellung des allgemeinen Chemikalienexperimentdesigns und der cDNA-Microarraytechnik. Im letzten Teil dieses Abschnittes werden die bioinformatischen und statistischen Grundlagen erläutert, auf denen die Methoden zur Qualitätskontrolle und Auswertung basieren. Als letzes wird erklärt, woher die für die Analyse verwendeten Datensätze stammen und weshalb diese ausgewählt wurden.

2.1 Gewinnung der experimentellen Daten

2.1.1 Modellorganismus Zebrabärbling

Der Zebrabärbling *Danio rerio* ist ein aus Nordostindien stammender Zierfisch. Der Zebrabärbling wird ca. 4,5 Zentimeter groß. Durch seine einfache Aufzucht und hohe Reproduktionsrate hat er sich schnell in Forschung und Wirtschaft als Modellorganismus etabliert.
Durch ihre geringe Größe kann man sie in großen Mengen in Laboren halten. Unter optimalen Bedingungen (konstante 26 °C und Licht) können Zebrabärblingweibchen einmal pro Woche ca. 300 Eier ablegen, welche vom Männchen befruchtet werden. Durch das Umsetzen der befruchteten Eier in 28,5 °C warmes Wasser entwickeln sich die Embryonen innerhalb von fünf Tagen zur Larve [1].

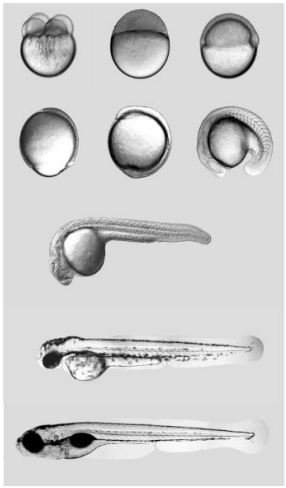


Abbildung 2.1: Entwicklungsphasen des Zebrabärblings [1]

Ein weiterer Vorteil neben dem kurzen Embryonalstadium ist die Beschaffenheit der Embryonen, denn diese sind komplett durchsichtig (Abb. 2.1). Somit ist es möglich die embryonale Entwicklung exakt zu beobachten. Bach 24 Stunden sind schon fast alle Organe gebildet und der Embryo beginnt sich zu bewegen.
Durch diese schnelle, leicht zu verfolgende Entwicklung, hohe Reproduktionsraten und einfache Haltung kann eine hohe Menge an Experimentaldaten erzeugt werden. Es erleichtert das Arbeiten mit künstlich erzeugten Mutanten des Zebrabärbling, da man in der großen Menge der Nachkommen, häufiger Mutanten mit gewünschten Veränderungen vorfindet. Dadurch kann man die Auswirkung der gezielten Veränderung einzelner Gene genauer erforschen. Die Durchsichtigkeit der Embryonen ist auch für Chemikalienversuche von Vorteil, da man nun genau sehen kann, welche Organe beeinflusst und wie diese von Chemikalien verändert werden (Abb.2.2) [1].

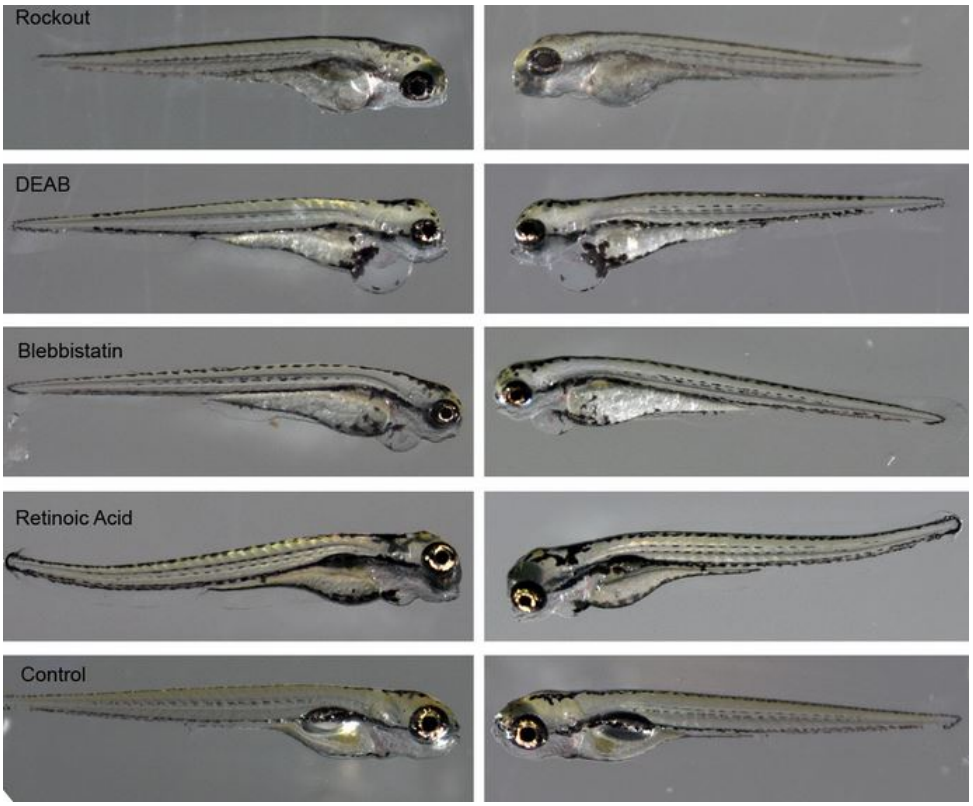


Abbildung 2.2: Schädigender Einfluss von terratogenen Chemikalien während der embryonalen Entwicklung des Zebrabärblings [taken from: http://people.hsc.edu/faculty-staff/edevlin/ edsweb01/ E-2-5dpf-whole-embryo-compo.jpg]

1996 wurde eine Studie (Hafter et al. 1996 [1]) veröffentlicht, welche 600 genetische Defekte im Zebrafisch beschreibt. Mittlerweile ist der Zebrabärbling vollständig sequenziert und annotiert sowie seine biologische Entwicklung im embryonalen Stadium gut untersucht und beschrieben [6].

2.1.2 Chemikalienexperimente mit verschiedenen Substanzen am Zebrabärbling *Danio rerio*

Chemikalienexperimente sind der Lieferant für die Daten der Arbeit. Bei ihnen wird der toxikologische Einfluss von Chemikalien bestimmt. Hierbei wird ein Modellorganismus und dessen Embryo einer Chemikalie für einen bestimmten Zeitraum exponiert. Als erstes wird, falls nicht bekannt, der LC50 Wert bestimmt. Dieser gibt die Konzentration an, bei der die Hälfte aller Versuchstiere gestorben ist. Aus der dabei entstehenden sig-

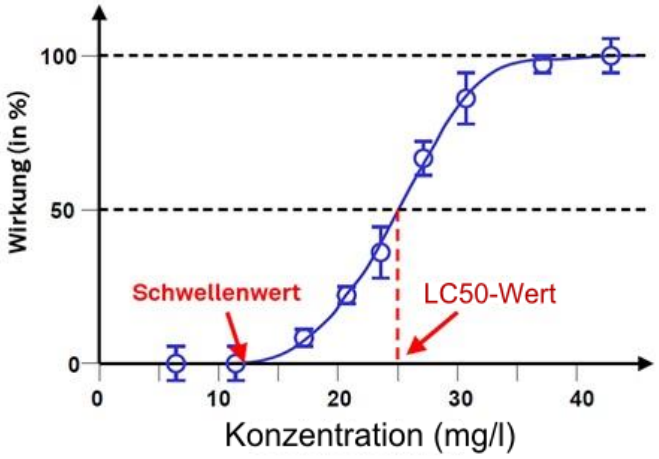


Abbildung 2.3: Allgemeine Dosis-Wirkungskurve [taken from: http://www.novo-argumente.com/images/ uploads// pic/140428_kraemer_abb1.jpg]

moidale Kurve (Abb.2.3) werden Chemikalienkonzentrationen und Expositionszeiträume bestimmt, bei denen alle Organismen überleben. Dies soll sicher stellen, dass man alle Effekte unverfälscht untersuchen kann [].
Die Versuchstiere für den eigentlichen Versuch werden in Kontrolle und Behandlung unterteilt. Sie sollten für den Versuch aus einem Ansatz ('batch') sein, d.h. die Organismen stammen aus der gleichen Generation. Im Falle des Zebrabärblings wäre dies eine Eiablage. Die Behandlungsgruppe wird der Chemikalie in einer Kontrolllösung ausgesetzt; die Kontrollgruppe wird nur der Kontrolllösung ausgesetzt. Aus beiden Gruppen wird danach mRNA für die Genexpressionsanalyse extrahiert.

2.1.3 Genexpressionsanalyse und cDNA-Microarrays

Bei der Genexpressionsanalyse wird die Aktivität eines Gens, die Häufigkeit wie oft ein Genprodukt hergestellt wird, bestimmt. Dabei werden in Experimenten äußere Einflüsse (Chemikalen, Medikamente usw.) betrachtet, die auf einen Organismus wirken. So gibt es eine Kontrollgruppe und eine Behandlungsgruppe zwischen denen der Unterschied in der Genexpression bestimmt wird. Die bisherige Methode ist die Untersuchung mit cDNA-Microarrays, alternativ ist auch Sequenzierung möglich.

cDNA-Mircoarrays sind eine Möglichkeit die Expression von Genen zu bestimmen. Ein Microarray besteht aus einem Trägermaterial (Glas oder Silizium) mit Sonden und ist auf die Gene eines Organismus ausgelegt. Die Sonden bestehen aus Einzelstrangoligonucleotiden, welche eine Teilsequenz ausgewählter Gene des zu untersuchenden Orgnismus sind. Die Sonden befinden sich in einem rechteckigen Feld auf dem Trägermaterial angeordnet [7]. Neben den Genen des Organismus gibt es, abhängig vom Design des Microarrays, noch Kontrollsonden (controlspots) auf dem Array, die jeweils die dunkle und helle Referenz für ein Fluoreszenzspektrum angeben.

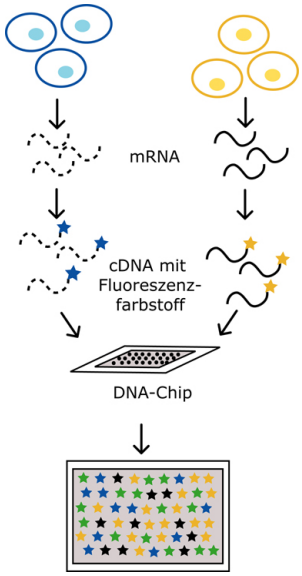


Abbildung 2.4: Schritte für die Genexpressionsanalyse mittels Microarray [taken from: http:// www. biosicherheit.de/ data/ media/ 644/ 382x738.png]

Für den Microarrayversuch muss zuvor mRNA aus dem zu untersuchen Organsimus isoliert und aufbereitet werden (Abb. 2.4). Durch PCR und dem Enzym Reverse Transkriptase wird aus der mRNA cDNA (complementary DNA) synthetisiert [8]. Die cDNA wird mit Fluoreszenzfarbstoff (Cy3, Cy5) markiert. Für die Analyse wird die gesammelte cDNA auf den Mircroarray gegeben. Die cDNA hybridisiert zu ihren komplementären Oligonucleotidgegenstücken auf dem Microarray. Nach der Hybridisierung wird der Array in einer Fluoreszenzsmessung abgescannt und die Leuchtintensität der Sonden gemessen. Durch die Positionsdaten und Intensitätsmesswerte können jetzt Unterschiede zwischen den Genexpressionen bestimmt werden [8].
Die Scandaten müssen als nächstes statistisch ausgewertet werden. Diese Auswertung erfolgt dabei nicht nach festen Regeln, jedoch verwendet sie ein allgemeines Schema zur Analyse von statistischen Daten. So verfolgen alle Fachartikel, deren Daten für diese Arbeit ausgewählt wurden, andere Herangehensweisen, sei es die Auswahl des Programms zu Analyse (R, Matlab, Gerätinterne Programme) oder der statistischen Methode (ANOVA, MAANOVA, Clustering, t-Test, f-Test, SOM usw.)[4][9][10][11][2][5]. Dieser Teil der cDNA-Microarrayanalyse wird in dieser Arbeit einheitlich durchgeführt.

2.2 Bioinformatische und statistische Grundlagen

2.2.1 Geneexpressiondatenbanken

Die Daten aus Genexpressionsanalysen werden in Geneexpressiondatenbanken abgelegt. Dabei handelt es sich um Informationen der gescannten Microarrays, die nach Experiment geordnet werden. Für diese Daten haben die Hersteller von Mircroarrays jeweils ihre eigenen Dateiformate ('.cel' für Affymetrix, '.txt' für Agilent).
Es gibt zwei große Datenbanken für Genexpressionsdaten:

- Gene Expression Omnibus - NCBI
- ArrayExpress - EBI

Die Arraydaten bei beiden Datenbanken müssen dabei dem MIAME (Minimum Information About A Microarray Experiment) Standard [12] entsprechen, welcher ein Minimum an Experimentinformation fordert, um so eine angemessene Qualität der Einträge zu wahren. Die Richtlinien für diese Informationstandards sind:

- die Rohdaten für jede Hybridisierungsreaktion (Microarray)
- die prozessierten, normalisierten Daten für jeden Microarray
- Beschreibung der experimentbeeinflussenden Faktoren (Konzentration, Expositionszeiträume, Chemikalien)
- Aufbau der Experimente, Unterteilung in Replikate und Behandlungen
- Annotation des/der für das Experiment(e) verwendeten Microarray(s) (diverse Genidentifikatoren (Ensemble-, Entrez-, external gene -ID), Sequenzen der Oligonucleotide)
- Informationen über den Ablauf im Labor und die Prozessierung der Daten

2.2.2 Lineare Modelle für die cDNA-Microarrayanalyse

Der Kern der cDNA-Micorarrayexperimente ist die statistische Auswertung dieser. Sie sind der Hauptbestandteil dieser Arbeit und der wichtigste Ansatzpunkt, die Auswertung zu vereinheitlichen, damit eine allgemeine Analyse möglich wird.
Zur Identifikation von differentiell expremierten Genen in cDNA-Micorarraydaten werden lineare Modelle verwendet. Bei einem linearen Modell geht man davon aus, dass ein Zusammenhang zwischen den Daten und einer bekannten Einflussvariable/Koeffizienten vorhanden ist. Im diesem Falle wäre dies die Auswirkung der Chemikalien auf die Genexpression des Zebrabärblingembryos.
Für die Arbeit wird das Statistikprogramm `R` verwendet, bei dem mit `limma` ein Programmpaket für die Untersuchung von Genexpressionsdaten auf Basis linearer Modelle vorhanden ist. Die folgenden Ausführung beruhen auf der Funktionsweise dieses Paketes.

Die für das lineare Modell verwendeten Daten müssen die logarithmierten Intensitäten sein. Die Skalierung der Daten durch den Logarithmus erfolgt zur Basis zwei. Die Daten werden logarithmiert, damit sie später besser vergleichbar sind. So ist es dann möglich direkt anzugeben, wie viel höher die Expression im Vergleich zur Kontrolle ist (logFold-Change) [21].
Diese Daten sollten ebenfalls bereits zueinander normalisiert sein. Bei der Normalisierung wird für die Signalintensität einer Sonde ein Mittelwertssignal über alle Microarrays, die zum gleichen Replikat gehören, gebildet. Durch dieses Vorgehen wird das Signal einer Sonde auf allen Microarrays auf einen Wert gebracht. Dies ist notwendig, um die natürlich streuenden Intensitäten der Sonden besser vergleichbar zu machen [13].
Die für die Genexpressionsanalyse verwendete Methode für die Normalisierung ist die quantilweise Normalisierung nach B. Bolstad [13]. Diese erfolgt in mehreren Schritten und geht von der Annahme aus, dass es eine gemeinsame zugrundeliegende Verteilung der Intensitäten für alle Microarrays gibt [13]. Für die Normalisierung benötigt man einen Einheitsvektor $\vec{d}$:

$$\vec{d} = (\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}) \tag{2.1}$$

Dieser Vektor ermöglicht es entlang einer Diagonalen die Intensitäten von n Microarrays von der Größe p quantilweise zu ordnen [13]. Die Quantile der Intensitäten eines Microarray werden durch den Vektor

$$\vec{q}_k = (q_{k1}, ..., q_{kn}) \;\; k = 1, ..., p \tag{2.2}$$

bestimmt. Damit die einzelnen Quantile jetzt entlang von $\vec{d}$ geordnet werden, wird $\vec{q}$ auf $\vec{d}$ projeziert:

$$P_{\vec{d}}\vec{q} = (\frac{1}{n}\sum_{j=1}^{n} q_{kj}, ..., \frac{1}{n}\sum_{j=1}^{n} q_{kj}) \tag{2.3}$$

Die Normalisierung erfolgt dann in vier Schritten:

- Es gibt n Microarrays mit der Größe p. Sie bilden eine Matrix X der Größe $p \times n$ in der jede Spalte einem Microarray entspricht.
- Die Spalten der Matrix X werden mittels $P_{\vec{d}}\vec{q}$ sortiert um X_{sort} zu erhalten.
- Für jede Zeile wird der Mittelwert bestimmt und ersetzt nun alle Elemente dieser Zeile. Die erhaltene Matrix wird als X'_{sort} bezeichnet.
- Im letzten Schritt werden die Spalten von X'_{sort} so angeordnet, dass sie wieder der ursprünglichen Reihenfolge der Matrix X entspricht.

Dies wird gemacht um nichtbiologische Unterschiede zu minimieren, die durch Microarrays entstehen [14].
Nachdem die Normalisierung abegeschlossen ist, wird für das weitere Vorgehen zwei Matrizen benötigt. Die eine ist die 'Designmatrix', welche Informationen über den zu untersuchenden Microarray enthält. Die zweite Matrix ist die 'Kontrastmatrix'. Sie trägt die Informationen über die Unterteilung in Kontrollarrays und Behandlungsarrays, damit diese in den zu untersuchen Kontrast zueinander kombiniert werden können [15].
Der erste Schritt der Analyse passt ein lineares Modell auf die Daten an. Jede Zeile der 'Designmatrix' entspricht einem Array des Experiements und jede Spalte einer Einflussvariable. Im zweiten Schritt wird durch die 'Kontrastmatrix' einer der zu untersuchenden Kontraste festgelegt. Doch was bedeutet dies genauer:
Es wird eine lineares Modell angenommen:

$$E[y_j] = X\alpha_j \tag{2.4}$$

$E[y_j]$ ist der Erwartungswert für y_j. y_j hält die Expressionsdaten für das Gen j, womit der Erwartungswert berechnet wird. X ist die 'Designmatrix', in der die Zugehörigkeit (Behandlung, Kontrolle) der Microarrays mathematisch beschrieben ist. α_j ist ein Vektor von geschätzten Koeffizienten für das Gen j. Der zu untersuchende Kontrast mit der Matrix C wird durch:

$$\beta_j = C^T \alpha_j \tag{2.5}$$

angegeben. C steht dabei für die Kontrastmatrix, die die zu untersuchende Behandlung mit ihrer Kontrolle verknüpft. Durch die Anwendung dieses Kontrastes auf das Modell werden die geschätzten Koeffizienten α_j durch β_j ersetzt. Das lineare Modell hat für jedes Gen j eine Varianz σ_j^2, die empirische Varianz s_j^2 (Schätzung für σ_j^2) und den Freiheitsgrad f_j [15].
Nachdem das lineare Modell erstellt ist, wird die empirische Bayesmethode genutzt, um eine stabile Analyse möglich zu machen. Bei Experimenten mit Microarrays gibt es oftmals nur wenige Replikate. Aufgrund der geringen Anzahl an Replikaten wird Bayes genutzt, damit die statitstische Analyse stabiler ist. Die empirische Bayes- Methode nimmt dabei an, dass es einen inversen Chiquadrat prior für σ_j^2 mit einem Mittelwert von s_0^2 und Freiheitsgraden f_0 gibt. Die Posteroir Werte für die Varianzen sind durch:

$$\tilde{s}_j^2 = \frac{f_0 s_0^2 + f_j s_j^2}{f_0 + f_j} \tag{2.6}$$

gegeben. f_0 steht für die zusätzlichen Freiheitsgrade, die durch das 'Borgen' der Informationen des gesamten Arrays erlangt wird [15].
Die resultierende angepasste t- Statistik sieht wie folgt aus:

$$\tilde{t}_{jk} = \frac{\tilde{\beta}_{jk}}{u_{jk}\tilde{s}_j} \tag{2.7}$$

k steht für den k- ten Kontrast. u_{jk} gibt die unskalierte Standartabweichung für das Gen j im Kontrast k an. Das Ergebniss der empirischen Bayes- Methode enthält die t- Statistik mit den korrespondierenden angepassten p- Werten [15]. Diese müssen im nächsten Schritt ausgewertet werden.

2.2.3 Statistische Methoden zur Analyse des linearen Modells für Genexpressionsdaten

Für die weitere Analyse der linearen Modelle liefert `limma` die Funktion `topTable()`. Diese erzeugt eine Ergebnisstabelle für das lineare Modell, fasst dessen Eregebnisse zusammen, ermöglicht Hypothesentests und passt den p-Wert für multiples Testen (Formel 2.7) an.
Die Basis Methode dafür ist der angepasste t- Test, welcher auch als einheitliche Methode in dieser Arbeit zum Einsatz kommt. Der t- Test wird deshalb als angepasst bezeichnet, da der Standartfehler durch das Bayessche Modell der empirischen Bayesmethode über alle Gene eines Arrays angepasst wird [16].
Für jede Gensonde j eines Kontrastes k wird getestet, ob $\beta_{jk} = 0$ ist. In der Ergebnistabelle des `topTable` werden Ergebnisse aufsteigend nach dem p-Wert geordnet. Niedrige p- Werte geben dabei eine hohe Wahrscheinlichkeit an, dass das Gen differentiell exprimiert ist [15]. In der Analyse von cDNA-Microarraydaten wird der Cut- Off (Abschneiden des Ergebnisses) für einen FDR- Wert, und damit die signifikanten Ergebnisse, auf 0,05 oder 0,1 festgesetzt [16][17].
Für das multiple Testen wird die 'False Discory Rate' FDR nach Benjamini und Hochberg angewendet. Diese Methode hat sich für die Auswertung von biologischen Daten etabliert [17]. Für einen einfachen Hypothesentest wird eine Nullhypothese H_0 gegen eine Alternativhypothese H_1 basierend auf einer Statistik X, in diesem Fall das lineare Modell, getestet. Es wird eine Region für die Zurückweisung Γ definiert [18]. Ist $X \in \Gamma$, dann wird die Nullhypothese H_0 abgelehnt. Soll X kein Element aus Γ sein, wird H_0 angenommen [18]. Dabei kann es aber zu zwei Arten von Fehlern kommen:

- Typ I Fehler: $X \in \Gamma$, aber H_0 ist wahr
- Typ II Fehler: $X \notin \Gamma$, aber H_1 ist wahr

Es wird eine Fehlerrate α festgelegt, die auf dem Typ I Fehler beruht, um die Region Γ zu bestimmen. Es wird angenommen das alle Regionen, für das Annehmen oder Abweisen einer Hypothese, nun eine Typ I Fehlerrate gleich oder geringer als α haben. Als nächstes wird der Rückweisungsbereich mit der geringsten Typ II Fehlerate ausgewählt. Dieser Bereich wird ausgewählt um die Typ I Fehlerrate zu kontrollieren. Dadurch kann der Typ II so gering wie möglich und Typ I auf dem α-Niveau gehalten werden [18].
Für multiples Testen treten Typ I und Typ II Fehler in jedem Test auf. Man geht von m Nullhypothesen aus, von denen m_0 wahr sind. R gibt die Anzahl der zurückgewiesenen

Hypothesen an. R ist dabei eine beobachtbare zufällige Variable; fn, tp, fp und tn können nicht beobachtet werden und sind ebenfalls zufällig (Tab. 2.1).

	Nicht Signifikant	**Signifikant**	**Total**
H_o korrekt	False negative (fn)	True positive (tp)	m_0
H_1 korrekt	True negative (tn)	False positive (fp)	$m-m_0$
	$m-R$	R	m

Tabelle 2.1: Tabelle zu Übersicht der Anzahl der Fehler für den Test von m Nullhypothesen wie von Benjamini und Hochberg eingeführt [19]

Typ I Fehler können durch die zufällige Variable $\mathbf{Q}$ betrachtet werden. Sie setzt sich aus:

$$\mathbf{Q} = \frac{tp}{tp+fp} \tag{2.8}$$

einem Verhältnis der wahren signifikanten Nullhypothesen zu der Summe aller Nullhypothesen, ob wahr oder nicht, die als signifikant eingeteilt wurden zusammen [19]. Die FDR wird als der Erwartungswert von $\mathbf{Q}$ definiert:

$$FDR_e = E\{\frac{tp}{tp+fp}\} = E(\frac{tp}{R}) \tag{2.9}$$

Als nächstes muss bestimmt werden wie hoch die Wahrscheinlichkeit eine falsch deklarierten Nullhypothese ist. Es wird davon ausgegangen, dass die p- Werte für die Nullhypothese einer Statistik gleichmäßig verteilt sind [17]. Diese müssen für die Wahrscheinlichkeit der falschen Zuordnung geordnet werden:

$$p_i \leq \frac{i}{m+1-i}\alpha \;\; i = 1,2,...,k \tag{2.10}$$

Die Ordnung erfolgt mit aufsteigender Wahrscheinlichkeit für Fehler vom Typ I, welche durch α repräsentiert wird. i gibt den k- ten Perzentilrang für einen p- Wert an [17]. In `limma` legt man durch den p- Wert den Cut- Off für die Ergebnisstabelle fest [19][15].
Dies sind die Schritte, die als einheitliche Methode für die Auswertung der cDNA- Microarraydaten gegangen werden, damit eine allgemeine Metaanalyse der Experimentaldaten möglich ist. Wie diese Daten sich zusammensetzen wird im nächsten Abschnitt erklärt.

2.3 Herkunft der verwendeten Microarray-Datensätze

Im Rahmen des Praktikums zur Vorbereitung der Masterarbeit wurde eine Datenbank erstellt, in der Informationen zu Zebrabärblingsexperiment auf Basis von Genexpressionanalyse mit Microarrays enthält. Aus dieser Datenbank wurden für eine allgemeine Analyse geeignete Genexpressionsdatensätze (Abb. 2.2) ausgewählt. Diese beinhalten

Herkunft	Chemikalien	Replikate	Datenbankeintrag
Büttner et al. 2012 [4]	Gant, Cyclopamin	5	GSE29357
Klüver et al. [11]	Azinphosmethyl	4	GSE27680
Chen et al. [9]	TCDD, Retinsäure	3	GSE9020
Hermsen et al. 2012 [10]	Flusilazol	6	E-MTAB-832
Park et al. 2012 [2]	Fluoxetin, Sertralin	3	GSE31712
Busch et al. [unpub.]	PCE	2	-
Schiller et al. 2013 [5]	Genistein, Linuron, ...	4	GSE34616, GSE44263
Xu et al. 2014 [3]	PCP	3	GSE37019

Tabelle 2.2: Microarraydaten, verwendete Chemikalien, Anzahl der Replikate sowie Kennziffer für Datenbankeintrag

20 verschiedene Chemikalien (Abb. 2.5), die verschiedene Punkte in der embryonalen Entwicklung abdecken, den Einfluss der Lösemittel zeigen und die Wirkung verschiedener Dosen. Die vollständige Tabelle mit allen Datensätzen, die in Betracht für eine Analyse gezogen wurden, befindet sich im Anhang [Anhang/ Datensätze/ Datenbank-MicroarrayExperimente.xls].

Primär wurden Datenreihen aus Experimenten gewählt, die in den ersten 24 bis 48 Stunden nach der Befruchtung (hpf) liegen. Die Daten wurden bei Arrayexpress (https://www.ebi.ac.uk/arrayexpress/) und GEO (http://www.ncbi.nlm.nih.gov/geo/) heruntergeladen. Für die Konzentration wurde die höchste gewählt. Nur für Flusilazol [10] wurde die zweithöchste Konzentration (28 μM) gewählt.

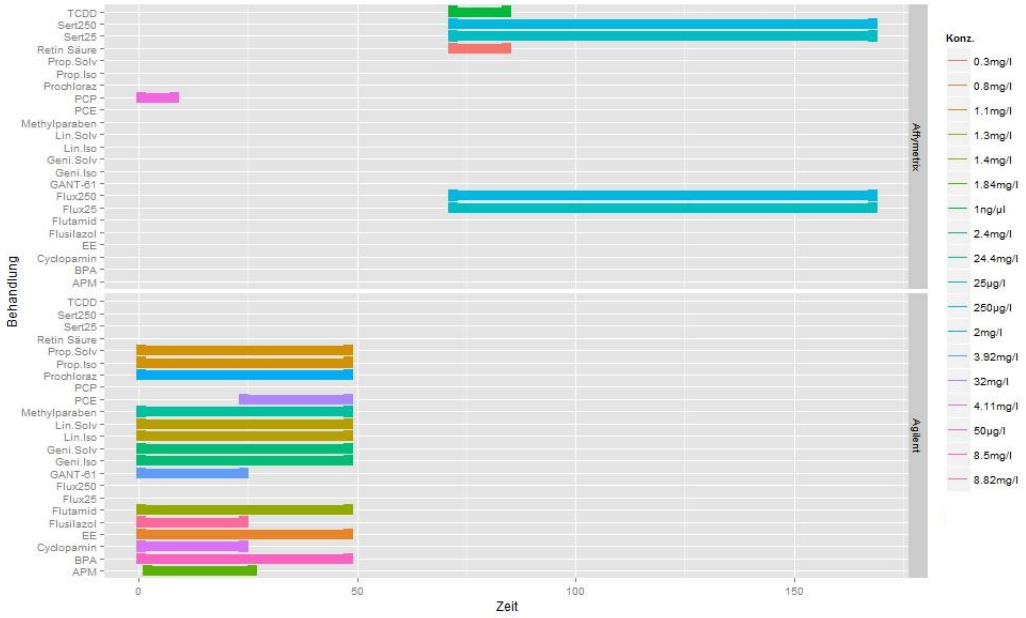


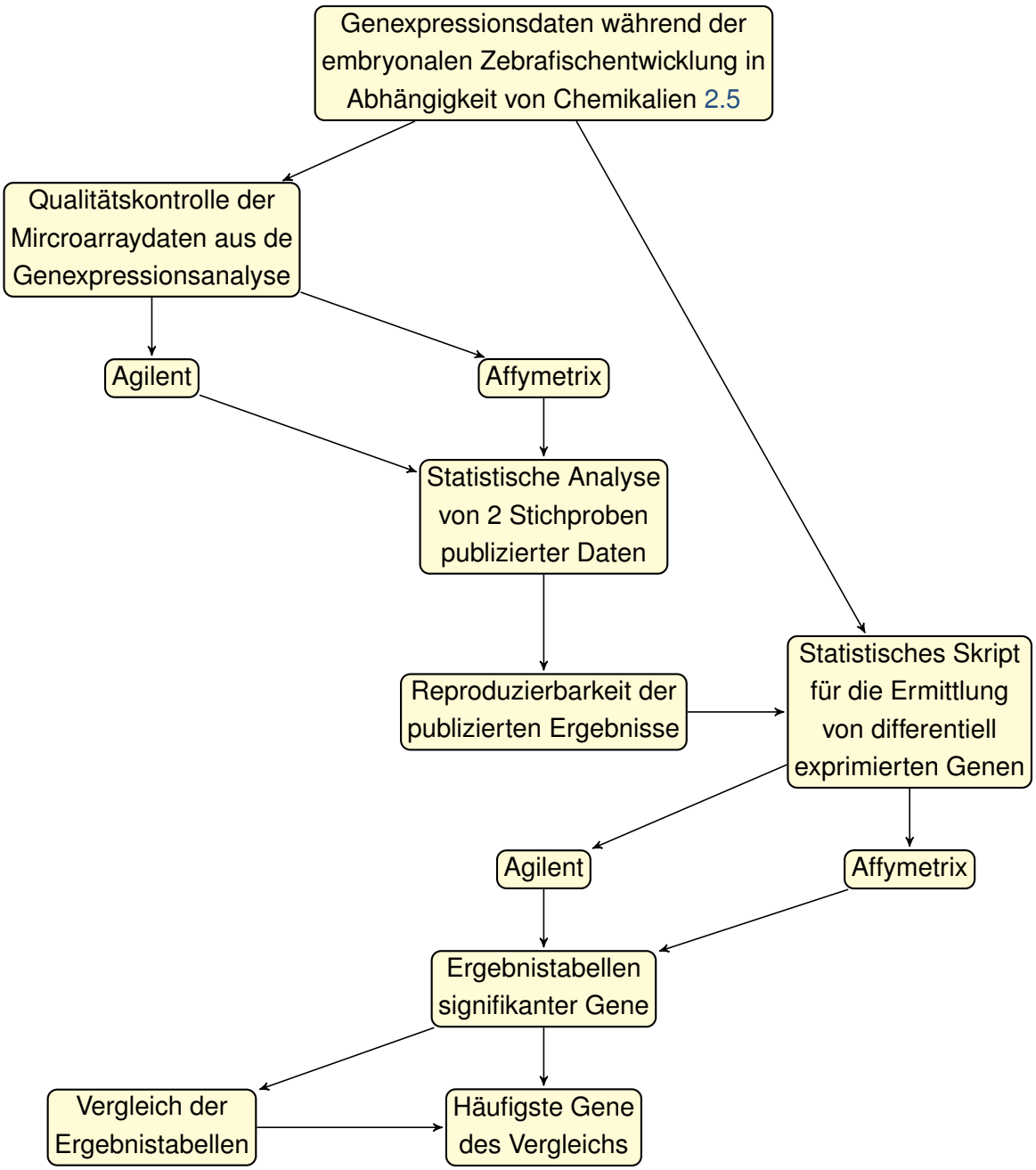
Abbildung 2.5: Die 23 für die Analyse verwendeten Datensätze. Zeit in 'hours past fertilisation' (hpf).

Die gesammelten Genexpressionsdaten wurden in ihren ursprünglichen Experimenten

mit verschiedensten Methoden untersucht und ausgewertet. Das Ziel der Masterarbeit ist es eine einheitliche Methodik für diese Genexpressionsdaten zu schaffen und jeden Datensatz mit den gleichen statistischen Parametern zu untersuchen. Hierbei muss aber beachtet werden, dass die Daten von verschiedenen Microarrayplattformen (Abb. 2.5) stammen und diese in den gleichen Ausgangspunkt versetzt werden müssen. In den Abschnitten **3.3 Analyse der differentiellen Genexpression** und **3.6 Vergleich aller Datensätze** wird das genaue Vorgehen zum Vereinheitlichen der Daten und Auswertung erklärt.

3 Methoden

Ziel ist es zuerst die Qualität der vorliegenden Microarraydaten 2.5 zu prüfen und nach der vollständigen Prüfung eine einheitliche, vergleichende statistische Analyse der Daten untereinander durchzuführen. Hierbei wurde die freie Statistiksoftware R gewählt. Dieser Abschnitt beschreibt die Auswahl der Pakete, Qualitätskontrolle der Daten, dem statistischen Skript und der Korrektheitskontrolle des Skriptes. Folgende Übersicht zeigt die Verknüpfung der einzelnen Schritte.



3.1 Material

3.1.1 R Software

Die Skripte wurden mit R-Studio erstellt. Die Skripte nutzen Pakete, die auf die für die Auswertung von cDNA- Microarraydaten entwickelt wurden. Das Hauptpaket für die Analyse bildet Bioconductor (http://www.bioconductor.org/). Dies ist ein freies Paket für R, das auf die Analyse und den Vergleich genomischer Daten ausgelegt ist. Entworfen wurdes das Paket im Herbst 2001 und erfährt seitdem regelmäßige Updates. Es unterstützt eine große Anzahl an biologischen Analysefunktionen, unter anderem Analysen für Affymetrixarrays, Genom Annotation und umfangreiche graphische Auswertung [20]. Alle verwendeten Pakete und ihre Bedeutung sind in der Tabelle (Tab. 3.1) einzusehen.

Pakete	**Beschreibung**
limma	Lineare Modelle zur Auswertung von Microarraydaten
genefilter	Filter für bessere Suche nach differntielle exprimierten Genen
marray	Paket zum Einlesen von Mehrfarbarrays
affy	Basis zur Analyse von Affymetrixarrays
affyplm	Methoden zum Anpassen von Probe-Level Modellen
simpleaffy	Einlesen von .cel- Dateien und einfache Analysen wie t- Tests
affyQCReport	Qualitätskontrolle für Affymetrixdaten
arrayQualityMetrics	Diverse Qualitätsmesswerte für Microarrays
gplots	Paket für diverse graphische Darstellungen

Tabelle 3.1: Tabelle zur Übersicht und Beschreibung verwendeter R Pakete

3.2 Qualitätskontrolle

3.2.1 Microarraykontrollen

Ziel der Qualitätskontrolle ist es, frühzeitig Microarrays von schlechter Qualität auszusortieren. Arrays mit schlechter Qualität beeinflussen die statistische Auswertung der einzelnen Arrays negativ und verschlechtern letztendlich das Gesamtbild für die allgemeine Analyse. Hierbei muss aber zwischen den Microarrayplattformen unterschieden werden, da für Affymetrix und Agilent R verschiedene Funktionen bereit hält. So wird als erstes auf die für Affymetrix verwendeten Methoden eingegangen und im Anschluss auf Agilent.
Die Methoden nutzen dabei einmal Kontrollregionen, die von der Herstellern auf den

Microarrays angebracht wurden. Neben diesen Kontrollen betrachtet man auch das Gesamtbild aller Microarrays. Durch diese kann man äußere physische Beeinflussung (Artefakte) des Arrays sehen. Als letztes stützt man sich auf 'einfache' statistische Methoden, um die Gesamtheit der Leuchtintensitäten der Sonden zu betrachten und diese sich in einem vorgegebenen Rahmen befinden.

3.2.2 Skript für die Qualitätskontrolle von Affymetrixarrays

Als erstes werden über eine Targetdatei alle Dateien mit der `readAffy`- Funktion eingelesen (Abb. 3.1). Es überträgt .cel-Dateien in ein `R Objekt` der Klasse `Affybatch`. `readAffy` ist automatisch in der Lage die verschiedenen Generationen von Microarraydaten zu lesen. Nach dem Einlesevorgang tragen die Datensätze noch die Dateinamen, weswegen diese mit denen aus der Targetdatei ersetzt werden. Diese Schritte dienen der Vorbereitung und wurden immer wieder für das Einlesen von Affymetrixdaten, auch bei der statistischen Auswertung, vollzogen [16].

```
targets <- readTargets("targetbon15.txt", row.names = "Name")
targets
```

```
##                          File    Name   Label       Type
## C1_15   GSM1129268_15_CA.CEL   C1_15   C1_15 control_15
## C2_15   GSM1129269_15_CB.CEL   C2_15   C2_15 control_15
## C3_15   GSM1129270_15_CC.CEL   C3_15   C3_15 control_15
## MBT1_15 GSM1129277_15_HA.CEL MBT1_15 MBT1_15     MBT_15
## MBT2_15 GSM1129278_15_HB.CEL MBT2_15 MBT2_15     MBT_15
## MBT3_15 GSM1129279_15_HC.CEL MBT3_15 MBT3_15     MBT_15
```

```
ab <- ReadAffy(filenames = targets$File)
sampleNames(ab) <- targets$Name
eset <- rma(ab)
```

Abbildung 3.1: Auschnitt des Skriptes für das Einlesen von Affymetrixarrays

Die nächsten Abschnitte sind die eigentliche Qualitätskontrolle für Affymetrixarrays, welche man nutzt, um die qualitativ schlechten Microarrays zu finden. Sollten Microarrays vorhanden sein, die qualitativ ungenügend sind, werden sie für die Analyse entfernt.
Die Methoden dafür setzen sich zusammen aus einer allgemeinen Qualitätskontrolle für Affymetrixarrays, Clustering, N.U.S.E., R.L.E., Boxplots, Log2- Intensities, RNA- Degradation und Bildern der einzelnen Arrays. Die allgemeine Qualitätskontrolle beinhaltet Parameter die von Affymetrix vorgeschlagen wurden [16] [21] und mit der Funktion `qc` abgerufen werden. Sie wurde in erster Linie für den human HGU95A- Array entwickelt, weswegen diese Kontrolle für den Zebrabärbling mit Vorsicht bewertet werden muss [21].
Für diese Abbildung (Abb. 3.2) werden die von Affymetrix empfohlenen GAPDH- Werte und Actin3/Actin5- Verhältnis gemessen. Die prozentuale Angabe neben den Arraynamen steht für die durchschnittlichen Hintergundintensität. Dieser Wert kann zwischen

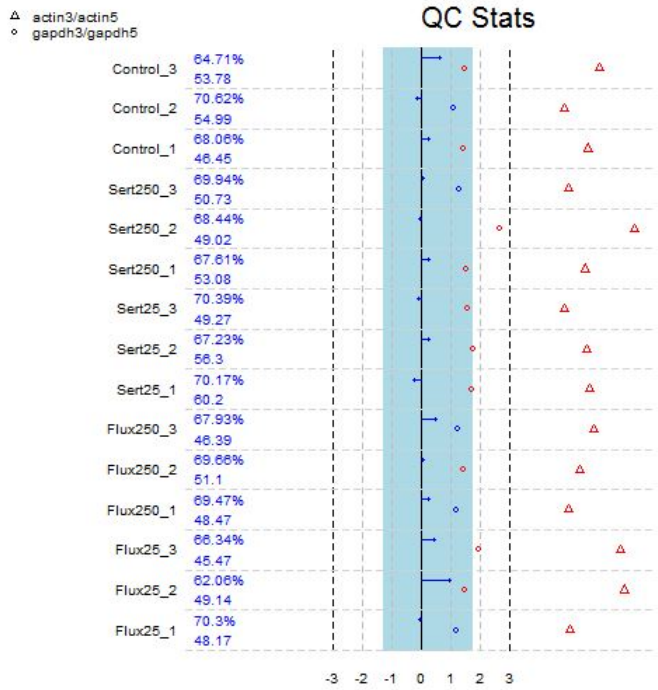


Abbildung 3.2: Interne Qualitätskontrollen von Affymetrixarrays für Park et al. 2012 [2]

den Arrays variieren. Die gestrichelten Linien sind die vorgegebenen Grenzwerte, die nicht überschritten werden sollten, mit Ausnahme des GAPDH- Wertes. Die blaue Fläche beschreibt das dreifache Varianzlevel ausgehend vom Mittelwert aller Arrays. Alle Arrays sollten in diesen Bereich fallen, was durch die blauen Linie mit Kreis gekennzeichnet wird. Die Kreise beschreiben die GAPDH- Werte, dabei stehen blaue Kreise für Werte, die unter dem empfohlenen Wert liegen; rote Kreise liegen darüber. Da der empfohlene Wert vom HGU95A- Microarray ausgeht, ist es sehr wahrscheinlich, dass die Werte höher sind. Allerdings sollten sie nicht zu stark streuen und im gleichen Bereich liegen [21].
Die Dreiecke 3.2 stellen die Actin-Werte dar; für sie gilt das gleiche wie bei GAPDH. Allerdings geht man bei ihnen vom dreifachen Intensitätswert aus. Verlässt einer der Faktoren die Grenzwerte, ist es ein Hinweis auf einen qualitativ schlechten Microarray, der einer genaueren Betrachtung in den nächsten Schritten mit unterzogen werden sollte. Sollten aber alle Microarrays in diesem Wert im selben Bereich außerhalb des vorgegebenen liegen, ist dies auf den Mircroarraytypen zurückzuführen [21].
Im nächsten Schritt wird über alle Arrays eine Heatmap erstellt. Diese zählt zu den Methoden des unüberwachten maschinellen Lernens. Dabei wird in einem Datensatz nach ähnlichen Daten gesucht, die sich zueinander gruppieren, d.h. Gruppen bilden, ohne zuvor feste Zielgruppen (supervised) vorzugeben. Es wird die Differenz (Median) zwischen zwei Microrrays ermittelt. Dies erfolgt mit dem `dist2`- Befehl und der Logarithmierung zur Basis Zwei des Affybatchobjektes. Bei der Heatmap wird die euklidische Distanz der Intensitäten zwischen zwei Microarrays bestimmt.

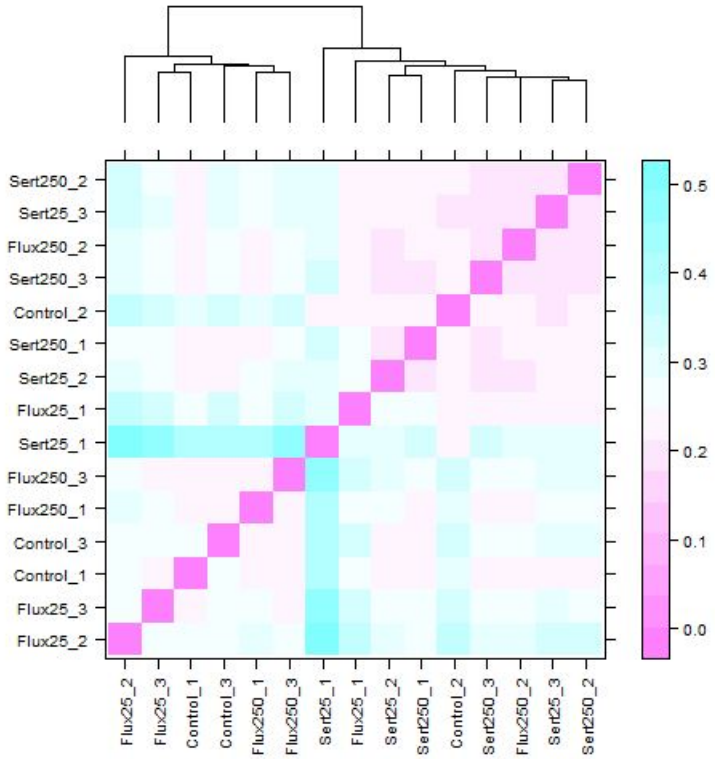


Abbildung 3.3: Heatmap und Clustering aller Arrays für den Datensatz Park et al. [2]

In der Beispielabbildung (Abb. 3.3) sieht man wie sich die eingelesenen Microarrays nach Kontrollen und Behandlungen für den Datensatz Park et al. 2012 [2] gruppieren . Durch sie kann man Batcheffekte (Gruppierung nach Experimentansatz) ermitteln, sowie sehen ob die Proben aus verschiedenen Geweben stammen.
Als nächstes wird vom Paket `affyPLM` Gebrauch gemacht. Mit diesem lassen sich die beiden Abbildungen N.U.S.E. (Abb. 3.4) und R.L.E. (Abb. 3.5) erstellen. Bevor man beide Abbildungen erzeugt , muss mit `fitPLM` aus dem Affybatch ein probe-level Model (Formel 3.1) geschaffen werden. Dies ist Expressionsmesswert, der aus robusten Regression mittels M-Schätzung entsteht [16]. M- Schätzer sind eine Verallgemeinerung von Maximum-Likelihood-Methoden, welche robuster gegen Ausreißer sind.

$$log(Y_{gij}) = \theta_{gi} + \omega_{gj} + \varepsilon_{gij} \tag{3.1}$$

g steht für eine Gensonde, i für den Microarray auf dem sie sich befindet und j das einzelne Replikat dieser Gensonde [21]. θ_{gi} ist die logarithmisch skalierte Intensität (Expression) der Gensonde g auf dem Microarray j. ω_{gj} ist der Effekt den das spezielle Replikat j der Gensonde i auf die Intensität hat. ε_{gij} steht für den sich ergebenden Messfehler [21].
N.U.S.E. (Abb. 3.4) ist der 'Normalized Unscaled Standard Error'. Hierbei wird der Standardfehler für alle Arrays auf Eins gesetzt. Die Boxplots für qualitative schlechte Microarrays im N.U.S.E.- Plot sind dabei höher zentriert und streuen mehr als qualitativ gute Microarrays. Wenn ein Boxplot deutlich über 1.1 liegt ist dies ein eindeutiges Zeichen

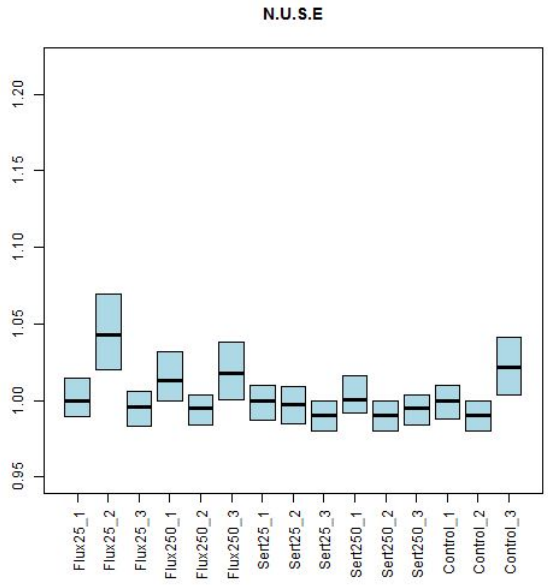


Abbildung 3.4: N.U.S.E. Boxplot für den Datensatz Park et al. 2012 [2]

für schlechte Qualität und sollte auch in allen anderen Qualitätskontrollen auffallen [21]. Die zweite Abbildung (Abb. 3.5) ist die 'Relative Log Expression'. R.L.E.-Plots entstehen aus logarithmisch skalierten Schätzungen für die Expression jeder Gensonde auf über alle Arrays. Es wird das Verhältnis zwischen der Gensonde und dem Mittelwert der Gensonden diesen Typs (Probe-Set) auf allen Arrays betrachtet. Diese relativen Expressionswerte werden als Boxplot dargestellt [16].

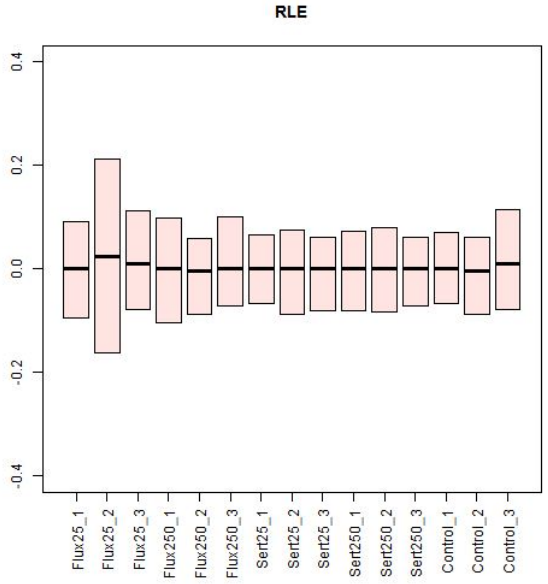


Abbildung 3.5: R.L.E.-Boxplot für den Datensatz Park et al. 2012 [2]

Die meisten Boxplots (Abb. 3.5) sollten um den Nullwert zentriert sein, da auf den meisten Arrays nur wenige Gene differentiell expremiert sind. So sollten sie auch ähnlich in

ihrer Varianz sein [21].
Eine weitere Kontrollinstanz bilden die Bright- und Darkcontrolspots, die die helle und dunkle Referenz definieren [16][21].

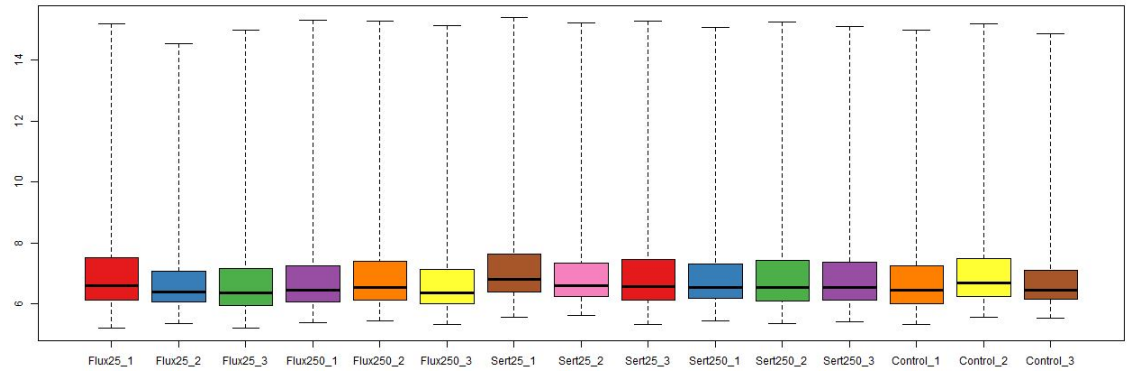


Abbildung 3.6: Boxplot für die Kontrolle der hellen und dunklen Referenz für Park et al. 2012 [2]

Dabei sollten die Boxplots (Abb. 3.6) jeweils auf den dunklen und hellen Referenz zentriert sein und auch nicht zu stark um diesen streuen. Abweichende Arrays weisen wiederum auf Fehler hin.
Bei 'RNA degredation Plots' (Abb. 3.7) werden alle Sonden vom 5'- Ende des Transkripts aus geordnet. Da der RNA- Verdau vom 5'- Ende ausgeht, sollte sich dies in einem Trend in Richtung niedriger Intensitäten für Gensonden, die näher an diesem Ende liegen. Es entsteht ein Graph mit verschiedenen Anstiegen [21].

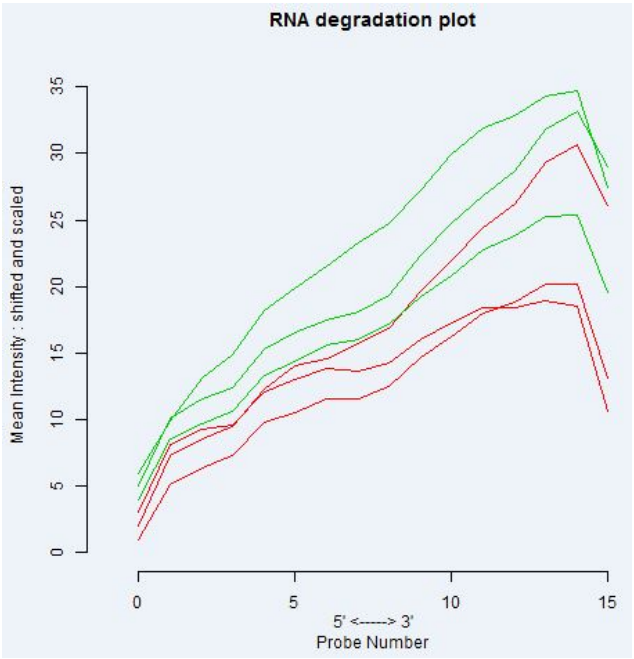


Abbildung 3.7: Grafik die den Verlauf des RNA- Abbaus für den Datensatz Xu et al. 2014 [3] darstellt

Je stärker dabei der Anstieg des Graphes ist, desto stärker war deren RNA-Abbau. Allerdings sind diese Anstiege für jeden Organismus sowie Array verschieden, weshalb man betrachtet wie ähnlich die Anstiege der Graphen untereinander sind. Sie sollten alle den selben Verlauf aufweisen; Graphen (Microarrays) deren Verlauf stark von der Masse abweicht sind, als qualitativ schlechter zu bewerten [21].
Als letztes wird ein Bild des Arrays (Abb. 3.8) erzeugt, um direkt zu sehen, ob sich Artefakte (Blasen, Kratzer usw.) auf dem Arrays zeigen [16].

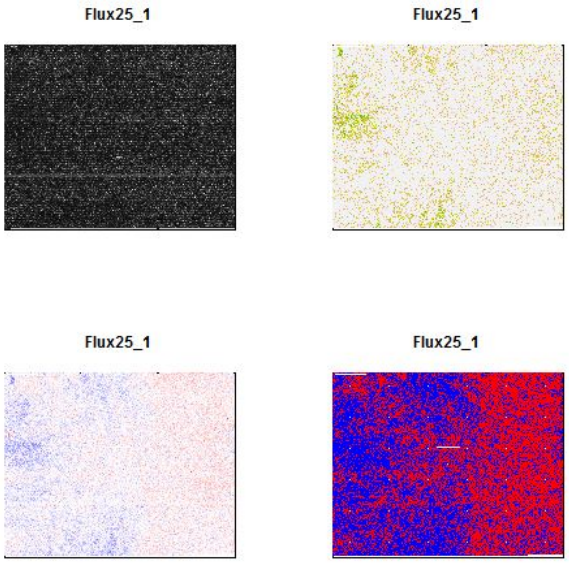


Abbildung 3.8: Bilder eines einzelnen Microarrays `Flux25_1` für den Datensatz Park et al. 2012 [2]

Damit sind die Qualitätskontrollen für Affymetrix abgeschlossen. Nun werden die Qualitätskontrollen für die andere cDNA- Microarrayplattform von Agilent beschrieben.

3.2.3 Skript für die Qualitätskontrolle von Agilentarrays

Für Agilentmicroarrays stehen nicht alle Qualitätskontrollen zur Verfügung, da die meisten speziell für Affymetrix entwickelt wurden. Deswegen wird bei Agilent auch kein probe- level Modell erzeugt.
Wie bei Affymetrix wird eine Targetdatei eingelesen. Aus dieser wird ein `marrayInfo` Objekt erzeugt. Mit `read.Agilent` werden die Arraydaten eingelesen. Das `marrayInfo` Objekt vergibt dabei die Namen und zeigt auf die Dateien. Durch zusätzliche Parameter werden die Farbkanäle, grün (`gProcessedSignal`) und rot (`rProcessedSignal`), festgelegt. Diese Rohdaten müssen als nächstes logarithmiert werden. Dies erfolgt zur Basis Zwei.
Den ersten eigentlichen Schritt der Qualitätskontrolle für Agilentdaten stellen die Bilder (Abb. 3.9) der Microarrays dar. Hierbei wird ein Bild des Arrays aus dem grünen Kanal erzeugt. Bei Zweifarbarrays werden beide Farbkanäle betrachtet.

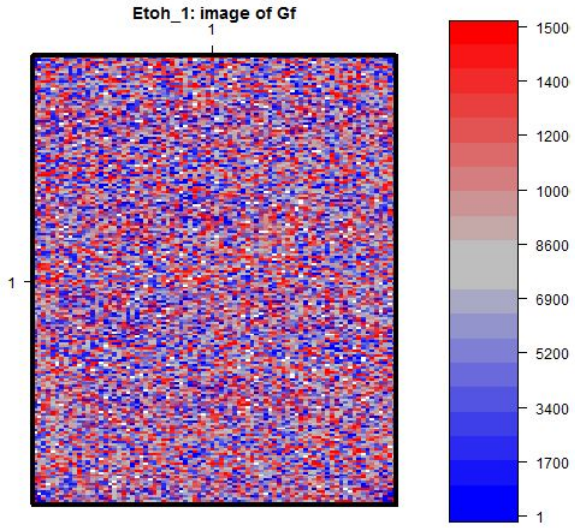


Abbildung 3.9: Bild des ersten Kontrollarrays `Etoh_1` des Datensatzes Büttner et al. 2012 [4]

Rote Punkte signalisieren hohe Intensitäten, graue mittlere und blaue geringe Intensität. An den Ecken des Microarrays zeigen sich die 'Dark Corners' in denen sich die Kontrollsonden für die dunkle Referenz befinden. Analog zu Affymetrix kann man so Artefakte auf dem Microarray identifizieren, die dort eigentlich nicht hingehören. Sollten diese vorhanden sein, sollte man aber noch nicht die Daten des Microarray verwerfen und erst die restlichen Kontrollen durchführen.
Für Agilentdaten werden ebenfalls die Boxplots (Abb. 3.10) der Kontrollsonden erzeugt. Diese sollten analog zu Affymetrix, zentriert zu der hellen und dunklen Referenz sein und nicht zu stark um diese streuen. Dabei wird jeweils wieder auf den grünen und roten Kanal zugegriffen.

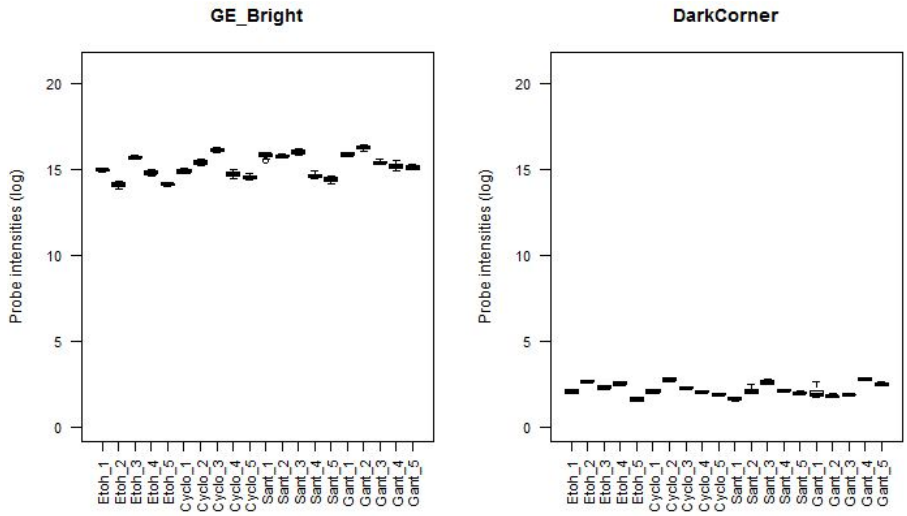


Abbildung 3.10: Boxplots für die helle (links) und dunkle (rechts) Referenz für Büttner et al. 2012 [4]

Es zeigen sich wieder die typischen Boxplots (Abb. 3.10) für die hellen und dunklen Referenzen. Zusätzlich können die Boxplots für die negativ Kontrollen des Arrays, 3xSLv [Anhang/Qualitätskontrollen/XXXXQualitätskontrolle.pdf], gemacht werden. Diese sollten auf keinen Fall exprimiert sein.
Nach den Boxplots kommen zwei Methoden zum Einsatz, die dem maschinellen unüberwachtem Lernen (unsupervised clustering) zugeordnet werden.
Die erste Methode ist, analog zu Affymetrix, die 'Heatmap' (Abb. 3.11). Sie wird mit dem Befehl `dist2` erzeugt. Es wird wieder die euklidische Distanz zwischen zwei Microrrays ermittelt wird.

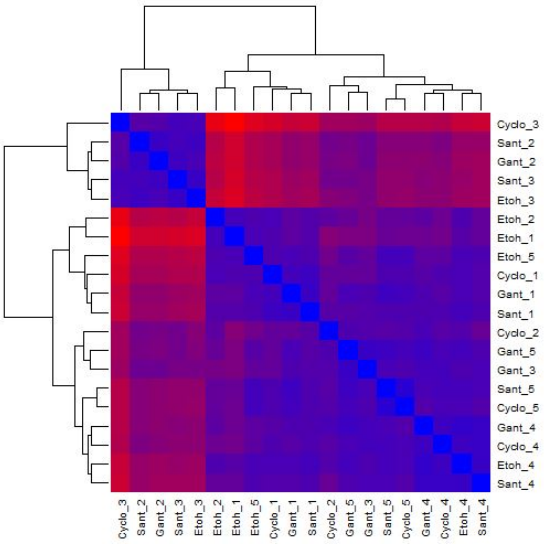


Abbildung 3.11: 'Heatmap' für den Datensatz Büttner et al. 2012 [4] erzeugt mit der `dist2`-Funktion

So gruppieren sich die Arrays untereinander und man kann so Kontrollen sowie Batcheffekte ermitteln. Die Unterschiede zwischen den Microarrays werden durch die Farbcodierung sowie ein Phylogramm gezeigt (Abb. 3.11). Sollte hier ein Micorarray völlig aus dem Rahmen fallen bildet er eine Außengruppe.

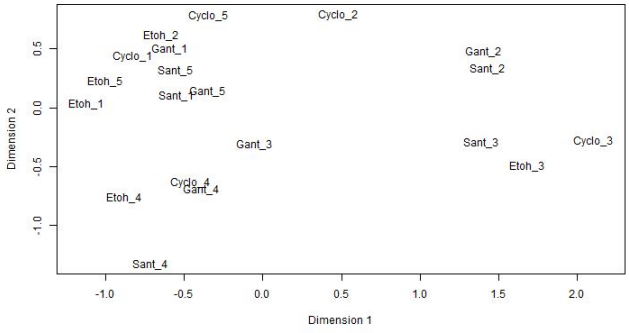


Abbildung 3.12: MDS- Plot für Büttner et al. 2012 [4]

Eine Alternative des unüberwachten Lernens zur 'Heatmap' stellt der MDS- Plot (Abb. 3.12) dar. Die Multidimensionale Skalierung (MDS) ist ein statistisches Verfahren, das Daten räumlich nach ihrer Ähnlichkeit zueinander ordnet. Je näher Daten in diesem Verfahren beieinander liegen, desto ähnlicher sind sie. Hierbei wird wieder auf den grünen oder roten Kanal, je nach Microarraytyp, zugegriffen. Weit von der Kerngruppe entfernte Microarrays müssen wieder genauer in der Kontrolle betrachtet werden. Im Falle der dargestellten Abbildung (Abb. 3.12) sind sich alle Microarrays qualitativ sehr ähnlich, weswegen sie eine gemeinsame Gruppe darstellen. Dadurch wirkt es, als wären einige Datensätze weiter entfernt.
Als letztes werden die Signalintensitäten der Gensonden betrachtet. Dabei werden zuerst die unprozessierten Intensitäten der Kanäle betrachtet.

```
library("limma")
par(mfrow=c(1,2))
gf <- maGf(data.raw.log)
plot(density(gf[,1]), ylim=c(0,0.4), xlab="Probe intensities", main="Raw intensities")
for(i in 2:ncol(gf)){ lines(density(gf[,i])) }
data.norm <- data.raw.log
data.norm@maGf <- normalizeBetweenArrays(data.norm@maGf, method="quantile")
gf <- maGf(data.norm)
plot(density(gf[,1]), ylim=c(0,0.4), xlab="Probe intensities", main="Normalized intensities (quantile)")
for(i in 2:ncol(gf)){ lines(density(gf[,i])) }
```

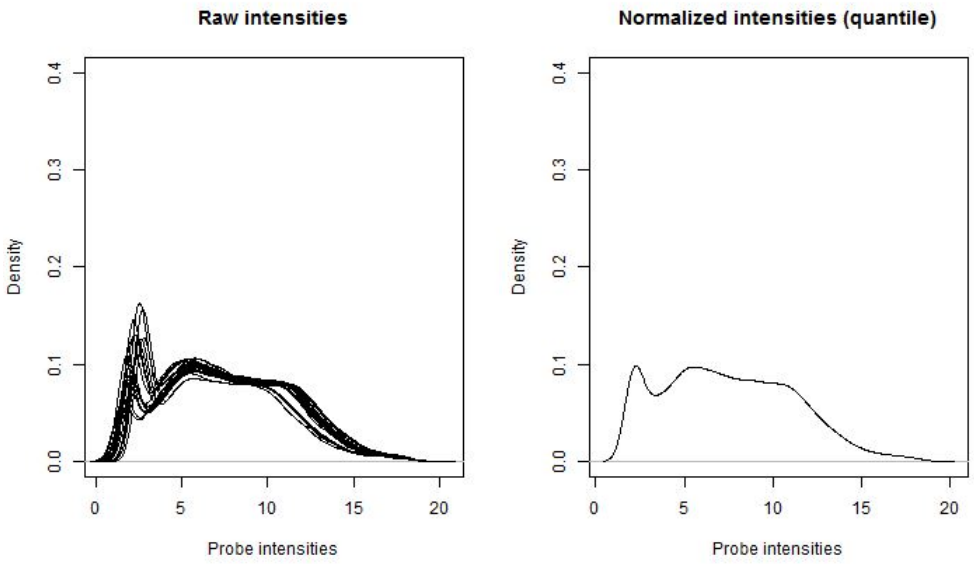


Abbildung 3.13: Skriptausschnitt für die Dichtefunktionen zur Darstellung der Sondenintensitäten. Links Graphen aller Microarrays für den Datensatz Büttner et al. [4]. Rechts die zueinander normalisierten Daten.

Die entstandene Dichtefunktion (Abb. 3.13) sollte einen Verlauf mit einem hohen Peak zu Beginn zeigen und danach ein langsam abfallende Kurve. In dieser Form wird die Dichtefunktion für alle Microarrays einzeln erzeugt. Im zweiten Schritt werden Microarrays quantilweise zueinander normalisiert, sodass die Abbildung dieser normalisierten Signale (Abb. 3.13) nur noch eine einzige Dichtefunktion zeigt, da das Signal aller Gensonden der gleichen Art, zu einem Signal zusammengefasst wurde. Diese Intensitätenverteilung kann dann quantilweise geordnet werden. Ausgehend von einem Datensatz

werden die Quantile der Datensätze transfomiert [14]. Dies wird gemacht um nichtbiologische Unterschiede zu minimieren.
Um nun Microarrays für Agilent als auch Affymetrix aus der Analyse zu entfernen, sollten diese jeweils in allen Kontrollen stark auffallen. Allerdings sollte man auch hier noch nicht vorschnell Datensätze entfernen, denn die Fehler können auch nur einen kleinen Teil des Microrrays betreffen, welcher in der statistischen Analyse nicht mehr ins Gewicht fällt. Sollten grobe Fehler vorhanden sein, können diese Arrays bedenkenlos entfernt werden.

3.3 Analyse der differentiellen Genexpression

3.3.1 Annotation der Datensätze

Bei der Annotation werden den auf den cDNA- Microarrays befindlichen Sonden, die in den Rohdaten nur eine Seriennummer tragen, die zu dieser Seriennummer korrespondierenden Gendatenbankidentifikatoren zugeordnet. Die Annotation für Affymetrix und Agilentarrays verwendet den gleichen Vorgang. Für die Annotation werden die `GPL.soft` Dateien der Arrayplattformen verwendet, welche den Datensätzen in der GEO- Datenbank beiliegen. In dieser Datei befindet sich eine Tabelle, die den Sonden Seriennummern Gennamen, RefSeq- IDs und Ensembl- IDs zugeordnet hat.
Sie wird über `read.table` eingelesen. Aus den Arrayrohdaten werden die Sondennummern, die sich in `data.raw@maGnames@maInfo$ProbeName` für Agilentarrays und in `rownames(eset)` für Affymetrix befinden, auf einen neuen Vektor `array_ids` übertragen. Die `merge` Anweisung vergleicht und verknüpft die eingelesene Tabelle und den neuen Probenamenvektor über die Sondenseriennummer. Hieraus wird ein neuer Dataframe `t1` erstellt. Dieser enthält alle zusammengehörigen Identifikatoren (Abb. 3.14).

```
array_ids = data.raw@maGnames@maInfo$ProbeName
#t <- getBM(attributes=c('agilent_g2518a','external_gene_id'), filters = 'agilent_g2518a' , values = array_ids,
#t1<-data.frame(probe_id=t$agilent_g2518a, external_gene_id=t$external_gene_id)
t<-read.table("GPL.soft", header=TRUE, sep="\t", fill = TRUE)
m <- merge(array_ids, t, by.x="x", by.y="NAME")
t1 <- data.frame(probe_id=m$x, external_gene_id=m$GENE_SYMBOL, entrez_id=m$REFSEQ, ensemble_id = m$ENSEMBL_ID)
```

Abbildung 3.14: Annotation eines Datensatzes mittels `GPL.soft` aus der GEO Datenbank

Sobald die Rohdaten die statistischen Methoden durchlaufen haben, kommt die Annotation aus `t` dazu. In `x` werden die Sondenseriennummern aus der Ergebnisstabelle `tp` mit `t1` verknüpft. Mit dieser Verknüpfung als Bedingung werden die Annotationsinformation an die Ergebnistabelle angehängt.
Durch die Annotation ist es nun möglich, in den Ergebnisstabellen genau zu sagen, wie stark welches Gen exprimiert ist, da die Seriennummern nun eindeutig identifiziert sind.

3.3.2 Allgemeines Skript für die statistische Analyse der Agilentarrays

Das Skript für die statistische Auswertung läuft bis zum Logarithmieren gleich. Es wird mit der Targetdatei und dem `marray`- Objekt eingelesen. Nach dem die Rohdaten gelesen sind werden sie logarithmiert (Abb. 3.15). Im Anschluss wird quantilweise zwischen den Microrrays normalisiert. Danach erfolgt der Annotationsschritt, welcher zuvor erklärt wurde.

```
metadata <- read.table("targetbuett.txt", header=TRUE)
data.info <- new("marrayInfo", maNotes="targetbuett.txt", maLabels=as.character(metadata$Label), maInfo=as.data.frame(metadata))
data.raw <- read.Agilent(targets=data.info, name.Rf = "gProcessedSignal", name.Rb = "gBGMedianSignal", name.Gf = NULL, name.Gb = NULL, sep="\t", quote="\"")
```

```
## Reading ...  ./GSM725686.txt
## Reading ...  ./GSM725687.txt
## Reading ...  ./GSM725688.txt
## Reading ...  ./GSM725689.txt
## Reading ...  ./GSM725690.txt
## Reading ...  ./GSM725691.txt
## Reading ...  ./GSM725692.txt
## Reading ...  ./GSM725693.txt
## Reading ...  ./GSM725694.txt
## Reading ...  ./GSM725695.txt
## Reading ...  ./GSM725696.txt
## Reading ...  ./GSM725697.txt
## Reading ...  ./GSM725698.txt
## Reading ...  ./GSM725699.txt
## Reading ...  ./GSM725700.txt
## Reading ...  ./GSM725701.txt
## Reading ...  ./GSM725702.txt
## Reading ...  ./GSM725703.txt
## Reading ...  ./GSM725704.txt
## Reading ...  ./GSM725705.txt
```

```
#####
```

```
#####
data.raw.log <- data.raw
data.raw.log@maRf <- log2(data.raw.log@maRf)
data.raw.log@maRb <- log2(data.raw.log@maRb)
colnames(data.raw.log@maRf) <- data.raw.log@maTargets@maLabels
colnames(data.raw.log@maRb) <- data.raw.log@maTargets@maLabels
data.norm <- data.raw.log
data.norm@maRf <- normalizeBetweenArrays(data.norm@maRf, method="quantile")
```

Abbildung 3.15: Skript für das Einlesen und Logarithmieren der Daten für Agilent. Dieser Daten benutzt nur `Single Channel` bei dem das grüne Signal gelesen wird. Im mittleren Abschnitt sind alle erfoglreich gelesenen Arrays.

Nach diesen vorbereitenden Schritten findet die statistische Auswertung der Arraydaten statt. Als erstes muss die Hilfskript `loadFunc.R` geladen werden. Dessen Funktion ist es, aus den Rohdaten ein Biobase Expressionset- Objekt zu erzeugen. Dabei kombiniert das Skript die Intensitäten gleicher Gensonden zu einer mittleren Intensität für jeden Microarray. Das neue Expressionset kann danach von `genefilter` und `limma` verwendet werden [Anhang/Skripte/ `loadFunc.R`].
Vor der weiteren Auswertung wird das neue Set mit einem unspezifischen Filter bearbeitet. Dabei entfernt man Gensonden die keine Änderung in ihrer Expression aufweisen, denn sie würden das Problem des multiplen Testens erhöhen [21]. Zusätzlich werden die Sonden der Negativkontrolle (3xSlv) entfernt. Dieser Filter ist die zweite Funktion des Hilfskriptes `loadFunc.R`.
Das gefilterte Set wird in im Abhängigkeit vom 'Type', der in der Targetdatei festgelegt wurde, in eine 'Designmatrix' übertragen (`model.matrix`) (Abb. 3.16). Diese weisst allen Arrays ein Informationslabel zu um Kontrollen und Behandlungen zu trennen. Sie bildet den Rahmen für die statistische Auswertung der Datensätze. Die Spaltennamen

der Matrix werden korrigiert (gsub). Mit lmfit wird das Expressionset mittels der 'Designmatrix' auf ein lineares Modell angepasst (fit).

```
source("loadFunc.R")
library("Biobase")
library(genefilter)
data.norm.eS <- agilent.marray2exprset.median(data.norm)
#####
#gleiches Problem
p=1
design <- model.matrix(~0+factor(data.norm.eS$Type))
colnames (design) <- gsub("factor\\(data.norm.eS\\$Type\\)", "", colnames (design))
fit <- lmFit (data.norm.eS, design)
contM <- makeContrasts(dif=Cyc-control, levels=design)
fit2 <- contrasts.fit(fit, contM)
fit2 <- eBayes(fit2)
tp <- topTable(fit2, coef="dif", adjust="BH", number=nrow(fit2), p.value=p)
tp$ID <- rownames(tp)
#rownames(tl)<-tl$probe_id
x<-match(rownames(tp),tl$probe_id)
tp$external_gene_id <-tl$external_gene_id[x]
tp$entrez_id <- tl$entrez_id[x]
tp$ensemble_id <- tl$ensemble_id[x]
```

Abbildung 3.16: Skriptabschnitt für die statistische Auswertung der Agilentdaten. Dieses Skript ist auf den Datensatz Büttner et al. 2012 [4], im speziellen die Behandlung mit Cyclopamin gegen Kontrolle angepasst.

Die Variable p definiert den Cut- Off der FDR. Als nächste wird der Kontrast contM festgelegt. Der Kontrast (Intensitätsverhältnis Kontrolle gegen Behandlung) beschreibt wie sich die Genexpression für eine Chemikalie ändert, d.h. welche Anstiege des linearen Modells miteinander verglichen werden. Diese Änderung wird im Ergebnis als Log-Fold- Change (logFC) angegeben. Er zeigt an wie viel höher (positiver Wert) oder niedriger (negativer Wert) die Expression ist. Der Kontrast wird auf das bisherige Lineare Modell angepasst (contrasts.fit).
Im nächsten Schritt wird mit dem Befehl eBayes die empirische Bayes-Methode abegrufen, die sich Informationen von alle Microarrays borgt, damit die Statistik robuster wird. Durch die Anweisung tp wird ein t- Test durchgeführt, der prüft ob die Anstiege im linearen Modell ungleich null sind. Damit tp weiß, dass es für das multiple Testen den t- Test verwendet wird, muss im Befehl auf den Kontrast coef=‘‘dif‘‘ verwiesen werden. Für das multiple Testen wird für die FDR die Methode von Benjamini-Hochberg (adjust="BH‘‘ [19]) mit einem Cut- Off von p. Das Ergebnis wird in einer Ergebnistabelle in Form einer Textdatei gespeichert.
Zum Schluss wird eine 'Heatmap' aus dem Expressionset erstellt.

3.3.3 Allgemeines Skript für die statistische Analyse der Affymetrixdatensätze

Das Affymetrixskript ist in seinem Aufbau analog zum Agilentskript. Das Einlesen erfolgt nach derselben Methode wie bei der Qualitätskontrolle für Affymetrixdaten. Aus den Rohdaten wird mit dem Befehl rma ein Expressionset erzeugt. Dieser Befehl beinhaltet alle Zwischenschritte des Agilentsskripts, die dort für das Expressionset notwen-

dig waren [16]. Affymetrixdatensätze benötigen somit nicht das `loadFunc.R` Hilfsskript. Als nächste kommt der Annotationsschritt (siehe **3.3.1 Annotation der Datensätze**). Die statistische Methode ist die gleiche wie bei dem Agilentscript, jedoch wird die Designmatrix aus dem `strain`- Vektor (Abb. 3.17) erzeugt, der in Behandlungen und Kontrolle einteilt. Dieser Vektor ist dabei auf die Experimentdaten individuell angepasst d.h. er muss jedes mal neu gestaltet werden.

```
p=0.05
strain <- c("Flux250","Flux250","Flux250","control","control","control")
design <- model.matrix(~0+factor(strain))
colnames(design)<- c("control", "Flux250")
fit <-lmFit(eset, design)
contM <- makeContrasts(dif=Flux250-control, levels=design)
fit2 <- contrasts.fit(fit, contM)
fit2 <- eBayes(fit2)
tp<-topTable(fit2, coef="dif", adjust.method="BH", number=nrow(fit2), p.value=p)
tp$ID <- rownames(tp)
x<-match(rownames(tp),tl$probe_id)
tp$external_gene_id <-tl$external_gene_id[x]
tp$entrez_id <- tl$entrez_id[x]
tp$ensemble_id <- tl$ensemble_id[x]
head(tp)
```

Abbildung 3.17: Statistische Auswertung von Affymetrixarrays. Der verwendete Datensatz ist Park et al. 2012 [2] für Fluoxetin 250 μg/l

Die Ergebnistabelle `tp` (Abb. 3.17) wird mit der Annotation versehen, eine Heatmap des Expressionsets `eset` erzeugt und die Ergebnisse in einem Textdokument gespeichert. Durch beide allgemeinen Skripte ist es nun möglich eine Auswertung der Microarraydaten auf Basis von linearen Modellen und t- Tests durchzuführen, die die Datenbasis für die eigentliche allgemeine Analyse bilden.

3.4 Spezifität und Sensitivität - Vergleich der Ergebnisse mit denen der Fachartikel

Bevor man aber mit den Methoden für die allgemeinen Analyse fortfährt, muss noch die Korrektheit der statistischen Skripte gewährleistet sein, damit keine falschen Ergebnisse, etwa durch fehlerhaftes Einlesen, erzeugt werden.
Um die Korrektheit der beiden allgemeinen Skripte (**3.3.2** und **3.3.3**) für Agilent und Affymetrix zu prüfen, wurden diese genutzt um stichprobenartig die Ergebnisse aus zwei Publikationen (Büttner et al. 2012 [4], Hermsen et al. 2012 [10]) zu rekonstruieren. Dazu wurde für die Ergebnistabelle der FDR-Wert auf eins gesetzt, um alle Ergebnisse, egal ob signifikant oder nicht, zu sehen. Die Ergebnisstabelle wurde in einer Datei abgespeichert. Die beiden Datensätze nutzen einen Agilent-Microarray, allerdings wurde bei Büttner et al. 2012 [4] nur ein Farbkanal genutzt und bei Hermsen et al. zwei Farbkanäle. Die Korrektheit von Affymetrix ist soweit schon gegeben, da sie eine eigene `R`-Funktion zum Einlesen der Daten besitzt. Bei Agilent ist dies nicht der Fall, weswegen die in **3.3.2** verwendete Methode geprüft werden muss. Die statistische Auswertung läuft für beide Methoden (**3.3.2** und **3.3.3**) gleich, weswegen hier der Test mit den Agilentdaten genügt.
Diese Datei musste nun mit der unterstützenden Material (Ergebnistabellen) der Publikation verglichen werden. Hierzu wurde eine Skript entwickelt. Als erstes werden

```
Vergleich <- read.table("VergleichCyclo.txt", header=TRUE, sep="")
myTp <- read.table("GPLAnnoTP.txt", header=TRUE, sep="\t")


myResults <- data.frame(my_logFC=myTp$logFC, my_FDR=myTp$adj.P.Val, my_avreps=myTp$AveExpr, my_genename = as.character(myTp$external_gene_id), my_probe_id=myTp$ID,
my_entrez=myTp$entrez_id, my_ensembl=myTp$ensemble_id)
myResults$my_contrast <- "Cyclo"
#myResults$my_Nm <- tolower(myResults$my_Nm)
#Vergleich$nm_id <- tolower(Vergleich$nm_id)

newTg <- merge(Vergleich, myResults, by.x = "nm_id", by.y = "my_entrez", all.x = TRUE)
newTg2 <- merge(Vergleich, myResults, by.x = "nm_id", by.y = "my_ensembl", all.x = TRUE)
View(newTg)
write.table(newTg, file="BuettnerCycloEntrez.txt", sep="\t", quote = FALSE)
View(newTg2)
write.table(newTg2, file="BuettnerCycloEnsemble.txt", sep="\t", quote=FALSE)
S1<-which(newTg2$my_FDR < 0.05, TRUE)
S2<-which(newTg$my_FDR < 0.05, TRUE)
S3<-which(myResults$my_FDR < 0.05, TRUE)
length(S1)
```

```
## [1] 8
```

```
length(S2)
```

```
## [1] 41
```

```
length(S3)
```

```
## [1] 96
```

```
length(Vergleich$Gene_ID)
```

```
## [1] 57
```

Abbildung 3.18: Vergleich zwischen den Hilfsdaten und der erzeugten Ergebnistabelle des Datensatzes Büttner et al. 2012 [4]

das unterstützende Datenmaterial (`Vergleich`) und Ergebnistabelle (`myTp`) eingelesen (Abb. 3.18. Dies geschieht mit dem `read.table`-Befehl. Im zweiten Schritt überträgt das Skript die Ergebnistablle in einen neuen `dataframe`. Die Spalten werden umbenannt, damit sie eindeutig für die Ergebnistabelle (z.B `my_FDR`) des vereinheit-

lichten Skript stehen. Als letztes wird in das neue Objekt (`my_Results`) eine Spalte eingetragen, die den Kontrast, mit dem die Ergebnisstabelle zugeordnet ist, angibt (`my_contrast`).
Nach diesen vorbereitenden Schritten erfolgt der Vergleich über `merge`, welcher auch gleichzeitig die beiden Objekte in einer Tabelle verbindet (Abb. 3.18). Durch das Attribut `all.x=TRUE` werden alle Zeilen aus dem unterstützende Datenmaterial aufgeführt. Die kombinierten Tabellen werden in einer Textdatei abgespeichert. Als letztes wird durch `length` und `which` die Anzahl der differentiell exprimierten Gene bestimmt. Dabei wird der Wert FDR auf kleiner als 0.05 gesetzt um die signifikanten Ergebnisse zu filtern.
Die Ergebnisse werden auf ihre Sensitivität und Spezifität untersucht. Dabei gilt:

$$Sens = \frac{tp}{tp + fn} \tag{3.2}$$

$$Spez = \frac{tn}{tn + fp} \tag{3.3}$$

- tp = true positiv - Gene die in beiden Ergebnistabellen differentiell exprimiert sind
- fp = false positiv - Gene die nur im allgemeinen Skript differentiell exprimiert sind
- tn = true negativ - Gene die in beiden Ergebnistabellen nicht differentiell exprimiert sind
- fn = false negativ - Gene die im allgemeinen Skript nicht differentiell exprimiert sind

Für Büttner et al. 2012 [4] (Cyclopamin) konnten alle differentiell exprimierten Gene aus der Publikation wiedergefunden werden, allerdings waren nur 49 (tp) der 57 Gene aus dem Hilfsmaterial in der Ergebnistabelle signifikant (FDR 0,05). Insgesamt waren in der Ergebnistabelle 96 Gene signifikant (Abb. 3.18) [4]. Dies entspricht einer Sensitivität von 85,9% und einer Spezifität von 99,7%, da mit dem verallgemeinerten Agilentskript 96 differentiell exprimierte Gene gefunden wurden.
Das gleiche wurde für Hermsen et al. 2012 [10] ebenfalls durchgeführt, jedoch mit einem f-Test (kein Kontrast spezifiziert für das multiple Testen). Der f- Test prüft, ob unter der Annahme der Nullhypothese, grundlegende Unterschiede in der Varianz zwischen zwei Populationen (Kontrolle gegen Behandlung) bestehen. So vergleicht der Test in diesem Fall alle neun Chemikalienkonzentrationen paarweise. Ebenfalls waren alle Gene (239) in der Ergebnistabelle vorhanden und 238 (tp) wurden als differentiell exprimiert (FDR 0,05) ausgegeben [10]. Für diesen Datensatz bedeuted dies, dass das allgemeine Skript eine Sensitivität von 99,5% hat, allerdings wurde mit der ursprünglichen Methode [10] ein Gen mehr gefunden, das im differentiell exprimierten Bereich liegt.
Somit konnte die korrekte Arbeitsweise der allgemeinen Skripte bestätigt werden, da sie fast 100% der Gene bei Hermsen et al. 2012 [10] fanden und mehr als 80% der Gene aus der Publikation von Büttner et al. 2012 [4].

3.5 Vergleich der Kontrolldatensätze

Als erster Schritt der allgemeinen Analyse wurde allen Kontrollen, der zur Analyse stehenden Datensätze verglichen. Um sie zeitlich besser einzuordnen, wurde eine Übersicht geschaffen (Abb. 2.5). Zu sehen sind alle Datensätze/Chemikalien und der zugehörige Zeitraum in dem der Versuch durchgeführt wurde. Desweiteren sieht man die Arrayplattform und eingesetzte Chemikalienkonzentration.
Zum Vergleich wurden alle Kontrollen in eine Targetdatei aufgenommen. Diese wurden nochmals in Agilent und Affymetrix aufgeteilt. Nach dem Einlesen wurden die Daten logarithmiert, zueinander normalisiert und eine Heatmap (Abb. 3.19) erzeugt, genau so wie im Qualitätskontrollskript.

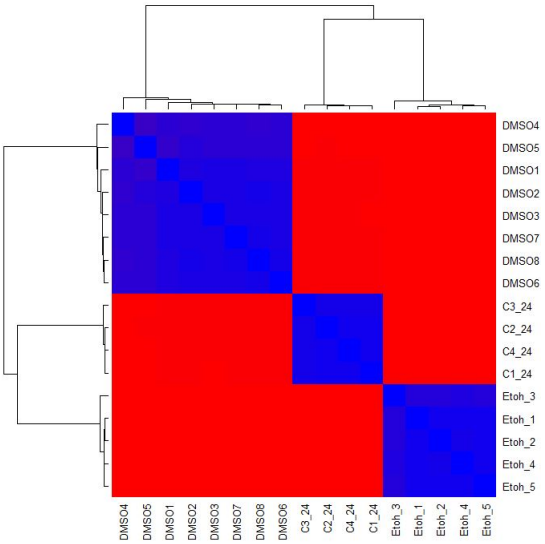


Abbildung 3.19: Heatmap für alle 24 h Agilentkontrollen

Zu sehen sind alle Agilentdaten für 24 hpf (hours past fertilisation). Die Daten ordnen sich jeweils nach ihren zugehörigen Experimenten an. Zusätzlich gruppieren sie sich auch nach den jeweiligen Einrichtungen. Es war allerdings zu erwarten, dass die Kontrolldaten sich unterscheiden. Die zu sehenden Effekte (Abb. 3.19) sind so stark, dass für den Vergleich zwischen den Datensätzen, die Daten innerhalb einer Publikation sich stärker überschneiden.

3.6 Vergleich der differentiellen Expression aller Datensätze

Für erste Ergebnisse wurden die Ergebnistabelle aller Datensätze untereinander verglichen. Die Ergebnistabelle wurde mit einer FDR von 0,05 erstellt. Das Vergleichskript (Abb. 3.20) liest als erstes die Ergebnistabelle der Chemikalie ein, die verglichen werden soll. Leere Zellen werden mit 'NA'- Einträgen aufgefüllt. Es wird der p- Wert, der Genname (`external_gene_id`), die RefSeq-ID (`entrez_id`) und die Ensembl- Transkript- ID (`ensemble_id`) in einen neuen Dataframe übergeben. Danach werden die Ergebnistabellen aller anderen Chemikalien eingelesen. Für sie werden die gleichen Werte in einen neuen `Dataframe` übertragen. Sämtliche Daten werden jetzt noch von NA-Einträgen bereinigt. Der Vergleich zwischen zwei Ergebnistabellen erfolgt über den `merge` Befehl, dieser Vergleicht wahlweise Genname, RefSeq- ID oder Ensemble-Transkript- ID.

```
c<-read.table("TpWiebke.txt", header=TRUE, sep="\t",na.strings=c("", "NA"))
C<-data.frame(c$ensemble_id, c$adj.P.Val)
C <- na.omit(C)
q<-read.table("ParkSert250.txt", header=TRUE, sep="\t", na.strings=c("", "NA"))
Q<-data.frame(q$ensemble_id, q$adj.P.Val)
Q <- na.omit(Q)
Vg1<-merge(C,Q,by.x="c.ensemble_id",by.y="q.ensemble_id")
Vg1<-unique(unlist(Vg1$c.ensemble_id))
View(Vg1)
write.table(Vg1, file="XvsSert250.txt", sep="\t", quote=FALSE)
#####
w<-read.table("ParkFlux250.txt", header=TRUE, sep="\t", na.strings=c("", "NA"))
W<-data.frame(w$ensemble_id, w$adj.P.Val)
W <- na.omit(W)
Vg2<-merge(C,W,by.x="c.ensemble_id",by.y="w.ensemble_id")
Vg2<-unique(unlist(Vg2$c.ensemble_id))
View(Vg2)
write.table(Vg2, file="XvsFlux250.txt", sep="\t", quote=FALSE)
```

Abbildung 3.20: Vergleichsskript für Ergebnistabellen

Für jeden dieser Werte gibt es ein eigenes Skript, aber für den Vergleich wurde die RefSeq- ID verwendet. Die daraus entstehenden verknüpften Tabellen werden noch von doppelten Einträgen, die aus Mehrfachsonden resultieren, bereinigt. Diese Gene werden in einem Textdokument abgespeichert.
So wird schrittweise jede Chemikalie miteinander verglichen. Die Menge an überlappenden Genen wird gezählt und in eine Tabelle eingetragen. Zusätzlich wird das Auftreten der Gene über die gesamte Tabelle gezählt, um die absoluten Häufigkeiten zu ermitteln. Diese Zählung erfolgt dabei mit der `rle`- Funktion von `R`. Die 23 am häufigsten im Vergleich auftretende Gene sind genauer untersucht worden. Überprüft wurde, bei welchen Behandlungen sie differentiell exprimiert waren sowie den Stoffwechselsystemen, denen sie zu geordnet werden können. Die Einordnung der Gene wurde mit Biosystems (http://www.ncbi.nlm.nih.gov/biosystems/) gemacht. Es wurden die GO-Terms für die Gene übernommen. Diese stehen für Gen- Ontologien und ordnen ein Gen einem oder mehreren Zellbestandteilen oder Prozessen zu.

4 Ergebnisse

4.1 Ergebnisse

Die Ergebnisse teilen sich in vier Bereiche auf. Als erstes werden die Ergebnisse der Qualitätskontrolle gezigt, die darüber entscheiden, welche Microarrays für die differentielle Expression verwendet werden. Der zweite Teil der Auswertung befasst sich mit den Ergebnissen der differentiellen Expression. Der dritte Bereich betrifft die Überlappungen zwischen den Ergebnissen der differentiellen Expression, also die Anzahl der Gene die in mehreren Ergebnistabellen enthalten sind. Als letztes wird die absolute Häufigkeit des Auftretens der Gene präsentiert.

4.2 Ergebnisse der Qualitätskontrolle

Die Qualitätskontrolle der Arrays zeigte nur wenige stark auffällige Microarrays. Jedoch waren deren defekte nur lokaler Natur, sodass sie nicht das Gesamtbild der Microarrayanalyse beeinflussten. Somit gingen wir mit allen vorhanden Replikaten in die Analyse, damit auch die Anzahl der Replikate möglichst groß ist. Die vollständige Qualitätskontrolle für jeden einzelnen Microarray ist im Anhang einzusehen [Anhang/Qualitätskontrolle/XXXQualitätskontrolle.pdf].
Hier sind in einer Übersicht (Tab. 4.1) alle Microarrays aufgelistet, die in der Qualitätskontrolle aufgefallen sind und warum:

Microarray	**Art des Fehlers**
`Flus4` (Hermsen et al. 2012 [10])	Artefakt links (Bild des Arrays)
`DMSO3` (Hermsen et al. 2012 [10])	Artefakt mitte (Bild des Arrays)
`DMSO2` (Chen et al.2008 [9])	Artefakt rechts (Bild des Array)
`TCDD1` (Chen et al.2008 [9])	Diagonale Farbverteilung (Bild des Arrays)
`A3_24` (Klüver et al. 2011 [11])	Artefakt unten rechts (Bild des Arrays)
`Flux25_2` (Park et al. 2012 [2])	Linie untere Hälfte (Bild des Arrays)
`control_3` (Xu et al. 2014 [3])	N.U.S.E. und R.L.E. Abweichung
`EEs4` (Schiller et al. 2013 [5])	Artefakt unten links (Bild des Arrays)
`Flus4` (Schiller et al. 2013 [5])	Artefakt unten links (Bild des Arrays)
`Pro4` (Schiller et al. 2013 [5])	Artefakt rechts (Bild des Arrays)
`Lin3` (Schiller et al. 2013 [5])	Breiter Boxplot für Helle Kontrolle (GE_Bright)
`Lin4` (Schiller et al. 2013 [5])	Breiter Boxplot für Helle Kontrolle (GE_Bright)
`Met1` (Schiller et al. 2013 [5])	Breiter Boxplot für Helle Kontrolle (GE_Bright)

Tabelle 4.1: Microarrays die in der Qualitätskontrolle aufgefallen sind und weshalb.

Im Abschnitt der Diskussion der Qualitätskontrolle, werden einige der auffälligen Microarrays genauer erläutert. Der nächste Abschnitt beschäftigt sich mit den Ergebnissen der differentiellen Expression

4.3 Ergebnisse der differentiellen Expression

Für das bessere Verständnis der Ergebnisse werden die einzelnen Behandlungen im Kontext ihrer Publikation nochmals vorgestellt. Dies wird gemacht, um zu sehen welche Ergebnisse zu erwarten sind und worin sie sich von den Ergebnissen der verallgemeinerten Skripte (**3.3.2** und **3.3.3**) unterscheiden. Die Behandlungen (Chemikalien) stammen aus acht verschiedenen Publikationen (Tab. 4.2). Für die Tabelle (Tab. 4.2) sind alle Publikationen farblich markiert. Schwarze Publikationen haben nur eine Behandlung in ihrem Datensatz beigesteuert. Desweiteren ist schon mit aufgeführt, wie viele Gene in der differentiellen Expression zu sehen sind und wie oft sie sich mit den anderen Behandlungen überschneiden. Desweiteren sind für die vollständige Erfassung die Anzahl der Replikate einer Behandlung, die Konzentration und der Expositionszeitraum, unterteilt in Start der Behandlung und Dauer, angegeben. Die letzten beiden Spalten der Tabelle (Tab. 4.2) zeigen die Microarrayplattform (Agi=Agilent, Affy=Affymetrix) und die Chemikaliengruppe der sich die Behandlung zuordnen lässt.

In Hinblick auf ihre chemische Struktur gehören alle Chemikalien zu den organischen Verbindungen. Unter ihnen gibt es zwei besondere Fälle. Der eine ist Flusilazol, das zu den siliziumorganischen Verbindungen zählt. Der andere Fall ist Perchlorethen (PCE) die einzige Verbindung, die keine Ringstrukturen im Molekül aufweist. Durch diese grundlegenden strukturellen Unterschiede sind deutliche Unterschiede im Vergleich mit den anderen Chemikalien zu erwarten. Für einige der Chemikalien ist bekannt, dass sie eine östrogenartige Wirkung haben, weswegen sie in der Tabelle (Tab. 4.2) als Östrogen eingetragen sind. Für diese ist zu erwarten, dass sie in ihrer differentiellen Expression sehr ähnlich zueinander sind. Die Behandlungen, die in der Spalte Gruppe mit magenta markiert sind, enthalten Chlor oder Fluor. Es sind damit halogenhaltige Verbindungen. Es sollten sich auch, übergreifend für diese Behandlungen, gemeinsame/ähnliche Effekte zeigen.

Für die Chemikalien Propanil, Genistein und Linuron gibt es eine Lösemittelkontrolle (Solventcontrol) und Wasserkontrolle (Isocontrol). Dies wird gemacht, um differentiell exprimierte Gene zu sehen, die nur durch das Lösemittel exprimiert werden. D.h. für die differentielle Expression sollten sich für den Kontrast mit der Lösemittelkontrolle weniger exprimierte Gene zeigen.

Wie in **3.5** gezeigt, gruppieren sich die Kontrollen stark innerhalb ihrer Datensätze. Dies sollte sich dann später für den Vergleich zeigen, dass die Datensätze innerhalb der gleichen Gruppe ähnlicher zueinander sein werden. Einen vergleichbaren Einfluss können die verschiedenen Microarrayplattformen haben.

In der Publikation Büttner et al. [4] zeigten sich für Cyclopamin 57 differentielle exprimierte Gene und für GANT- 61 nur 20 differentielle exprimierte Gene. Die verein-

Chemikalie	Gene	Rep.	Konz.	Überl.	Start	Dauer	Array	Gruppe
Geni.Iso	693	4	2.4 mg/l	1251	0.5hpf	48h	Agi	Östrogen
Geni.Solv	393	4	2.4 mg/l	875	0.5hpf	48h	Agi	Östrogen
APM	14	4	1.84 mg/l	53	2hpf	24h	Agi	Triazin
TCDD	28	3	1 ng/μl	46	72hpf	12h	Affy	Dibenzodioxin
Retinsäure	6	3	0.3 mg/l	25	72hpf	12h	Affy	Vitamin-A/ Terpen
Cyclopamin	96	5	4.11 mg/l	141	0.5hpf	24h	Agi	Alkaloid
GANT- 61	17	5	3.92 mg/l	16	0.5hpf	24h	Agi	Benzolamin
PCP	149	3	50 μg/l	156	0.5hpf	8h	Affy	Phenol
Sert25	33	3	25 μg/l	109	72hpf	96h	Affy	Amin
Sert250	36	3	250 μg/l	93	72hpf	96h	Affy	Amin
Flux25	139	3	25 μg/l	297	72hpf	96h	Affy	Amin
Flux250	252	3	250 μg/l	407	72hpf	96h	Affy	Amin
Lin.Iso	4793	4	1.3 mg/l	6108	2hpf	48h	Agi	Harnstoffderivat
Lin.Solv	5553	4	1.3 mg/l	5592	2hpf	48h	Agi	Harnstoffderivat
Methylp.	1757	4	24.4mg/l	2229	0.5hpf	48h	Agi	Östrogen
BPA	194	4	8.5 mg/l	406	0.5hpf	48h	Agi	Östrogen
Prop.Iso	62	4	1.1 mg/l	158	0.5hpf	48h	Agi	Anilid
Prop.Solv	35	4	1.1 mg/l	104	0.5hpf	48h	Agi	Anilid
EE	70	4	0.8 mg/l	181	0.5hpf	48h	Agi	Östrogen
Flutamid	1	4	1.4 mg/l	8	0.5hpf	48h	Agi	Anilid/Antiandrogen
Prochloraz	88	4	2 mg/l	238	0.5hpf	48h	Agi	Imidazol/Amid
Flusilazol	2185	6	8.82 mg/l	1128	0.5hpg	24h	Agi	Silizium-Triazol
PCE	1225	2	32 mg/l	1248	24hpf	24h	Agi	Chlorethen

Tabelle 4.2: Übersicht über die Anzahl der als signifikant eingestuften differentiell exprimierten Gene, Anzahl der Replikate, Konzentration, Überlappung, Expositionszeitraum, Microarrayplattform und Art der Chemikalie der jeweiligen Behandlung.

heitlichte Methode zeigt ein ähnliches Bild mit 96 differentiell exprimierten Genen für Cyclopamin und 17 für GANT- 61. In der Publikation zeigte sich, dass zwischen den differentiell exprimierten Genen von GANT- 61 und Cyclopamin keine Überschneidung stattfand. Die Funktion der differentiell exprimierten Gene bei GANT- 61 sind noch nicht bekannt [4] und gehören zu keinem zusammenhängenden Stoffwechselsystem. Somit ist es auch später für den Vergleich nicht zu erwarten, dass große Überlappungen mit den anderen Behandlungen stattfinden.
Schiller et al. [5] fanden für Ethinylestradiol (EE) und Flutamid 33 bzw. 44 differentiell exprimierte Gene. Die anderen Behandlungen zeigten wesentlich mehr regulierte Gene:

- Linuron 1095
- Genistein 881
- Methylparaben 735
- Bisphenol A 569
- Propanil 260

- Prochloraz 120

Die Ergebnisse mit dem vereinheitlichten Skript sind sehr ähnlich (Tab. 4.2), besonders im Verhältnis zueinander. Allerdings wurden für Linuron und Methylparaben viel mehr regulierte Gene gefunden und für Flutamid nur 1 reguliertes Gen. Flutamid ist im Experimentdesign der Publikation eine Referenzchemikalie, die nur eine schwache Reaktion hervorrufen sollte [5]. Dies zeigt sich auch wieder für das vereinheitlichte Skript. Bei einer Gensetanalyse der Behandlungen zeigte sich für Schiller et al. [5], dass EE und Flutamid sich allein gruppierten. Genistein war die einzige Chemikalie, die eine Überlappung in den Stoffwechselsystemen mit allen Behandlungen hatte.
Die differentielle Expression zeigt in der Publikation von Park et al. 2012 [2] für die Behandlungen mit Sertralin und Fluoxetin:

- 288 Gene - Fluoxetin 25 μg/l
- 131 Gene - Fluoxetin 250 μg/l
- 33 Gene - Sertralin 25 μg/l
- 52 Gene - Sertralin 250 μg/l

Dies ist sehr ähnlich zu der Menge an differentiell exprimierten Genen (Tab. 4.2), die mit dem allgemeinen Affymetrixskript gefunden wurden. Allerdings zeigt es für die niedrige Konzentration Fluoxetin weniger regulierte Gene als bei Park et al. 2012 [2] und mehr regulierte Gene für die hohe Konzentration. Die Anzahl an differentiell exprimierter Gene ist für Sertalin 25 μg/l sogar identisch. Das markanteste Gen in den Ergebnissen von Park et al. 2012 [2] ist *fkpb5*. Es wurden auch Überlappungen der differentiell exprimierten Gene zwischen den Behandlungen untersucht [2]:

- Fluoxetin 25 μg/l mit Fluoxetin 250 μg/l - 81 Gene
- Fluoxetin 250 μg/l mit Sertralin 250 μg/l - 9 Gene
- Sertralin 25 μg/l mit Sertralin 250 μg/l - 9 Gene
- Sertralin 25 μg/l mit Fluoxetin 25 μg/l - 23 Gene

Hermsen et al. 2012 [10] untersuchten Änderungen in der Genexpression zwischen verschiedenen Konzentrationen von Flusilazol. So sind diese Ergbenisse mit denen des allgemeinen Agilentskriptes schwer vergleichbar, da ein völlig anderer Ansatz verfolgt wurde [10]. So zeigt das allgemeine Agilentskript 2185 differentiell exprimierte Gene. Wie bereits zuvor erwähnt ist Flusilazol eine siliziumorganische Verbindung, wodurch es sich grundlegende von den anderen Chemikalien in seiner chemischen Struktur unterscheidet. Neben dem zentralen Siliziumatom besitzt Flusilazol drei Ringe, von denen zwei Phenylringe mit Fluor sind. So ist die restliche Struktur wiederum sehr ähnlich zu den anderen Chemikalien. Bei der Untersuchung von Hermsen et al. 2012 [10] trat das Gen *cyp26a1* hervor. Es war besonders stark in seiner Expression reguliert.
Die Publikation von Klüver et al. 2011 [11] untersuchte drei Chemikalien. Unter diesen war Azinphos-methyl (APM), das ein AChE-Inhibitor ist [11]. Für APM zeigten sich Gene

differentiell exprimiert, die in erster Linie mit der Immun- (*socs3a*) und der Stressantwort (*hspb11*) in Verbindung stehen. Es wird erwartet, dass sich für die anderen Behandlungen auch Gene überlappen, die diesen beiden System zugeordnet werden können.
Die Publikation von Chen et al. 2008 [9] hatte das Ziel, die Wirkung von Retinsäure und TCDD auf die Entwicklung des Herzen zu untersuchen. Für maximale Wirkung beider Chemikalien wurde der Zeitpunkt von 72 hpf als Expositionsbeginn gewählt, da in diesem Zeitraum die Herzentwicklung beginnt [9]. So zeigte sich nach 12 h 278 regulierte Gene. Im Gegensatz dazu brachte das allgemeine Affymetrixskript nur sechs differentiell exprimierte Gene hervor (Tab. 4.2). Unter diesen befindet sich *cyp26a1*, das auch bei Chen et al. 2008 [9] differentiell exprimiert war. Die zweite Chemikalie TCDD hatte keine Angaben über eine genaue Anzahl an differentiell exprimierten Genen [9]. Das allgemeine Affymetrixskript zeigt hier 28 differentiell regulierte Gene. Allerdings hat man eine große Überlappung zwischen den regulierten Genen nach einer Expositionsdauer von 12 h gefunden.
Pentachlorphenol war die von Xu et al. 2014 [3] für ihre Publikation verwendete Chemikalie. Für die Konzentration von 50 μg/l wurden 826 differentiell exprimierte Gene gefunden. Im Gegensatz dazu brachte das allgemeine Affymetrixskript nur 149 Gene hervor. Eine Gensetanalyse zeigte, dass die meisten stark regulierten Gene in der Glykolyse involviert sind [3]. Einige ausgewählte Gene wurden genauer betrachtet [3], die der Glykolyse und dem Wachstum zugeordnet sind. Dies sind potentielle Gene die im Vergleich der Behandlung hervortreten können.
Der Datensatz und die zugehörige Publikation aus denen PCE stammt, ist noch nicht veröffentlicht. Die Ergebnistabelle wurde für diese Arbeit zu Verfügung gestellt. Sie zeigt 1225 differentiell exprimierte Gene (Tab. 4.2). PCE ist die einzige organische Verbindung ohne Ringstrukturen im Molekül, dies könnte sie von den anderen Behandlungen differenzieren. Allerdings besitzt sie vier Chloratome, für die Überlappungen mit den anderen halogenhaltigen Behandlungen zu erwarten sind.
Sämtliche Ergebnistabellen wurden mit den allgemeinen Skripts für Agilent und Affymetrix erstellt. Die False Discovery Rate (FDR) ist auf 0,05 gesetzt und sämtliche Ergebnistabellen in einer eigenen Textdatei gespeichert [Anhang/Datensätze/ ErgebnistabellenXXXXX].
Die Ergebnistabellen wurden nochmals zur besseren Übersicht in einer vergleichenden Abbildung (Abb. 4.1, Tab. 4.2) zusammengefasst. Da die Anzahl der in der differentiellen Expression als signifkant eingestuften Gene von unter 30 Genen bis 5553 reicht, wurden diese Zahlen logarithmiert (Abb. 4.1). Dunklere Balken stehen somit für niedrige Werte und helle Balken für eine große Anzahl an signifikanten Genen. Die meisten differentiell exprimierten Gene hat dabei Linuron aus dem Datensatz Schiller et al. 2013 [5] mit 5553. Die wenigsten differentiell exprimierten Gene hat Flutamid mit genau einem Gen: *fos*.
Für den Aussetzungszeitraum sieht man (Abb. 4.1), dass Behandlungen die zu einem späteren Zeitpunkt (72 hpf Start 4.1) stattfanden, im Vergleich zu den früheren Zeitpunkten weniger Effekte zeigten. Hier sieht man aber den Einfluss des Expositionszeitraum, der bei längerer Dauer mehr differentiell exprimierte Gene hervorbrachte. Bei Agilentmi-

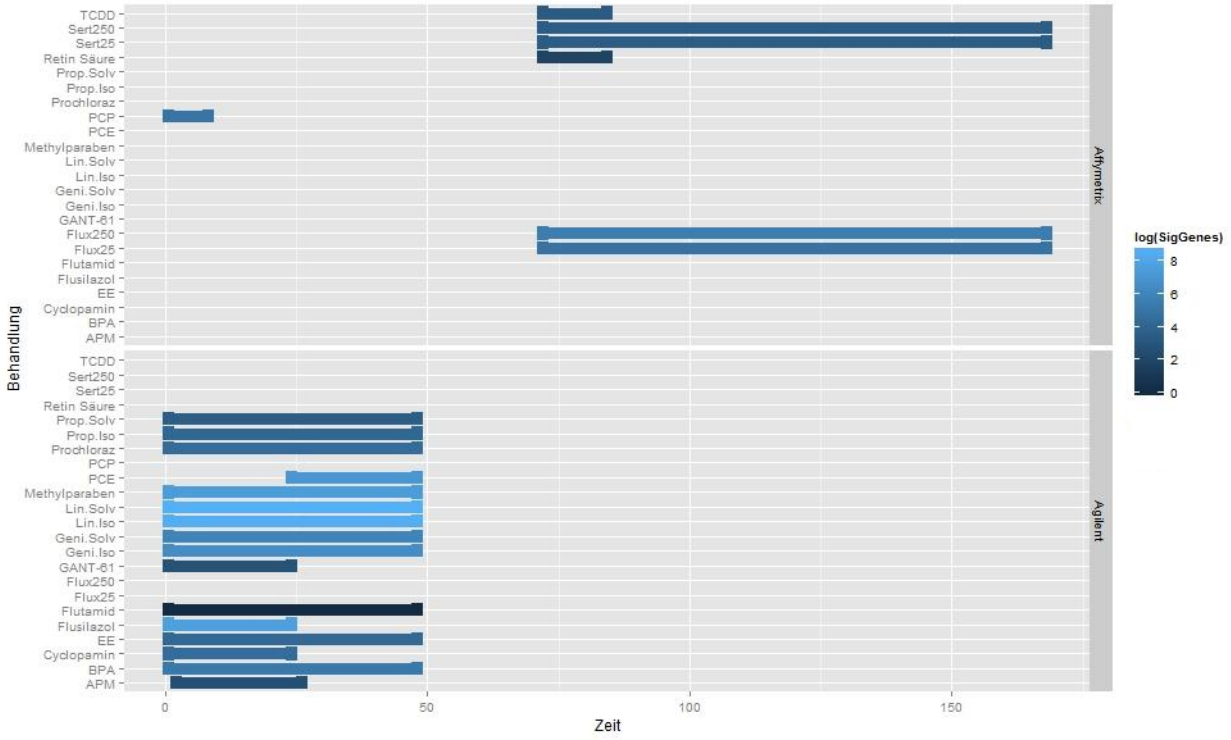


Abbildung 4.1: Verhältnis der logarithmierten Anzahl der signifikanten Gene zueinander. Zeit in 'hpf'

croarrays fanden sich deutlich mehr differentiell exprimierte Gene, aber die Anzahl der gefundenen Gene korrespondiert mit den Ergebnissen der Publikationen, wie bereits zuvor beschrieben. Ein direkter Einfluss den die Anzahl der Replikate (Tab. 4.2) einer Behandlungen hätten, ist nicht ersichtlich.
Ein übergreifendes Muster für die Ergebnisse der differentielle Expression mit den beiden allgemeinen Skripten für Agilent und Affymetrix ist nicht ersichtlich. Die Ergebnisse entsprechen weitesgehend, denen der Publikationen. Auch die restlichen Faktoren wie die verschiedenen Kontrollgruppen, Microarrayplattform und Expositionszeiträume zeigen die zu erwartenden Ergebnisse.
Als nächstes werden die Ergebnisse des Vergleiches zwischen den Behandlungen vorgestellt.

4.4 Ergebnis des Vergleichs der differentiellen Expression

Aus dem Vergleich der Ergebnistabellen ging folgende Übersicht (Abb. 4.3 und 4.4) hervor. Sie zeigt wie viele Gene zwischen den Behandlungen mit den Chemikalien sich überlappen. Aufgrund ihrer Größe wurde sie in zwei Teile gegliedert und an das Ende des Kapitels Ergebnisse gesetzt.
Die Zahlen in der Tabelle 4.3 und 4.4 geben die sich überschneidenden differentiell exprimierten Gene zweier verschiedener Chemikalien an. Die Prozentwerte geben die Höhe der jeweiligen anteiligen Überlappung an. Aus der Zählung dieser Überlappungen ging hervor, dass insgesamt 4548 differentiell exprimierte Gene, für alle 23 Behandlungen, sich überlappen (Abb. 4.5).
Methylparaben und die beiden Genisteinvarianten aus dem Datensatz Schiller et al. 2013 [5] überlappen sich mit allen Chemikalien. Alle anderen Chemikalien überlappen sich mindestens mit einer Chemikalie nicht, d.h. es gibt für alle Überschneidungen an differentiell exprimierten Genen (Abb. 4.3 und 4.4). Dieses Ergebnis geht mit den Beobachtungen der Publikation konform, die bereits bei der differentiellen Expression beschrieben wurden. Gegensätzlich dazu ist Methylparaben, das bei Schiller et al. 2013 [5] die wenigsten Überlappungen mit den anderen Behandlungen der Östrogengruppe zeigte. Bedingt durch nur ein einziges als differentiell eingestuftes Gen bei Flutamid, hat es immer eine 100% Übereinstimmung mit den Chemikalien, bei denen dieses Gen auch vorhanden ist (Abb. 4.3 und 4.4). Flutamid überlappt sich insgesamt mit sieben anderen Behandlungen. Bisphenol A, Ethinylestradiol, Prochloraz und Propanil zeigen 14 (Propanil) bis 21 (BPA) Überlappungen mit den anderen Chemikalien (Abb. 4.3 und 4.4). Da sie alle strukturähnlich sind und östrogenartige Wirkungsweisen besitzen [5], spiegelt dies der Vergleich wieder. Linuron hat zwar in der differentiellen Expression die meisten Gene, aber es überschneidet sich nicht mit Flutamid, das die wenigsten hat. Durch die große Anzahl von regulierten Genen, war es sehr wahrscheinlich, dass es Überlappungen mit allen anderen Behandlungen zeigt. Zwischen Genistein, Linuron und Propanil, für die es eine Kontrolle mit und ohne Lösemittel gab, bestehen hohe Überlappungen zwischen den Kontrollen, aber keine die mehr als 75% prozentualen Anteil hätten. Somit hat das Lösemittel einen deutlichen EInfluss auf die differentielle Expression.
Cyclopamin und GANT- 61 überschneiden sich 11 und 7 mal. Auch der Anteil der Gene bei einer Überlappung ist nur sehr gering. Es war bereits bekannt, dass die Überlappung für GANT- 61 so gering ausfallen würde. Wie bei Büttner et al. 2012 [4] ist keine Überlappung zwischen beiden Chemikalien vorhanden. Allerdings zeigt auch Cyclopamin wenige Überlappungen mit den restlichen Behandlungen (Abb. 4.3 und 4.4)
Die Behandlungen des Datensatzes Park et al. 2012 [2] zeigen viele Überlappungen mit den Behandlungen des Datensatzes Schiller et al. 2013 [5]. Mit den anderen Datensätzen hingegen fast keine. Ausnahmen sind hierbei Pentachlorphenol, PCE und Flusilazol. Der Anteil der überlappenden differentiell exprimierten Gene ist aber nur sehr gering.

In Sertralin 250 μg/l sind fünf der sechs signifikant regulierten Gene der Retinsäure enthalten. Für die Überlappungen des Datensatzes untereinander zeigte sich:

- Fluoxetin 25 μg/l mit Fluoxetin 250 μg/l - 69 Gene
- Fluoxetin 250 μg/l mit Sertralin 250 μg/l - 12 Gene
- Sertralin 25 μg/l mit Sertralin 250 μg/l - 11 Gene
- Sertralin 25 μg/l mit Fluoxetin 25 μg/l - 16 Gene

Diese Überlappungen sind sehr ähnlich, fast identisch, zu denen, die bereits von Park et al. 2012 [2] beobachtet wurden. Für die höheren Konzentrationen waren mehr Überlappungen mit den anderen Chemikalien zu finden. Fluoxetin zeigte im Vergleich mit Sertralin mehr Überlappungen (Abb. 4.3 und 4.4).
Retinsäure zeigt eine starke prozentuale Überlappung mit Sertralin und Linuron (Solvent), aber es hat nur sechs differentiell exprimierte Gene. Insgesamt überschneidet es sich nur mit 11 Behandlungen. TCDD teilt sich mit 16 Behandlungen differentiell exprimierte Gene, allerdings immer nur zu einem geringen Anteil (Abb. 4.3 und 4.4).
APM überlappt sich mit 14 anderen Behandlungen und hat damit verhältnismäßig wenige Überschneidungen. Auch der prozentuale Anteil ist nur gering. Diese wenigen Gene könnten aber die Gene sein, die auf die übergreifende Immun- und Stressantwort hindeuten (siehe 4.3). Das sich bei APM nur mit 14 Behandlungen überschneidet, deutet aber nicht unbedingt daraufhin (Abb. 4.3 und 4.4).
Flusilazol überschneidet sich mit 19 der Behandlungen. Die Zahl der überlappenden Gene ist relativ hoch, was eine direkte Folge der großen Menge an Genen ist die für Flusilazol differentiell exprimiert ist. PCE die andere Chemikalie, die neben Flusilazol so eine Sonderstellung hat, besitzt 21 Überlappungen mit anderen Behandlungen. Dabei ist das Überlappungsprofil von PCE und Flusilazol sehr ähnlich, was die Menge an überlappenden Genen mit anderen Chemikalien betrifft. Allerdings teilen sie sich nur 96 Gene (Abb. 4.3 und 4.4), bei über tausend differentiell exprimierten Gene für beide Chemikalien.
Die letzte Chemikalie Pentachlorphenol (PCP) zeigt 14 Überlappungen, die meisten mit dem Datensatz Schiller et al. 2012 [5]. Für TCDD mit dem PCP eine starke Wechselwirkung hat [3] gibt es keine Überlappung für differentiell exprmierte Gene.
Um die Summe der Überlappungen zu bewerten müssen für alle Behandlungskombinationen die überlappenden Gene gezählt werden. Die Ergebniss dieser Zählung werden als nächste vorgestellt.

4.5 Die häufigsten differentiell exprimierten Gene

Die Zählung der am häufigsten auftretenden Gene (Abb. 4.5 am Ende des Kapitels Ergebnisse) brachte hervor, dass *fkbp5* 55 mal im Vergleich auftaucht. Es gehört zum Östrogensignalweg und ist auch bei den östrogenartigen Chemikalien, sowie Sertralin und Fluoxetin zu finden.

Gen	**Vollständiger Name**	**Behandlungen**
fkbp5	FK506 binding protein 5	11
nfil3-6	nuclear factor, interleukin 3 regulated, member 6	10
pfkfb4l	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4, like	9
fos	v-fos FBJ murine osteosarcoma viral oncogene homolog	9
klf11a	Kruppel-like factor 11a	8
pik3r3a	phosphoinositide-3-kinase, regulatory subunit 3a	8
cyp24a1	cytochrome P450, family 24, subfamily A, polypeptide 1	8
socs3a	suppressor of cytokine signaling 3a	8

Tabelle 4.3: Die acht häufigsten Gene des Vergleichs mit der Anzahl an Behandlungen in denen sie gefunden wurden.

Das zweithäufigste Gen ist *nfil3-6* und tritt 45 mal auf. Es ist für die Regulierung von Interleukin verantwortlich.
Für das dritthäufigste Auftreten gibt es zwei Gene: *fos* und *pfkfb4l*. Beide treten 36 mal auf und sind für das Binden von heterocyklischen organischen Verbindungen mitverantwortlich. *fos* ist zusätzlich in der RNA-Regulation involviert.
Die meisten dieser Gene findet man bei Linorun für die differentiell Expression mit der Lösemittelkontrolle als Kontrast (Lin.Solv Abb. 4.5). Für das Auftreten der Gene zeigen sich zwei große Gruppen:
Die erste Gruppe hat großen Anteil an den 22 häufigsten Genen mit einem Anteil von zehn oder mehr. Die zweite große Gruppe hat fast keinen Anteil (<3) an dieser Gruppe von Genen. Flutamid zählt nicht zur zweiten Gruppe, da sich in der differentiellen Expression nur ein Gen als signifikant zeigte. Dieses war aber *fos*, welches mit am dritthäufigsten Auftritt.
Die zu Schiller et al. 2013 gehörigen Behandlungen (Abb. 4.2 grün) haben den größten Anteil an den häufigsten Genen (Abb. 4.5). Da dieser Datensatz die meisten Chemikalien stellt, musste ein gewisser Ergebnisschwerpunkt für die häufigsten Gene bei ihm liegen. Propanil zeigt aber in beiden Varianten nur geringen Anteil an den häufigsten Genen. Ethinylestradiol und Flutamid zeigen wieder das zu erwartende, welches aus der Publikation ([5]) bekannt ist. Methylparaben hingegen, hat gegenteilig zu den Ergebnissen der Publikation großen Anteil (14 von 21) an den häufigsten Genen (Abb. 4.5). Besonders deutlich wird der Anteil an Genen, die im Östrogenstoffwechsel aktiv sind (Abb. 4.5 & Abb. 5.1).
Flusilazol und PCE haben ebenfalls einen großen Anteil an den häufigsten Genen. Besonders in der Gruppe der Gene, die 21 mal Auftreten, besitzen beide die gleichen

differentiell exprimierten Gene. Für PCE findet sich auch das zweithäufigste Gen *nfil3-6*. *cyp26a1* findet sich für Flusilazol und PCE unter den häufigsten Genen wieder.
Die Behandlungen des Datensatzes Park et al. 2012 (Abb. 4.2 gelb), Sertralin und Fluoxetin fallen relativ unterschiedlich aus. Bereits im Vergleich (Abb. 4.3 und 4.4) hatte Fluoxetin mehr Überlappungen als Sertralin. Für die häufigsten Gene ist dies noch deutlicher (Abb. 4.5). Beide haben Anteil an den Überlappungen von *fkbp5*, dass schon in der Publikation ([2]) differentiell expremiert war.
Azinphos-methyl (APM) hat nur zwei der häufigsten Gene: *socs3a* und *socs3b*. Beide sind mit der Immunantwort verbunden und sind ein Hinweis auf eine Gegenreaktion. Sie decken sich mit dem Ergebnis der Publikation.
Die beiden Chemikalien TCDD und Retinsäure (Chen et al. 2008 [9]) haben nur einen geringen Anteil, aber es treten die beiden Gene *socs3a* und *cyp26a1* auf, die bereits in der Publikation prägnant waren.
Cyclopamin und GANT- 61 haben keinerlei Überlappungen in der Menge der häufigsten Gene. Eine individuelle Zählung für die Überlappungen dieser beiden Chemikalien zeigte, dass sie für ihre differentiell exprimierten Gene immer nur einmal zu finden sind in den anderen Behandlungen. Der Status von GANT- 61 als 'Außenseiter' bestätigt sich ein weiteres mal und passt in die bisherigen Ergebnisse.
PCP hat nur mit einem Gen Anteil an den Überlappungen. Dieses Gen *pfkfb4l* zählt zu der Gruppe von Genen die von Xu et al. [3] genauer betrachtet wurde. So deckt sich dieses Ergebnis mit den zu erwartenden Ergebnissen.

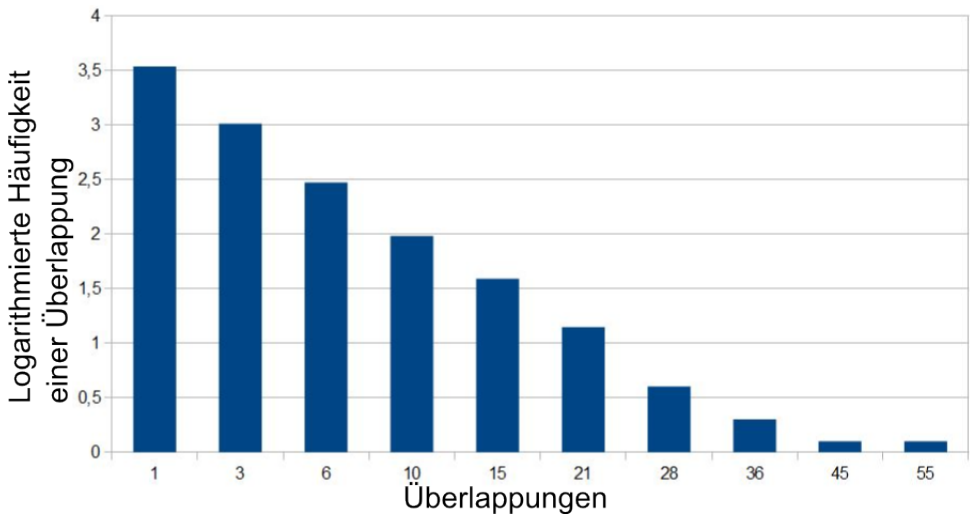


Abbildung 4.2: Verteilung der Überlappungen für ein Gen in Bezug auf die Häufigkeit ihres Auftretens

Die Verteilung in Bezug auf die Häufigkeit des Auftretens einer Überlappung zeigt, dass die meisten nur bei einer Überschneidung liegen und danach die Häufigkeit des Auftretens sinkt. Für eine bessere Darstellung wurden die Werte logarithmiert (Abb. 4.2).

Behandlung	Geni.Iso	Geni.Solv	APM	TCDD	RA	Cyclopamin	GANT.61	PCP	Sert25	Sert250	Flux25	Flux250
Geni.Iso	x	314	5	3	2	11	1	9	6	8	13	23
Geni.Solv	45.0% / 79.8%	x	5	2	1	7	1	6	5	6	8	17
APM	0.7% / 35.7%	1.2% / 35.7%	x	1	0	2	0	0	0	3	1	0
TCDD	0.4% / 10.7%	0.5% / 7.1%	7.1% / 3.5%	x	1	0	1	0	0	0	1	6
RA	0.2% / 33.3%	0.2% / 16.6%	-	3.5% / 16.6%	x	0	0	0	0	5	0	1
Cyclopamin	1.5% / 11.5%	1.7% / 7.3%	14.2% / 2.1%	-	-	x	0	0	0	3	0	0
GANT-61	0.1% / 5.8%	0.2% / 5.8%	-	3.5% / 5.8%	-	-	x	0	0	0	0	0
PCP	1.4% / 6.0%	1.5% / 4.0%	-	-	-	-	-	x	1	0	7	7
Sert25	0.8% / 18.1%	1.2% /15.1%	-	-	-	-	-	0.7% / 3.0%	x	11	16	11
Sert250	1.1% / 22.2%	1.5% / 16.6%	21.4% / 8.3%	-	83.3% / 13.8%	3.1% / 8.3%	-	-	33.3% / 30.5%	x	3	12
Flux25	1.8% / 9.3%	2.0% / 5.7%	7.1% / 0.7%	3.5% / 0.7%	-	-	-	4.7% / 5.0%	48.4% / 11.5%	8.3% / 2.1%	x	69
Flux250	3.3% / 9.1%	4.3% / 6.7%	-	21.4% / 2.3%	16.6% / 0.3%	-	-	4.7% / 2.7%	33.3% / 4.3%	33.3% / 4.7%	49.6% / 27.3%	x
Lin.Solv	40.6% / 4.6%	40.2% / 2.5%	64.2% / 0.1%	39.2% / 0.1%	83.3% / 0.1%	46.3% / 0.08%	17.6% / 0.7%	32.9% / 0.05%	36.3% / 0.8%	38.9% / 0.1%	38.1% / 0.2%	34.9% / 1.4%
Lin.Iso	34.1% / 4.4%	33.6% / 2.4%	35.7% / 0.09%	17.8% / 0.09%	66.6% / 0.07%	36.8% / 0.6%	17.6% / 0.05%	27.5% / 0.7%	36.3% / 0.2%	22.2% / 0.1%	29.4% / 0.7%	20.2% / 0.9%
NP	17.8% / 7.0%	19.3% / 4.3%	64.2% / 0.5%	10.7% / 0.1%	50.0% / 0.1%	12.6% / 0.6%	29.4% / 0.2%	5.3% / 0.4%	24.2% / 0.4%	19.4% / 0.3%	17.9% / 1.4%	13.1% / 1.8%
Prop.Iso	0.8% / 14.5%	1.0% / 6.4%	7.1% / 1.6%	3.5% / 1.6%	-	-	-	0.7% / 1.6%	3.0% / 1.6%	-	1.4% / 3.2%	1.9% / 8.0%
Prop.Solv	0.7% / 14.2%	0.2% / 2.8%	-	-	-	-	-	-	3.0% / 2.8%	-	1.4% / 5.7%	1.2% / 8.5%
BPA	3.0% / 10.8%	3.6% / 7.2%	7.1% / 0.5%	14.2% / 2.0%	16.6% / 0.5%	4.2% / 2.0%	11.7% / 1.0%	2.0% / 1.5%	12.1% / 2.0%	5.5% / 1.0%	6.4% / 4.6%	4.4% / 5.6%
EE	2.3% / 22.8%	2.2% / 12.8%	7.1% / 1.4%	3.5% / 1.4%	-	1.1% / 1.4%	-	0.7% / 1.4%	12.1% / 5.7%	8.3% / 4.2%	4.3% / 8.5%	3.6% / 12.8%
Flutamid	0.1% / 100%	0.2% / 100%	-	-	-	-	-	-	3.0% / 100%	-	-	0.3% / 100%
Prochloraz	2.7% / 21.5%	2.7% / 12.5%	7.1% / 1.1%	7.1% / 2.2%	-	-	-	1.3% / 2.2%	21.1% / 7.9%	5.5% / 2.2%	10.8% / 17.0%	4.7% / 13.6%
Flusilazol	7.7% / 1.7%	8.6% / 1.1%	-	10.7% / 0.1%	33.3% / 0.1%	9.4% / 0.2%	-	8.7% / 0.4%	6.0% / 0.06%	2.7% / 0.03%	7.9% / 0.3%	7.9% / 0.6%
PCE	14.4% / 8.1%	17.8% / 5.7%	14.2% / 0.1%	14.2% / 0,3%	50.0% / 0.2%	5.2% / 0.4%	-	4.7% / 0.5%	6.0% / 0.1%	11.1% / 0.3%	5.7% / 0.6%	5.1% / 1.0%

Abbildung 4.3: Übersichtstabelle der Ergebnistabellenvergleiche Teil 1

Behandlung	Lin.Solv	Lin.Iso	MP	Prop.Iso	Prop.Solv	BPA	EE	Flutamid	Prochloraz	Flusilazol	PCE	Sig.Gene
Geni.Iso	282	237	124	9	5	21	16	1	19	49	93	693
Geni.Solv	158	132	76	4	1	14	9	1	11	31	66	393
APM	9	5	9	1	0	1	1	0	1	4	5	14
TCDD	11	5	3	1	0	4	1	0	2	3	4	28
RA	5	4	3	0	0	1	0	0	0	3	3	6
Cyclopamin	44	35	12	0	0	4	1	0	0	13	9	95
GANT-61	3	3	5	0	0	2	0	0	0	0	0	17
PCP	49	41	8	1	0	3	1	0	2	13	8	149
Sert25	12	12	8	1	1	4	4	1	7	4	5	33
Sert250	14	8	7	0	0	2	3	0	2	1	5	36
Flux25	53	41	25	2	2	9	6	0	15	14	12	139
Flux250	88	51	33	5	3	11	9	1	12	27	21	252
Lin.Solv	x	3572	822	23	14	89	31	0	35	411	383	5553
Lin.Iso	58.3% / 67.1%	x	685	19	9	73	24	0	29	302	305	4793
MP	13.4% / 46.7%	12.8% / 38.9%	x	25	19	77	28	1	33	138	158	1757
Prop.Iso	0.37% / 37.0%	0.35% / 30.6%	1.4% / 40.3%	x	25	8	4	0	5	13	12	194
Prop.Solv	0.22% / 40%	0.1% / 25.7%	1.0% / 54.2%	40.3% / 71.4%	x	5	2	0	2	7	9	62
BPA	1.4% / 45.8%	0.16% / 37.6%	4.3% / 39.6%	12.9% / 4.1%	14.2% / 2.5%	x	15	0	16	16	31	35
EE	0.5% / 44.2%	1.3% / 34.2%	1.5% / 40%	6.4% / 5.7%	5.7% / 2.8%	7.7% / 21.4%	x	1	12	6	7	70
Flutamid	-	-	0 / 100%	-	-	-	1.4% / 100%	x	1	0	1	1
Prochloraz	0.5% / 39.7%	0.5% / 32.9%	1.8% / 37.5%	8.0% / 5.6%	5.7% / 2.2%	8.2% / 18.1%	17.1% / 13.6%	100.0% / 1.1%	x	16	18	88
Flusilazol	6.2% / 12.3%	5.3% / 9.1%	7.3% / 4.1%	20.9% / 0.4%	20.0% / 0.2%	10.8% / 0.6%	11.4% / 0.2%	-	22.7% / 0.6%	x	93	2185
PCE	5.6% / 28.4%	5.3% / 23.2%	9.1% / 13.1%	17.7% / 0.8%	25.7% / 0.7%	15.9% / 2.5%	11.4% / 0.6%	100.0% / 0.08%	22.7% / 1.6%	3.1% / 7.8%	x	1225

Abbildung 4.4: Übersichtstabelle der Ergebnistabellenvergleiche Teil 2

Chemiekalie	Anteil	fkbp5	nfil3.6	pfkfb4l	fos	klf11a	pik3r3a	cyp24a1	socs3a	ankrd9	sesn2	pcp4l	meis2b	cyp26a1	per2	agt	rcv1	higd1a	ndrg1b	slc3a2b	ctsla	hsd11b2	foxg1b
Geni.Iso	14	x	x	x	x	x	-	x	x	-	-	x	x	-	-	x	x	x	x	-	-	-	x
Genis.Solv	12	-	x	x	x	x	-	x	x	-	-	x	x	-	-	-	x	x	x	-	-	-	x
Lin.Iso	15	x	x	-	-	x	-	-	-	x	x	x	x	-	-	-	x	x	x	x	x	x	x
Lin.Solv	18	x	x	-	-	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	-
MP	13	x	x	-	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	-	-	-	x
BPA	7	x	-	x	-	x	x	-	-	-	-	-	-	x	-	x	x	-	-	-	-	-	-
EE	8	x	x	x	x	x	x	x	-	-	-	-	-	-	x	x	-	-	-	-	-	-	-
Flutamid	1	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Prochloraz	12	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	-	-	-	-	-	x	-
Prop.Iso	3	-	-	-	-	-	-	-	-	-	x	-	-	x	-	-	-	-	-	-	x	-	-
Prop.Solv	2	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-	-
APM	1	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TCDD	2	-	-	-	-	-	-	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
RA	2	-	-	-	-	-	-	-	-	-	-	x	x	-	-	-	-	-	-	-	-	-	-
PCP	1	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sert25	7	x	-	x	x	-	x	-	-	-	-	-	-	-	-	-	-	-	x	x	-	x	x
Sert250	2	x	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Flux25	10	x	x	x	-	-	x	-	-	x	-	-	-	x	-	-	-	x	x	x	-	-	x
Flux250	12	x	x	x	x	-	x	x	-	-	-	-	-	x	-	-	-	x	x	x	x	-	-
Flusilazol	13	-	-	-	-	x	-	-	-	-	x	-	x	x	x	x	x	x	-	x	x	x	-
PCE	14	-	x	-	x	-	x	x	x	x	x	x	x	-	-	-	x	-	-	x	x	x	x
Cyclopamin	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GANT-61	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ueberlapp.	-	55	45	36	36	28	28	28	26	21	21	21	21	21	21	21	21	21	21	21	21	21	21

Abbildung 4.5: Die am häufigsten auftretenden Gene aus dem Vergleich der differentiellen Expression. Kreuze kennzeichnen Behandlungen bei denen dieses Gen auftrat. Anteil gibt an wie viele dieser Gene in der jeweiligen Behandlung vorhanden waren.

5 Diskussion

5.1 Diskussion

In der Diskussion soll die Frage beantwortet werden, ob es eine Chemikalien übergreifende Antwort- und Gegenreaktion gibt. Dazu werden zuerst nochmal einige der auffälligen Microarrays der Qualitätskontrolle vorgestellt und genauer auf die einzelnen Kontrollen eingegangen.
Als nächstes wird anhand der Ergebnisse von differentiellen Expression, Vergleich und Genzählung die Frage beantwortet. Der letzte Abschnitt befasst sich mit einem Ausblick auf weitere Analysen der Daten, Herausforderungen bei der Analyse und Verbesserungsmöglichkeiten des experimentellen Aufbaus.

5.2 Auffällige cDNA- Microarrays der Qualitätkontrolle

Da die Kontrolle der Qualität der Microarraydaten für die Ergebnisse von entscheidender Bedeutung ist, werden hier zwei Beispiele für auffällige cDNA- Microarrays vorgestellt. Die Beispiele unterteilen sich dabei in Agilent und Affymetrix, sodass für beide Plattformen die Bewertung gezeigt wird.
Der erste Microarray stammt aus dem Datensatz Schiller et al. 2013 [5]. Es ist ein Microarray für Propanil, das Replikat Nummer 4 (Abb. 5.1):

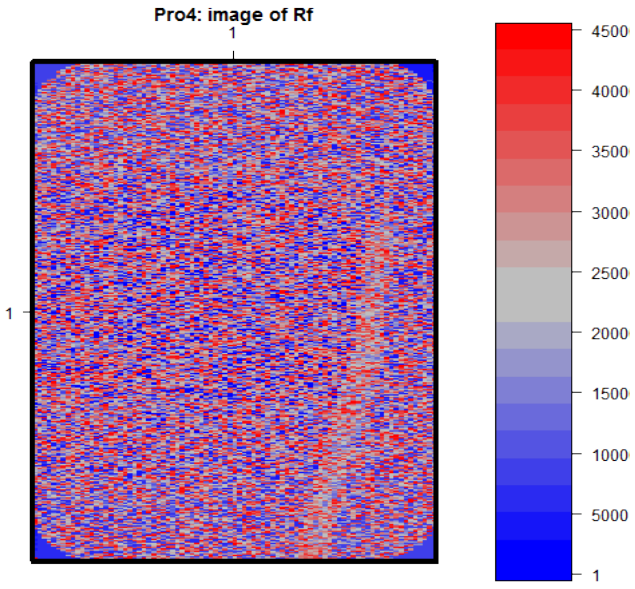


Abbildung 5.1: Bild des Microarrays Propanil Replikat 4 aus Schiller et al. 2013 [5]

Der Microarray zeigt eine längere, hellere Spur auf der rechten Seite - ein Artefakt. Jetzt muss bestimmt werden, ob und wie stark dieses Artefakt sich auf die Gesamtqualität des Arrays auswirkt. Aufgrund der Größe der restlichen Abbildungen für diese Methoden, befinden sie sich im Anhang [Anhang/Qualitätskontrollen/ SchillerQualitätskontrolle.pdf].
Die Boxplots für die dunkle und helle Referenz zeigen für Propanil Replikat 4 keine Auffälligkeiten. Die 'Heatmap' für den Datensatz zeigt für Propanil Replikat 4 keine Besonderheiten, es befindet sich sogar in einer Gruppe mit den restlichen Propanilreplikaten. Auch die andere Methode des unüberwachten Lernens (MDS-Plot) zeigt, dass sich dieser Microarray in einer Gruppe (Cluster) mit den anderen befindet. Somit weist auch diese Methode in zwei Varianten daraufhin, dass der Microarray nicht weiter dadurch beeinflusst wird.
Da dieser Mircoarray in allen anderen Kontrollen so unauffällig war, wurde er auch weiterhin in der Analyse behalten.
Der zweite Microarray stammt aus dem Datensatz Park et al. 2012 [2]. Es ist das zweite Replikat für die niedrige Konzentration von Fluoxetin (interne Bezeichnung `Flux25_2`).

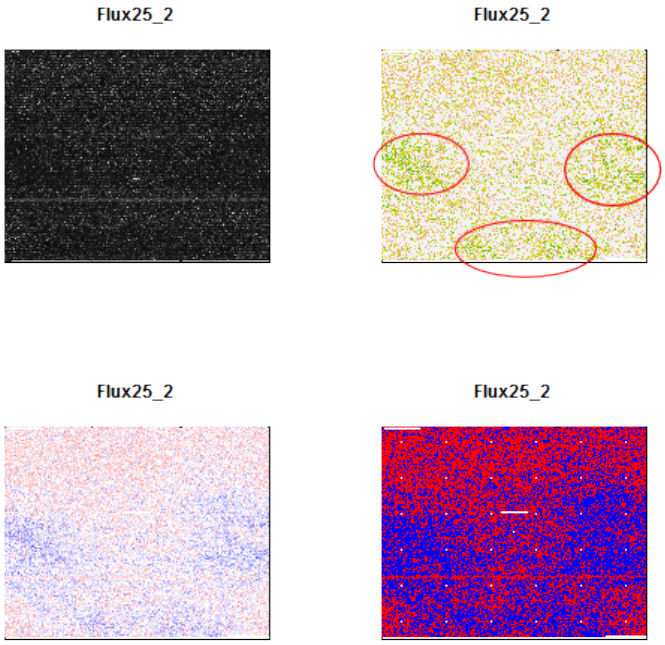


Abbildung 5.2: Bild des Microarrays Fluoxetin Replikat 2 niedrige Dosis $25\mu g/l$ aus Park et al. 2012 [2]

Das Bild des Microarrays (Abb. 5.2) zeigt unten in der Mitte, am rechten Rand und am linken Rand auffällige Punkte. Analog zu dem ersten Microarray sind die restlichen Abbildungen für die Kontrolle im Anhang einzusehen [Anhang/ Qualitätskontrolle/ ParkQualitätskontrolle.pdf]. Bei der internen Affymetrixkontrolle `QC Stats` fällt dieser

Microarray nicht weiter auf. Er bewegt sich im gleichen Rahmen wie die anderen Microarrays. Das einige der Kontrollen mit rot markiert sind hängt damit zusammen, dass diese Kontrolle auf dem HGU95A- Array basiert, da die Microarrays in einem Bereich liegen, deuten sie für diese Microarrayplattform nicht auf einen Fehler.
Die 'Heatmap' zeigt auch keine weitere Auffälligkeit für diesen Microarray, dass es starke Unterschiede zwischen den Microarrays gibt. Allerdings zeigen sich zwei große Gruppen. Ihnen gehören die Replikate für eine Behandlung und der Kontrolle sehr gemischt an. Dies deutet darauf hin, dass die Qualität aller Microarrays für diesen Datensatz sehr ähnlich ist und keine großen Abweichungen vorhanden sind. Durch einen stark Abweichenden Microarray könnten beide Gruppen zu einer zusammengefasst werden.
Für den N.U.S.E- Plot ist das Replikat höher zentriert als die restlichen Replikate, allerdings niederiger als 1.05. Wie bereits im Abschnitt Methode für die Qualitätskontrolle von Affymetrixdaten erklärt, muss der Wert aber deutlich über 1.1 liegen um auf schlechte Qualität hinzuweisen. Auch bei dem R.L.E- Plot liegt das Replikat 2 leicht höher und hat eine größere Varianz. Für den Boxplot fällt `Flux25_2` gar nicht auf und geht uniform mit den restlichen Microarrays.
Die Abbildungen für die logarithmierten Intensitäten sowie den RNA- Verdau zeigen keinerlei Auffälligkeiten für eine oder mehrere Ausreißer. Der Verlauf der Dichtefunktion für die Intensitäten ist bei allen Microarrays sehr ähnlich. Auch die Anstiege der Graphen des RNA- Verdaus sind zueinander gleicher Natur.
Da der Microarray für das zweite Replikat von Fluoxetin 25 μg/l zwar in einigen Kontrollen auffällt, aber sich trotzdem innerhalb der Parameter, z.B für den N.U.S.E.- Plot, bewegt, wird dieser für die statistische Analyse behalten.
Nach dem gleichen Prinzip wurde mit allen auffälligen Microarrays verfahren und letztendlich kein Microarray für die allgemeine Analyse entfernt.

5.3 Vergleich der differentiellen Expression

Das Ziel dieser Arbeit war es zu zeigen, dass es Gene gibt, die immer auf eine subakute Chemiekalienkonzentrationen reagieren und ob es auch Gene gibt, die auf eine allgemeine Gegenreaktion deuten. Dies sollte durch eine vereinheitlichte und vergleichende statistische Analyse von Genomexpressionsdaten des Zebrabärblings beantwortet werden.
Die Ergebnistabellen für die Analyse der differentiellen Expression zeigen für alle untersuchten Chemikalien differentiell exprimierte Gene. Der Vergleich dieser Tabellen untereinander beweist, dass die Ergebnistabellen sich überschneiden und die differentiell exprimierten Gene in mehreren Behandlungen vorhanden sind.
Allerdings ist die Spannweite an Genen sehr groß für die einzelnen Tabellen an sich. Dies zeigt sich besonders bei der Flutamidtabelle, die nur ein signifikantes Gen enthält. Im Gegensatz dazu hat Linuron im Kontrast zur Lösemittelkontrolle 5553 Gene (LinSolv).
Genistein, in beiden Lösemittelvarianten, und Methylparaben teilen sich differentiell ex-

Gen	Stoffwechsel-/Signalweg
fkbp5	Östrogensignalweg
nfil3-6	Interleukin 6 Regulierung
pfkfb4l	Binden heterozyklischer Verbindungen
fos	RNA-Regulierung/ Binden heterozyklischer Verbindungen
klf11a	Binden heterozyklischer Verbindungen / Binden von organischen Verbindungen
pik3r3a	FoxO Signalweg (Apoptose, oxidative Stressresistenz)
cyp24a1	Binden heterozyklischer Verbindungen / Binden von organischen Verbindungen
socs3a	Interleukin 6 Signalweg/ Regulationsantwort auf zellulären Stress
ankrd9	-
sesn2	Regulation der Stressantwort/ p53-Signalweg
pcp4l1	-
meis2b	Binden heterozyklischer Verbindungen/ Regulation RNA-Synthese
cyp26a1	Binden heterozyklischer Verbindungen/ Abbau organischer Verbindungen
per2	Gefäßentwicklung/ Runterregulierung von Entwicklungsprozessen
agt	Endokrine Prozesse/ Hormonregulierung
rcv1	Ionenbindung
higd1a	Membrankomponente
ndrg1b	-
slc3a2b	Membrantransport/ Metabolismus für org. Substanzen
ctsla	Metabolismus für org. Substanzen
socs3b	Interleukin 6 Signalweg/ Regulationsantwort auf zellulären Stress
foxg1bl	Binden heterozyklischer Verbindungen/ Hoch- und Runterregulierung RNA-Synthese

Tabelle 5.1: Die 22 häufigsten Gene die 21 mal oder öfter während des Vergleichs auftraten. Die zweite Spalte zeigt die zugehörige Signalwege und Stoffwechselsysteme. Die Daten stammen aus Biosystems (NCBI)

primierte Gene mit allen untersuchten Chemikalien. Methylparaben zeigte bei Schiller et al. 2013 [5] eher geringe Überlappung in Bezug auf die beeinflussten Stoffwechselsysteme. Möglicherweise hat es jedoch einen breiten Effekt auf verschiedene Gene, für die noch keine Funktion im Stoffwechselsystem bekannt ist, so wie bei GANT- 61 [4]. Aus diesem Grund könnte es sich auch mit GANT- 61 überschneiden. Das bei Flutamid (Schiller et al. 2013 [5]) vorhandene Gen fos ist ein Onkogen. Dieses steht im direkten Bezug zur tumorwachstumhemmenden Wirkung von Flutamid [22]. Dadurch, dass nur dieses eine Gen im signifikanten Bereich lag, ist die Menge an Überlappungen mit den anderen Chemikalien natürlich kleiner.
Bei den Chemikalien, aus dem Datensatz Schiller et al. 2013 [5], für die zwei Kontrollen, einmal mit Lösemittel und ohne, vorhanden sind zeigen bereits die Lösemittel der Chemikalie Effekte auf die Expression. Dies ist aber ein normaler Umstand, der so sein sollte, weswegen man in der Ergebnisbetrachtung für die Lösemittelvariante weniger Effekte sieht, da diese bereits selbst einen expressionsverändernden Einfluss haben. Die abweichende Menge an differentiell exprimierten Genen für den gesamten Datensatz Schiller et al. 2013 [5] ist in der statistischen Analyse (one- way ANOVA [5]) begründbar. Sie testet wie ein f-Test die Varianz zwischen zwei Datenreihen. Im Gegensatz dazu kann der t- Test für Flutamid auch nur ein differentiell exprimiertes Gen zeigen.

Mit den höheren Konzentrationen von Sertralin und Flouxetin wurden, wie zu erwarten, mehr Gene beeinflusst. Allerdings waren bei Park et al. 2012 [2] für die niedrige Konzentration von Fluoxetin weniger Gene zu sehen. Dies könnte eine Ursache in der Methode haben, da bei Park et al. 2012 [2] ein modifizierter unspezifischer Filter und GC- Gehaltanalyse verwendet wurden [2]. Die generelle Überlappung, dieser zu einem Datensatz (Park et al. 2012 [2]) gehörenden Chemikalien, ist relativ hoch. Beide Chemikalien gehören zu den Aminen und werden als Medikament gegen Depressionen eingesetzt. Diese Übereinstimmung im Einfluss auf die Expression wurde auch schon bei Park et al. 2012 [2] beobachtet und wird auch mit der verallgemeinerten Methode zu Analyse von Genexpressionsdaten in ähnlicher Form gezeigt. Somit sind die Gene, die innerhalb des Datensatzes überlappen (Abb. 4.1), von beiden Methoden gefunden wurden.
Cyclopamin und GANT- 61 sind die beiden Chemikalien mit den wenigsten Überlappungen und haben keinerlei Anteil an den häufigsten Genen (Abb. 4.2). Für GANT- 61 ist bereits bekannt, dass seine Wirkung als Onkogen bei der Expressionsregulation Gene mit unbekannten Funktionen reguliert. Diese gehören auch keinem einheitlichen Stoffwechselsystem an [4]. Cylopamin fällt aus dem Gesamtbild. Die molekulare Struktur ist der der restlichen Behandlungen sehr ähnlich, da es ein organisches Molekül mit mehreren Ringen ist. Möglicherweise ist seine Wirkung auf eine Änderung der Expression nur für Gene des 'hedge-hog'-Signalweges beschränkt [4]. Die zweite Möglichkeit könnte ein zu kurzer Expositionszeitraum sein, bei dem noch keine allgemeine Antwort- oder Gegenreaktion stattfinden konnte.
Andere Chemikalien wie TCDD (Dioxin), APM (Azinphosmethyl), Retinsäure, Propanil und und PCP (Pentachlorphenol) zeigen nur geringen Anteil an den häufigsten differentiell exprimierten Genen und nur wenige Überlappungen. Aber die Gene, die sich auch bei ihnen in der Ergebnistabelle der differentiell Expression finden, sind die die auf eine allgemeine Stressantwort bzw. Gegenreaktion schließen lassen (Abb. 4.2 & 5.1). Somit stützen auch diese Chemikalien das Ergebnis.
Die geringe Überlappung von Pentachlorphenol könnte ein direktes Ergebnis des kurzen Expositionszeitraums von nur acht Stunden sein. So konnten noch keine Mechanismen greifen, die eine Gegenreaktion auf genomischer Ebene auslösen. Die Diskrepanz in der Menge der exprimierten Gene für PCP, lässt sich nur auf den von Xu et al. 2014 [3] verwendeten Filter zurückführen, da der Rest der Methoden für die differentielle Expression identisch ist.
Für Flusilazol und PCE konnte viele Überlappungen mit den anderen Behandlungen festgestellt werden. Diese gehen vermutlich auf die Halogene in der Molekülstruktur zurück, da die Überlappungen von beiden sehr ähnlich sind (Abb. 4.3 & 4.4). Ein Vergleich für die differentielle Expression mit der Publikation ist für PCE nicht möglich, da die Daten nicht veröffentlicht sind. Bei Hermsen et al. 2012 [10] weichen die Daten aufgrund der verschiedenen Methoden (f- Test vs. t- Test) ab. Allerdings konnte mit dem allgemeinen Agilentskript für Flusilazol *cyp26a1* als differentiell exprimiertes Gen gefunden werden. Dieses Gen ist auch in der Publikation von Hermsen et al. 2012 [10] stark exprimiert gewesen. So konnte dies durch beide Methoden gezeigt werden.
Im großen Rahmen zeigen sich für die gesamte Analyse, dass gleichartige Chemikalien

auch gleichartige Wirkungen auf genomischer Ebene erzielen, was man durch die vielen Überschneidungen der differentiell exprimierten Gene sieht. Dadurch wird bereits die erste Fragestellung zum Teil beantwortet: Ja ähnliche Chemikalien zeigen Gene, die wiederholt auftreten. Jedoch muss die Frequenz des Auftretens einzelner Gene betrachtet werden.
Diese Analyse erfolgte durch die Zählung (Abb. 4.5). Für eine genauere Analyse wurden die Gene betrachtet, die 21 mal und häufiger Im Vergleich erschienen. Am häufigsten trat dabei das Gen *fkbp5* auf. Es ist in den Behandlungen mit Genistein, Linuron, Methylparaben, Bisphenol A, Ethinylestradiol, Prochloraz sowie bei Fluoxetin und Sertralin, in beiden Konzentrationen, vorhanden. Auch zeigt sich diese Gen unabhängig vom Expositionszeitraum, der bei Fluoxetin und Sertralin erst viel später begonnen hat und länger (96 h) dauerte (Tab. 4.2). Das Produkt des Gens findet im Östrogensignalweg seinen Zweck, was sich mit den endokrinen Wirkweisen dieser Stoffe deckt [5].
Das zweithäufigste differentiell exprimierte Gen ist *nfil3-6* (Abb. 4.5). Es hat eine Rolle in der Regulierung von Interleukin 6. Interleukin 6 wird gebildet um Gewebsentzündungen entgegen zu wirken [23], welche ein unmittelbarer Effekt der Chemikalien sind. Damit zeigt sich das erste Gen, das auf eine allgemeine Gegenreaktion des Organismus auf die Einflüsse der Chemikalie schließen lässt.
Die dritthäufigsten Gene die im Vergleich auftreten sind *fos* und *pfkfb4l*. Ersteres ist für die Regulierung von RNA verantwortlich (GO:2001141) und hat damit unmittelbaren Einfluss auf die Expression. Desweiteren hilft es beim Binden von heterozyklischen, organischen Verbindungen (GO:1901363). Diese zweite Funktion trifft auch auf *pfkfb4l* zu. Dies spiegelt auch die Häufigkeitstabelle wieder. Hier erscheint *fos* bei Perchlorethen (PCE), der einzigen Chemikalie die keine Ringe aufweist. Das andere Gen ist bei PCE nicht differentiell exprimiert, weswegen *fos* vermutlich bei dieser Chemikalie (PCE) nur RNA- regulierende Wirkung hat. Damit sind zwei weitere Gene gefunden, die auf eine allgemeine Reaktion gegen die Chemikalien schließen lassen, auch wenn sie sich nur in neun Ergebnistabellen der 23 Behandlungen finden lassen.
Auch unter den restlichen Genen, die sehr oft zu finden waren, findet sich vor allem solche wieder, die mit der Regulation (GO:0071704) und Bindung (GO:1901363) von organischen Molekülen verknüpft sind. *socs3a* ist direkt mit der Stressantwort (GO:0044767) verknüpft - wieder ein Gen das ein Hinweis für die allgemeine Gegenreaktion darstellt. *cyp24a1*, das 28 mal auftritt, ist mit dem Rezeptorsignalweg zur Erkennung von giftigen Substanzen verknüpft.
Den größten Anteil an diesen Genen hat Linuron im Kontrast zu seiner Lösemittelkontrolle. Es hat Anteil an 18 der 22 Topgene. Die wenigsten davon haben Cyclopamin und GANT- 61, die keinerlei Anteil haben. In beiden trat keins der Topgene auf. Die Überlappungen die für beide Chemikalien zu sehen waren, lagen bei genauerer Auswertung dieser, nur bei einer Überlappung je differentiell exprimiertem Gen.
Nach dieser globalen Betrachtung werden jetzt die einzelne Chemikalien nocheinmal genauer vorgestellt.
Genistein ist ein Östrogen, hat damit eine hormonelle Wirkung, und inhibiert eine Reihe von Systemen (COMT, uPA, Tyrosinkinase, PAI-1 [24]). Die endokrine/hormonelle Wir-

kung wird auch im Vergleich deutlich. Es hat einen sehr hohen Anteil an den häufigsten Genen.
Schiller et al. 2013 [5] hatten im experimentellen Design Flutamid als Gegenprobe zu den östrogenartigen Chemikalien angelegt. D.h. zwischen beiden Chemikaliengruppen sollte nur geringe Übereinstimmung in der differentiellen Expression sein. Beide Behandlungsgruppen überschneiden sich in unserer Analyse für *fos*, was auf die RNA-regulierende Wirkung hinweist (GO:2001141).
Flusilazol (Hermsen et al. 2012 [10]) ist eine siliziumorganische Verbindung, d.h. Silizium ist das zentrale Atom der Verbindung. Dadurch unterscheidet es sich grundlegend in seiner Struktur von den anderen Chemikalien. Neben dem zentralen Siliziumatom besitzt es mehrere aromatische Ringe, dadurch ist bei der Chemikalie *cyp26a* wahrscheinlich differentiell exprimiert, das bei der Bindung von heterozyklischen Verbindungen beteiligt ist. Auffällig für Flusilazol ist, das fasst alle Überlappungen bei den Genen zu finden sind, die 21 mal Auftreten. Nur *ndrg1b* ist nicht zu finden. Da es für die Regulierung von organischen Molekülen verantwortlich ist (GO:0071704), reagiert es wahrscheinlich auf das Silizium nicht.
PCE ist die einzige Chemikalie ohne aromatische Ringe. Es ist ein einfaches Ethen mit vier Chloratomen. Seine karzinogene und hautreizende Wirkung sind durch *fos* (Onkogen) und *nfil3-6* (Interleukin 6, Entzündung) wiedergespiegelt. Auch finden sich bei PCE die Gene wieder, die für eine Stressantwort verantwortlich sind (Abb. 4.5). Somit zeigt auch diese Chemikalie, die in seiner Struktur sehr verschieden zu den Restlichen ist, die Gene für eine Gegenreaktion auf zellularen Stress.
Pentachlorphenol (PCP) ist eine weitere der Chemikalien, die nur einmalig bei den häufigsten Genen auftritt. Von der Struktur ist es ein aromatischer Ring mit fünf Chloratomen. Dies spiegelt sich auch im Vergleich wieder. Hier ist PCP nur einmal in den häufigsten Genen vertreten durch *pfkfb4l*. Dieses Gen ist mit dem Binden von zyklischen Verbindungen (GO:1901363) verknüpft und solch eine Verbindung ist PCP. Möglichweise wurden in der kurzen Zeit, wie bereits zuvor beschrieben, die restlichen Gene, die sich bei den anderen Chemikalien unter den häufigsten Gene finden, noch nicht stark genug beeinflusst.
Abschließend kann man sagen, dass es Gene gibt, die vorwiegend auf Chemikalien reagieren. Ebenfalls gibt es Gene, die auf eine allgemeine Gegenreaktion auf zellulären Stress deuten (Tab. 5.1). Somit konnten die beiden ursprünglichen Fragestellungen, die den Ausgangspunkt für diese allgemeine Analyse bildeten, positiv beantwortet werden. Allerdings ist diese Studie damit noch nicht abgeschlossen.

5.4 Ausblick auf die weitere Analyse

Für die weitere Untersuchung muss noch eine genaue Analyse der Stoffwechselwege, in denen die einzelen Gene wirken, erfolgen. Hier ist vor allem die große Anzahl an Genen interressant, die sich nur einmalig überlappt. So enthalten die beiden Behandlungen mit Cyclopamin und GANT- 61 nur Gene, die in diese Gruppe des Ergebnistabellenvergleichs fallen (Abb. 4.2).
Desweiteren muss die Methodik und die Auswahl der Daten optimiert werden. Die Auswahl geeigneter Daten und deren Qualitätkontrolle stellte die größte Herausforderung dar. Es gibt in `R` mit `Bioconductor` eine Reihe von Methoden und Paketen, die auf Qualitätskontrolle von cDNA-Microrarrays ausgelegt sind. Zusätzlich sind die meisten Methoden auf Affymetrix ausgelegt, was die Kontrollmöglichkeiten für andere Plattformen wie Agilent und Nimblegene einschränkt. So musste zuerst eine einheitliche Methode für Agilent entwickelt werden. Die zweite große Herausforderung stellte die Auswahl der Daten dar. Hierbei mussten geeignete Experimente gefunden werden, die Genexpressionsdaten enthalten, die auf cDNA-Microarraybasis entstanden sind. Diese mussten noch auf Chemikalien gefiltert werden, um eine breites Spektrum in Expositionszeiträumen, Konzentrationen und Substanzklassen zu haben. Im Moment stammen die Daten aus acht verschiedenen Experimenten mit verschiedenen Microarrayplattformen von zwei Herstellern (Agilent, Affymetrix). Ebenso stammen die Daten auch nur aus dem Fischei. Auch wenn die erste allgemeine Analyse Hinweise auf die Frage, ob es eine allgemeine Reaktion auf subakute Dosen gibt, sollten die Daten besser vereinheitlicht werden:

- eine einzige Microarrayplattform von nur einem Hersteller
- Konzentrationen und Expositionszeiträume vereinheitlichen - LC10 als Richtwert
- mehr Chemikalien, die zu einer Gruppe gehören
- einheitliche Anzahl an Replikaten

Es zeigte sich, dass in der Analyse die Anzahl der Replikate und die Plattform keinen großen Einfluss auf die Ergebnisse hatten, aber eine höhere Anzahl an Replikaten gibt der statistischen Grundlage mehr Gewicht. Die Vorbereitung für die Auswertung folgt so auch nur noch einer Methode und muss nicht wie in diesem Fall auf Agilent- und Affymetrixmicroarrays individuell angepasst werden. Die nachfolgende statistische Analyse sollte bei dem Standard mit t-Test und Benjamini- Hochberg-Verfahren erhalten bleiben. Die Vereinheitlichung der Microarrayplattform würde auch der Qualitätskontrolle zu Gute kommen, da nur noch eine Methode Verwendung findet. Idealerweise sollte die Plattform Affymetrix sein, da man hier mehr angepasste Methoden in `R` findet.
Desweiteren ist eine Optimierung des Experiments, das zur Erzeugung der Expressionsdaten dient notwendig, damit man die Konzentrationen und Expositionszeiträume für optimale Effekte kennt. Dies wurde bereits für TCDD und Retinsäure schon gemacht. Für sie hat man ermittelt, dass ein Expositionsstart bei 72 hpf der optimale Startpunkt

ist [9]. So konnten sie für ihren Exposiotionszeitraum einen maximalen Effekt erzielen. Diese Bestimmung sollte für alle anderen Chemikalien ebenfalls angewendet werden.
Die letzte Möglichkeit die Methodik zu verallgemeinern wäre auf mRNA-Sequenzierung des gesamten Genoms für die Genexpressionsanalyse umzusteigen. cDNA-Microarrays haben immer nur Platz für eine begrenzte Anzahl an Genen und so sieht man in der Genexpressionsanalyse nur diese Gene. Durch Sequenzierung hätte man alle Produkte der regulierten Expression.

Literaturverzeichnis

[1] P. Haffter, M. Granato, M. Brand, M.C. Mullins, M. Hammerschmidt, D.A. Kane, J. Odenthal, F.J. van Eeden, Y.J. Jiang, C.P. Heisenberg, R.N. Kelsh, M. Furutani-Seiki, E. Vogelsang, D. Beuchle, U. Schach, C. Fabian, and C. Nuesslein-Volhard. The identification of genes with unique and essential functions in the development of the zebrafish, danio rerio. *Development*, 123(1):1–36, Dec 1996.

[2] June-Woo Park, Tze Ping Heah, Julia S. Gouffon, Theodore B. Henry, and Gary S. Sayler. Global gene expression in larval zebrafish (danio rerio) exposed to selective serotonin reuptake inhibitors (fluoxetine and sertraline) reveals unique expression profiles and potential biomarkers of exposure. *Environmental Pollution*, 167:163–170, Aug 2012.

[3] Ting Xu, Jing Zhao, Ping Hu, Zhangji Dong, Jingyun Li, Hongchang Zhang, Daqiang Yin, and Qingshun Zhao. Pentachlorophenol exposure causes warburg-like effects in zebrafish embryos at gastrulation stage. *Toxicology and Applied Pharmacology*, 277(2):183–191, Jun 2014.

[4] Anita Buettner, Wibke Busch, Nils Kluever, Athanassios Giannis, and Stefan Scholz. Transcriptional responses of zebrafish embryos exposed to potential sonic hedgehog pathway interfering compounds deviate from expression profiles of cyclopamine. *Reproductive Toxicology*, 33(2):254–263, Apr 2012.

[5] Viktoria Schiller, Arne Wichmann, Ralf Kriehuber, Christoph Schaefers, Rainer Fischer, and Martina Fenske. Transcriptome alterations in zebrafish embryos after exposure to environmental estrogens and anti-androgens can reveal endocrine disruption. *Reproductive Toxicology*, 42:210–223, Dec 2013.

[6] Kerstin Howe, Matthew D. Clark, Carlos F. Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E. Collins, Sean Humphray, Karen McLaren, Lucy Matthews, and et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498–503, Apr 2013.

[7] N. Campbell, J. Reece, Urry, Cain, Wasserman, Minorsky, and Jackson. *Biologie*, volume 8. Pearson Studium, 2009.

[8] W.J. Thiemann and M. A. Palladino. *Biotechnologie*, volume 1. Pearson Studium, Mar 2007.

[9] J. Chen, S. A. Carney, R. E. Peterson, and W. Heideman. Comparative genomics

identifies genes mediating cardiotoxicity in the embryonic zebrafish heart. *Physiological Genomics*, 33(2):148–158, Feb 2008.

[10] S. A. B. Hermsen, T. E. Pronk, E.-J. van den Brandhof, L. T. M. van der Ven, and A. H. Piersma. Concentration-response analysis of differential gene expression in the zebrafish embryotoxicity test following flusilazole exposure. *Toxicological Sciences*, 127(1):303–312, Feb 2012.

[11] Nils Kluever, Lixin Yang, Wibke Busch, Katja Scheffler, Patrick Renner, Uwe Straehle, and Stefan Scholz. Transcriptional response of zebrafish embryos exposed to neurotoxic compounds reveals a muscle activity dependent hspb11 expression. *PLoS ONE*, 6(12):e29063, Dec 2011.

[12] Alvis Brazma. Minimum information about a microarray experiment (miame) - successes, failures, challenges. *The Scientific World JOURNAL*, 9:420–423, 2009.

[13] B.M. Bolstad, R.A Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[14] Ben Bolstad. Probe level quantile normalization of high density oligonucleotide array data. *Division of Biostatistics, University of California, Berkeley/ unpublished manuscript*, -(-):8, Dec 2001.

[15] G.K. Smyth. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 1(1):397–420, Jan 2005.

[16] R. Gentleman, F. Hahne, W. Huber, and S. Falcon. *Bioconductor Case studies*, volume 1. Springer, Jan 2008.

[17] William S Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, Dec 2009.

[18] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, Aug 2002.

[19] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[20] R. Gentleman, V. J. Carey, and D. M. Bates. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:80, 2004.

[21] R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit. Bioinformatics and computational biology solutions using r and bioconductor. *Statistics for Biology and Health*, 123:1–36, Dec 2005.

[22] Fernando Pazos, Franco Sánchez-Franco, José Antonio Balsa, Javier Escalada, Nuria Palacios, and Lucinda Cacicedo. Mechanisms of reduced body growth in the pubertal feminized male rat: Unbalanced estrogen and androgen action on the somatotropic axis. *Pediatric Research*, 48(1):96–103, Jul 2000.

[23] P. Chakrabarty, K. Jansen-West, A. Beccard, C. Ceballos-Diaz, Y. Levites, C. Verbeeck, A. C. Zubair, D. Dickson, T. E. Golde, and P. Das. Massive gliosis induced by interleukin-6 suppresses a-beta deposition in vivo: evidence against inflammation as a driving force for amyloid deposition. *The FASEB Journal*, 24(2):548–559, Oct 2009.

[24] L. Lehmann, L. Jiang, and J. Wagner. Soy isoflavones decrease the catechol-o-methyltransferase-mediated inactivation of 4-hydroxyestradiol in cultured mcf-7 cells. *Carcinogenesis*, 29(2):363–370, Jan 2008.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, 15 Dezember 2014