Statistische Auswertung von Microarray-Daten

Von der Naturwissenschaftlichen Fakultät

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades

Doktorin der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

von

M.Sc. Cornelia Repenning
Geboren am 31.05.1979 in Kiel

2010

Referent:	Prof.Dr.Thomas Scheper
Korreferent:	Prof. Dr. Bernd Hitzmann
Tag der Promotion:	03.12.2010

Danksagung

Herrn Prof. Thomas Scheper, unter dessen Leitung diese Dissertation durchgeführt wurde, gilt mein besonderer Dank. Ihm danke ich für die Vergabe des interessanten Promotionsthemas und für die Möglichkeit diese Arbeit am Institut für Technische Chemie durchführen zu können.

Herrn Prof. Bernd Hitzmann danke ich für Unterstützung während der Arbeit. Die wertvollen Anregungen und Ratschläge habe ich immer sehr geschätzt.

Herrn Dr. Frank Stahl danke ich für die immer ausgesprochen angenehme Atmosphäre, die wertvollen Diskussionen und die produktive wissenschaftliche Zusammenarbeit.

1 Abstract

Microarray technology accelerates the way scientists study gene expression by its high parallelization degree. Unfortunately, several drawbacks to this form of experiments and the resulting data have been noted, including the inability to relate ratios to absolute expression levels. Thus, microarray results are not straightforward to interpret and coming along with the generation of complicated data sets and the difficulty to interpret them, the success with microarray approaches requires firstly a sound experimental design and particularly a coordinated and appropriate use of statistical methods. In this study a method for overcoming the obstacles associated with microarray data is demonstrated. To ensure the accurate measurement of sample reference intensities, which may vary by several orders of magnitude, a linear regression algorithm is implemented. This approach intends to combine the linear ranges of multiple scans, which are taken at several different scanner sensitivity settings onto an extended linear scale. The method is validated using data from a microarray experiment, which allows for an evaluation of absolute abundances occurring on the chip. For this aim, an experiment has been designed which rendered possible the inference on the absolute quantification of targets on the microarray, allowing thus the best possible valid appraisement of the applicability of the newly implemented method. Thereafter, the newly implemented software constitutes a promising method for computing microarray experiments which contain a broader range of intensities on one single chip. This newly tool is inserted into a an overall chip analysis program allowing for the efficient integration and interpretation of small (*low-density* chips) as well as large datasets (*whole-genome* chips). Thus, this dissertation introduces a very user-friendly microarray analysis program, which overcomes one of the main problems with microarray data by integrating a promising multi-scan approach.

Key words: Microarray, Chip-Experiment, Analysis, *Within-Array*-Normalisation, Multiple Scans

2 Kurzfassung

Mit der Einführung der Microarray-Technologie konnten die Möglichkeiten von Wissenschaftlern im Bereich der Genexpressionsanalyse aufgrund des diesen Experimenten innewohnenden hohen Parallelisierungsgrades erheblich beschleunigt werden. Leider weisen diese Experimente und die daraus hervorgehenden Daten einige Hindernisse auf: so erlauben die resultierenden relativen Angaben keine direkten Rückschlüsse auf absolute Expressionswerte. Aus diesem Grunde können Microarray-Ergebnisse nicht unmittelbar interpretiert werden. Einhergehend mit der Generation enormer Datenmengen und den Schwierigkeiten bei der Interpretation dieser Daten kann der Erfolg dieser Technologie nur durch ein einwandfreies experimentelles Design und vor allem koordinierte und angemessene statistische Methoden für die Auswertung der Daten gewährleistet werden. In dieser Studie wird eine Methode vorgestellt, die einige bei diesen Experimenten auftretende fundamentale Probleme zu lösen vermag. Um eine korrekte Messung der Probenwerte, die einen breiten Intensitätsbereich umspannen können zu garantieren, wurde Algorithmus implementiert, der eine lineare Regression beinhaltet. Mit diesem Ansatz werden lineare Bereiche multipler Scans kombiniert, die bei unterschiedlichen Scannerempfindlichkeiten aufgenommen wurden. Die Methode wurde anhand eines Microarray-Experiments validiert, das eine Bestimmung der absoluten Häufigkeiten der Genexpressionen auf den Chips gestattet. Dieses Experiment ermöglicht demzufolge Rückschlüsse auf die absolute Quantifizierung von Ziemolekülen auf den Microarrays, um so eine bestmögliche Qualitätsbestimmung der neuen Auswertemethode zu ermöglichen. Den Ergebnissen dieser Untersuchung zufolge stellt die neu implementierte Software ein vielversprechendes Instrument zur Analyse von Microarray-Experimenten dar, die ein breites Spektrum an Intensitäten abdecken. Dieses neue Analyse-Instrument wurde in ein globales Auswerteprogramm integriert, welches eine umfassende Auswertung sowohl von *low-density*- wie auch von *whole-genome*-Chips ermöglicht. In dieser Dissertation wird nun ein benutzerfreundliches Microarray-Analyse-Programm vorgestellt, das eins der größten Schwierigkeiten im Umgang mit Microarray-Daten durch Integration eines erfolgsversprechenden Multiscan-Ansatzes zu lösen vermag.

Schlagworte: Microarray, Chip-Experiment, Auswertung, *Within-Array*-Normalisierung, Multiple Scans

Inhaltsverzeichnis

1 Abstract **4**

2 Kurzfassung **5**

3 Abkürzungsverzeichnis **11**

4 Einleitung **12**

5 Theorie **13**
5.1 Anwendung . . . 13
5.2 Materialien und Herstellung . . . 15
5.3 Funktionsweise . . . 17
5.4 Limitierungen der Microarray-Technik . . . 19
5.5 Auswertung . . . 21
5.5.1 Experiment-Design . . . 21
5.5.2 Scannen . . . 23
5.5.3 Daten-Prozessierung . . . 24
5.5.4 Daten-Analyse . . . 27
5.5.5 Klassifizierung . . . 29
5.6 Externe Datenbanken . . . 30
5.6.1 Kyoto Encyclopedia of Genes and Genomes (KEGG) . . . 31
5.6.2 *Gene Expression Omnibus (GEO)* . . . 31

6 Ergebnisse **32**
6.1 *Gal*-File erstellen . . . 33
6.1.0.1 Anforderungen . . . 33
6.1.0.2 Durchführung und Ergebnisse . . . 33
6.1.0.3 Fazit . . . 34
6.2 Softwaretool . . . 34
6.2.1 Benutzeroberfläche . . . 34
6.2.1.1 Anforderungen . . . 34
6.2.1.2 Durchführung und Ergebnisse . . . 34
6.2.1.3 Fazit . . . 44
6.2.2 Qualitätsanalyse der Microarrays . . . 44
6.2.2.1 Hintergrund und Anforderungen . . . 44
6.2.2.2 Durchführung und Ergebnisse . . . 45
6.2.2.3 Fazit . . . 62
6.2.3 Vorverarbeitung der Daten . . . 63
6.2.3.1 Hintergrund und Anforderungen . . . 63
6.2.3.2 Durchführung und Ergebnisse . . . 64
6.2.3.3 Fazit . . . 76
6.2.4 *Within-Array*-Normalisierung . . . 76

6.2.4.1 Hintergrund und Anforderungen 76
6.2.4.2 Durchführung und Ergebnisse - Lineare Regression 77
6.2.4.3 Fazit - Lineare Regression 100
6.2.4.4 Durchführung und Ergebnisse - Lowess-Regression 100
6.2.4.5 Fazit - Lowess-Regression 105
6.2.5 *Between-Array*-Normalisierung 106
6.2.5.1 Hintergrund und Anforderungen 106
6.2.5.2 Durchführung und Ergebnisse 106
6.2.5.3 Fazit 115
6.2.6 Endauswertung 116
6.2.6.1 Hintergrund und Anforderungen 116
6.2.6.2 Durchführung und Ergebnisse 116
6.2.6.3 Fazit 128
6.2.7 Clusteranalyse und regulatorische Pfade 128
6.2.7.1 Hintergrund und Anforderungen 128
6.2.7.2 Durchführung und Ergebnisse 129
6.2.7.3 Fazit 140

7 Zusammenfassung **141**

A Liste der Angaben in der Ergebnisdatei **163**

B Liste der Experimenten-Informationen in den Datenblättern **163**

C Summen von Intensitätswerten der Quartale auf drei Microarrays **166**

D Häufigkeiten der sechs charakteristischen Spotform aus Abbildung 6.5 **168**

E Validierung der Hintergrund-Korrektur **169**

F Beispiel möglicher Scan-Einstellungen **170**

G Material und Methoden zur Absoluten Quantifizierung **173**

H Relative Standardabweichungen der Gene **175**

I Korrelation zwischen Erwartungswerten und Signalintensitäten **178**

J Microarray-Experiment mit Schwannzellen **181**
J.1 Materialien für Zellkulturexperimente 181
J.2 Zellkultivierung 181
J.2.1 iSZ 181
J.2.2 Beschichtungsprozeß 181
J.2.3 Zellkultivierung für Microarray-Experimente 182
J.3 Microarray-Experimente 182

J.3.1 Entwicklung der Ratten-spezifischen neuronalen Microarrays 182
J.3.2 RNA-Isolation 182
J.3.3 cDNA-Synthese und Reinigung 183
J.3.4 Hybridisierung 183
J.3.5 Waschen und Detektion 183
J.3.6 Scannen 183

K Clusteranalysen mit *E.coli*-Microarrays **185**

L Lebenslauf **189**

Abbildungsverzeichnis

5.1 Zellbiologische Ansätze von Microarray-Experimenten 14
5.2 Anwendungsgebiete 15
5.3 Verschiedene Microdispenser 17
5.4 Funktionsweise von Microarrays 18
5.5 Microarray-Designtypen 23
5.6 Fließdiagramm zur Normalisierung von Microarray-Daten 26
6.1 Übersicht über die Auswertung von Microarrays 32
6.2 Datenblatt 1 zur Beschreibung der Microarray-Experimente 36
6.3 Verifizierung der eingegebenen Informationen zum Experiment 38
6.4 Benutzeroberfläche für die Datenbank-Recherche 42
6.5 Benutzeroberfläche des Programms *Findspot* 46
6.6 Charakteristische Spotform vierer *low-denstiy*-Microarrays 50
6.7 Mittelwert/Median-Verhältnis als Qualitätsmerkmal 57
6.8 Signalintensitäten eines *low-density*-Microarrays der Ratte 66
6.9 Validierung der Hintergrund-Korrektur 71
6.10 Beispiele für *MA-Plots* aus der Literatur und aus eigenen Experimenten . . . 75
6.11 Auffinden von Sättigungseffekte - Beispiel 1 81
6.12 Auffinden von Sättigungseffekten - Beispiel 2 82
6.13 Relative Standardabweichungen der Quotienten aus Basenpaarlänge und Signalintensitäten 93
6.14 Relative Standardabweichungen der Quotienten aus der Anzahl dCTPs und Signalintensitäten 94
6.15 Korrelation zwischen Signalintensitäten und dCTP-Gehalt 98
6.16 Signalintensitäten von Sig2 gegen die Verdünnungsfaktoren (logarithmische Auftragung) 99
6.17 *MA-Plot* eines *E.coli whole-genome*-Microarrays 102
6.18 Ausreißertest nach Mandel in einem Fenster der Lowess-Regression 104
6.19 Vergleich eines *MA-Plots* vor und nach der Applikation einer Lowess-Regression 105
6.20 Darstellung eines *Box-Plots* nach Reimann *et al.*[186] 108

6.21 *Box-Plots* der Chips eines Microarray-Experiments vor einer *Between-Array*-Normalisierung 110
6.22 *Box-Plots* der Chips eines Microarray-Experiments nach der Zentrierung der Datensätze 111
6.23 *Box-Plots* der Chips eines Microarray-Experiments nach der Skalierung der Datensätze 112
6.24 *Box-Plots* der Chips eines Microarray-Experiments nach der Normalisierung der Verteilung der Datensätze 114
6.25 *MA-Plot* eines *whole-genome*-Microarrays mit Kennzeichnung der „*Fold-Change*“-Grenzwerte 118
6.26 Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (Lam gegen uo) 122
6.27 Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (CA gegen uo) 123
6.28 Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (CA gegen Lam) 127
6.29 Benutzeroberfläche zum Cluster der Microarray-Ergebnisse 131
6.30 Hilfe zu einer Distanz-Maß-Option beim Clustern 132
6.31 Clusterbeispiel anhand des dritten Clusters 137
6.32 Clusterbeispiel anhand des fünften Clusters 140
F.1 Datenblatt zur Abfrage der Scan-Einstellungen eines Chips 170

Tabellenverzeichnis

5.1 Unterscheidungsmerkmale von DNA-Mikrorarrays 16
5.2 Einige veröffentlichte Beispiele für Microarray-Design-Typen. (Lit = Literaturangaben) . . 23
6.1 Gen-Bezeichnung in Tabellenstrukur 41
6.2 Anordnung der Spots auf dem low-density-Microarray 47
6.3 Vergleich der aufsummierten Intensitätswerte der Blöcke eines Microarrays . . 48
6.4 Beispiel eines Ausschnitts aus der Ergebnistabelle des Programms *Validate QA-Parameter* 58
6.5 Anteil geflaggter Spots an der Gesamtzahl der Spots von geringer Güte - *low-density*-Chips 60
6.6 Anteil geflaggter Spots an Gesamtzahl der Spots von geringer Güte - *whole-genome*-Chips 61
6.7 Flag-Validierung anhand eines *low-density*-Microarrays mit Ratten-Genen . . 67
6.8 Validierung der Hintergrund-Korrektur 72
6.9 Zusammenfassung des Microarray-Experiments zur absoluten Quantifizierung 87
6.10 Vergleich der relativen Standardabweichungen der Gene - 1. Verdünnungsreihe 90
6.11 Vergleich der relativen Standardabweichungen der Gene - 2. Verdünnungsreihe 91
6.12 Vergleich der relativen Standardabweichungen der Gene - 3. Verdünnungsreihe 92

6.13 Relative Standardabweichungen der Korrelation zwischen Erwartungswerten und Ergebnissen 95
6.14 Ausreißertest nach *Nalimov* für die Quotienten aus $\frac{dCTP-Gehalt}{Intensitätswerte}$ 144 96
6.15 Reststandardabweichungen der Wertepaare aus normalisierten Signalintensitäten und dCTP-Gehalten 98
6.16 Standardabweichungen der vor und nach der *Between-Array*-Normalisierung (Zentrierung) 113
A.1 Liste der Tabellenblätter in der Ergebnisdatei 163
C.1 Visuelle Auswertung der Quartale dreier Microarrays 166
D.1 Häufigkeiten der sechs charakteristischen Spotform in % sowie die Anzahl der vorkommenden Spots entsprechender Formen auf den untersuchten Microarrays 168
E.1 Validierung der Hintergrund-Korrektur - entsprechend der 1. Verdünnungsreihe in Tabelle 6.8 und der in Abbildung 6.9 gezeigten Graphik 169
F.1 Beispiel möglicher Scan-Einstellungen für einen Chip eines Microarray-Versuchs 171
F.2 Sortierung der Scan-Einstellung für einen Microarray 172
H.1 Relative Standardabweichung der Genreplikate - 1. Verdünnungsreihe 175
H.2 Relative Standardabweichung der Genreplikate - 3. Verdünnungsreihe 176
H.3 Relative Standardabweichung der Genreplikate - 3. Verdünnungsreihe 177
I.1 *Within-Array*-Normalisierung: Quotienten aus $\frac{Molekülzahl}{Signalintensität}$ 179
I.2 *Within-Array*-Normalisierung: Quotienten aus $\frac{AnzahldCTPs}{Signalintensität}$ 180
K.1 Clustergene 186
K.2 Clustergene 186
K.3 Clustergene 187
K.4 Clustergene 187
K.5 Clustergene 188

3 Abkürzungsverzeichnis

PCR	Polymerasekettenreaktion (engl.: polymerase chain reaction)
CA	Colomin-Säure
cDNA	komplementäre DNA (engl.: complementary DNA)
dCTP	Desoxycytidintriphosphat
ECM	Extrazellulärmatrix
E.coli	Escherichia coli
EST	engl.: Expressed Sequence Tags
FPR	engl.: False-Positive Rate
GEO	Gene Expression Omnibus
iSZ	Immortalisierte Schwannzellen
KEGG	engl.: Kyoto Encyclopedia of Genes and Genomes
Lam	Laminin
Lowess	Locally Weighted Polynomical Regression
MAD	Mittlere absolute Abweichung (engl.: median absolute deviation)
MIAME	engl.: Minimum Information About a Microarray Experiment
NCAM	neurales Zelladhesionsmolekül
NCBI	engl.: National Center of Biotechnology Information
PMT	engl.: Photomultiplier
PNS	Peripheres Nervensystem
PSA	Polysia-Säure
SAGE	engl.: Serial Analysis of Gene Expression
T	Temperatur
S	Zucker (engl.: sugar)
uo	unbeschichtete Oberfläche
ZNS	Zentrales Nervensystem

4 Einleitung

Mit der Erforschung von Krankheiten rückte im Laufe des letzten Jahrhunderts auch die detaillierte Untersuchung ihrer molekularen Mechanismen (und der damit verbundenen Interaktion molekularer *Targets*) in den Fokus der Forschung. Erste Forschungsansätze konzentrierten sich auf die Sequenzierung des menschlichen Genoms und die Identifizierung der mit spezifischen Krankheiten assoziierten Gene. Ausgehend von der medizinischen Diagnostik hat mittlerweile die Bestimmung der Expression von Genen, die in spezifische organismische Abläufe involviert sind, in vielen interdisziplinären Forschungsbereichen stark an Interesse gewonnen. Eine wichtige Rolle spielt in diesem Zusammenhang die Entwicklung der Microarray-Technologie, durch die die Geschwindigkeit der Expressions-Analyse der Gene deutlich beschleunigt werden konnte.

Microrarray-Experimente ermöglichen die direkte Erforschung der Expressionsmustern von großen Genverbänden bis hin zu vollständigen Genomen innerhalb kurzer Zeit. Ihre vielfältigen Anwendungsmöglichkeiten machen sie zu einer der gebräuchlichsten Methoden zur Bestimmung der Genexpression. Oligonukleotid-Microarrays werden unter anderen zur Erforschung von mRNA-[1] und Protein-Mengen[2] eingesetzt, aber auch um das Verständnis von Protein-DNA-Interaktionen[3] zu erweitern sowie für die Analyse von DNA-Kopiezahlen[4] oder um methylierte Sequenzen aufzufinden[5] sowie in vielen weiteren Bereichen. Solche zahlreichen Einsatzmöglichkeiten setzen eine komplexe, viele Arbeitsschritte umfassende Technik voraus, die letztlich als klassische *"precision in-precision out"*-Technologie hochwertige Daten generiert.[6]

Doch trotz des vielfach beschriebenen Potentials dieser Methode, sind einige wichtige Fragestellungen der Auswertung noch immer ungeklärt.[7] Dies ist auf die zahlreichen möglichen Variabilitäten zurückzuführen, die mit jedem einzelnen experimentellen Schritt verbunden sind und die das Zentrum vielfacher statistischer Berechnungen darstellen.[8] Statistiker zeigen daher großes Interesse an der umfangreichen Menge quantitativer Informationen, die im Laufe der Microarray-Experimente entstehen. Es wurden Programme entwickelt, die eine Vielzahl verschiedener statistischer Methoden beinhalten, um den vielen Einflussgrößen während eines Microarray-Experiments Rechnung tragen zu können. Einer formellen Unterteilung der Microarray-Analyse zufolge werden sie der Primärauswertung, der Nachbearbeitung der Daten (Sekundärauswertung) und der funktionellen Analyse (Tertiäranalyse) untergeordnet. Die Primärauswertung folgt unmittelbar auf den letzten experimentellen Teil eines Microarray-Experiments, das Scannen der Microarrays mit einem konfokalen Laserscanner. Die mögliche Erfassung der gesamten Bandbreite an Intensitäten durch mehrfaches Scannen der Microarrays bei unterschiedlichen Scaneinstellung wurde hierbei bisher wenig berücksichtigt.

Die beim Scannen generierten Bilder (TIFF-Dateien) werden mittels der Primärauswertung in numerische Daten umgewandelt und können anschließend einem Datenbereinigungs-

schritt unterworfen werden. Die anschließende Sekundärauswertung umfasst verschiedene Normalisierungsmethoden, mit deren Hilfe systematische Fehler auf einem Microarray (*"within-array-Normalisierung"*) sowie zwischen den unterschiedlichen DNA-Microarrays (*"between-array-Normalisierung"*) ausgeglichen werden sollen und liefert schließlich in der Datenanalyse Informationen über die Regulation der im Experiment untersuchten Gene. Letztlich können in einer Tertiäranalyse die Gene unterschiedlichen Clustern zugeordnet werden, um so eine übergeordnete Analyse der Daten im Kontext von beispielsweise metabolischen Vorgängen zu ermöglichen. Im Rahmen dieser Arbeit wurden existierende Microarray-Analysemethoden miteinander verglichen und auf der Grundlage dessen neue Auswertemethoden entwickelt, die eine Erweiterung und Optimierung bisher vorhandener Auswerteverfahren zum Ziel haben. Als Basis für eine möglichst valide Analyse der Daten wurde dabei die zuverlässige Erfassung des gesamten Intensitätsbereichs vorausgesetzt. Für diesen Zweck wurden die in bisherigen Auswertemethoden wenig in Betracht gezogenen Mehrfachscans eingesetzt. Das (in Matlab programmierte) neue Auswerteverfahren umfasst alle standardmäßig angewandten Auswerteschritte und beinhaltet zusätzlich neuartige Methoden, welche auf die Korrektur empirisch beobachteter Variabilitäten abzielen. Es soll unterschieden werden, ob einem statistisch unerfahrenen Experimentator eine anwenderfreundliche Auswertung mittels einer Benutzeroberfläche, welche eine Vielzahl verschiedener Optionen sowie die graphische Anzeige der Zwischen- und Endergebnisse gewährt werden kann.

5 Theorie

5.1 Anwendung

Der Begriff „*Microarray*“ beschreibt moderne Untersuchungssysteme, welche auf den grundlegenden Prinzipien der Molekularbiologie beruhen und den Vorteil einer höheren Durchsatzrate sowie einer größeren Genauigkeit gegenüber herkömmlichen Filter- und Blottingtechniken aufweisen[9,10] . Neben den vielen klassischen biochemischen und genetischen Studien in der Molekularbiologie, mittels derer Faktoren gefunden werden konnten, die in Genexpressionsvorgänge involviert sind, bietet die Entwicklung Genom-basierter Analysewerkzeuge zudem fundamentale neue Einsichten in den Systemaufbau genregulatorischer Abläufe. So werden DNA-Microarrays im großem Maßstab eingesetzt, um *steady-state* RNA-Mengen zwischen unterschiedlichen Zelltypen und -zuständen zu vergleichen[11,12] . Die Integration dieser Daten ermöglicht die Beschreibung von komplexen regulatorischer Netzwerken auf Transkriptionsebene sowie der darin involvierten Gene, die kohärente globale Antworten in physiologischen und entwicklungstechnischen Abläufen kontrollieren.[13] Dieser globale Ansatz wird vor allem im Bereich der Zellbiologie verfolgt. Die Zellbiologie untersucht andererseits auch auf lokaler Ebene die durch spezifische Gene ausgelöste Veränderung des Phänotyps (siehe Abbildung 5.1).[14]

Zusätzlich zu den explorativen Studien in der Zellbiologie, der Biotechnologie, der Landwirtschaft und anderen Bereichen, die die Erforschung differentiell exprimierter, co-exprimierter und interagierender Gene zum Ziel haben, werden Microarrays auch für prognostische Untersuchungen eingesetzt.[16] Die entstehenden Genexpressionsprofile bieten hierbei umfassend Ein-

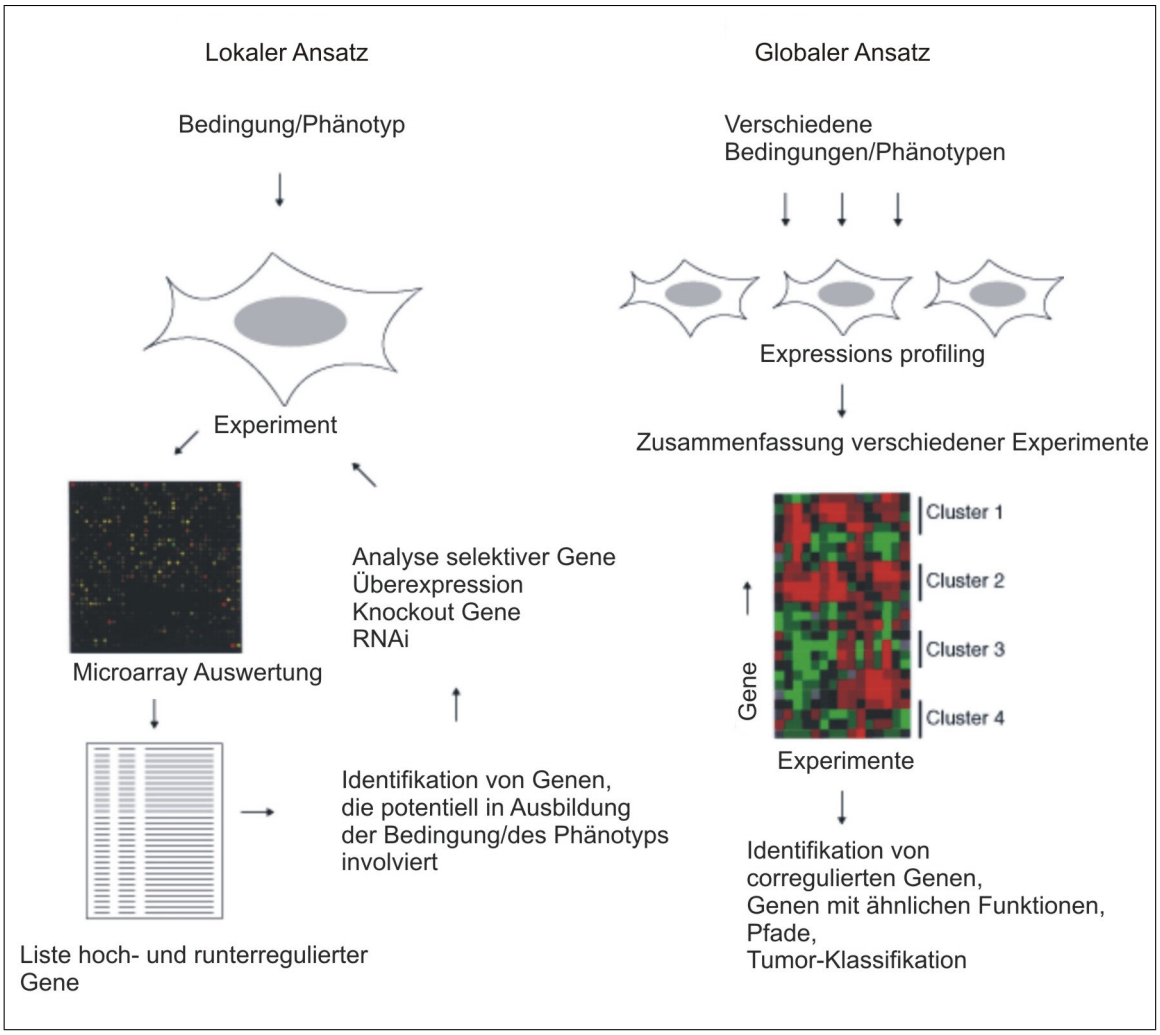


Abbildung 5.1 – *Zellbiologische Ansätze von Microarray-Experimenten.*[15]

blicke in die molekularen Funktionsweisen unterschiedlicher Krankheiten (wie u.a. von Krebs und Krankheiten, die das zentrale Nervensystem oder das kardiovaskuläre System betreffen). Bei diesen, vorwiegend im klinischen Sektor vorgenommenen Studien steht die Krebsforschung bisher im Mittelpunkt. So trugen Genexpressions-Daten bereits entscheidend zur erfolgreichen Unterscheidung verschiedener Krebserkrankungen wie Leukämie, Lymphomen, Melanomen und Brustkrebs bei. Vielfach Anwendung finden Microarrays aber auch in der pharmazeutischen Industrie, um Medikamente für definierte Krankheitserreger zu finden, die Sicherheit und Effektivität von Arzneimitteln zu prüfen (*Pharmacogenomics*) und die Toxizität von Medikamenten zu testen (*Toxicogenomics*).[17,18] Auch die schnelle Identifikation und Sub-Typisierung von Bakterien, Viren und Parasiten infektiöser Krankheiten mittels Microarrays spielt eine wichtige Rolle in der Pharmaforschung.[19] Neben den genannten Forschungsschwerpunkten halten Microarrays in jüngster Zeit aber auch in zahlreichen anderen, vorwiegend medizinischen, Bereichen Einzug: Sie werden unter anderem in der Zahnmedizin,[20] der Plazentaforschung,[21] der Neurowissenschaft,[22] der Pankreasforschung,[23] der Allergologie,[24] und anderen Forschungsgebieten eingesetzt (siehe Abbildung 5.2).

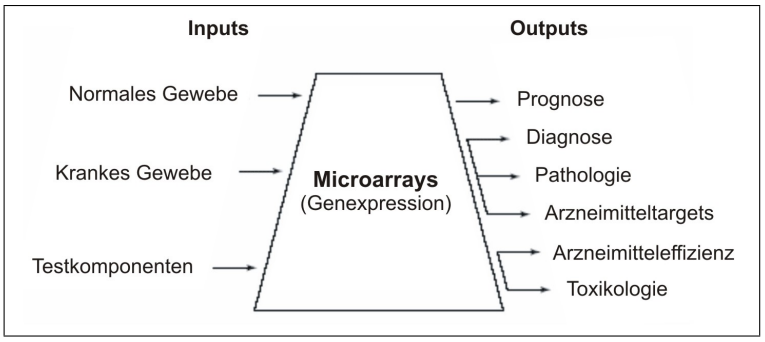


Abbildung 5.2 *– Microarrays bieten eine integrierte Plattform für die funktionelle Analyse von ganzen Genomen.[25] Die aus der mRNA spezieller Proben erhaltene, markierte cDNA (Input) wird auf den Chip aufgetragen und analysiert. Die resultierenden Genexpressionsdaten stoßen in den unterschiedlichsten Forschungsgebieten auf großes Interesse.*

Der Vorteil des hochparallelen Ansatzes der Microarrays führte also in der molekulargenetischen Diagnostik in kurzer Zeit zu zahlreichen Anwendungen dieser Technologie. Das breite Spektrum von Einsatzbereichen ist derzeit noch nicht erschöpft, was sich in den zahlreichen Artikeln zu dieser Technologie (über 28000, aktueller Stand Dezember 2008, NCBI) in den unterschiedlichsten Fachgebieten niederschlägt.

5.2 Materialien und Herstellung

Die Weiterentwicklung vom Einzelexperiment, als Grundelement des Microarrays, zur parallelen Bestimmung mehrerer Hybridisierungszustände auf einer Matrix (engl. *array*) wurde erstmal von E.M. Southern 1975 vorgenommen.[26] Bei den von Southern entwickelten Macroarrays, den sog. *Southern Blots*, werden durch Restriktionsenzyme geschnittene DNA-Fragmente elektrophoretisch in einem Gel aufgetrennt und schließlich auf eine Nitrozellulosemembran als spezifische Sonde (sog. *target*) übertragen. So konnten erstmals DNA-Sequenzen spezifisch identifiziert werden.

Das dieser DNA-Array-Technologie zugrunde liegende Prinzip der Hybridisierungsreaktionen zweier einzelsträngiger zueinander komplementärer Nucleinsäuren wurde seitdem entsprechend dem allgemeinen Trend der Miniaturisierung und Parallelisierung stetig weiterentwickelt, so dass auf heutigen Microarrays bereits bis zu mehrere Millionen unterschiedlicher DNA-Fragmente (sog. *probes*) immobilisiert werden können.[27] Die Vielzahl der mit den Anwendungen und dem teils hohen Ausmaß der Miniaturisierung dieser Multiparameteranalyse verbundenen Fragestellungen hat weltweit zu den unterschiedlichsten Lösungs- sowie Systemansätzen geführt. So musste beispielsweise das Problem der herstellungstechnischen Realisierung hoher Dichten mit räumlich getrennten, verschiedenen DNA-Sonden auf einem Sensorsubstrat gelöst werden. In Tabelle 5.1 werden einige Beispiele von Unterscheidungsmerkmalen von DNA-Micorarrays aufgeführt.

Die aufgezeigten Beispiele können mit nur wenigen Einschränkungen fast beliebig zu einem bestehenden Experimentalansatz kombiniert werden. Zu den notwendigen Systemblöcken

Tabelle 5.1 – *Unterscheidungsmerkmale von DNA-Mikrorarrays*

Merkmal	**Beispiel**
Sensorsubstrat	Glas, Silizium, Polymer
Sonde auf dem Chip	cDNA, synthetische Oligonucleotide
Synthese der Sonde	Externe Synthese, *On-Chip*-Synthese
Aufbringen der Sonde	Fotolitographie, *Wet Printing*, *Spotting*
Spotzahl / Dichte pro Chip	einige zehn bis mehrere 100.000 pro cm^2
Markierung	Fluoreszenz, Lumineszenz, markierungsfrei
Detektion	optisch, elektrisch

zählen die folgenden sechs Gerätschaften:[28]

1. Gerät zur Herstellung des Microarrays
2. Werkzeuge zur Extraktion des Analyts aus biologischem Material
3. Funktionalisierter Sensorchip mit Sonden
4. Fluidsystem zur Durchführung der Hybridisierung
5. Detektionssystem zum Nachweis der Hybridisierung
6. Softwaretool zur Quantifizierung und statistisch gesicherten Interpretation der Ergebnisse.

Da eine Unterscheidung der Microarrays in erster Linie anhand des jeweiligen Herstellungsverfahrens vorgenommen wird, werden im Folgenden kurz die standardmäßig angewandten Herstellungsmethoden vorgestellt. Sie unterscheiden sich in der Art der Auftragung der Sonden-DNA auf der Microarray-Oberfläche. Die *In-situ-* oder *On-Chip*-Synthese bezeichnet die schrittweise Synthese der Oligonukleotid-Sonden auf der Oberfläche des Microarrays. Sie basiert v.a. auf fotolitographischen Verfahren. Die fotolitographische *On-Chip*-Synthese wurde in Anlehnung an mikroeletronische Fertigungsprozesse entwickelt. Eine mit fotoempfindlichen Schutzgruppen belegte Oberfläche wird lokal durch Belichtungsprozesse entschützt, wodurch die anschließende Bindung ebenfalls geschützter Nukleotide ermöglicht wird.[29,30] Mittels dieses zyklischen Prozesses können zwar hohe Targetdichten erreicht werden,[31] der hohe technische Aufwand der Maskentechnik jedoch führte zur Entwicklung von Mikrospiegeln zur alternativen Fokussierung des Lichtkegels.[32] Insgesamt bergen beide *On-Chip*-Synthese-Verfahren das Problem der durch die Vielzahl an Syntheseschritten bedingten hohen Fehlerquoten in den synthetisierten Oligonukleotiden, die darüber hinausgehend nur kurze Sequenzen aufweisen können.[33] Diese Einschränkungen werden zum Teil durch neue *Ex-situ*-Synthese-Verfahren der Sonden-Moleküle aufgehoben.[34,35] Die getrennte Synthese gewährleistet dem Anwender eine höhere Flexibilität durch die freie Wahl des Chipformats sowie der Sonden-DNA. Die

beiden prominentesten vorkommenden Systeme der *Ex-situ*-Synthese sind die Pin-basierten fluiden Transfer- einerseits und die Piezo-basierten Inkjet-Systeme andererseits.[36] Mit Hilfe unterschiedlicher Mikropipetierverfahren kann hierbei verschiedenartige Sonden-Molekül auf die Microarray-Oberfläche aufgebracht werden.

Limitierende Faktoren dieser Methode sind vorwiegend auf den Mikrodispenser (siehe Abbildung 5.3) zurückzuführen: Die Dosierungsgenauigkeit des Druckkopfes und die teilweise unzureichende Justagegenauigkeit führen zu einer eingeschränkten Reproduzierbarkeit der Spots auf dem Chip. Das so genannte *Wet-printing*-Verfahren ist weder der *On-Chip*- noch der externen Synthese zuzuordnen und stellt eher einen Kompromiss dieser Verfahren dar.[37] Die Reaktionsflüssigkeit - ganze Sonden-Moleküle aber auch kurze Oligonukleotide - wird hierbei durch mikrofluidische Kanäle auf definierte Positionen auf den Microarrays lokalisiert.

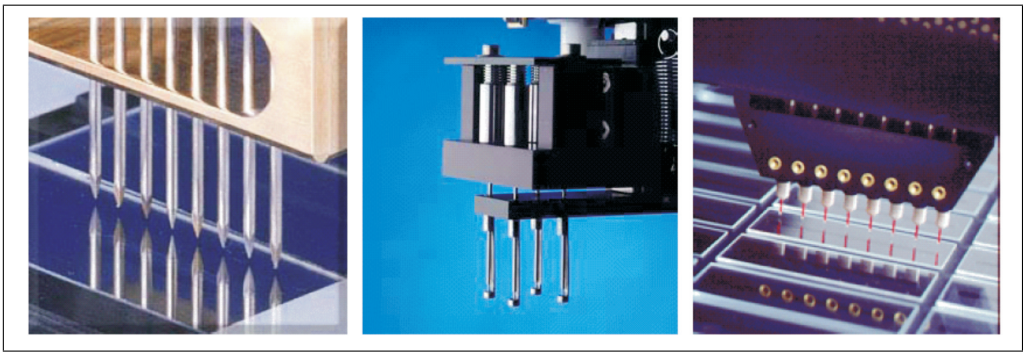


Abbildung 5.3 – *Die Abbildung zeigt verschiedene Microdispenser. (links: PinArray^{TM} von Cartesian Technologies mit acht Transfernadeln Mitte: Pin-and-Ring-Spotter GMS417 von Genetic Microsystems rechts: SpotArray^{TM} von Packard Instruments mit acht Piezopipetten)*

Neben den hier vorgestellten Methoden existiert eine Vielzahl weiterer, seltener angewandter Verfahren zur Herstellung von Microarrays.[33] Insgesamt stellt die *Off-Chip*-Methode mittlerweile die marktbescherrschende Methode zur Synthese von Microarrays dar, da hierbei zum Einen die an der Oberfläche gebundenen DNA-Sonden strukturell zu 100 % bekannt sind, was auch längere DNA-Sonden ermöglicht ohne die Gesamtausbeute gekoppelter Oligonukleotide zu beeinflussen. Zum Anderen können im Gegensatz zu *On-Chip*-Methoden auch PCR-Produkte an die Oberfläche gebunden werden, womit diese Verfahren universeller einsetzbar sind.

5.3 Funktionsweise

Die Herstellungsart der Microarrays beeinflusst das weitere Vorgehen der Experimente insofern, als dass nur *Off-Chip*-Synthese-Verfahren bestimmte Sonden und damit auch definierte Experimentdesign-Typen (siehe oben) zulassen. Die im Laufe der Microarray-Herstellung kovalent gebundenen Sonden dienen als hochspezifische Reaktionspartner und Fängermoleküle (engl. *probe* oder *capture*-DNA) für die Ziel- oder Proben-DNA (engl. *target*-DNA), deren relativer Expressionsgrad in Microarray-Experimenten ermittelt werden soll. Es kann demnach ausschließlich dann zur Ausbildung einer Doppelhelix aus zwei jeweils zueinander komplemen-

tären DNA-Strängen kommen, wenn die zu den auf den Chips befindlichen komplementären Gegenstränge in der Hybridisierungslösung vorhanden sind[38] (siehe Abbildung 5.4).

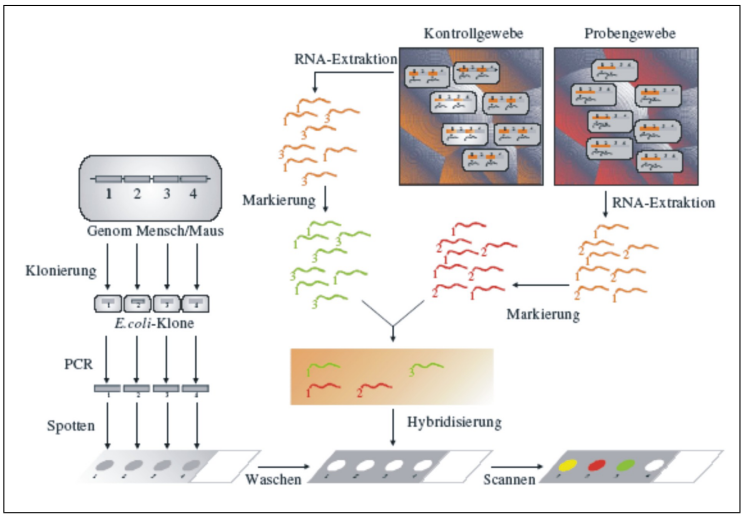


Abbildung 5.4 *– Die Produktion eines Arrays auf der Grundlage von hypothetischen Gensequenzen ist links dargestellt. Die cDNA-Sequenzen stammen aus Plasmid-Inserts in Escherichia coli(E.coli)-Klonen und werden mit Hilfe der Polymerase-Kettenreaktion (PCR) amplifiziert. Die aufgereinigten Produkte werden auf den Objektträger aufgebracht. Um ein solches Arrays herzustellen, wird aus verschiedenen Geweben RNA extrahiert (rechts oben). Die extrahierten RNA-Proben werden unterschiedlich markiert und auf das Microarray hybridisiert (Mitte). Entsprechend der Genexpression in beiden Geweben lassen sich abschließend für die Gensonden auf dem Array folgende Signale detektieren: gleich hohe Intensität in beiden Farbkanälen für das erste Gen (gelb: in beiden Geweben exprimiert), ein Signal in nur einem der beiden Kanäle für die zweite und dritte Sonde (rot bzw. grün: Gen nur im Gewebe mit roter bzw. grüner Markierung der RNA exprimiert) und schließlich kein Signal für das vierte Gen, das in keinem der Gewebe exprimiert ist.*[39]

Da es unmöglich ist, aus einem DNA-Strang die Expressionsstärke eines Gens abzulesen, liegt die Ziel-DNA als so genannte cDNA (*copy-* oder *complementary*-DNA) vor, die während der Reversen Transkription aus der mRNA hervorgeht. Dieser Vorgang wird zur gleichzeitigen (i.d.R. Fluoreszenz-)Markierung der cDNA genutzt. Da die Position der einzelnen Sondenmoleküle auf dem Microarray bekannt ist, ist so ein ortsaufgelöster Nachweis der Hybridisierung möglich. Die Effizienz jedes einzelnen dieser Schritte wird von einer Vielzahl interagierender Parameter beeinflusst, was eine absolute Quantifizierung der Genexpression erschwert. Aus diesem Grunde werden Microarray-Experimente in der Regel als so genannte Zweikanal-Experimente durchgeführt, wobei je zwei unterschiedlich markierte Zustände auf einem Chip verglichen werden können. Aufgrund des Fluoreszenz-Ratios kann so auf den relativen Expressionsgrad eines bestimmten Gens in der Probe geschlossen werden.[40] Um die dabei vorliegenden Farbstoffunterschiede ausgleichen zu können, werden die jeweiligen Zustände unter Verwendung weiterer Chips mit dem entsprechend anderen Farbstoff markiert.[41] Im Gegensatz zu dieser, *(multiple) Dye-swap* genannten Technik, die dem so genannten *Loop*-Design ähnelt (siehe unten), wird beim Referenz-Design jeder Zustand einer vergleichenden Arbeitsprobe gegenübergestellt.[42]

Von den auf diese Weise designten Microarrays werden abschließend überschüssige Probensequenzen abgewaschen und die zurückbleibende, markierte cDNA der Sonden-Spots durch Scannen der Fluoreszenzsignale nachgewiesen. Die gemessene Intensität jedes Spots korreliert

hierbei mit der in der Probe enthaltenen Menge an für die Sequenz spezifischem Transkript. Durch Zuweisung der Fluoreszenzsignale zu den korrespondierenden Spotpositionen kann abschließend ein Expressionsprofil für die durch die Sonden repräsentierten Gene erstellt werden. Eine verlässliche Aussage über die Expression beziehungsweise Regulation der Gene erfordert jedoch eine zuverlässliche Auswertung der Rohdaten unter Berücksichtigung der diesen Experimenten innewohnenden Variabilitäten.

5.4 Limitierungen der Microarray-Technik

Auch wenn die Microarray-Technik prinzipiell die Betrachtung komplexer Zusammenhänge bei der Expression erlaubt, unterliegt diese recht junge Technik noch einigen Limitierungen. Dies ist größtenteils durch die meist hohe Varianz der Messwerte bedingt, wodurch schwache Effekte nur geringfügig erfasst werden können.[43] So kann die Qualität der Daten durch eine Vielzahl unterschiedlicher Faktoren beeinträchtigt werden, die einerseits auf molekularer/zellulärer Ebene und andererseits auf systembedingter Ebene wirken:

1. Molekulare und zelluläre Ebene

   - Gensonden für verschiedene Abschnitte eines Gens: In der Regel misst eine Sonde die Expression eines einzelnen Gens. Enthält ein Chip jedoch verschiedene Gensonden für unterschiedliche Abschnitte desselben Gens, so können für diese Gensonden durch alternatives Spleißen, alternative poly(A)-Stellen oder durch Fehler auf dem Array unterschiedliche Expressionswerte resultieren.[44] Dies gilt vor allem für Affymetrix-Chips.
   - Fehler in Sequenzdatenbanken: Fehler in Sequenzdatenbanken können zu fehlerhaften Annotationen der Daten führen, die letztlich zu einer falschen Zuordnung der Sondensequenz zum jeweiligen Gen führen. Dies gilt vor allem für Microarrays, deren Annotation auf vorläufigen Versionen noch nicht bestätigter Genomsequenzen basieren.[45,46]
   - Kreuzhybridisierungen: Bei bis zu 10 % der Gensonden auf einem Chip führen Spleißvarianten zu Kreuzhybridisierungen, so dass verfälschte Expressionswerte entstehen.[47]
   - Minimale Änderungen der Genexpression in heterogenen Systemen: Mit Hilfe von Microarrays können Unterschiede der Expression von Subkulturen zwar mit hoher Sensitivität erfasst werden. Da jedoch nur 5 % der Totalpopulation solche Änderungen aufweisen, können diese in nur einem sehr geringen Teil der Zellen einer heterogenen Zellkultur oder Gewebe auftreten.[48]

2. Systemabhängige Ebene

   Beim Down-Scaling von Experimenten entstehen üblicherweise externe und interne Fluktuationen.[49] Die folgende Liste enthält die größten zu erwartenden (weitestgehend experimentell bedingten) Quellen dieser Fluktuationen:[50]

- Transkription: Bei der Reversen Transkription entstehen cDNA-Fragmente unterschiedlicher Längen.
- Labeling: In Abhängigkeit von der Nukleotidzusammensetzung kann es zu zufälligen und systematischen Schwankungen kommen. Des Weiteren führt der Einsatz unterschiedlicher Farbstoffe zu einer unterschiedlichen Einbaurate.
- Amplifikation: Die Verwendung von Klonen setzt eine oft schwer quantifizierbare PCR-Amplifikation voraus.
- Systematische Variationen in der Pin-Geometrie: Die unterschiedlichen Oberflächenbeschaffenheiten verschiedener Pins erschwert den Transport einheitlicher cDNA-Mengen.
- Zufällige Fluktuationen im Target-Volumen: Die Menge an übertragenen Targets fluktuiert sogar bei einem Pin stochastisch.
- *Target*-Fixierung: Der Anteil chemisch gebundener cDNA an die Chipoberfläche ist noch weitestgehend unbekannt.
- Hybridisierungs-Parameter: Die Effizienz der Hybridisierung wird durch viele experimentelle Parameter wie beispielsweise die Temperatur, die Zeit, die Pufferbedingungen und die Gesamtmenge an Probenmolekülen beeinflusst.
- Chip-Inhomogenitäten: Die Probe kann ungleich über dem Chip verteilt werden. Außerdem kann es zu Hybridisierungsreaktionen unterschiedlicher Vollständigkeit kommen.
- Nicht-spezifische Hybridisierung: Es erfolgt eine nicht-spezifische Bindung von markierten Targets auf der Oberfläche.
- Nicht-spezifischer Hintergrund: Oftmals treten nicht-spezifische Signale von benachbarten Spots auf.
- Bild-Analyse: Die generierten Bilddateien der gescannten Microarrays weisen nicht-lineare Transmissions-Charakteristiken, Sättigungseffekte und Variationen in den Spot-Formen auf. Die nicht-lineare Transmissionseffekte machen sich sogar beim mehrfachen Scannen der Microarrays mit gleicher Lasereinstellung bemerkbar.

Die genannten Schwankungen treten sowohl in Form von zufälligen als auch systematischen Fehler auf.[51,52] Sie können sich entsprechend dem Experiment-Design als Variabilitäten zwischen:

- Gen-Replikaten auf dem gleichen Microarray,
- zwei unterschiedlich gelabelten Proben auf demselben Chip ,
- Proben auf unterschiedlichen Chips oder
- unterschiedlichen Individuen einer Population, die auf den selben Chip hybridisiert wurden,

bemerkbar machen.[53] Es steht also insgesamt eine verhältnismäßig geringe Probenmenge (z.B. Zelllinien, Patienten, u.s.w.) einer vergleichsweise großen Anzahl an Variablen gegenüber.[19] Traditionelle statistische Methoden können demnach den Anforderungen an eine statistisch signifikante Auswertung, die die Minimalanforderung für eine biologische Signifikanz erfüllt, nicht genügen. Um die immer wieder gestellte Frage einer ausreichenden Validität der Microarray-Ergebnisse positiv beantworten zu können, wurden bereits eine Vielzahl an statistischen Methoden entworfen. Letztlich ist es dennoch nach wie vor empfehlenswert, die aus Microarray-Experimenten erhaltenen Daten stichprobenartig mit Hilfe anderer Messmethoden wie dem *Northern Blot*, der Polymerase-Kettenreaktion (PCR), der Nuklease Protektion, mit *Expressed Sequence Tags* (ESTs), der *Differential Display*-Analyse oder der *Serial Analysis of Gene Expression* (SAGE)-Methode[54] zu verifizieren.

5.5 Auswertung

Die Ursachen für die genannten Abweichungen sind theoretisch mittels Qualitätsanalysen identifizierbar. Praktisch jedoch sind die unterschiedlichen fehlerhaften Effekte oft weder orthogonal noch linear. Dieser Umstand erschwert die Quantifizierung der spezifischen Fehlerquellen. Nichtsdestotrotz konnte mit Hilfe eines verbesserten experimentellen Designs und robusten statistischen Ansätzen, welche bei der Analyse von Hochdurchsatz-Experimenten wie der hier vorgestellten Microarray-Technologie notwendig sind, beträchtlich zur Korrektur der Fehler beigetragen werden.[50,55] Im Folgenden werden experimentelle Ansätze und unterschiedliche Auswertemethoden vorgestellt.

5.5.1 Experiment-Design

Ein angemessenes Experiment-Design wird benötigt, um sicherzustellen, dass die Fragestellung des Experiments *präzise* und unter Berücksichtigung experimenteller Einschränkungen, wie z.B. Kosten und Verfügbarkeit der mRNA, beantwortet werden kann.

Die rapide steigende Menge an durchgeführten Microarray-Experimenten erforderte allgemein gültige Richtlinien, welche 2001 in den so genannten *MIAME* (Minimum Information About a Microarray Experiment)-Standards von Brazma *et al.* formuliert wurden und deren Befolgung mittlerweile in vielen Zeitschriften vorausgesetzt wird[56] . Bei dem Design eines Experiments sollten diese Richtlinien befolgt werden. Sie erfordern Informationen zum experimentellen Design, zur Probenherstellung und -markierung, zum Hybridisierungsprozess, zu Parametern, Messdaten und Spezifikationen sowie zur Microarray-Produktion.
Die Möglichkeit, mit Hilfe von Microarrays Aussagen über die Häufigkeit hybridisierter *Targets* zu machen, hängt entscheidend von den an die Oberfläche gebundenen Sonden ab. Gespottete Microarrays wurden traditionell mit Produkten aus experimentell abgeleiteten Bibliotheken oder PCR-Produkten hergestellt.[57] Aus den beispielsweise durch kontaminierte Klone hervorgehenden Variationen und Fehlern der PCR-Produkte resultieren jedoch Kreuzhybridisierungen auf dem Chip,[58] so dass neuerdings längere synthetische Oligonukleotide eingesetzt werden, die diese Einschränkungen nicht aufweisen.[59] Die Länge der jeweiligen Oligonukleoti-

de sollte dabei einen Kompromiss aus ausreichender Sensitivität einerseits und unerwünscht auftretenden Homologieregionen mit steigender Basenpaarlänge andererseits darstellen.[60] Des Weiteren sollte die Sequenz auf die Einzigartigkeit im Genom, eine geeignete Hybridisierungstemperatur, mögliche Sekundärstrukturen und die Nähe zum 3'-Ende vom Gene überprüft werden.[61]

Auf molekularer Ebene spielen neben den Sondenmolekülen vor allem die *Targets* eine große Rolle. Diesbezüglich gehört zu den wichtigsten Entscheidungskriterien die Frage nach der Anzahl der Replikate einerseits und der Notwendigkeit, die Probe zu poolen, andererseits. Prinzipiell gilt, dass eine höhere Anzahl an Replikaten die Sensitivität und Signifikanz der Experimente erhöht.[43] Diese können in der Realität jedoch aus Kostengründen selten beliebig hoch gewählt werden. Replikate können ihrem Charakter zufolge zwei Kategorien zugeordnet werden: Den biologischen Replikaten einerseits und den technischen Replikate andererseits.[62] Biologische Replikate können als RNA-Proben von unabhängigen Präparationen derselben Quelle oder von Präparationen unterschiedlicher Quellen (wie zum Beispiel aus unterschiedlichen Organismen oder Zelllinien) vorliegen. Eine Art von technischen Replikaten ist die Duplikation von Spots auf einem Chip. Auch die Replikation von Chips mit identischen Targets ist technischer Natur.[63] Typicherweise nutzen Forscher biologische Replikate, um Verallgemeinerungen von Schlussfolgerungen zu validieren, und technische Replikate zur Reduzierung der Variabilitäten dieser Schlussfolgerungen. Das Poolen der RNA aus ähnlichen Quellen ist oft unvermeidbar, um ausreichend Material für eine einzelne Hybridisierung zur Verfügung zu stellen. Eine Möglichkeit, das Problem unzureichenden Startmaterials zu umgehen, ist die RNA-Amplifikation.[64] Mit dem Poolen von Proben kann auch die Anzahl an benötigten Arrays reduziert werden, so dass zusätzliche Kosten gespart werden können.[65] Ein einziger Pool aus vielen Proben ermöglicht jedoch keine Abschätzung der technischen und biologischen Variabilität. Shih *et al.*[66] konnte statistisch nachweisen, dass Poolen die Freiheitsgrade vermindert, so dass Pools unterschiedlicher Proben nur eingesetzt werden sollten, wenn die Anzahl verschiedener Pools nicht zu klein und die der Individuen zur Kompensation ausreichend groß ist.[67]

Bezüglich des Designs des gesamten Experiments wird zwischen drei etablierte Optionen unterschieden, die jeweils unterschiedliche Vorteile aufweisen: Beim klassischen direkten Vergleich werden jeweils zwei unterschiedlich markierte Proben auf zwei Chips miteinander verglichen. Dieses Verfahren wird auch *Dye-swap*-Design genannt (siehe Abbildung 5.5, Typ A). Beim dem *Referenz*-Design werden unterschiedlichen Proben mit einer Kontrolle verglichen, was dieses Verfahren im Vergleich zu den anderen Methoden besonders flexibel macht (siehe Abbildung 5.5, Typ B). Dahingegen werden beim *Loop*-Design mit Hilfe gegensätzlicher Markierung alle Proben innerhalb eines multiplen Vergleichs untereinander verglichen (siehe Abbildung 5.5, Typ C).

Das balancierte Block-Verfahren stellt eine Mischform der unterschiedlichen Designtypen dar (siehe Abbildung, Typ D). In Tabelle 5.2 sind bereits veröffentlichte Beispiele der Design-Typen zusammengetragen.

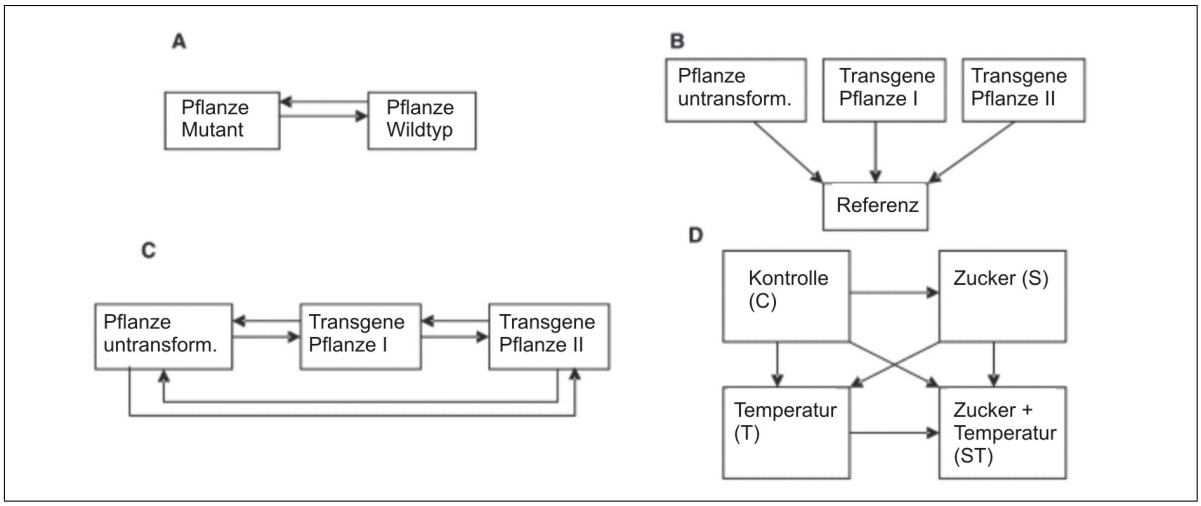


Abbildung 5.5 *– Diagramm mit vier Beispielen für Microarray-Designtypen. Jeder Chip wird durch einen Pfeil repräsentiert. Der Pfeilkopf steht für die mit Cy5-gelabelte Probe, während das Pfeilende für die Cy3-gelabelte Probe steht. A: Direkter Vergleich zwischen einer mutanten und wild-type Arabidopsis-Pflanze. B: Indirekter Vergleich über das Referenz-Design. C: Loop-Design, mit dem unterschiedlich exprimierte Gene in transgenen Pflanzen untersucht werden. D: Block-Verfahren, mit dem die Interaktion zwischen zwei Einflussfaktoren (z.B. Temperatur [T] und Zucker [S]) untersucht werden soll.*[63]

Tabelle 5.2 *– Einige veröffentlichte Beispiele für Microarray-Design-Typen. (Lit = Literaturangaben)*

Design	**Untersuchung**	**Replikation**	**Lit**
Direkter Vergleich	1 Einflussfaktor	2 Spots/Gen/Chip, *Dye-swap*	68
Loop	2 Einflussfaktoren	2 Spots/Gen/Chip, *Dye-swap*	69
Balanced block	1 Einflussfaktor	2 Spots/Gen/Chip, kein *Dye-swap*	70
Referenz	1 Einflussfaktor	2 Spots/Gen/Chip, kein *Dye-swap*	71

5.5.2 Scannen

Nach der Durchführung des Microarray-Versuchs entsprechend dem experimentellen Design werden alle Spots eines Chips mit einem konfokalen Lasermikroskop gescannt. Die Verstärkung der Signale hängt dabei von der Spannung zwischen der Foto-Anode und der -Kathode ab und sollte so gewählt werden, dass ein gutes Verhältnis zwischen Signalintensität und Rauschen resultiert. Außerdem sollte der dynamische Bereich des Scanners möglichst gut ausgenutzt werden. In Abhängigkeit des Scanners können einige Einstellungen variiert werden, um die Sensitivität des resultierenden Bildes zu verbessern. Üblicherweise wird eine einzelne Einstellung gewählt, bei der die *Photomultiplier* (PMT)-Spannung so angepasst wird, dass der hellste Pixel gerade unterhalb des Sättigungsbereichs liegt.[72] Aufgrund der großen Intensitätsspanne bei Microarray-Experimenten wird ein einzelner Scan der Anforderung an eine detaillierte Erfassung dieses Datenbereichs jedoch nicht gerecht. Daher eignet sich die Auswahl mehrerer Scan-Einstellungen in Kombination mit einer entsprechenden Auswertemethode zur besseren Berechnung der exakten Intensitätswerte.[73]

Im Anschluss an das Scannen der Microarrays werden die dabei erzeugten Bilder (typischer-

weise Paare aus 16-bit *tagged image file format* [TIFF]-Dateien - eins für jeden Fluoreszenz-Farbstoff) in der Regel mit kommerziellen Programme wie beispielsweise GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) in der Primäranalyse in numerische Daten umgewandelt. Auf diese Weise werden neben den Intensitätswerten für die einzelnen Spots auch Daten geliefert, die für eine Qualtitätsanalyse der Microarrays genutzt werden können. Bei diesen Daten handelt es sich beispielsweise um die Ratios der Mediane bzw. Mittelwerte aller Pixel eines Spots, aber auch um die Mediane bzw. Mittelwerte der Ratios. Ferner können Angaben über die Zirkularität ($Ausdehnung/Perimeter^2$) der Spots sowie Regressionsratios und viele andere Angaben verwendet werden.[74] Durch statistische Lernprozesse im Rahmen einer umfangreichen optischen und statistischen Auswertung einer Vielzahl von Microarrays können diese Daten dann kombiniert werden, um Auswahlkriterien für eine ausreichende Güte von Microarrays zu definieren, um so fehlerhafte Microarrays von der weiteren Auswertung zu entfernen. Beispielsweise werden häufig Spots mit einem besonders kleinen oder großen Durchmesser im Vergleich zur Mehrzahl der Spots oder einem auffällig hohen Hintergrund aussortiert. Auch Spots, die nicht ausreichend kreisförmig sind, deren Signal-Rausch-Verhältnis zu gering oder deren Hintergrund sehr heterogen ist, können vor der weiteren Datenprozessierung eliminiert werden.[75,76]

5.5.3 Daten-Prozessierung

Die Auswertung der Daten aus den Microarray-Experimenten erfordert neue Ansätze bei den statistischen Aspekten der Analyse. Die derzeit seitens der Statistik verfügbaren Auswerteverfahren für biologische Fragestellungen eignen sich nur bedingt für Experimente, bei denen pro Probe mehrere tausend Messwerte generiert werden können und die gleichzeitig zahlreiche Variabilitäten beinhalten. Daher muss auf stabilere Verfahren zurückgegriffen werden.
Grundsätzlich lässt sich die Datenanalyse von Microarrays in zwei Bereiche unterteilen: Die Vorverarbeitung der Daten und die funktionelle Analyse der Daten.

Bei der Vorverarbeitung (Filterung) werden unzulässige Daten aus dem Datensatz entfernt.[77] Die extrahierten Daten werden transformiert (Normalisierung) und replizierte Datenpunkte gemittelt.[78] Bei der Filterung der Daten werden Qualitätsmerkmale wie die Intensität und Homogenität einzelner Spots einerseits und die Reproduzierbarkeit der Messwerte andererseits berücksichtigt. So können beispielsweise Spots, die aufgrund technischer Mängel (z.B. unzureichende Flüssigkeitsübertragung beim Spotten) geringe Intensitäten aufweisen oder aber hohe Varianzen zwischen den Pixelwerten eines Spots, aus dem Datensatz entfernt werden. Da jedoch gleichzeitig eine Eliminierung zu vieler Spots vermieden werden soll, stellt die Auswahl geeigneter Kriterien für die zuverlässige Identifizierung nicht verwertbarer Datenpunkte bei diesen Filterungsverfahren die größte Schwierigkeit dar. Da einige Substrate einen beträchtlichen Hintergrund aufweisen, der die Intensitätswerte erhöht, gehört die Korrektur des Hintergrunds zu den gängigen Methoden während der Daten-Prozessierung.[79] Diese Korrektur kann beispielsweise durch Subtraktion des Hintergrundes vom Signalwert erfolgen. Hierfür sind sowohl globale als auch lokale Ansätze in der Literatur zu finden sind.[80] Auch nach der Filterung der Daten muss der resultierende Datensatz noch nicht über eine

verwertbare Güte der Einzeldaten verfügen. Die Notwendigkeit der Normalisierung lässt sich auch aus Selbst-Hybridisierungsversuchen ableiten, bei denen jeweils die gleichen Proben mit unterschiedlichen Farbstoffen markiert auf den selben Chip aufgebracht werden.[81] Entgegen der Erwartungen weist der rote Kanal prinzipiell geringere Intensitätswerte auf als der grüne Kanal, was beispielsweise auf unspezifische Hybridisierungen und emittierte Fluoreszenz von anderen Chemikalien auf dem Chip zurückgeführt wird. Dieses Ungleichgewicht ist zudem nicht konstant über die Spots innerhalb eines und im Vergleich zwischen Chips und kann variieren je nach Gesamt-Spotintensität, Position auf dem Microarray, Chipherkunft, und möglicherweise anderen Variablen. Ziel der Normalisierung ist es also, die Intensitätswerte so aufeinander abzugleichen, dass die resultierenden Expressionsgrade vergleichbar werden. Insgesamt dienen Normalisierungsverfahren der Entfernung systematischer Fehler. Die hierzu im Rahmen der Statistik angewandte Methodik basiert auf verschiedenen Grundannahmen, die sich für *Low-density*- und *Whole-genome*-Microarrays durchaus unterscheiden können. So gilt nach Huber *et al.* (2002) für *Whole-genome*-Microarrays, dass die Expression aller Gene im Mittel gleich bleibt und dass eine Normalverteilung über die logarithmisch transformierten Quotienten exisitieren muss.[82] Bezüglich der beiden Proben auf einem Microarray gelten eine Reihe gemeinsamer externer Einflüsse wie z.B. die Hintergrundintensität und die Variation des Chips. Normalisierungsverfahren müssen in diesem Fall also bei den relativen Fluoreszenz-Intensitäten in jedem der beiden gescannten Kanäle ansetzen und somit Unterschiede in der Quantität der Ursprungs-RNA der beiden Proben sowie zusätzlich Differenzen im Labeling und der Detektionseffizienz des Fluoreszenzlabels angleichen.[83] Zu diesem Zweck wurden eine Reihe verschiedener Normalisierungs-Algorithmen entwickelt, von denen die drei gängigsten Ansätze vorgestellt werden sollen Die vorgestellten Algorithmen gelten nach Yang *et al.* für konstant exprimierte Gene und Kontrollgene.[84]

- Globale Normalisierung[83]

  Diese Technik geht davon aus, dass die eingesetzte Anfangsmenge an mRNA und die hybridisierte cDNA-Menge auf dem Chip identisch ist. Resultierend daraus sollte die integrierte Gesamtintensität der Proben ebenfalls gleich sein. Deshalb werden die Daten so transformiert, dass die Summen der Spotintensitäten pro Probe identisch sind.

- Intensitäts-abhängige lineare Normalisierung[84]

  Mit Hilfe eines so genannten MA-Plots, bei dem die logarithmierten Einzelquotienten der beiden Proben eines Microarrays gegen die logarithmierten Einzelsummen dieser Proben aufgetragen werden, kann eine Ausgleichsgerade berechnet werden. Die normalisierten Daten werden durch Subtraktion der berechneten theoretischen Werte der Geraden von den logarithmierten Einzelquotienten erhalten.

- Intensitäts-abhängige nicht-lineare Normalisierung[85]

  Wie bei der Intensitäts-abhängigen linearen Normalisierung wird auch hier der Intensitäts-abhängige Verlauf der Datenpunkte über einen *Scatterplot* dargestellt, der jedoch einen nicht-linearen Zusammenhang aufweist. Als Ausgleich kann daher die so genannte *Locally weighted linear regression* (Lowess)-Analyse als Normalisierungsmethode herange-

zogen werden, die Intensitäts-abhängige Effekte der logarithmisch transformierten Quotienten auszugleichen vermag.[86,87]

Als mögliche Quellen für systematische Abweichungen können neben unterschiedlichen Farbstoffen auch die beim Drucken der Microarrays verwendeten verschiedenen Druck-Köpfe (engl. *print-tips*) einen Einfluss auf die Signifikanz der Daten haben (siehe Abbildung 5.6). Da dieser Einfluss von geringerem Ausmaß ist, wird eine Normalisierung zur Korrektur der unterschiedlichen *print-tips* nur dann angewandt, wenn er im vorangegangenen Test auch nachgewiesen werden konnte. Bei der Bereinigung *print-tips*-abhängiger systematischer Fehler hat sich die so genannte *print-tips*-Lowess-Regression sich als leistungsstarkes Werkzeug erwiesen.[88]

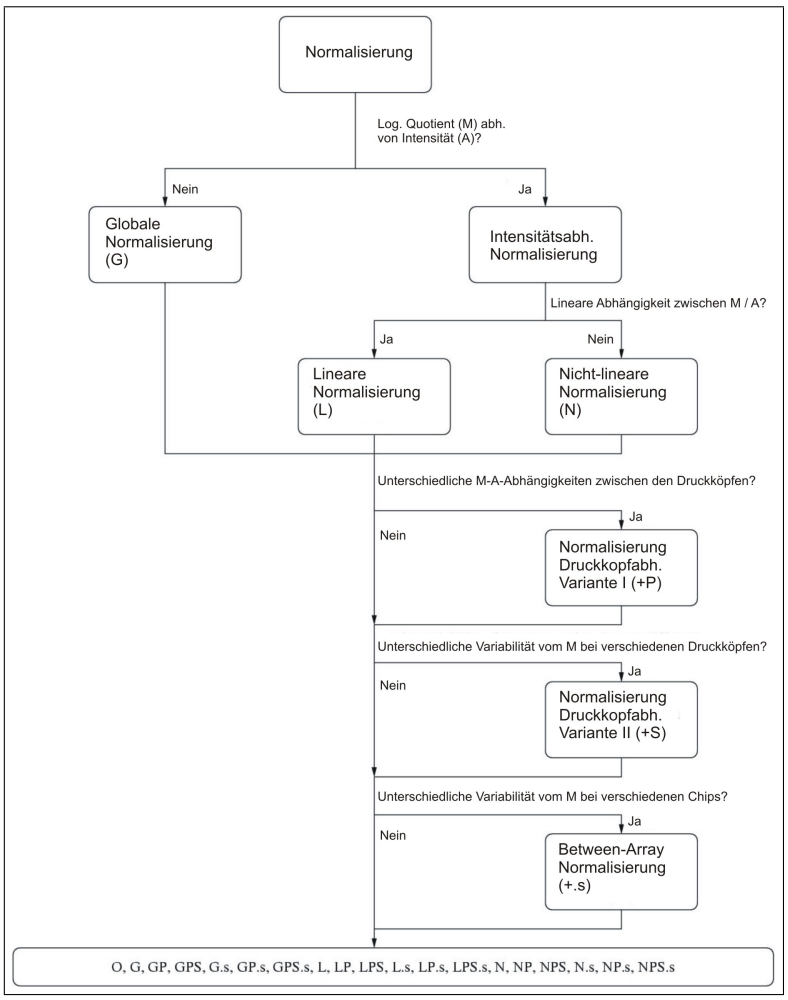


Abbildung 5.6 – *Fliess-Diagramm zur Normalisierung. O:Originaldaten, G:Globale Median-Normalisierung, L:Intensitäts-abhängige lineare Normalisierung, N:Intensitäts-abhängige nicht-lineare Normalisierung (Lowess), P:Print-tip-Normalisierung, S:Variante der Print-tip-Normalisierung, .s:Between-array-Normalisierung. Buchstabenkombinationen stellen Kombinationen der Methoden dar.*[83]

Wie in dem Fliess-Diagramm dargestellt, wird der Datensatz also zunächst auf bekannte systematische Fehlerquellen hin untersucht. In Abhängigkeit der daraus resultierenden Kombinationen an Fehlerquellen wird dann ein Algorithmus gewählt, der die Bereinigung der bekannten, im Datensatz vorhandenen Fehler durch eine Verbindung der unterschiedlichen

Kombinationen vorsieht.

Diese Normalisierungsmethoden gelten größtenteils nur für *Whole-genome*-Microarrays, bzw. wenn angenommen werden kann, dass die Expression über den ganzen Chip in beiden Zuständen gleich ist. Diese Annahme muss jedoch nicht für alle *Whole-genome*-Chips gelten, da beispielsweise Effekte, die den Transkriptionsapparat betreffen oder von einem malignen Tumor ausgelöst werden, weitreichende Auswirkungen auf das ganze Genom haben können.[77] Auch für *Low-density*-Microarrays müssen andere Ansätze entwickelt werden, die beispielsweise auf der Grundlage von *Dye-swap*-Experimenten die Hybridisierung zweier gleichartiger Proben mit je unterschiedlicher Markierung nutzen.

Nachdem die Daten der einzelnen Microarrays durch Normalisierungsmethoden in Bezug auf ihre systematischen Fehler korrigiert worden sind, werden gegebenenfalls vorhandene Genreplikate auf einem Chip vereint. Die unterschiedlichen jeweils ein Gen repräsentierenden Datenpunkte werden in der Regel lokal so über den Chip verteilt, dass positionsabhängige Unebenheiten auf dem Chip erfasst werden können. So können mit Hilfe von Ausreißertests wie beispielsweise dem *Nalimov*-Test fehlerhafte Genreplikate erkannt und eliminiert werden. Die verbleibenden Datenwerte können optional zu einem neuen repräsentativen Genwert vereint werden. Die Notwendigkeit einer Zusammenfassung der Genreplikate zu diesem Zeitpunkt der Auswertung und der Spots hängt von der weiteren Datenanalyse ab.

5.5.4 Daten-Analyse

Der folgende Teil der Datenverarbeitung, die funktionelle Analyse, gestaltet sich normalerweise flexibler als die Daten-Prozessierung. Die Vorgehensweise hängt hier in der Regel von der Fragestellung des Experiments ab. Einzelexperimente beispielsweise sind häufig auf die Frage nach der Regulation bestimmter Gene ausgerichtet, während bei Zeitreihen der zeitliche Verlauf der Expression eines Genes interessiert. Bei Gruppenvergleichen wird nach Genen gesucht, die sich einer besonders guten Klassifikation zuordnen lassen. Unabhängig von der detaillierten Fragestellung des jeweiligen Experiments, stellt die Beantwortung der Kernfrage, welche Gene signifikant unterschiedlich exprimiert sind, eins der Hauptziele der Microarray-Datenanalyse dar. Dieses Ziel besteht aus zwei Teilen: Im ersten Teil werden die Gene entsprechend ihrer Signifikanz bezüglich der differentielle Expression (Regulation) sortiert. Im zweiten Teil wird ein kritischer Wert festgelegt, oberhalb dessen jeder Datenpunkt als signifikant reguliert bewertet wird.[72] Der erste Teil ist der weitaus bedeutendere Teil, da nur eine begrenzte Anzahl an Genen in einem typischen biologischen Experiment weiter untersucht werden können. Viele Microarray-Experimenten, vor allem *Whole-genome*-Experimente, zielen daher darauf ab, Kandidaten-Gene für nähere Untersuchungen zu selektieren. Die Auswahl einer begrenzten Anzahl zu untersuchender Gene is daher entscheidend. Die einfachste Möglichkeit, die Gene ihrer Regulation zufolge zu sortieren, besteht in der Anordnung der logarithmisch transformierten Quotienten entsprechend ihrer Größe. Da diese Sortierung jedoch keinerlei Variabilitäten der Daten berücksichtigt, erfüllt sie nur unzureichend die Anforderung an eine signifikante Aussage der widergegebenen Ergebnisse.[89] So können Gene mit hohen Variabilitäten einen

hohen Sortierungs-Rang erhalten, auch wenn keine Regulation vorliegt. Eine sinnvollere Methode der Sortierung, die die Variabilität der Genreplikate berücksichtigt, stellt die gemäß der t-Statistik erfolgte Anordnung dar:

$$t = \frac{\overline{M}}{\sigma/\sqrt{n}} \tag{5.1}$$

mit σ als Standardabweichung der Genreplikate, n als Anzahl der Replikat-Arrays und M als logarithmisch transformierter Quotient der beiden Zustände. Beziehungsweise

$$t = \frac{|\overline{y_{g1}} - \overline{y_{g2}}|}{\sigma_g} \tag{5.2}$$

mit $\overline{y_{gi}}$ als mittlerer Expressionswert vom Gen g und Experiment i und σ_g als der Standardabweichung der gemittelten Differenz der Expressionswerte der Probe. Wobei die Berechnung von σ_g davon abhängt, ob es sich bei den Experimenten um unabhängige oder abhängige Experimente handelt. Ein Vorteil der t-Statistik gegenüber den einfachen Quotienten ist der ihr innewohnende Schutz gegenüber Ausreißern unter den logarithmischen Quotienten (Ms) und Spots von geringer Qualität.[72] Im Gegensatz zur einfachen Sortierung nach den Quotienten, bei dem Werte mit hoher Standardabweichung mit größerer Wahrscheinlichkeit hohe Ränge in der Sortierung einnehmen, können beim t-Test Werte mit geringer Standardabweichung unverhältnismäßig hoch eingeordnet werden. Lönnstedt und Speed *et al.* schlagen daher einen parametrischen empirischen *Bayes*-Ansatz zum Auffinden differentiell exprimierter Gene vor.[90] Der *Bayes*-Ansatz dient als Äquivalent zur Sortierung der Gene gemäß einer gewichteten t-Statistik:

$$t = \frac{\overline{M}}{\sqrt{(a + \sigma^2)/n}} \tag{5.3}$$

mit der Wichtung a, die aus dem Mittelwert und der Standardabweichung der Probenvarianz σ^2 berechnet wird. Tusher *et al.* (2001) und Efron *et al.* (2001) hingegen benutzen für Oligonukleotid-Microarrays folgende empirisch begründete Variante einer gewichteten t-Statistik[88,91]

$$t = \frac{\overline{M}}{(a + \sigma)/\sqrt{n}} \tag{5.4}$$

zur Bestimmung der differentiell exprimierten Gene. Diese Methode unterscheidet sich in sofern geringfügig von der zuvor genannten Statistik, als dass sich die Wichtung auf die Stan-

dardabweichung σ der Probe statt auf die Varianz σ^2 der Probe bezieht. Während bei Tusher *et al.* a zur Minimierung des Variationskoeffizienten des absoluten t-Wertes dient,[88] wählt Eforn *et al.* seinen empirischen Studien folgend a als das 90^{th} Perzentil des s-Wertes. Die gewichtete t-Statistik kann durch viele Modi erweitert und somit an gegebene experimentelle Situationen angepasst werden. Liegen beispielsweise nur eine verminderte Anzahl an Werten für einige Microarrays vor (z.B. aufgrund einer hohen Anzahl qualitativ unzureichender Spots), so spiegelt der Wert n im Nenner statt der Gesamtzahl an Microarrays die Anzahl der Werte für jedes Gen wider. Auf ähnliche Art und Weise kann die Wichtungsmethode an komplizierte Experiment-Designs angepasst werden. Dazu sind jedoch immer eine Vielzahl an individuellen, empirischen Studien nötig, die wiederum Kosten verursachen.

Nach der Sortierung der Gene auf der Basis geeigneter statistischer Methoden besteht der nächste Schritt in der Auffindung eines adäquaten Grenzwertes, oberhalb dessen Gene als signifikant reguliert gekennzeichnet werden.

Die einfachste Methode einer Auswahlbestimmung besteht in der simplen Festlegung des Grenzwertes 2 bzw. -2 für die Quotienten ($y_{g12} = \frac{y_g 1}{y_g 2}$) der beiden Zustände (wobei alle Quotienten y_{g12} zwischen 0 und 1 anhand der Division $-1/y_{g12}$ in den negativen Bereich transformiert werden). Eine ähnlich simple Methode ist die graphische Festlegung eines Grenzwertes anhand der Auftragung der sortierten Genwerte in einer der Normal- oder t-Verteilung entsprechenden Wahrscheinlichkeitsdarstellung. Ausgehend von der Annahme, dass die Mehrzahl der Gene nicht differentiell exprimiert ist und somit auf einer Geraden liegen, können die deutlich abseits dieser Gerade liegenden Datenpunkte als signifikant reguliert betrachtet werden. Der Nachteil dieser Methode liegt in ihrem rein informellen Charakter, da nicht davon ausgegangen werden kann, dass die implizierte Annahme der Normalverteilung der M-Werte und der Unabhängigkeit zwischen den Genen zutrifft. Aus diesem Grund neigt diese Methode dazu, die Anzahl differentiell exprimierter Gene zu überschätzen. Weitere Methoden von Shaffer *et al.* (1995) und Dudoit *et al.* (2002) beruhen auf der Kontrolle der *family-wise error rate* (FWER), also der Wahrscheinlichkeit für wenigstens einen Fehler I.[81,92] Art. Eine weitere Möglichkeit ist die Kontrolle der *false-positive rate* (FDR),[93] also dem erwarteten Anteil Fehlern I. Art an der verworfenen Null-Hypothese. Tusher *et al.* (2001) und Efron *et al.* (2001) hingegen schätzen die *false-positive rate* einer zuvor ausgewählten Untermenge an Genen unter der Annahme einer Abhängigkeit dieser Gene ab.[88,92] In der Literatur sind eine Vielzahl weiterer Methoden zur Bestimmung des Grenzwertes beschrieben, die unter anderem Experiment-spezifische Unterschiede aufweisen.[89,94,95]

5.5.5 Klassifizierung

Die Microarraydaten werden im Wesentlichen dazu gebraucht, um die bei der Datenanalyse detektierten differentiell exprimierten Gene bestimmten Genexpressionsprofilen zuzuordnen. Anhand dieser Genexpressionprofile sollen einerseits Unterschiede zwischen unterschiedlichen Zelltypen oder Bedingungen (z.B. tumorales und gesundes Gewebe) aufgedeckt werden. Andere Ansätze zielen auf die Identifizierung neuer Zelltypen, Bedingungen, u.s.w. (z.B. neue Unterklassen einer bestimmten Tumorklasse) ab. Golub *et al.* (1999) bezeichnet diese Ziel-

setzung als *Gruppen-Vorhersage* oder *Gruppen-Entdeckung*,[96] in der statistischen Literatur hingegen sind die Benennungen *Diskriminierung* und *Clustering* zu finden. Bei vorab nicht definierten Klassen / Gruppen werden *Clustering*-Methoden angewandt, während *Diskriminierungs*-Analysen bei bereits existierenden Klassen angemessen sind.
Diskriminierungs-Methoden beinhalten unterschiedlichste Methoden wie zum Beispiel lineare Diskriminierungs-Analysen,[97] Klassifizierungsbäume,[98] aggregierende Klassifizierungen,[99] neuronale Netzwerke[100] und einige mehr. Die genannten Methoden weisen bezüglich ihrer Applikation sehr unterschiedliche Schwierigkeitsgrade auf und erfordern somit teilweise fachkundige Einblicke in die statistische Methodik.
Bei Cluster-Analysen werden die Daten entsprechend ihrer Ähnlichkeit Gruppen zugewiesen. Die Ähnlichkeiten können

- in Form von Bedingungen spezifiziert werden. Dies kann beispielsweise der Fall sein, wenn Experimente unter verschiedenen Bedingungen durchgeführt wurden, so dass Gene mit einem ähnlichen Expressionsmuster innerhalb der unterschiedlichen Experimenten zusammengefasst werden (beispielsweise die Aufdeckung unterschiedlicher Krebstypen).
- Die Ähnlichkeiten können aber auch Gene betreffen. In diesem Fall werden Gene entsprechend ihres individuellen Expressionsprofils über die Experimente gruppiert. Dabei wird angenommen, dass Gene mit unterschiedlichem Expressionsmuster durch den gleichen regulatorischen Mechanismus kontrolliert werden könnten.[14]

Die Cluster-Algorithmen lassen sich in zwei Kategorien unterteilen: Solche, die keine externen Eingaben benötigen und die Daten entsprechend ihres Daten-Musters gruppieren (*unsupervised cluster analysis*) und solche, die den Algorithmus mit kleinen Untereinheiten der Daten trainieren, welche mittels externer Informationen gruppiert werden. Die resultierenden Ergebnisse werden dann auf den Rest der Daten angewandt (*supervised cluster analysis*).

5.6 Externe Datenbanken

Für die weitere Prozessierung und Verwendung der ausgewerteten und klassifizierten Datensätze können öffentlich verfügbare Datenbanken genutzt werden. Einerseits kann mit Hilfe von Datenbanken (w.z.B *KEGG* - siehe unten) die möglichen Zuordnung bestimmter Gene zu regulatorische Pfaden näher untersucht werden. Andere Datenbanken hingegen dienen der Verwaltung öffentlich zugängiger Daten unter Einhaltung bestimmter Richtlinien beispielsweise für Publikationen (siehe *MIAME*).

5.6.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Die *Kyoto Encyclopedia of Genes and Genomes* (KEGG)-Datenbank[101] dient als Ressource zum Ermittlung der Funktionen von biologischen Systemen und bietet eine zugleich eine künstliche Repräsentation dieser Systeme. *KEGG* setzt sich aus vier eigenen Datenbanken zusammen:

1. *KEGG PATHWAY* bietet Abbildungen der biologischen Stoffwechselwege.
2. *KEGG GENES* beinhaltet einen Genkatalog und beschreibt orthologe Beziehungen zwischen Genomen.
3. *KEGG LIGAND* beschreibt chemische Verbindungen und Reaktionen. Und
4. *KEGG BRITE* enthält Funktionshierarchien biologischer Systeme (Ontologien).

Die *KEGG PATHWAY* Datenbank beschreibt molekulare Interaktionsnetzwerke in der Zelle sowie Varianten dieser Netzwerke für bestimmte Organismen. Dazu liegen manuell gezeichnete Karten aus den Bereichen *Metabolismus*, *genetische Informationsverarbeitung*, *Signalweiterleitung* und verschiedenen *zellulären Prozessen* und Krankheiten vor. Diese Informationen liegen verknüpft mit den zu den jeweiligen Genen bekannten Positionen in den regulatorischen Pfaden vor und können so detaillierte Auskunft über die im Microarray-Experiment untersuchten Gene liefern.

5.6.2 *Gene Expression Omnibus (GEO)*

Die Untersuchung der Genexpression mittels Hochdurchsatz-Ansätzen wie die Microarray- oder SAGE-Technologie hat in den letzten Jahren derart zugenommen, dass öffentlich zugängige Datenbanken nötig wurden, die nicht nur Richtlinien zur Durchführung von Microarray-Experimenten (MIAME) erforderten, sondern auch die Speicherung einheitlicher Ergebnis-Datensätze. Die 2000 vom *National Center for Biotechnolgoy Information* (NCBI) ins Leben gerufene Datenbank *GEO*[102] verwaltet mittlerweile knapp 260.000 Einträge (Stand September 2008). Sie ermöglicht die simultane Quantifizierung zehntausender Gentranskripte. Als frei zugänglicher Speicherort archiviert und verteilt die Datenbank öffentlich Genexpressions-Datensätze und gewährleistet so eine zentrale Plattform zur Veröffentlichung der in ein einheitliches Format gebrachten Daten aus Hochdurchsatzexperimenten der wissenschaftlichen Gesellschaft. Der abschließende Schritt eines erfolgreich durchgeführten Microarray-Experimentes sollte also die für eine Veröffentlichung zugängliche Platzierung des Datensatzes in *GEO* darstellen.

Insgesamt kommen in der Microarray-Technologie hochkomplexe Forschungsmethoden zur Anwendung, die sorgfältiger Analysen bedarf, um letztlich umfassende Genexpressionsinformationen zu liefern.

6 Ergebnisse

Die Notwendigkeit der Anwendung statistisch abgesicherter Auswertemethoden in Microarray-Experimenten ist vielfach durch fachspezifische Literatur beschrieben worden. Dies beruht wie erwähnt auf den zahlreichen, der sich aus den einzelnen experimentellen Schritt ergebenen Variabilitäten, deren Korrektur durch die Anwendung statistischer Methoden vorgesehen ist.[43,50,51,53,103,104] Da *whole-genome*- und *low-density*-Microarrays die selben technischen Prinzipien zugrunde liegen, kann diese Aussage für beide Experimenttypen gleichermaßen getroffen werden. Bezüglich der spezifischen Wahl der anzuwendenden Algorithmen für die Lösung von Teilprobleme bei der Auswertung der Microarray-Experimente sollte aber durchaus zwischen *whole-genome*- und *low-density*-Microarrays unterschieden werden, da für die jeweiligen Fragstellungen unterschiedliche Annahmen getroffen werden müssen.

Die folgende Abbildung zeigt zusammenfassend die Grundprinzipien einer Microarray-Auswertung, wie sie in dem hier vorgestellten Programm vorgenommen wurde.

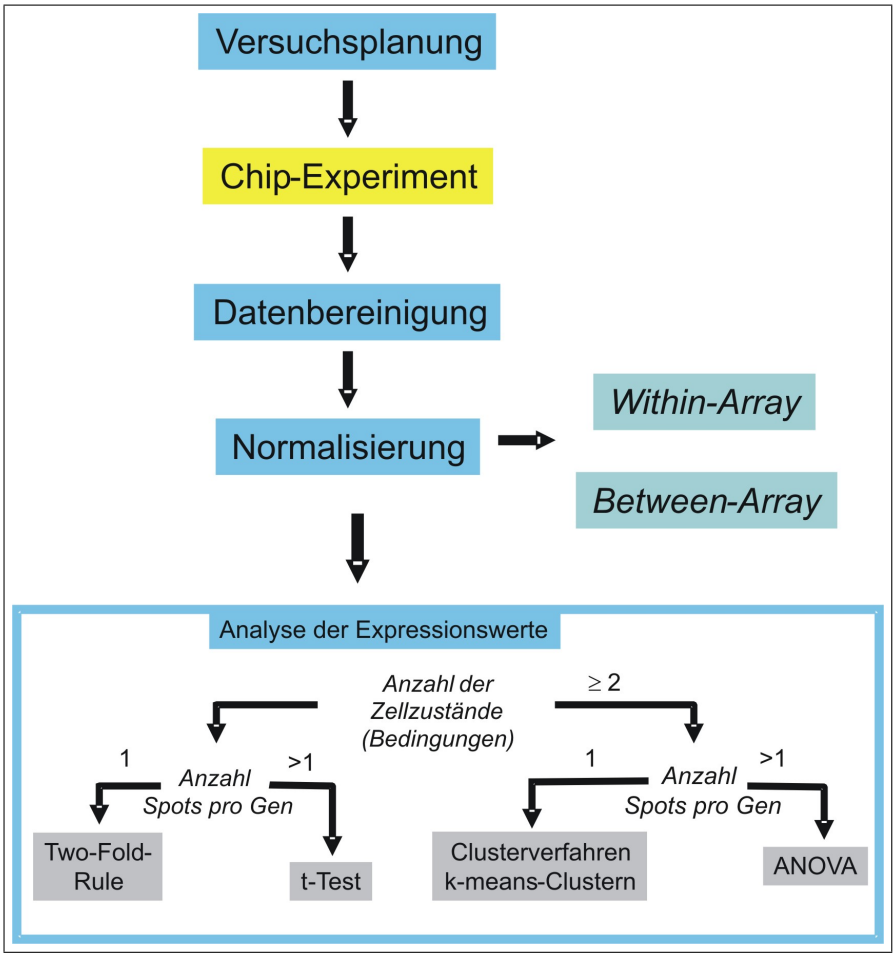


Abbildung 6.1 – *Übersicht über die Auswertung von Microarrays*

6.1 *Gal*-File erstellen

6.1.0.1 Anforderungen

Die Versuchsplanung und -durchführung eines Microarray-Experimentes kann grundsätzlich in zwei Teile gegliedert werden: den experimentellen einerseits und den bioinformatorischen Abschnitt andererseits. Um einen Experimentator mit keinen oder nur geringen Informatik-Kenntnissen also mit der Fähigkeit der kompletten Durchführung eines Microarray-Experiments auszurüsten, muss dieser über die Werkzeuge für die komplette informatorische Handhabung der Experimente verfügen.
Zu den Berechnungen im Rahmen von Microarray-Experimenten zählen jedoch nicht nur Algorithmen zur Auswertung der experimentell erhaltenen Daten. Die Bioinformatik setzt teilweise schon vor Beginn des ersten experimentellen Handgriffs ein: dies gilt für den Fall, dass ein selbst-gespotteter Microarray hergestellt werden soll. Die Software Affymetrix 417 Arrayer (Affymetrix, Inc., Santa Clara, CA) steuert zwar den Roboter während des Druckverfahrens, es bietet jedoch kein Computerprogramm zur Erstellung der notwendigen Dateien für die spätere primäre Bild-Auswertung in GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA), die auf den experimentellen Teil der Microarray-Experimente folgt. Aus diesem Grund soll dem Forscher ein Programm zur Verfügung gestellt werden, das die Erstellung einer Datei im sogenannten *Gal*-Format ermöglicht. Diese Datei muss also ein in GenePix Pro 6.0 einlesbar Format mit allen nötigen Angaben zu den selbst-gespotteten Chips besitzen, um den von diesem Programm definierten Vorgaben zu genügen.

6.1.0.2 Durchführung und Ergebnisse

Für die Erstellung der *Gal*-Dateien sind Angaben des Benutzers über die Anzahl der Genreplikate sowie die Positionen aller verwendeten 96-*well*-Microtiterplatten, in denen sich die Lösungen mit den Oligonukleotiden befinden, notwendig. Die entsprechenden Abfrage-Masken fordern zur Eingabe dieser Informationen auf. Derart spezifische Eingaben sind notwendig, um individuelle unterschiedliche Chip-Design-Typen zu gewährleisten.
Anschließend werden Excel-Dateien mit den Namen der Oligonukleotide aller Microtiterplatten eingelesen und die Gene den Positionen auf dem Microarray zugeordnet, die gemäß des durch den Affymetrix 417 Drucker vorgegebenen Druckmusters festgelegt sind. Es handelt sich dabei um eine Einteilung in Blöcke aus je 4 Zeilen und 6 Spalten.
Nach erfolgter Zuordnung der Positionen der Gennamen auf den Chips besteht die Möglichkeit, das komplette Spotting-Muster um 180 ° zu drehen. Die Notwendigkeit einer solchen Rotation hängt von dem verwendeten Gerät zum Scannen der Microarrays ab. In einem letzten Verarbeitungsschritt werden die Daten nach dem Index der Blöcke sortiert und die zum Einlesen in GenePix Pro benötigten Angaben zu den Positionen der Blöcke auf den Microarrays berechnet.
Das auf diese Weise berechnete Spotting-Muster wird in eine Textdatei geschrieben und kann direkt in GenePix Pro eingelesen werden.
Die auf diese Weise generierten Dateien bestehen also einerseits aus einem für das *gal*-Format spezifischen Kopf (englisch: *Header*) mit Druckerabhängigen Angaben wie zum Beispiel zu

der Anzahl und den Positionen der Blöcke, dem Durchmesser der Spots sowie dem Abstand zwischen den Spots. Außerdem enthalten die *gal*-Dateien für jeden Spot positionsspezifische (Block-, Reihen- und Spaltennnummer) und identitätsspezifische Informationen (GeneIDs und u.U. Gennamen).

6.1.0.3 Fazit

Um Experimente mit individuell designten *low-density*-Microarray zu ermöglichen, können mit Hilfe des im Rahmen dieser Arbeit erstellten Programms innerhalb weniger Minuten Dateien generiert werden, die einerseits das Scannen und andererseits die Primärauswertung dieser Chips durch das standardmäßig verwendete Softwaretool GenePix Pro ermöglichen.

6.2 Softwaretool

6.2.1 Benutzeroberfläche

6.2.1.1 Anforderungen

Die Reproduzierbarkeit von Experimenten bezüglich der Planung, Ausführung und Datenauswertung stellt eine Notwendigkeit in wissenschaftlichen Arbeiten dar.[110] Diese Bedingung kann bei der Konzipierung eines geeigneten Auswerteprogramms von Microarray-Experimenten durch die Integration elektronischer Datenblätter realisiert werden. In den Datenblättern sollten detaillierte Informationen zum Design und der Durchführung der Versuche gespeichert werden. Die Eingabe dieser versuchsbezogenen Informationen sollte nach dem experimentellen Teil der Microarray-Versuche erfolgen und an die Aufforderung zur Auswertung geknüpft werden. Die entsprechenden Ergebnisse zu diesen Experimenten sollten nach Vollendung der Daten-Analyse elektronisch abrufbar mit den Datenblätter verlinkt in einer internen Datenbank gespeichert werden. Auf diese Weise können differenzierte Folgeversuche unter gleichen Bedingungen geplant werden und Ergebnisse analoger Versuche im Nachhinein miteinander verglichen werden.
Sowohl die Dateneingabe als auch der spätere Abruf der Informationen zu vorangegangenen Experimenten sollte benutzerfreundlich erfolgen, so dass auch Experimentatoren ohne weitere Computer- und Programmierkenntnisse das Programm einfach bedienen können.

6.2.1.2 Durchführung und Ergebnisse

Die notwendigen Erläuterungen zum Experiment, die in den Datenblättern festgehalten werden sollen, beinhalten sowohl formelle Daten wie Angaben zur Person (Experimentator) und zum Datum der Durchführung als auch versuchsspezifische Daten beispielsweise bezüglich der verwendeten Organismen, des Extraktions-, Transkriptions- und Hybridisierungsprotokolls oder auch zu den Materialeigenschaften der Microarrays und dem Scanprotokoll.

Datenblätter

Die vielfältigen unterschiedlichen Annotationen lassen sich in zwei Gruppen unterteilen: Erstens die allgemeinen Angaben, die die eher technische Begebenheiten der Experimente re-

präsentieren. Dazu gehören unter anderem die Chip-Eigenschaften (Oberflächenbeschaffenheit, Beschichtung, usw.), die Beschreibung des Labeling- und Hybridisierungsprozesses, die Waschbedingungen der Chips und detaillierte Angaben zum Scanvorgang. Mit Ausnahme der individuell angepassten Scan-Einstellungen, setzen sich diese **technischen** Merkmale der Chipversuche im Vergleich mehrerer Experimente, aus einer begrenzten Anzahl an Möglichkeiten zusammen. Die zweite Klasse schließt die **Organismus-spezifischen** Annotationen ein, die den unterschiedlichen Anforderungen der spezifischen Forschungsgebiete genügen - wie zum Beispiel der *Tumortyp* unterschiedlicher Tumorproben oder die *Kultivierungseigenschaften* von Zellen auf verschiedenen Matrices -, und somit vielfältige Variationen zulassen.

Sowohl die technischen als auch die Organismus-spezifischen Kriterien sollten in einer einheitlichen Struktur gespeichert werden, so dass sie automatisiert ausgewertet werden können. Des Weiteren sollten die dafür notwendigen technischen Angaben unmittelbar der statistischen Auswertung übergeben werden. Da für die Auswertung auch das Einlesen der nach dem Scanvorgang generierten numerischen Informationen erforderlich ist, werden mit den technischen Informationen auch die Dateipfade für diese Primärdaten (*Gpr*-Dateien) übergeben.

Die Anwendung algorithmischer Befehlsketten in der Auswertung kann jedoch nur dann reproduzierbar erfolgen, wenn die Gesamtheit der relevanten technischen Informationen zu den Experimenten durch die Algorithmen erkannt werden kann. Daher sollten alle technisch möglichen Eventualitäten, die für die Auswertung von Bedeutung sind, vorab definiert und dem Benutzer in Form determinierter Angaben in den Abfragemasken der Datenblätter zur Verfügung gestellt werden. Auf diese Weise wird dem Benutzer erstens die Eingabe der experimentellen Informationen erleichtert und nur so kann zweitens die reliable und valide Anwendung multivariater Prozeduren zur Berechnung der differentiellen Genexpression Experiment-spezifisch sichergestellt werden.

Um alle relevanten experimentellen Details direkt für die statistische Auswertung zugänglich machen zu können, werden definierte Datentypen für die Annotation der Experimente zugelassen.
Dazu zählen zum einen **ganze Zahlen (`Integer`)**, zum Beispiel für die Anzahl der verwendeten Microarrays, die Menge der Zustandsvergleiche, die Anzahl der Genreplikate und die Einstellungen, die zum Scannen der Microarrays verwendet wurden. Zum anderen wurden **Wörter (`String`) aus vordefinierten Listen** verwendet, zum Beispiel für den Chiptyp (*whole-genome* oder *low-density*) oder aus dem frei editierbaren Text (Organismus-spezifische Annotationen) übernommene Zustandsnamen, die dann einer Liste übergeben wurden, aus der die anzustellenden Vergleiche ausgewählt werden können.
Abgesehen von diesen für die Auswertung unerlässlichen technischen Informationen, werden in den Datenblättern auch technische Details zu den Experimenten abgefragt, die zwar für Wiederholungsexperimente unerlässlich, für die Auswertung selbst jedoch nicht erforderlich sind (Angaben zur Hybridisierungsdauer, Reduktionsbedingungen, Waschpuffern, usw.). Diese

Informationen können Experiment-spezifisch variieren und sind somit teilweise frei editierbar einzugeben. Gleiches gilt für die Organismus-spezifischen Informationen (siehe Abbildung 6.2).

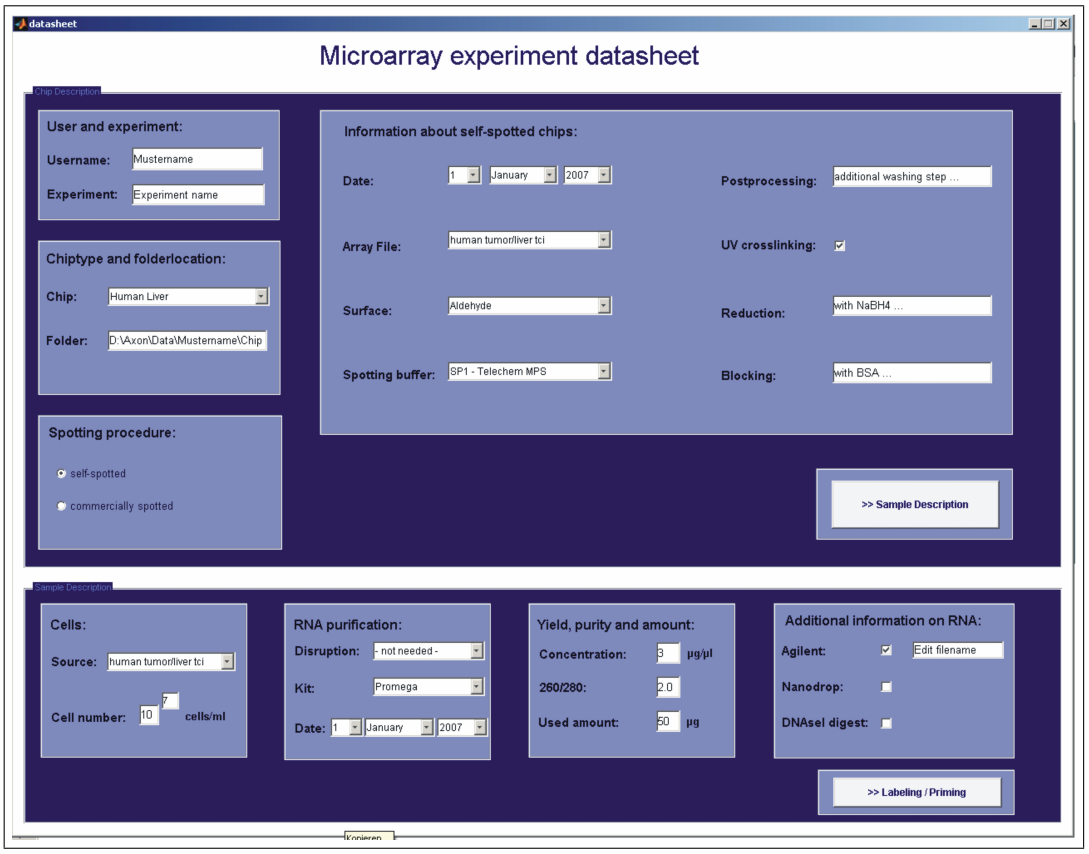


Abbildung 6.2 – *Datenblatt zur Eingabe der Experiment-spezifischen Informationen der Microarray-Experimente durch den Benutzer. Insgesamt liegen zwei Datenblätter mit Organismus-spezifischen und technischen Informationen zum Experiment vor, die für die Auswertung nicht direkt benötigt werden. Das dritte Datenblatt enthält technische Informationen, die einen Einfluss auf die zu verwendenden Algorithmen der Auswertung haben.*
Im oberen Eingabefeld des gezeigten Datenblattes werden die technischen Details abgefragt, im unteren Bereich die den verwendeten Organismus betreffenden Einzelheiten.

Zusammenfassend bedeutet das, dass die experimentellen Details inhaltlich in Organismus-spezifische und technische Merkmale unterteilt werden können. Von den technischen Details muss ein Teil wiederum der statistischen Auswertung zur Verfügung gestellt werden. Dieser Analyse-basierten Unterscheidung zufolge gliedert sich die Datenblatt-Abfrage in zwei Gruppen, von denen die eine Informationen zum Microarray-Experiment beinhaltet, die sowohl technischer als auch Organismus-spezifischer Natur sind und für die Auswertung keine Rolle spielen. Diese Gruppe wird durch die ersten beiden der insgesamt drei Datenblättern repräsentiert (von denen das erste in Abbildung 6.2 exemplarisch gezeigt ist). Innerhalb dieser beiden erstens Datenblätter erfolgt zusätzlich eine thematische Aufteilung, um eine Unterscheidung in technische und Organismus-spezifische Einzelheiten zu gewährleisten (siehe Abbildung 6.2). Die zweite Gruppe, repräsentiert durch das dritte Datenblatt, enthält die technischen, Auswerte-relevanten Informationen wie beispielsweise die Anzahl der Microarrays, die Art des Chips (z.B. *low-density*), die Anzahl der Chip- und Genreplikate.

In beiden Gruppen soll dem Benutzer eine möglichst vereinfachte und schnelle Eingabe der Daten geboten werden. Dieser Anspruch wird - wie erwähnt - einerseits durch vordefinierte Abfrageoptionen realisiert, wenn alle möglichen Optionen bekannt sind (für den Chiptyp sind beispielsweise nur die Optionen *whole-genome* oder *low-density* möglich). Dazu werden beispielsweise *Pop-up Menus*, *Checkboxes*, *Radiobuttons* oder Listen implementiert. Außerdem besteht eine Verknüpfung zwischen potentiell zusammenhängenden Abfragedetails, so dass zum Beispiel bei der Wahl eines bestimmten gedruckten Chips aus der vordefinierten Gruppe (zum Beispiel *Human Liver*) die entsprechend zu diesem Chip bereits vorhandene Chipdatei (*Arrayfile* - in diesem Fall *human tumor/liver tci*), die das entsprechende Spot-Muster bereitstellt, automatisch nach oben sortiert wird. Ähnliches gilt für die Wahl des Herstellungsverfahrens (*spotting procedure*): Nur wenn der Benutzer den Radiobutton für *self-spotted*-Chips aktiviert, öffnet sich das Feld mit weiteren Eingabeoptionen zum Spotting-Vorgang (*Information about self-spotted chips*), das bei kommerziell erhaltenen Chips überflüssig ist (siehe Abbildung 6.2).
Wie in dieser Abbildung ebenfalls zu sehen, wurden in die Angabemaske z.B. für die unterschiedlichen Microarrays (*Pop-up Menu* für *Escherichia coli Secco*-, *Human Tumor*-, *Arabidopsis*-Chip, usw.) bereits alle bisher im Insitut verwendeten Chiptypen aufgenommen. Eine Eingliederung neuer Details (z.B. die Verwendung eines neu-gespotteten Microarrays) in die jeweiligen vordefinierten Masken kann jederzeit vorgenommen werden.
Solche automatische Voreinstellungen sollen nicht nur die Bedienung des Auswerteprogramms vereinfachen, sondern auch zu einer Minimierung der Fehler während der Eingabe beitragen. Zu diesem Zweck erfolgt im Anschluss an die Eingabe der experimentellen Informationen in den Datenblättern eine Verifizierung der gemachten Angaben in den Datenblättern, in der alle vom Benutzer eingegebenen Informationen auf ihre Richtigkeit abgefragt werden (siehe Abbildung 6.3). Eine vollständige Liste aller Angaben in den Datenblättern ist im Anhang B zu finden.

Erst nachdem alle Angaben vom Benutzer bestätigt worden sind, erfolgt der Aufruf zur Analyse der Daten. Zugleich werden alle für spätere Experimente interessanten Angaben in einer Excel-Datei gespeichert. Die detaillierten Informationen zu den Microarray-Experimenten liegen demnach also schon vor der Durchführung der Auswertung vor, werden aber dennoch nach der Analyse der Experimente mit den Ergebnissen verknüpft. Auf diese Weise können nach erfolgreicher Durchführung der Experimente bequem weitere Experimente geplant werden. Die Excel-Datei bietet außerdem eine übersichtliche Zusammenfassung der für eventuelle Veröffentlichungen notwendigen experimentellen Details.

Verwaltung der internen Datenbank

Um die in den Excel-Dateien vorliegenden Experiment-Informationen zu einem späteren Zeitpunkt nutzen zu können, muss der Benutzer jedoch nicht nur den Speicherort und -namen der Datei kennen, sondern auch wissen, ob eine entsprechende Datei überhaupt existiert. Da

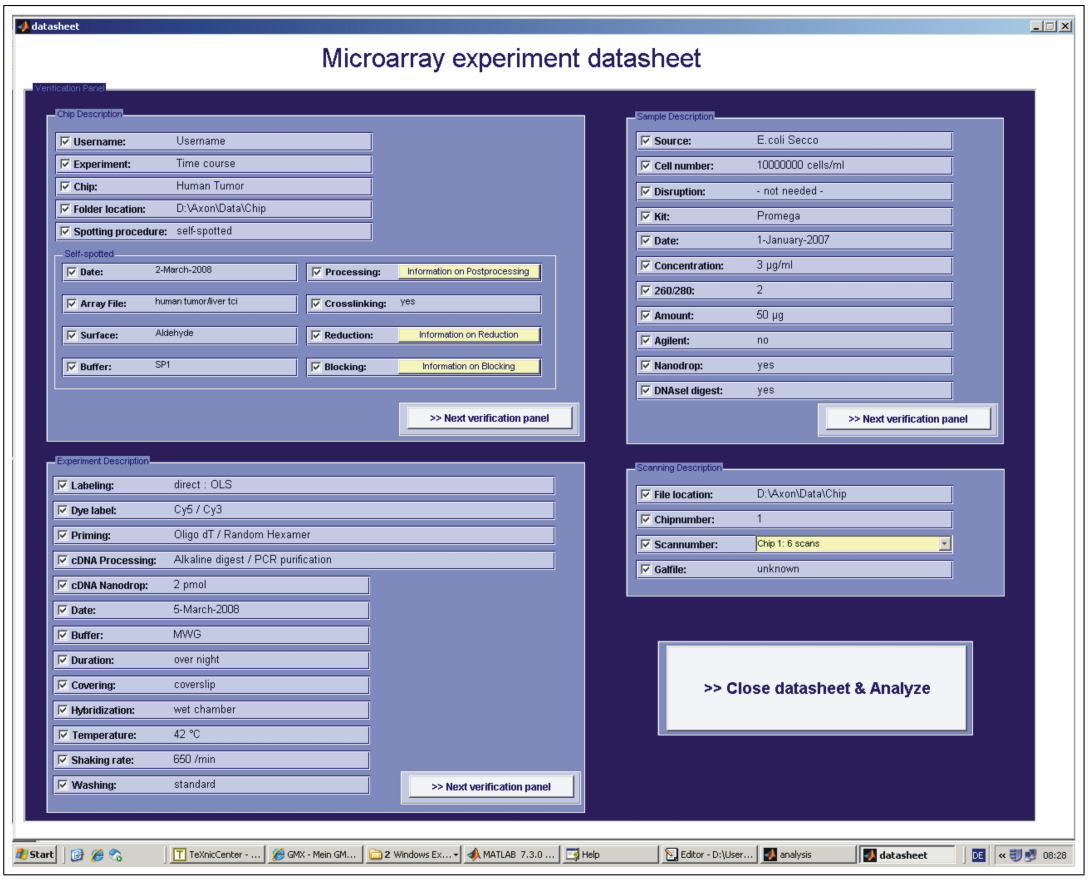


Abbildung 6.3 – *Datenblatt zur Verifizierung der vom Benutzer eingegebenen Informationen zum Experiment. Das Datenblatt fasst alle Experiment-Angaben zusammen. Wenn alle Angaben vom Benutzer durch Klicken der Checkboxes bestätigt worden sind, kann die Auswertung gestartet werden.*

aber nicht vorausgesetzt werden kann, dass jeder Experimentator über alle in seinem Interessenbereich liegenden, bereits durchgeführten Microarray-Experiment im Bilde ist, muss eine Suchfunktion für die angelegten Excel-Dateien zur Verfügung gestellt werden. Abgesehen von der Suche nach konkreten Dateien, kann die Nachforschung nach definierten Experimenten (und den dazugehörigen Ergebnissen) basierend auf einzelnen in den Dateien enthaltenen Informationen zum Experiment anhand verschiedenster Suchkriterien (und damit verbunden: aus unterschiedlicher Motivation heraus) vorgenommen werden. So kann ein Benutzer beispielsweise alle bisherigen Versuche eines auf seinem Gebiet arbeitenden Kollegen abfragen wollen (*username*). Es kann nach allen mit einem bestimmten Organismus (*source*) oder einem spezifischen Chip (*chip*) - z.B. einem *whole-genome*-Tomaten-Array - durchgeführten Microarray-Experimente gesucht werden. Ein Anwender kann auch die in einem bestimmten Zeitraum (*date*) durchgeführten Experimente erfragen wollen.
Aufgrund der Vielfalt an Variablen bei Microarray-Experimenten ist die Anzahl der Suchoptionen daher beträchtlich. Ohne Kenntnisse der Dateinamen müsste dazu jedoch jeweils die Gesamtheit aller zu den durchgeführten Experimenten vorhandenen Excel-Dateien durchgesehen werden, was bei einer entsprechend hohen Zahl an Microarray-Experimenten zu einem enormen Zeitaufwand führen kann.

Aus diesem Grund werden die zu einem Experiment gemachten Angaben nicht nur in Form von Excel-Files gespeichert, sondern auch in einer übergeordneten internen Datenbank abgelegt, die alle Einzelinformationen kategorisch speichert. Es hant sich gezeigt, dass eine übergeordneten Datenbank eine sinnvolle Ergänzung zur internen Datenbank darstellt, um auf einzelne Elemente dieser Datenbank schnell zugreifen zu können. Daher wurde eine zusätzliche Datenbank realisiert, die die Daten der interne Datenbank verwaltet. Neben dem Anspruch der multiplen Datenspeicherung, ermöglicht die interne Datenbank auch sinnvolle Querverknüpfungen innerhalb zweckmäßiger Kategorien.

Die Definition zahlenmäßig limitierter Kategorien erfüllt den Zweck einer Strukturierung der Daten in thematisch und mathematisch sinnvolle Untergruppen. Demgemäß können alle Daten (sowohl aus den drei Datenblättern als auch die Ergebnis-Verknüpfungen) über unterschiedliche Suchfoki erhalten werden (siehe oben).

In die Datenbank können je nach Versuchsdichte und -vielfalt Daten aus unterschiedlichen Experimenten mit verschiedenen Microarrays einfließen, die sich in den immobilisierten DNA-Fragmenten unterscheiden. Alle Daten, die mit identischen Microarrays erzeugt wurden, gehören einer Microarray-Familie an. Wenn die Chips im Rahmen unterschiedlicher Experimente verwendet wurden, stellen sie also multi-konditionale Experimente dar, die mit identischen Chips generiert wurden. Jede Microarray-Familie ist demnach mit einem spezifischen Satz an Gen-Annotationen ausgestattet, der das entsprechende Spotting-Schema widerspiegelt. Die oftmals interessierenden Gene sind also über die Microarray-Familie an das experimentelle Daten-System mit allen weiteren Einzelinformationen der Experimente gekoppelt. So bedingt die über das experimentelle Design definierte Auswahl des Chips einer Microarray-Familie die auf den Microarrays vorkommenden Gene. Die Suche nach einem bestimmten Gen setzt daher in der Regel die Kenntnis der zugehörigen Chip-Familie voraus, so dass Gen-Annotationen über die Auswahl der entsprechenden Microarray-Familie (*Chip*) abgerufen werden können. Nach der Auswahl einer Microarray-Familie können also einzelne Gene innerhalb dieser Familie angewählt werden.
Gen-Annontationen als Suchkriterium spielen eine herausragende Rolle, da experimentelle Daten aus unterschiedlichen Experimenten zu diesem Gen verglichen werden können und andererseits zusätzliche Informationen aus externen Datenbanken (je nach vorhandenen Informationen: *KEGG* und / oder *Multifun*) präsentiert werden, so dass - falls vorhanden - zusätzliche explizite Informationen zu den interessierenden Genen abgerufen werden können. Die Suche über die Microarray-Familie liefert also nicht nur alle der Microarray-Familie angehörigen Einzelexperimente der multikonditionellen Experimente, sondern über ein separates Excel-Datenblatt auch alle verfügbaren Informationen zu den Genen der Microarray-Familie. Diese übergeordnete Suche der Gene über die Chip-Familie verhindert durch Namensredundanz auftretende Missverständnisse bei der Auswahl eines Genes, das mit ähnlicher Annotation auf unterschiedlichen Chips (in unterschiedlichen Organismen) vorkommen kann.

Die Ergebnisse zu den Genen liegen als prozessierte Daten in Form von vereinfachten Regulationsangaben (differentielle Expression) und Expressionswerten vor. Sie enthalten aber auch weitere Informationen, die bezüglich der Bewertung der Experimente (z.B. *p*-Werte) und für spätere Veröffentlichungen von Bedeutung sind (z.B. *GEO*-Angaben).
Sowohl die experimentellen Annotationen als auch die Ergebnisse werden in einer gemeinsamen experimentspezifischen Kategorie registriert, die für jede Messung die Hybridisierungs- und sonstige experimentelle Bedingungen so wie alle weiteren multi-konditionellen Bedingungen speichert, im Rahmen derer die interessierende Messung vorgenommen wurde. Das experimentelle System wurde so programmiert, das sie die Schnittstelle und somit Hauptverwaltung der internen Datenbank darstellt. Sie verknüpft die Ergebnisse der Versuche mit den experimentellen Annotationen, so dass explizite Charakterisierungen der Messungen mit spezifischen Expressionsmustern vorgenommen werden können.

Für die Einzelinformationen zu den Genen werden also tabellarische Ordnungssysteme benötigt. Im Programm werden für die Speicherung unterschiedlichste Datentypen bereitgestellt, die einer Kategorie zugeordnet werden können, so genannte *structures* bereit. Diese *Arrays* umfassen „Datencontainern“ ähnliche Feldern, denen Namen zugeteilt werden können. So wurde zum Beispiel für jedes Gen eine *structure* implementiert, die alle relevanten Informationen bezüglich dieses Gens in Unterkategorien verschiedenster Datentypen speichert (siehe Tabelle 6.1).

Wie in Tabelle 6.1 gezeigt, enthält das jeweilige, nach der Gen-ID benannte *Array* Informationen, die bereits vor der Datenauswertung bekannt sind. Diese Daten werden aus den nach dem Scanprozess durch GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) generierten primären Dateien (Tabulator-getrennte Textdatei im *Gpr*-Format) extrahiert. Dazu gehören Auskünfte zu der Struktur des Microarrays (Lokalisation der Spots, Spotnummer und Anzahl der Genreplikate) und den darauf enthaltenen Genen (GeneID und Beschreibung). In die Gen-Annotationen fließen aber auch Informationen ein, die erst nach Durchlauf der Auswertung zur Verfügung stehen. So werden im Laufe der Analyse der Microarrays frei verfügbare Datenbanken wie *KEGG* (und die *E.coli*-spezifischen Datenbank *Multifun*, wenn es sich bei dem untersuchten Organismus um das Bakterium handelt) mit den Accessionnummern der Gene auf die Existenz entsprechender Einträge abgesucht und die betreffenden regulatorischen Pfade, in denen das Gen beteiligt ist, angegeben.

Der Anwender kann neben den Genen, die wegen ihrer zahlreichen Detailinformationen in separaten *structure* gespeichert werden, aber auch alle möglichen weiteren Informationen zum Experiment (und den Ergebnissen) erhalten. Das Programm zur internen Datenbank-Recherche berücksichtigt die entscheidenden Suchkriterien, die während einer übergeordneten Suche sinnvoll sind (siehe Abbildung 6.4).

Bei der Auswahl eines der in Abbildung 6.4 aufgeführten Suchkriterien, erscheint in dem unteren Eingabefeld automatisch alle zu diesem Begriff möglichen Auswahloptionen. Wenn die interne Datenbank beispielsweise auf alle Microarray-Experimente hin durchsucht werden soll,

Tabelle 6.1 *– Beispiel einer Gen-Bezeichnung für das Gen GFP aus Escherichia coli (E.coli). Die zu den Genen vorhandenen Informationen werden in der Datenbank in structures hinterlegt. Diese structures können unterschiedliche Datentypen einschließen, von denen hier einige beispielhaft dargestellt sind. Bei den hier gezeigten Geninformationen handelt es sich um die Gen-ID, den Organismus, den Gennamen, die Anzahl der Genreplikate und die Spotnummer auf dem Chip, eine ggf. vorhandene Beschreibung des Gens, die Accessionnummer und den KEGG-Eintrag.*

Gen: glcC		
Feldname	**Datentyp**	**Feldinhalt**
id	char()	glcC
organism	text	*E.coli*
genname	text	DNA-binding transcriptional dual regulator
replicates	int8	5
spotno	int8	10
description	text	GlcC transcriptional dual regulator
acc	char()	P52072
kegg	text	Ebene 1: Metabolism
		Ebene 2: Carbon compound utilization
		Ebene 3: Trehalose degradation

die von einem bestimmten Benutzer durchgeführt worden sind (Suchkriterium *Username*), so werden in dem unteren Feld die Namen all der Benutzer aufgeführt, die bereits Microarray-Versuche abgeschlossen haben und daher bereits in der Datenbank geführt werden.
Die Ergebnisse der Datenbank-Recherche werden dann unmittelbar unterhalb der Suchmaske in einem Ausgabefeld wiedergegeben. Sie befinden sich in Form der vollständigen Dateinamen der entsprechenden Versuche in einem *Pop-up Menu* und können entweder als Übersicht gespeichert werden oder direkt aus dem *Pop-up Menu* geöffnet werden.

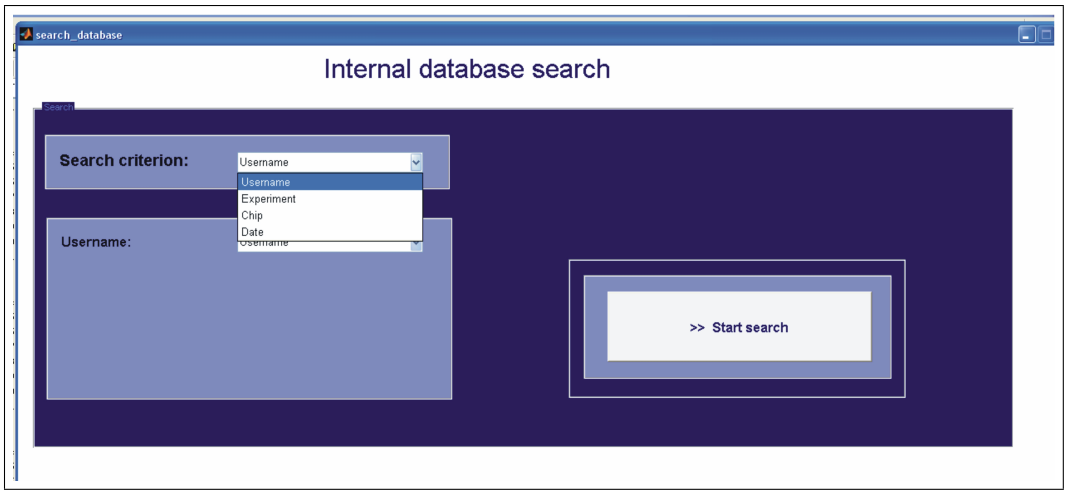


Abbildung 6.4 *– Benutzeroberfläche zur Datenbank-Recherche. Die Eingabe ermöglicht die Suche nach bestimmten Kriterien. Alle Versuche, die von einem Experimentator durchgeführt wurden oder einen bestimmten Experimentnamen tragen, bzw. die mit einer Microarray-Familie ausgeführt worden sind oder in einem bestimmten Zeitraum können in der internen Datenbank durchsucht werden.*

Bezüglich der experimentellen Angaben existiert also für jeden Versuch eine eigene Excel-Datei. Auf diese Weise sind alle zu einem Experiment vorhandenen Informationen konzentriert in einer Datei angegeben, um so einerseits eine bessere Systematik und andererseits einen erleichterten Datentransfer zu gewährleisten. Die aufgrund der Datenvielfalt notwendige Übersichtlichkeit über die Einzelinformationen soll durch die inhaltliche Gruppierung der Daten in

1. **Gen-Annotationen**
2. **Experiment-Design**
   - experimenteller Aufbau und Versuchsdurchführung
   - technische Angaben
3. **Ergebnisse**
   - Ergebnisse
   - GEO-Datenblätter
   - Chip-Qualitäten
   - Erklärung zu den Ergebnissen

erfolgen. Diese Informationen befinden sich in separaten Tabellenblättern, die entsprechend ihres Inhalts benannt sind.
Wie schon für die Gen-Annotationen erläutert, befinden sich auch die Daten für die Experimente sowie die Ergebnisse der Auswertung der Versuche, die an die Excel-Dateien übergeben werden, thematisch sortiert in *structures*. Auf diese Weise können gezielt, definierte Pakete innerhalb der Auswertung verschoben werden. So erfolgt beispielsweise die Übergabe der für die Analyse notwendigen technischen Angaben an die Auswerte-Software in Form einer *structure*.

Innerhalb der *structures* befinden sich nun also die umfangreichen Ergebnisse für die Suche. Um die Experimente jedoch verwalten und später ohne großen Zeit- und Speicheraufwand während der Datenbank-Recherche wiederfinden zu können und entsprechend dem jeweiligen Suchkriterium zuordnen zu können, werden die für die Recherche notwendigen Details in einem separaten Zwischenspeicher abgelegt. Dieser Zwischenspeicher ist mit den *structures* verlinkt und wird mit jedem neuen Experiment um die dazu erforderlichen Daten erweitert. Zu diesen Daten zählen alle bereits vorhandenen Angaben, die in den Suchkriterien definiert sind (Benutzername, Experiment, Chiptyp und Datum). Die Arrays mit diesen Daten enthalten zudem einen Verweis auf das Experiment, zu dem die Daten gehören. Um die Verweise eineindeutig anlegen zu können (da jedes Suchkriterium mehrfach vorkommen kann), werden die Experimente analog zur Reihenfolge, in der sie ausgeführt wurden, durchnummeriert. Der Verweis wiederum ist mit der Ergebnisausgabe der Suche, dem vollständigen Dateinamen für die Excel-Datei, in dem alle Daten gespeichert werden, verknüpft.

Wie bereits beschrieben, werden alle während der Erstellung der Datenblätter gemachten Angaben und die im Laufe der Auswertung berechneten Ergebnisse durch Übergabe des Dateinamens in die entsprechende Exceldatei geschrieben. Der Dateiname setzt sich aus dem Pfad und dem Namen der Datei „Results“ zusammen. Der Pfad setzt sich aus unterschiedlichen automatisch generierten Ordner zusammen: Mit der Anwendung des Programms wird die Angabe eines Zielordners erfragt. In diesen Zielordner werden automatisch dann Verzeichnisse für *low-density* und *whole-genome*-Experimente angelegt: jeweils zwei für die Ergebnisse und zwei für die Speicherung von Zwischenergebnissen zur Sicherung der Daten. In diesen Verzeichnissen werden mit jedem neuen Benutzernamen Unterverzeichnisse mit dem Namen des Benutzers angelegt. Die letzte Unterebene in den Namens-Verzeichnissen beinhaltet Ordner mit dem Datum, in denen die jeweilige Ergebnisdatei bzw. die Daten aus dem Zwischenspeicher wiederzufinden sind. Die Wahl des Datums als unterstes Verzeichnis innerhalb des Verzeichnisbaums sollte die Sortierung der Daten erlauben. Die Ergebnis-Zwischenspeicherdateien werden also in einen datierten Ordner geschrieben, der sich in einem nach dem Experimentator (Abfrage *Username* siehe Abbildung 6.2) benannten Ordner befindet, der wiederum in einem Ordner wiederzufinden ist, der den Chiptyp beschreibt.

Der Inhalt der Dateien wird also während des Durchlaufs des Programmes stetig um neue Tabellenblätter erweitert. Die zuletzt angelegten Tabellenblätter enthalten die in der Liste (siehe oben) angegebenen Ergebnisdaten. Dazu zählen einerseits Erklärungen zu den Ergebnissen, die die Auslegung und das Verständnis der Daten erleichtern und die Zuverlässigkeit der Ergebnisse beschreiben. Des Weiteren werden Tabellenblätter angelegt, die alle für die Veröffentlichung der Experimente in *GEO*[111] erforderlichen Daten enthalten sowie Angaben über die Qualität der einzelnen Chips.
Abschließend werden Tabellenblättern angelegt, die die eigentliche Fragestellung des Experiments nach der differentiellen Expression zwischen unterschiedlichen Zuständen beantworten.

Zusätzliche Daten, wie die Expressionsstärke der Einzelzustände, die Signifikanz der Regulation, die Anzahl ungeflaggter Spots pro Zustand, usw. ermöglichen ein tieferes Verständnis der Daten in Kombination mit Hintergrundinformationen zu diesen Größen. Eine allgemeine Zusammenfassung über die Regulation aller verglichenen Zustände mit farbiger Sortierung der Daten nach Signifikanz der differentiellen Expression soll zu einem schnellen Überblick der Ergebnisse verhelfen. Die Speicherung der Ergebnisse der Auswertung in der Excel-Datei stellt den letzten Schritt des entsprechenden Microarray-Experiment dar.
Diese Vielzahl an Daten, die während der Auswertung generiert werden, können letztlich alle über die Suchfunktion erhalten werden, so dass die Details und Ergebnisse von unterschiedlichen Experimenten miteinander vergleichen können.

6.2.1.3 Fazit

Vor dem Hintergrund der Notwendigkeit einer detaillierten Dokumentation des experimentellen Designs und der Durchführung von Microarray-Experimenten wurden elektronische Datenblätter entwickelt, die mithilfe automatisierter Abfragemasken die Eingabe aller Einzelheiten zu den Chipversuchen erleichtern. Die auf das multivariate Anforderungsprofil der Microarrays ausgerichteten Datenblättern werden einer internen Datenbank übergeben. Die für die Analyse der Experimente notwendigen technischen Informationen aus den Datenblättern werden zusammen mit den numerischen Ergebnissen aus dem Scanvorgang (*Primäranalyse*) an die weitere Datenauswertung überliefert.
Die Datenbank beinhaltet somit Details zu den Experimenten sowie die Ergebnisse aller durchgeführten Experimente und kann jederzeit anhand verschiedener Suchkriterien auf bereits durchgeführte Microarray-Experimente durchsucht werden.

6.2.2 Qualitätsanalyse der Microarrays

6.2.2.1 Hintergrund und Anforderungen

Prinzipiell kann mit den Angaben aus dem ersten Programm-Abschnitt, also allen versuchsrelevanten Informationen aus den elektronischen Datenblättern sowie den entsprechenden Pfaden zu den numerischen Daten aus der Bildanalyse, die Auswertung der Microarray-Experimente begonnen werden. Eine wissenschaftlich fundierte Analyse dieser multivariaten Experimente sollte jedoch nicht nur quantitativer, sondern auch qualitativer Natur sein. Dementsprechend sollten zunächst alle vorliegenden Daten eines Microarray-Experiments herangezogen werden, um eine Bewertung der an dem Versuch beteiligten Microarrays vornehmen zu können. Die zahlreichen, sich zu einem komplexen Qualifikationsprofil zusammensetzenden Einzelaspekte der Qualitätsanalyse vermitteln dem Benutzer eine erste Beurteilung der Güte der Einzel-Chips und der damit verbundenen Signifikanz der Ergebnisse. Zudem sollte die Festlegung bestimmter Gütekriterien die Eliminierung unzureichender Microarrays ermöglichen, um zu verhindern, dass die von solchen Chips stammenden Informationen die Ergebnisse verfälschen.
Die Zuordnung der Qualitätsanalyse innerhalb der Datenanalyse ist in der Literatur nicht

einheitlich beschrieben. Während einige Arbeitsgruppen die Qualitätsanalyse vor der Vorverarbeitung der Daten als separaten Schritt der Analyse bewerten, gliedern andere Arbeitsgruppen die Qualitätsüberprüfung der Chips in die Vorverarbeitung der Daten (siehe Kapitel 6.2.3) ein. In dieser Auswerte-Software erfolgt die Qualitätsanalyse während der Vorverarbeitung der Daten. Die Ermittlung der entsprechenden Gütekriterien für die Qualitätsanalyse, die dann in jeder weiteren Qualitätsüberprüfung appliziert werden, wird in diesem Kapitel separat vorgestellt.

6.2.2.2 Durchführung und Ergebnisse

Die Notwendigkeit einer Qualitätsanalyse vor der Durchführung einer quantitativen Datenauswertung wurde einerseits durch die visuelle Charakterisierung und Sondierung der Spots auf mehreren *low-density*-Microarrays überprüft und die Ergebnisse wurden andererseits mit der zu diesem Thema vorhandenen Literatur verglichen. Anhand der eigenen Beobachtungen und der in der Literatur beschriebenen Aspekte wurden Basis-Gütekriterien entwickelt, auf deren Grundlage die Chips beurteilt werden können (siehe Seite 58).

Obgleich sich die Mehrzahl der Veröffentlichungen im Bereich der statistischen Microarray-Analysen bisher vornehmlich mit der quantitativen Auswertung von *whole-genome*-Chips befasst, widmen sich einige Autoren der Probleme, die bei der Umwandlung der beim Scannen erzeugten Bilder (*tif*-Dateien) in numerische Informationen (Primärauswertung) entstehen. Dabei handelt sich sowohl um Variabilitäten einzelner Spots als auch teilweise daraus resultierende, sich auf den ganzen Chip erstreckende Variabilitäten. Der immense Einfluss der Variabilitäten (die bei der Herstellung der Microarrays, während der RNA Extraktion, der Amplifikation und des Labelings, des Hybridisierens und Scannens[112] entstehen) auf die Meßwerte der Gen-Expression (siehe Kapitel 5.4) ist verhältnismäßig oft beschrieben.[112–114] Methoden, die jedoch eine Aussage über die physikalisch messbare Qualität der einzelnen Microarrays erlauben, sind bis dato kaum verfügbar.
Sofern technische Replikate der Microarrays vorliegen, kann zwar eine erste vergleichende Aussage über die Güte der Microarrays getroffen werden,[115] da dies jedoch erstens kostenintensiv ist und zweitens überdies bei nur einem technischen Replikat eine objektive Entscheidung getroffen werden muss, welches der Microarrays die Gütekriterien besser erfüllt, sind zusätzliche Maßnahmen erforderlich. Es werden also erstens allgemein gültige Annahmen gesucht, die für den kompletten Chip gelten. Außerdem wird nach statistischen Faktoren gesucht, die Unregelmäßigkeiten der Microarrays beschreiben, um diese gegebenenfalls korrigieren zu können

Die Notwendigkeit einer Qualitätsanalyse wurde zunächst anhand von Abweichungen von allgemeingültigen Annahmen für Microarray-Signale validiert. Idealerweise sollten die Bandbreite aller Intensitätswerte über das Microarray zufällig verteilt sein ohne systematische oder räumliche Muster aufzuweisen. Die geographische Lage eines Spots sollte also nicht die Stärke der Intensität beeinflussen. Um die Richtigkeit dieser Annahme abzuwägen, wurden die Bilddateien (*tif*-Files) unterschiedlicher Microarrays aus *E.coli*-Experimenten untersucht. Bei

diesen Untersuchungen stehen immer die selbst-gespotteten Microarrays im Vordergrund, da die Auswertung der Microarray-Daten vorrangig auf diese *low-density*-Chips ausgerichtet ist. Zu diesem Zweck wurde das Programm *Findspot* implementiert, das der vereinfachten Visualisierung der Bilder anhand gezielter lokaler Such- und Vergleichsoptionen dient. Mithilfe von *Findspot* (siehe Abbildung 6.5) können ausgewählte Spots auf den Microarrays (beispielsweise alle Replikate eines Gens) oder größere zusammenhängende Flächen (beispielsweise ein Block bzw. ein Quartal, aber auch der ganze Chip) visuell analysiert werden. Um die Spots möglichst umfassend charakterisieren zu können, werden verschiedene Ansichtsoptionen zur Verfügung gestellt. So kann die Perspektive des Betrachters beliebig gewählt werden, was einerseits durch manuelles Drehen der Bilder, andererseits durch die digitale Eingabe der gewünschten Perspektive realisiert werden kann. Außerdem können die Spots wahlweise im Konturplot oder dreidimensional dargestellt werden. Auf diese Weise kann eine erste optische Übersicht über die Dateien erhalten werden, die bisherige visuelle Darstellungsoptionen der Spots (zweidimensional oder in Form von Histogrammen) übertrifft.

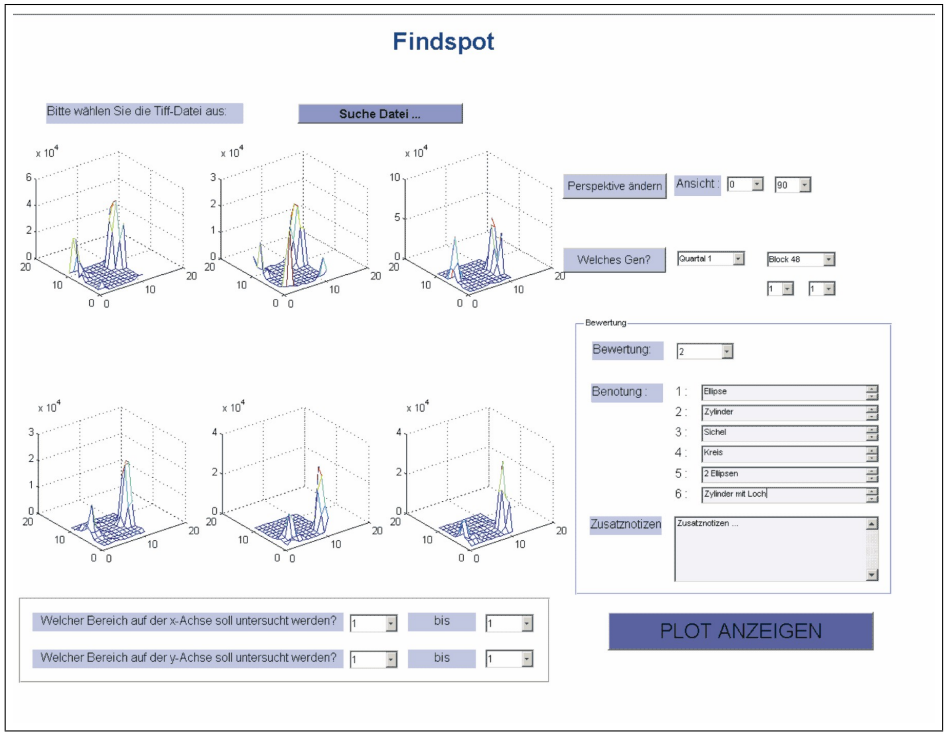


Abbildung 6.5 – *Benutzeroberfläche des Programms Findspot. Findspot dient der visuellen Analyse der tif-Dateien. Das Programm ermöglicht die globale Ansicht über den ganzen Chip. Darüberhinausgehend können - wie hier gezeigt - einzelne Abschnitte auf dem Chip angewählt werden, um so beispielsweise die Genreplikate nebeneinander zu stellen. Außerdem kann die Perspektive geändert werden und zwischen Kontur- und dreidimensionaler Einstellung gewechselt werden. Bei Bedarf kann das Bild gespeichert und mit einer selbst-definierten Bewertung versehen werden.*

Im Folgenden wird zunächst kurz der Aufbau und der Druckvorgang der vier untersuchten *low-density*-Chips beschrieben. Dieser für spezifische Sekretionsforschungen konzipierte Microarray trägt 109 Gene mit je fünf Genreplikaten (ein Original und fünf Genreplikate ergeben sechs Spots pro Gen). Die Spots auf den Microarrays sind in Quartalen angeordnet,

die wiederum in sechs identische Blöcke (entsprechend der fünf Genreplikate) unterteilt sind (siehe Tabelle 6.2).

Tabelle 6.2 – *Anordnung der Blöcke in Quartalen auf dem low-density-Microarray*

Quartal I	
Block 4	Block 1
Block 5	Block 2
Block 6	Block 3

Quartal II	
Block 4	Block 1
Block 5	Block 2
Block 6	Block 3

Quartal III	
Block 4	Block 1
Block 5	Block 2
Block 6	Block 3

Quartal IV	
Block 4	Block 1
Block 5	Block 2
Block 6	Block 3

Jeder identische Block eines Quartals auf den *low-density*-Chips setzt sich je aus 24 Genen in einer 4 x 6-Anordnung zusammen.
Ein Quartal wurde je durch einen Druckkopf bedient und jeder Spot wurde durch fünfmaliges Durchstechen des Meniskus und Aufsetzen der Nadel aufgetragen.
Um die Positionsabhängigkeit der Spots untersuchen zu können, sollte der Vorgang des Spottens bekannt sein: In einem Spotting-Prozess werden in der Regel mehrere Microarrays mit identischem Spotmuster zugleich gedruckt (Chips aus einer Microarray-Familie - siehe Kapitel 6.2.1) um Kosten zu sparen. Dabei werden immer drei Spots (spezifisches Gen und zwei Genreplikate) aus drei der sechs Blöcke („Triplets“) auf einem Microarrays aufeinanderfolgend gedruckt (zum Beispiel Position 1,1 aus der 4 x 6-Anordnung der Blöcke 1,2 und 3), bevor die identischen Genreplikate mit analoger Position auf den restlichen Microarrays gespottet werden (zum Beispiel zunächst je ein Gen und zwei Genreplikate aus Block 1-3). Dieser Vorgang (je drei Spots eines Blocktriplets auf allen Microarrays) wiederholt sich bis alle Spots gedruckt sind.

Für die Untersuchung der Annahme der Gleichverteilung der Intensitätswerte identischer Gene über den Chip wurde in einer ersten statistischen Auswertung der Bilddateien die numerischen Daten der entsprechenden Spots aus der Primärauswertung mit GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) berücksichtigt, um gegebenenfalls eine Intensitätsabhängige Aussage über die Spots treffen zu können. Diese Daten wurden zunächst für eine großflächigere Analyse der Intensitätsverhältnisse verwendet, in dem die Intensitätswerte aller Spots eines Blocks aufsummiert wurden. Aufgrund der lokalen Verteilung der Gen-Replikate in den Blöcken (je Quartal sechs Blöcke mit identischem Spottingmuster), können mittels der

Block-Summen, die innerhalb eines Quartals identisch sein sollten, Abweichungen von den erwarteten einheitlichen Intensitätssummen detektieren werden.
Tabelle 6.3 zeigt beispielhaft die Intensitätswerte eines Microarrays. Die Daten der anderen drei *low-density*-Microarrays sind in Anhang C hinterlegt.

Tabelle 6.3 *– Vergleich der aufsummierten Intensitätswerte der Blöcke eines Microarrays. Die jeweils dick markierten Summen repräsentieren die größere Summe der beiden nebeneinander gelegenen Blöcke.*

Quartal I		
90	<	**133**
109	<	**139**
137	<	**149**

Quartal II		
130	<	**132**
126	>	113
139	>	109

Quartal III		
57	<	**77**
45	<	**71**
35	<	**40**

Quartal IV		
81	>	77
77	>	70
68	>	64

Demzufolge tendieren die im Zentrum der vier *low-density*-Microarrays gelagerten Spots zu höheren Intensitätswerten (dies gilt für 12 der 16 ausgewerteten Quartale). Dieser Trend manifestiert sich in horizontaler Richtung (im Vergleich je zweier nebeneinander gelegener Block-Summen eines Quartals - wie durch die Pfeile in Tabelle 6.3 angedeutet), wird vor allem aber in in vertikaler Richtung deutlich.
Solche abnehmenden Tendenzen der Intensitäten zum Rand der Microarrays hin können durch Hybridisierungseffekte zustande kommen. Wie bereits vielfach in der Literatur beschrieben, spielen Verdunstungseffekte bei Microarrays eine große Rolle.[116,117] Dies hat bereits zu der Entwicklung neuer Techniken mit dem Ziel der Vermeidung dieser aus der Evaporation resultierenden lokalen Ungleichheiten geführt.[118] Die auf den *low-density*-Microarrays beobachteten lokalen Intensitätsschwankungen resultieren daher wahrscheinlich ursächlich aus eben diesen Verdunstungseffekten. Ebenfalls verantwortlich für die genannten Effekte könnten Unterschiede in der Größe der Radien der Fängermoleküle (*probes*) auf dem Microarray sein. Mit größerem Radius der Spots nimmt auch die Wahrscheinlichkeit größerer Intensitätswerte zu, da den Probenmolekülen (*targets*) aufgrund der größeren Radien eine größere Oberfläche und somit eine erhöhte Anzahl an Fängermolekülen zur komplementären Bindung (Hybridisierung) geboten wird. Das Programm *Findspot* wurde genutzt, um die Microarrays auf diese These hin zu analysieren. Die Untersuchung ergab, dass die Größe der Spots tendenziell zur Mitte des Microarrays hin zunehmen.[121]
Die Verteilung ungleich großer Spots über den Chip stellt ein bereits von Zhang *et al.* beschrie-

benes Phänomen der Microarray Experimente dar.[119] Die Annahme der lokalen Konstanz der Größe der Spots über einzelne Microarrays wurde auch durch Kim *et al.* widerlegt.[120] Als Ursachen für unterschiedliche Spotgrößen kommen beispielsweise Präzipitate, Unreinheiten und Ablagerungen in der zu druckenden Lösung in Frage. Als weitere Aspekte sind ein nicht adäquater Druckkontakt mit der Oberfläche des Chips, beschädigte oder dreckige Pins bzw. Microarrays oder eine nur unzureichende aufgenommene Druck-Flüssigkeit denkbar.[120] Unreine oder defekte Pins können in diesem Fall jedoch ausgeschlossen werden, da die Spots, die jeweils durch einen der vier Pins aufgetragen wurden, keine einheitlich mangelhafte Form aufwiesen. Große Spots, wie sie eher mittig der untersuchten Microarrays zu finden sind, können aufgrund erhöhter Feuchtigkeitsbereich oder einer unebenen Beschichtung zustande kommen.[119]
Die auf visuellen und numerischen Untersuchungen beruhenden Befunde der zur Mitte der *low-density*-Microarrays zunehmenden Spotgrößen und Intensitäten ergänzen einander daher insofern, als dass einander bedingende Feuchtigkeitseffekte (ein potenziell erhöhter Feuchtigkeitsbereich in der Mitte[119] und Verdunstungseffekte am Rand der Microarrays[116,117]) ursächlich für die beobachteten Resultate verantwortlich gemacht werden könnten. Diese beiden Effekt komplettieren einander also womöglich und resultieren daher in der beobachteten Abweichung vom Idealfall der geographisch konformen Intensitätsverteilung der Spots.
Im Allgemeinen dient eine uneinheitliche Größenverteilung der Spots vorrangig als Hinweis auf Probleme während der Herstellung der Microarrays.[119]

Es wurde also zunächst die globale (über den Chip verteilte) Intensitäts- und Größenverteilung mehrerer Spots untersucht. Wie erwähnt schlägt sich aber die Qualität des Chips bei Microarray-Experimenten auch auf mikroskopischer Ebene in der Qualität der Einzel-Spots selbst nieder, da über die entsprechenden Intensitätswerte die für die Experimente entscheidenden Expressionsdaten ermittelt werden. Daher müssen zunächst Qualitätsmerkmale der individuellen Spots beispielhaft überprüft werden, um daraus Rückschlüsse über die Notwendigkeit einer Qualitätsanalyse der *low-density*-Microarrays im Ganzen treffen zu können. Daher wurde das Aussehen der Einzelspots in Hinblick auf eine charakteristische einheitliche Form aller Spots ermittelt, die eine Voraussetzung für eine valide Auswertung ist. Dafür wurde die Gesamtheit aller Spots im Einzelnen mit *Findspot* betrachtet und verglichen.
Da bisher keine dreidimensionale Beschreibung typischer Spotformen vorliegt, wurden die Microarrays zunächst dahingehend untersucht. Der Vergleich individueller Spots ergab, dass keine für alle Spots charakteristische uniforme Spotform existiert. Vielmehr existiert eine breite Diversität an Formen, von denen die charakteristischsten Grundtypen selektiert wurden. Diese sechs Spotformen sind in Abbildung 6.6 gezeigt.

Die auf dem Microarray auftretenden Spotformen sind in Abbildung 6.6 können von oben nach unten wie folgt charakterisiert werden: umrandete Ellipse, Zylinder, umrandete Sichel, Donut, zwei umrandete Ellipsen und Zylinder mit Loch. Die Häufigkeiten der einzelnen Spotformen sind im Anhang D angegeben.
Die Suche nach beispielsweise Intensitätsabhängigen Tendenzen beim Auftreten unterschiedlicher Spotformen ergab keine einheitliche Aussage über bestimmte Spotformen in Abhängig-

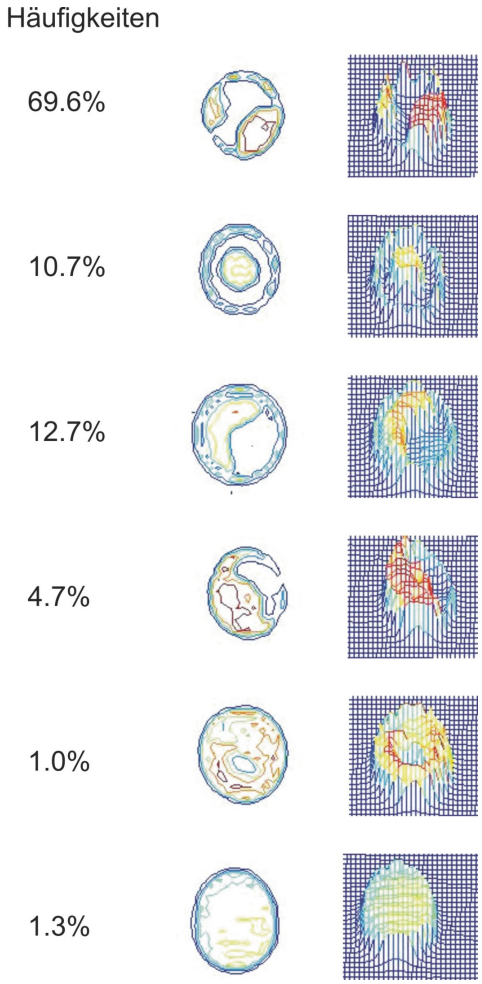


Abbildung 6.6 – *Beispiele für charakteristische Spotformen aus vier low-density-Microarrays mit je 109 E.coli-Genen und 5 Replikaten. Die Abbildung zeigt die Spotformen, die die Mehrheit der Spots auf den Microarrays repräsentieren. Auf der linken Seite der Abbildung sind die Häufigkeiten der einzelnen Spotformen angegeben. Es wurde bei einer Laserstärke von 100% und einer PMT-Verstärkung von 500 entsprechend der von Zhang et al. definierten Kriterien gescannt.[119] Die linke Abbildungen zeigen je den Konturplot, die rechten zeigen die dreidimensionale Darstellung der Spotformen.*

keit der Spot-Intensitäten. Des Weiteren stehen die gezeigten Spotformen stellvertreten für sechs Gruppen in sich wieder teilweise heterogener Unterformen. So variiert beispielsweise die Ausrichtung und Größe sowie die Dichte der Spots, die den in Abbildung 6.6 gezeigten Spot-Kategorien untergeordneten werden können.

Die mit 69,6% der 1434 untersuchten Spot dominierende Spotform ist die umrandete Ellipsenform. Sie tritt sowohl bei hohen, wie auch bei mittleren und niedrigen Intensitäten auf. Wie auch alle anderen charakteristischen Spotformen weisen sie einen eindeutigen Rand auf, was auf die Herstellungsmethode der Microarrays zurückzuführen ist. Das bei der Herstellung dieser Microarrays zum Einsatz kommende Ring & Pin-System verwendet solide Nadeln, welche nach dem Kontakt mit dem Trägermaterial einen Teil der Oligonukleotid-Lösung verlieren. Ein Teil der Druck-Lösung kann andererseits auch während des Aufsetzens der Nadeln nach außen verdrängt werden, was die Bildung eines Flüssigkeitsrings um die Nadel zur Folge hat. Dieser Bereich wird beim Abheben der Nadel nicht wieder abgetragen und kann somit zur

Bildung eines ausgeprägten Rands führen.
Obgleich alle Spots einen unterschiedlich ausgeprägten Rand aufweisen, sind die Formen in sich sehr heterogen. Es ist neben der erwähnten nicht erkennbaren Intensitätsabhängigkeit auch keine lokale Ansammlung bestimmter Formen zu erkennen. Auch die von Räke vorgenommene Untersuchung des Einflusses der Druckreihenfolge lieferte keine einheitlichen Daten.[121] In dieser Arbeit wurde auch eine visuelle Untersuchung eines kommerziell erworbenen *whole-genome*-Microarrays (ebenfalls für *E.coli*-Bakterien) vorgenommen und mit den *low-density*-Daten verglichen. Diese Analyse ergab eine im Vergleich zu den *low-density*-Microarrays noch höhere Inhomogenität der Spotformen. Diese Beobachtung kann vermutlich durch das unterschiedliche Herstellungsverfahren (Kontaktverfahren) erklärt werden, dass in der Literatur als weniger valide gegenüber dem Ring & Pin-System beschrieben wird.

Die Analyse der individuellen Spots deutet auf multiple Varianzen zwischen den auf einem Chip befindlichen Spots hin, die sich in unterschiedlichen Größen der Spots, Abweichungen von der zylindrischen Form der Spots im Inneren und einer Nicht-Zirkularität bezogen auf den äußeren Kreis manifestierten. Basierend auf einer mit dem Programm *Findspot* durchgeführten Analyse von vier *low-density*-Microarrays wurden übereinstimmend mit Bylesjö *et al.* technische Kernpunkte definiert, welche die Spotqualität signifikant beeinflussen. Dabei handelt es sich um die in der folgenden Liste wiedergegebenen Punkte.[122]

- **Signalintensität**:
  Vorkommende schwache Signal sollten aus einer physiologisch schwachen Expression resultieren, können aber auch mit der Oberflächenbeschaffenheit der Chips, Signalbleichen (*signal bleaching*), Scannerproblemen oder unvollständiger bzw. ungleichmäßiger Hybridisierung in Verbindung gebracht werden.

- **Intensitätsverteilung innerhalb der Spots**:
  Die Gleichförmigkeit der Spots bezüglich ihrer Intensitätsverteilung über die Pixel innerhalb eines Spots ist oft nicht gegeben. Starke Abweichungen einiger Pixel von dem durchschnittlichen Intensitätswert können die Konsequenz nicht-spezifischer Bindungen oder unregelmäßiger Verteilungen der gedruckten DNA auf dem Chip sein.

- **Morphologie**:
  Formabhängige Variationen des Vordergrundbereichs des Spots können in Form besonders großer oder kleiner Spotgrößen, geringer Spot-Zirkularität oder als Spot-Vermengungen auftreten. Hierfür können beispielsweise Präzipitate oder Unreinheiten der Drucklösung verantwortlich sein.

- **Hintergrundintensitäten**:
  Variierende lokale Hintergründe der Spots über den Chip kommen meistens durch Kontaminationen von unspezifischen Bindungen oder einem unvollständigem Waschvorgang zustande.

Eine nähere Untersuchung der lokalen Ungleichheiten wird in Kapitel 6.2.3 vorgenommen, da die Vorverarbeitung der statistische Auswertung der Microarray-Experimente den lokalen Hintergrund der Spots berücksichtigt.
Der Zusammenhang zwischen der ermittelten Spotqualität und der Qualität des gesamten Microarrays wird durch die von Kim *et al.* deklarierten fünf Qualitätsmaßgaben deutlich, die die Qualität des Microarray-Bildes repräsentieren und weitestgehend mit genannten Kriterien (siehe oben) für die Spotqualität übereinstimmen. Dabei handelt es sich um das Signal- (bedingt durch die oben genannten Einflussgrößen) und Hintergrundrauschen, die Einheitlichkeit der Größe und der Form der Spots sowie die Position der Spots auf den Chips.
Die Ergebnisse der Qualitätsüberprüfung sowohl auf der globalen Ebene (erster Teil der Untersuchung) wie auch auf der lokalen Spot-Ebene (im zweiten Teil der Untersuchung) weisen auf die unbedingte Notwendigkeit einer Qualitätsanalyse der Microarrays hin. Dieser Bedarf einer Qualitätsüberprüfung wird durch die dazu analog in der Literatur beschriebenen potentiellen Mängel von Microarrays unterstrichen. Als Konsequenz dieser Auswertung soll dem Benutzer eine Übersicht über die Qualität der verwendeten Microarrays vermittelt werden, um unter Umständen Microarrays unzureichender Güte von der weiteren Auswertung zu entfernen. Darüber hinausgehend sollte der Herstellungsprozess der Microarrays optimiert werden, so dass die durch dieses Verfahren auftretenden Schwankungen unter den Spots minimiert werden.

Der sicherste Weg jedoch, um fehlerhafte Ergebnisse der Microarray-Experimenten zu vermeiden, ist der Entstehung fehlerhafter Chips vorzubeugen, indem mögliche Ursachen für Qualitätsmängel der Microarrays im Voraus entfernt werden.
Da die Qualität der Microarrays vor allem technischer und prozeduraler Ursache ist, sollte dem Herstellungsverfahren der Microarrays besonderes Augenmerk gewidmet werden. Aufgrund der vielfältigen Einflussgrößen schon beim Herstellungsprozess, existieren bisher weitestgehend Ansätze zu Verbesserung von Einzelaspekten dieser Vorgänge. Draghici *et al.* haben versucht, diese Ansätze in einem Softwareprogramm zu vereinen, um diese dann in die Design-Entwicklung von *in silico*-Chips einfließen zu lassen.[123] Weitere Verbesserungsvorschläge für die Herstellung von Microarrays, die auch für die auf den hier untersuchten selbst-gespotteten *low-density*-Microarrays auftretenden Probleme gelten, können Zhang *et al.* und Schena *et al.* entnommen werden.[119,124] Diese Literatur befasst sich mit spezifischen Drucker-Eigenheiten, die in der Konsequenz einer bestimmten Bedienung bedürfen. Dazu gehört auch der für die Herstellung der hier verwendeten *low-density*-Microarrays gebräuchliche Axon 4000B Scanner (Axon Instruments, Foster City, CA). Weiterhin wird ein besonders für Drucker mit soliden Nadeln geeigneter, auf DMSO-basierender Puffer empfohlen, der Verdunstungseffekte erheblich vermindern soll. Außerdem sollte die Qualität der Nadeln unter dem Mikroskop untersucht werden, da diese zum Beispiel unter dem Drucken von Proteinmembranen erheblich leiden kann. Des Weiteren sollte die Sensitivität der Nadeln auf den Flüssigkeitspegel in den *Wells* überprüft werden, da manche Nadeln unterschiedliche Volumina an Oligonukleotid-Lösung benötigen. Des Weiteren kann die Oberfläche der Microarrays uneben sein, so dass vom Sensor

mehr Photonen registriert werden, was zu einem veränderten Hintergrundrauschen beiträgt. Demzufolge sollte die Oberfläche der Microarrays vor dem Druckvorgang auf ihre Planarität getestet werden.
Abgesehen von der Sensitivität der Nadeln, können auch Fänger-DNA-Moleküle (*probes*) unterschiedlich sensitiv auf die Oberfläche reagieren und entsprechend verschiedenartig haften, so dass unter Umständen so genannte *Drop-outs*, also nicht oder nur unzureichend bedruckten Spots, entstehen können. Daher sollte der Druckvorgang evaluiert werden, was einerseits durch einfaches Anpusten der Chip-Oberfläche geschehen kann (der resultierende Flüssigkeitsfilm zeigt bedruckte Spots als eindeutige Spots) oder durch das verlässlichere Färben der Chips mit einem Fluoreszenz-Farbstoff vor Beginn des Experiments.
Durch präzises Arbeiten und unter Befolgung der in der Literatur beschriebenen Verbesserungsvorschläge bezüglich des Herstellungsverfahrens sollten also möglichst viele der beobachteten Qualitätsmängel vermieden werden. Dennoch bleiben die Spots auch bei idealen Bedingungen aufgrund der vielen physikalischen und chemischen Einflussfaktoren hochgradig variabel, so dass in Abhängigkeit der Bild-Auswertesoftware bestimmte Formänderungen eingeführt werden oder unkorrigiert bleiben.[119]

Die Eliminierung fehlerhafter Daten kann aber nur auf der Basis einer verlässlichen Identifizierung unzureichenden Spots bzw. Microarrays durchgeführt werden. Dabei sollte die Eliminierung potentiell noch auswertbarer Spots/Microarrays vermieden werden, um einen unnötigen Datenverlust zu vermeiden.
Oftmals als fehlerhaft gekennzeichnete Spots sind solche mit geringen Intensitätswerten.[125] Die Klassifizierung dieser Spots als fehlerhaft hängt ursächlich mit den stärkeren Schwankungen niedriger Intensitätswerte zusammen. Dies ist primär auf das verstärkte Hintergrundrauschen zurückzuführen. Solche niedrigen Signalwerte werden daher oftmals molekularen und physikalischen Eigenschaften (Unreinheit der DNA, unzureichende Bindung an die Glasoberfläche, usw.) des DNA-Spots, ungleichmäßigem Labeling der Proben mit den Fluorophoren, einer uneinheitlichen Hybridisierung, einem Signalbleichen (*signal bleaching*) oder einer geringen Scannersensitivität zugeschrieben.[119] Der Ausschluss all dieser Spots unter der Annahme, dass es sich dabei um derartig zustande gekommene, fehlerhafte Datenpunkte handelt, könnte aber schwerwiegend sein. So ist zumindest ein Teil der niedrigen Intensitätswerte nicht fehlerhaften Ursprungs, sondern spiegelt tatsächlich niedrige Expressionswerte der jeweilige Gene wider und ist somit für die Ermittlung der differentiellen Expression der Gene von immenser Bedeutung.
Um also nicht versehentlich valide Datenpunkte mit geringer Intensität von der weiteren Auswertung auszuschließen, fließen solche Spots in die Auswertung ein und werden unter Umständen dann erst in der weiteren Analyse eliminiert. Die Implementierung des in Kapitel 6.2.4 vorgestellten Algorithmus zielt vor allem auf die Ermittlung der wahren Werte von extremen (niedrige und hohe) Intensitäten ab. Daher wird die teilweise in der Literatur vorgeschlagene Eliminierung Intensitäts-schwacher Spots während der Qualitätsanalyse in dieser Auswertung nicht befolgt, sondern erfolgt erst im Verlaufe späterer Auswertealgorithmen.[126]

Um eine Auswahl möglicher Ausschlusskriterien definieren zu können, wurden unterschiedliche Microarrays aus verschiedenen Microarray-Familien ausgewählt, die zu unterschiedlichen Zeitpunkten mit Oligonukleotiden aus unterschiedlichen Organismen gedruckt worden waren: dabei handelte es sich um drei Microarrays zweier Microarray-Familien mit Genen des Bakteriums *E.coli* mit je zwei bzw. fünf Genreplikaten, zwei Microarrays einer Microarray-Familie mit Genen aus *Saccharomyces cerevisiae* und zwei humanen Leber-Tumor-Microarrays. Die jeweiligen *tif*-Bilder unterschiedlicher Scan-Einstellungen wurden in GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) geöffnet und mit den Darstellungen der Bildern von den selben Microarrays in *Findspot* verglichen. Die parallele Verwendung dieser beiden Programme ermöglicht eine gegenseitige Ergänzung: Während *Findspot* dreidimensionale Darstellungen, Konturplots und die Auswahl unterschiedlicher Perspektiven bereitstellt und die direkte Gegenüberstellung von Genreplikaten erlaubt, können in GenePix Pro 6.0 in der so genannten *Feature mode* (Spot-Modus) Graphiken angewählt werden, die eine Darstellung aller zu einem Spots sowie dem Hintergrund gehörigen Pixel beinhalten.
Zunächst wurden die Spots mit *Findspot* optisch in qualitativ befriedigende und unzureichende Spots (kein Signal oder *Artefakte*) unterteilt. Kriterien für diese Unterteilung wurden auf der Basis der Einschätzung zweier Experimentatoren, die jahrelange Erfahrung mit Microarray-Experimenten aufweisen konnten, entworfen und mit detaillierten Darstellungen aus der Literatur verglichen.[127–131] Diese Kriterien beruhen auf starken Abweichungen bzw. Ähnlichkeit der Spotformen zu idealen Kreisen (zweidimensional) und zu idealen Zylindern (dreidimensionalen).[131] Anschließend wurden die Spots mit schlechter Qualität eingehend auf charakteristische Merkmale untersucht, die kennzeichnend für Abweichungen von guten Spots sind. Dazu dienten die Pixelplots im Spotmodus, der für jeden Spot separat aufgerufen werden kann und eine Auftragung aller Intensitätswerte (Pixeldarstellung) der beiden (mit Cy5 und Cy3 markierten) Zustände sowohl für den Signal- als auch für den lokalen Hintergrundbereich auf dem Microarrays enthält. Da diese Pixelwerte der Spots und des Hintergrunds sonst nur über die *tif*-Dateien abgerufen werden können, liefert der Pixelplot übersichtlich dargestellte Hintergrundinformationen. Zusätzlich zu den Intensitätswerten werden weitere statistische Größen des dargestellten Spots (wie beispielsweise der Median, der Mean und der Regressionsquotient) angegeben und teilweise optional abrufbar in der Graphik dargestellt. Da das GenePix Pro 6. (Axon Instruments, Inc., Union City, CA) mit Ausnahme der Einzelintensitäten aller Pixel eines Spots eine Vielfalt an Daten, die auch im Pixelplot abrufbar sind, in den Ergebnisdateien (*Gpr*-Dateien) der Primärauswertung mitliefert, können diese Angaben für eine Qualitätskontrolle der Spots genutzt werden. Dieses Vorgehen der Nutzung der aus der Primärauswertung hervorgehenden Datenvielfalt ist allgemein gebräuchlich und wird auch in dem hier vorgestellten Programm befolgt.[72] Diese direkte und ausschließliche Verwendung der Daten aus der Primärauswertung von GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) beschleunigt erstens die Vorverarbeitungsschritte der Microarray-Auswertung und spart vor allem eine erhebliche Menge an Speicherplatz, da die für Implementierung einer zusätzlichen Bildauswertung erforderliche Speicherung der *tif*-Dateien auf dem auswertenden Computer umgangen werden kann: Da jeder Chip mehrfach gescannt werden sollte (siehe Kapitel 6.2.4), und ein Experiment aus mindestens zwei Chips besteht, kann daher bei jedem

Experiment eine große Datenmenge durch die *tif*-Dateien erzeugt werden.
Bei den von der Primärauswertung ausgegebenen relevanten Parameter für eine statistische Güteprüfung handelt es sich um folgende Größen:

- **Flags**:
  Dabei handelt es sich um eine von GenePix Pro vorgenommene Kennzeichnung der Qualität der Spots. Diese können entweder gut oder unzureichend (*geflaggt*) sein. Die geflaggten Spots wiederum sind in unterschiedliche Kategorien entsprechend ihrer Qualitätsmängel eingeordnet und sollten gemäß der Empfehlung der Betreiber von der weiteren Auswertung ausgeschlossen werden. Es handelt sich dabei um Spots mit sehr geringem Signal (niedriger als der Hintergrundsignalwert): „negative Spots“. Zu den geflaggten Spots zählen auch solche, deren Pixel-Standardabweichung im Vergleich zum Mittelwertsehr hoch ist. Sie werden von GenePix Pro als „schlechte Spots“ bezeichnet.[53]
  Obgleich geflaggte Spots ohnehin eliminiert werden, kann der Vergleich der Werte der entsprechenden Gütekriterien für diese Spots mit denen der nicht-geflaggten Spots hilfreiche Informationen bezüglich deren Einfluss auf die unzureichende Qualität bieten.
  Auf die Validität der empfohlenen Entfernung dieser Spots wird in Kapitel 6.2.3 näher eingegangen.

- **Zirkularität**:
  Die Zirkularität dient als Maß für die Rundheit der Spots und wird in Prozentwerten angegeben.

- **Median und Mittelwert der Signalpixel beider Wellenlängen**:
  Die Mediane und Mittelwerte aller Pixel, die den lokalen Signalbereich der Spots ausmachen.

- **Hintergrundcharakteristiken**:
  Der Prozentsatz aller Signalpixel, die Intensitäten aufweisen, die mit einer zweifachen Standardabweichung oberhalb der Hintergrundpixelintensitäten liegen. Als lokaler Hintergrund wird ein Ring rund um den Spot mit variablem Durchmesser verwendet. Diese Angaben werden wiederum für beide Wellenlängen aller Spots auf einem Chip gespeichert.

- **Regressionsquotient**:
  Der Regresssionsquotient spiegelt das Verhältnis der Steigungen wieder, die aus den einzelnen Pixeln der beiden Wellenlängen berechnet werden. Es werden Pixel verwendet, die sich in einem Kreis mit zweifachen Signaldurchmesser um das Zentrum des Spots befinden.

- **Quotient der Mediane beider Wellenlängen**:
  Der Quotient der Mediane entspricht dem Verhältnis der Median-Intensitäten der beiden Wellenlängen, die aus allen zu einem Spot gehörigen Pixeln berechnet wurden nachdem der Median der Hintergrundpixel subtrahiert wurde.

Mithilfe dieser Daten und der dazugehörigen Pixelbilder wurden Prüfkriterien für die Qualität der Spots definiert, die später durch eine statistische Analyse umfangreichen Datensätze validiert wurde (siehe unten). Diese Auswahlkriterien wurden mit bereits in der Literatur beschriebene Prüfkriterien abgeglichen. Auf dieser Basis wurden vorab spezifische Anforderungen festgelegt, die durch anschliessende Untersuchungen in zahlenmäßigen Qualitätsmerkmalen resultieren (siehe Seite 58):

- Das Maß der Zirkularität der Spots soll ein Mindestmaß nicht unterschreiten. Je runder ein Spot, umso signifikanter ist die Primäranalyse (abgesehen von geflaggten Spots, die von der Auswerte-Software mit einer Zirkularität von 100% notiert werden).[132,133]

- Hohe Abweichungen zwischen dem Median und dem Mittelwert aller Signalpixelwerte deuten auf Bereiche mit vielen Ausreißern auf dem Chip hin, da der Median im Gegensatz zum Mittelwert robust gegenüber Ausreißern ist. Es werden also Spots mit möglichst wenig variablem Signal in beiden Wellenlängen gesucht. Dabei soll ein Mindestwert der Mediane der Intensitätswerte vorliegen, da niedrige Werte zu geringen Abweichungen zwischen den Vergleichswerten führen können, was die fälschliche Schlussfolgerung eines guten Quotienten zulässt. Die Definition dieses Gütekriteriums zielt besonders auf die Detektion von hellen Spots ab, die durch Verunreinigungen auf dem Chip in Form von Schlieren (statt realer hoher Expressionswerte) beziehungsweise von Spots mit unregelmäßigen Formen ab.
  Abbildung 6.7 verdeutlicht die Relevanz des Verhältnisses zwischen Median und Mittelwert für eine qualitative Beurteilung der Spots auf den Microarrays. Dargestellt sind drei Spots mit einem hohen Quotienten und drei Spots mit einem Quotienten nahe eins. Ein niedriger Wert korreliert sehr gut mit Homogenität und Rundheit (Zirkularität) der Spots.[39]

- Valide Spots sollten einen gewissen Prozentsatzes an Pixeln aufweisen, deren Intensitätswerte in beiden Kanälen mehr als zwei Standardabweichungen oberhalb des jeweiligen Hintergrundwertes liegen.[8,134]

- Stark differierende Quotienten der Mediane und Regressionsquotienten verweisen auf nicht-uniforme Spots. Daher sollte ein Grenzwert gefunden werden, der im Vergleich der beiden Größen nicht unterschritten wird.[135–139]

Der Einfluss dieser Qualitätsmerkmale auf die Güte der Spots bzw. Microarrays wurde mit einer Datenmenge von 25 Primärdatendateien (*Gpr*-Dateien) mit unterschiedlichen Scaneinstellungen von 12 Microarrays getestet (sechs *low-density*-Microarrays, von denen zwei Chips mit Genen aus dem Organismus *E.coli*, ein Chip mit Genen aus dem Organismus *Rattus Norvegicus* und zwei Chips mit Genen aus dem Organismus *Saccharomyces cerevisia* gespottet waren und sechs weitere *whole-genome*-Microarrays, von denen zwei Chips Gene aus dem Organismus *Saccharomyces cerevisiae* und die verbleibenden vier Chips das *E.coli*-Genom trugen). Dafür wurde das Programm *Validate QA-Parameter* implementiert, das die Auswahl und das anschließende Einlesen verschiedener *Gpr*-Dateien aus der Primärauswertung zur

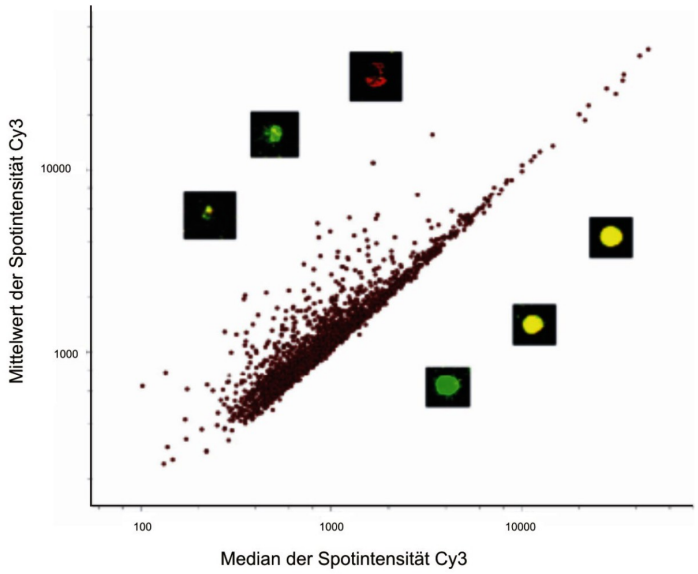


Abbildung 6.7 – *Mittelwert/Median-Verhältnis als Qualitätsmerkmal. Die Grafik verdeutlicht die Bedeutung des Quotienten von Intensitätsmittelwert zu -median eines Spots. Spots, bei denen dieser Wert stark von eins abweicht (Punkte oberhalb der Diagonalen mit einer Steigung die deutlich von eins abweicht), sind deutlich inhomogener als Spots (beispielhafte Spots mit solchen Charakteristiken sind oberhalb der Diagonalen gezeigt), bei denen Mittelwert und Median sehr ähnlich sind (Punkte auf der Diagonalen; Spotbeispiele, die entsprechende Charakteristiken aufweisen sind unterhalb der Diagonalen dargestellt).*[39]

gleichzeitigen Validierung der Prüfoptionen ermöglicht. Das Programm stellt die Möglichkeit zur Auswahl unterschiedlicher Grenzwerte für die Prüfkriterien bereit, so dass die empirischen Grenzwerte feiner justiert und gegebenenfalls bei Bedarf runter- oder hochgesetzt werden können (um beispielsweise eine strengere oder weniger strenge Qualitätskontrolle vornehmen zu können). Auf diese Weise können alle Spots aller zu testenden Microarrys auf die Relevanz der Einzelkriterien sowie deren potentielle Bedeutung für die Qualitätsanalyse im Zusammenspiel mit den anderen definierten Kriterien kalkuliert werden. Auf diese Weise kann jeder Spot der unterschiedlichen Microarrays auf die unterschiedlichen Kriterien hin untersucht werden. Für jedes potentielle Qualitätskriterium wird eine boolsche Antwort (Ja / Nein) zugelassen, die den jeweiligen Kriterien Werte (Ja = Ausschluss-Kriterium erfüllt = 1, Nein = Ausschluss-Kriterium nicht erfüllt = 0) zuweist. Als Startwerte für die Untersuchung der vordefinierten Kriterien (siehe oben) wurden die von der Firma GenePix Pro im Handbuch vorgeschlagenen Grenzwerte eingesetzt. Tabelle 6.4 zeigt einen Ausschnitt aus der Ergebnistabelle der Qualitätsüberprüfung der definierten Kriterien.

Anhand dieser Tabelle wurden die entsprechenden Spots, bei denen mindestens 2 der Kriterien erfüllt waren, optisch untersucht, in dem das Programm *Findspot* verwendet wurde. Dabei stellte sich heraus, dass vor allem die Kriterien 2, 3 und 5 stark mit der Form der Spots korrelieren: je stärker die entsprechenden Kriterien zutreffen (je höher bzw. niedriger also die Grenzwerte), umso geringer ist die Homogenität der Spots. Das vierte Kriterium hingegen weist auf einen verhältnismäßig starken Hintergrund im Vergleich zum Signalwert hin, wodurch eine valide Extraktion der Signalwertpixel erschwert wird. Diese Kriterien erscheinen

Tabelle 6.4 *– Beispiel eines Ausschnitts aus der Ergebnistabelle des Programms* Validate QA-Parameter. *Gezeigt sind die ersten zehn Spots eines* whole-genome-Microarrays *mit Genen aus dem Organismus* Saccharomyces cerevisiae. *Die erste Spalte gibt die Gennamen sowie den dazugehörigen Chip an (C1 = Chip 1) der eingelesenen Microarrays. Die darauffolgenden Spalten geben die definierten Gütekriterien an, deren Grenzwerte (im folgenden kursiv gedruckt) im Programm verändert werden können. Die folgenden Zeilen geben für jedes Kriterium an, ob es erfüllt ist (1 = Ausschlusskriterium erfüllt, 0 = Ausschlusskriterium nicht erfüllt). Kriterium 1 = Flag, Kriterium 2 = Zirkularität < 50%, Kriterium 3 = maximal 70 Pixel in beiden Kanälen weisen Intensitätswerte auf, die mehr als zwei Standardabweichung oberhalb der Hintergrundwerte liegen, Kriterium 4 = Die Mediane aller Signalpixel beider Kanäle > 200 und die Quotienten zwischen Median und Mittelwert aller Signalpixel > 1.2, Kriterium 5 = Der Quotient der Mediane der Einzelpixel aus beiden Kanälen befindet sich nicht im Intervall (0.8; 1.2) vom Regressionsquotienten.*

Genname	Krit. 1	Krit. 2	Krit. 3	Krit. 4	Krit. 5
YLR347c (C1)	1	1	1	1	1
YLR105c (C1)	0	1	0	0	0
YLR363c (C1)	1	1	1	1	1
YLR272c (C1)	0	0	0	0	0
YLR292c (C1)	1	1	1	1	1
YLR288c (C1)	1	1	1	1	1
YLR203c (C1)	0	0	0	0	0
YLR220w (C1)	0	0	0	0	1
YKL219w (C1)	0	0	0	0	0
YAL063c (C1)	0	0	0	1	0

somit durchaus sinnvoll für die Valdierung der Güte der Microarrays. Von besonderem Interesse für die Beurteilung der Gesamtgüte der Microarrays sind hierbei die zusätzlich zu den automatisch ermittelten Flags definierten Kriterien 2 - 5, da geflaggte Daten (Kriterium 1) per definitionem Spots von unzureichender Güte repräsentieren (siehe Kapitel 6.2.3):

- Kriterium 2:
  Zirkularität kleiner als *50%*
- Kriterium 3:
  maximal *70* Pixel in beiden Kanälen weisen Intensitätswerte auf, die mehr als zwei Standardabweichung oberhalb der Hintergrundwerte liegen
- Kriterium 4:
  Die Mediane aller Signalpixel beider Kanäle sind größer als *200* und die Quotienten zwischen Median und Mittelwert aller Signalpixel sollte größer sein als *1.2*
- Kriterium 3:

Der Quotient der Mediane der Einzelpixel aus beiden Kanälen ist nicht im Intervall (*0.8*; *1.2*) vom Quotient der Regressionskoeffizienten

Die Analyse der Spots in Bezug auf die untersuchten Gütekriterien ergab signifikante Unterschiede im Vergleich der *low-density*-Microarrays mit den Ergebnissen der *whole-genome*-Chips: Der Anteil an Flags an der Gesamtspotzahl ist bei den *whole-genome*-Chips erwartungsgemäß wesentlich höher (teilweise über 90%) als bei den *low-density*-Chips, deren Flag-Anteil bei den untersuchten Microarrays zwischen 1% und 70% schwankt. Da *whole-genome*-Microarrays Spots für das ganze Genom eines Organismus tragen, werden in der Regel bei Experimenten mit unterschiedlichen Behandlungen nur verhältnismäßig wenige Gene des Genoms reguliert, was sich in einer hohen Anzahl von negativen Flags widerspiegelt. Daher erscheint auf diesen Microarrays eine geringe anteilige Expression und damit eine hohe Anzahl an Flags (in Abhängigkeit der Versuchsplanung) durchaus realistisch.
Die Verwendung von *low-density*-Microarrays hingegen zielt vor allem darauf ab, spezifische (oftmals durch Vorscreenen mit *whole-genome*-Experimente ausgewählte) Gene, bei denen eine erhöhte Expression in Folge einer Behandlung angenommen wird, detaillierter zu untersuchen. Daher sollte der Anteil der exprimierten Gene auf diesen Chips hoch sein und mit einem geringen Anteil negativer Flags einhergehen. Aus diesem Grund deutet ein hoher Anteil an Flags bei *low-density*-Chips auf Chip-bedingte Qualitätsmängel hin.
Für nähere Untersuchungen wurde die Schnittmenge an Spots gesucht, die einerseits die oben genannten einzelnen Kriterien erfüllen und andererseits geflaggt sind. Bei diesem Vergleich zwischen Anzahl der Spots, die die definierten Auswahlkriterien erfüllen, und der geflaggten Spots, muss die Zirkularität gesondert betrachtet werden, da die Zirkularität geflaggter Spots von der Primärauswertesoftware mit 100% (ideal runder Spot) angegeben wird und irreführenderweise zu der Annahme verleitet, dass es sich dabei - bezügliches dieses Kriteriums - um auswertbare, da besonders repräsentative (runde) Spots handelt. Die Ergebnisse dieser Vergleiche sind in den Tabellen 6.5 (für die *low-density*-Microarrays) und 6.6 (für die *whole-genome*-Microarrays) angegeben.

Wie den Tabellen 6.5 und 6.6 zu entnehmen ist, ist der Anteil der geflaggten Spots an den Spots, die unter die definierten Kriterien fallen, bei *whole-genome*-Microarrays wesentlich höher und steigt mit dem Prozentsatz der Flags auf den Microarrays insgesamt. Dieser Zusammenhang ergibt sich aus den Eigenschaften der geflaggten Spots, deren Güte bezüglich weiterer Auswertungen mangelhaft ist, so dass sie neben den für das Flaggen definierten Anforderungen auch die oben erwähnten Kriterien 3 - 5 erfüllen. Dieser Zusammenhang bestätigt einerseits die Relevanz der negativen Qualitätsmerkmale. Andererseits bedeutet dies, dass je geringer der Anteil der Spots, die die definierten Kriterien erfüllen, an den geflaggten Spots ist, umso größer ist der über die Flags hinausgehende Anteil mangelhafter Spots auf den Microarrays.

Zusammenfassend sollte bei der Beurteilung der Güte von Microarrays insgesamt also zunächst der Anteil der Flags überprüft werden. Dazu muss individuell nachvollzogen werden, ob der Prozentsatz der Flags auf den Microarrays mit den Erwartungen übereinstimmt.

Tabelle 6.5 *– Anteil geflaggter Spots an Gesamtzahl der Spots von geringer Güte (gemäß den in Tabelle 6.4 definierten Kriterien 3 - 5). In der letzten Spalte ist der Anteil der Flags an der Gesamtspotzahl angegeben.*

Primärdatei	Kriterium 3	Kriterium 4	Kriterium 5	% Flags
C1, Scan 1	52%	71%	57%	30%
C1, Scan 2	53%	70%	56%	25%
C1, Scan 3	52%	67%	57%	22%
C2, Scan 1	90%	93%	89%	5%
C2, Scan 2	89%	94%	90%	3%
C2, Scan 3	88%	93%	91%	2%
C2, Scan 4	91%	90%	91%	2%
C3, Scan 1	49%	63%	53%	12%
C3, Scan 2	50%	64%	55%	10%
C3, Scan 3	49%	63%	56%	10%
C4, Scan 1	56%	71%	57%	15%
C4, Scan 2	55%	73%	56%	14%
C4, Scan 3	54%	76%	58%	12%
C5, Scan 1	61%	70%	55%	19%
C5, Scan 2	63%	70%	57%	21%
C5, Scan 3	63%	71%	56%	22%
C6, Scan 1	51%	72%	56%	66%
C6, Scan 2	53%	72%	54%	70%

Wie oben anhand des Unterschiedes zwischen *whole-genome*- und *low-density*-Microarrays erläutert, müssen dabei spezifische experimentelle Unterschiede berücksichtigt werden, die vor allem auf der Basis der erwarteten Expression beruhen. Da Spots vor allem dann Flags zugewiesen werden, wenn keine Expression vorliegt und somit keine signifikanten Signale ermittelt werden können, sollten bei *low-density*-Microarrays in der Regel nur sehr geringe Anteile an Flags vorliegen. Je stärker jedoch der Flaganteil von den Erwartungen (Annahme einer hohen Gesamtexpression) abweicht, umso größer ist die Wahrscheinlichkeit eines unverhältnismäßig starken Hintergrundes, der in einer geringeren Differenz zwischen Signal- und Hintergrundwert resultiert. In diesem Fall kann die Expressionsstärke einzelner Gene weniger signifikant ermittelt werden. Hohe Flaganteile deuten hier besonders deutlich auf Chips von unzureichender Qualität hin.

Tabelle 6.6 *– Anteil geflaggter Spots an Gesamtzahl der Spots von geringer Güte (gemäß den in Tabelle 6.4 definierten Kriterien 3 - 5). In der letzten Spalte ist der Anteil der Flags an der Gesamtspotzahl angegeben.*

Primärdatei	Kriterium 3	Kriterium 4	Kriterium 5	% Flags
C7, Scan 1	80%	89%	81%	88%
C7, Scan 2	81%	89%	82%	90%
C8, Scan 1	82%	88%	80%	93%
C8, Scan 2	79%	87%	80%	92%
C9, Scan 1	75%	86%	79%	78%
C9, Scan 2	74%	86%	80%	77%
C9, Scan 3	73%	86%	82%	78%
C9, Scan 4	74%	87%	80%	80%
C10, Scan 1	82%	91%	83%	82%
C10, Scan 2	83%	90%	85%	84%
C11, Scan 1	82%	88%	80%	83%
C11, Scan 2	81%	92%	79%	90%
C12, Scan 1	80%	88%	82%	91%
C12, Scan 2	81%	91%	82%	93%

In diesem Zusammenhang kann der Vergleich unterschiedlicher Microarrays eines Experiments hilfreich sein, da innerhalb eines Experiments ähnliche Flaganteile auf unterschiedlichen Chips zu erwarten sind. Aus diesem Grund werden die Ergebnisse der Qualitätsanalyse unterschiedlicher Chips in einem Tabellenblatt untereinander gespeichert.
Ebenfalls aufgelistet werden die weiteren oben aufgelisteten Kriterien. Hierbei spielt die Zirkularität der Spots eine gesonderte Rolle, da sie nicht in direkten Bezug zu den Flags gesetzt werden kann (siehe oben). Um jedoch signifikante Hinweise aus diesem Kriterium ziehen zu können, wurde der Grenzwert für die Zirkularität niedrig angesetzt (50%). Spots, die unterhalb dieser Grenze liegen deuten auf sehr inhomogene Spots von geringer Güte hin, so dass der Anteil dieser Spots auf dem Chip entsprechend gering sein sollte ($< 5\%$).
Betreffend der drei verbleibenden Kriterien gilt, dass die Spots, die diese Kriterien erfüllen einen möglichst großen Bereich der geflaggten Spots abdecken sollten. Wenn der Anteil der Flags auf den Chips den Erwartungen entspricht, darüber hinausgehend aber viele Spots von mangelhafter Güte sind, da ihre Signalwerte beispielsweise starke Abweichungen zwischen Median und Mittelwerten ausweisen und somit deutlich inhomogen sind (siehe Abbildung 6.7), so deutet dies insgesamt auf Chips von mangelhafter Güte hin.
Die untersuchten Grenzwerte erscheinen hierbei insofern sinnvoll, als dass die Anzahl der se-

lektierten Spots bei Chips von hoher Güte nahezu mit der Anzahl der Flags übereinstimmt (signifikante Unterschiede zwischen Chips von hoher / geringer Güte erkennbar), andererseits aber dennoch Spots von abweichender Qualität zusätzlich zu den Flags ermittelt werden können (hohe Spezifität).
Bezüglich der Aussage der Kriterien spielen vor allem die Zirkularität, der Unterschied zwischen Median und Mittelwert und der Vergleich zwischen und Medianquotienten und Regressionsquotienten eine Rolle in Bezug auf die Form der Spots. Ähnliche hohe Werte dieser Kriterien deuten also auf inhomogene Spots hin und sind daher insofern von großer Bedeutung, als dass sie keine eindeutige Aussage bezüglich der Expression von Genen zulassen (siehe oben). Hier machen sich vor allem Einschränkungen der Primärauswertungesoftware bemerkbar.
Der Vergleich der Signalwerte mit der zweifachen Standardabweichung der Hintergrundpixel hingegen liefert einen Eindruck von der Stärke des Hintergrunds. Je schwächer der Hintergrund ist, umso sensitiver kann die Expression der Gene erfasst werden. Aus diesem Grund weisen Spots, die das entsprechende negative Qualitätskriterium erfüllen, darauf hin, dass exprimierte Spots während der Primärauswertung nur unzureichend ermittelt werden können.
Die im Fokus dieser Kriterien liegende Qualitätsanalyse des gesamten Microarrays wurde also anhand empirischer Untersuchungen durchgeführt. Basierend auf einer Literaturrecherche wurden Kriterien definiert, die die Güte der Chips beeinträchtigen. Diese Kriterien wurden überprüft, in dem Microarrays auf diese Kriterien überprüft wurden und die entsprechenden Spots visuell auf ihre Güte überprüft wurden (siehe Tabelle 6.4). Die Ergebnisse der Qualitätsanalyse werden in einem separaten Tabellenblatt der gesamten Ergebnis-Exceldatei angegeben. Dabei werden pro Primärauswertedatei die Anzahl der Spots angegeben, die diese Kriterien erfüllen, wobei jeweils ein Bezug zu den Flagangaben der Chips angestellt wird. Eine detaillierte Definition der Kriterien und ihrer Bedeutung befindet sich unterhalb der Daten.
Diese Qualitätsangaben dienen als Richtlinie für Experimentatoren bei der Beurteilung der Güte der verwendeten Microarrays. Absolute Angaben bezüglich der Güte werden allerdings aufgrund der starken individuellen Unterschiede der Microarrays vermieden.

6.2.2.3 Fazit

Da Microarray-Experimente auf einem hochgradig variablen Verfahren beruhen, ist eine Qualitätsanalyse der verwendeten Microarrays besonders im klinischen Bereich unabdingbar.[140]
Aus diesem Grund wurden unterschiedliche *low-density*-Chips mithilfe des dafür implementierten Programs *Findspot* sowie den aus der Primärdatenanalyse hervorgehenden Intensitätsangaben untersucht. Die dabei beobachteten Variabilitäten stimmen mit den in der Literatur beschriebenen Mangelerscheinungen von Microarrays überein.
Es wurde eine Entwicklung von Gütekriterien zur Beurteilung der Qualität von Microarrays angestrebt. Dazu wurde der Einfluss potentiell relevanter statistischer Größen, die während der Primäranalyse entstanden sind, unter der Verwendung des Programms *Validate QA-Parameter* auf die Qualität der Microarrays getestet. Mithilfe der daraus resultierenden finalen Gütekriterien kann die Qualität unterschiedlicher Microarray vor der Auswertung eines

Experiments evaluiert und unzureichende Microarrays gegebenenfalls entfernt werden. Um absolute Aussagen bezüglich der Güte von Microarrays treffen zu können, muss eine Vielzahl weiterer Microarrays getestet und miteinander in Relation gesetzt werden. Dafür besteht vor allem die Notwendigkeit Microarray-Chargen-spezifischer Vergleiche, da Microarrays unterschiedlichen Typs (*low-density*/*whole-genome*, selbst-gespottet/kommerziell, usw.) nicht direkt miteinander in Relation gesetzt werden können. Die angegeben Gütekriterien dienen damit in erster Linie als Richtlinie, können aber im Vergleich unterschiedlicher Microarrays eines Experiments Auskunft über die mangelhafte Güte von Einzelchips liefern.

6.2.3 Vorverarbeitung der Daten

6.2.3.1 Hintergrund und Anforderungen

Nachdem eine Qualitätsanalyse implementiert wurde, die vor Beginn der eigentlichen Auswertung die Nutzbarkeit aller Microarrays für die Datenanalyse evaluiert, kann der erste Teil der Datenanalyse angegangen werden: die Präprozessierung der Messwerte aus den Microarray-Experimenten.
Die Vorverarbeitung der Daten beinhaltet analytische und transformatorische Prozeduren, die vor der Ermittlung der Expressionsgrade auf die Daten angewandt werden müssen. Dieser Filterungsprozess fungiert somit als Schnittstelle zwischen der Aufnahme der Rohdaten der Microarray-Experimente und fortgeschrittenen statistischen Algorithmen, die für eine Genklassifikation und -identifikation gebraucht werden.

Die Daten müssen daher vor einer weitergehenden Analyse der Vorverarbeitung unterzogen werden, um unzuverlässige Werte aus dem Datensatz zu entfernen.[77] Auf diese Weise soll die Anzahl System-bedingter Fehler reduziert bzw. nach Möglichkeit weitestgehend eliminiert werden. Die verbleibenden Daten werden daraufhin in einen gleichartigen Kontext transformiert.[78]
Die Vorverarbeitung der Daten orientiert sich einerseits an Qualitätsmerkmalen wie der Intensität und Homogenität einzelner Spots aber auch an der Reproduzierbarkeit der Meßwerte. Gleichzeitig beeinflusst auch die Anzahl der Daten nachhaltig deren Güte, so dass eine unnötige Eliminierung verwertbarer Datenpunkte vermieden werden soll.
Im Anschluss an diese standardmäßig definierten Aufgaben der Datenvorverarbeitung wurde eine Option eingeführt, die automatisch die für die Veröffentlichung in *GEO* (*Gene Expression Omnibus*[102]) notwendigen Daten berechnet und in ein Tabellenblatt in der Ergebnisdatei speichert.
Zusammenfassend kann die Vorverarbeitung inhaltlich in die folgenden Kategorien gegliedert werden:

- Einlesen der Daten aus der Primärauswertung
- (Qualitätsanalyse - siehe Kapitel 6.2.2)

- Entfernen von Flags
- Subtraktion des Hintergrundes
- graphische Darstellung und Transformation der Daten
- Schreiben von GEO-Dateien

Im Rahmen dieser Arbeit wurden folglich Ansätze bei der Präprozessierung der Primärdaten in die Auswertung integriert und ein standardisiertes Verfahren für die Datenvorverarbeitung etabliert.

6.2.3.2 Durchführung und Ergebnisse

Auswahl der Auswerteoptionen und Laden der Daten

Um die Daten aus der Primärauswertung der Vorverarbeitung übergeben zu können, müssen die Daten dem Programm zunächst verfügbar gemacht werden. Zu den während der Auswertung benötigten Informationen gehören aber nicht nur die Rohdaten der Primärauswertung, die in Form von *Gpr*-Dateien vorliegen, sondern auch die während der Eingabe der Datenblätter übergebenen Informationen zu auswerterelevenaten Details der Experimente.
Diese Informationen sowie die Pfade zu den *Gpr*-Dateien werden direkt aus der Datenblatteingabe an die Auswertung übertragen, nachdem der Befehl zum Start der Auswertung gegeben wurde. Dem Anwender steht die Möglichkeit offen, in einer elektronischen Abfrage optional zu wählen, die graphische Darstellung einiger Zwischenergebnisse und die Erstellung der *GEO*-Tabellenblätter (für eine etwaige spätere Veröffentlichung) anzuwählen. Daraufhin beginnt das Programm mit dem Einlesen der als *tab-stop* getrennte Textdateien vorliegenden *Gpr*-Dateien.
Um Speicherplatz zu sparen und den Einlesevorgang zeitlich zu reduzieren, werden von den Rohdaten nur die für die oben beschriebene Qualitätsanalyse (Kriterien der Qualitätsanalyse siehe Kapitel 6.2.2) und Datenauswertung benötigten Daten extrahiert und entsprechend ihrer Funktion (Qualitätsanalyse / Datenauswertung) in unterschiedlichen *Arrays* hinterlegt.

Die Daten der Microarrays werden im Folgenden der in Kapitel 6.2.2 erläuterten Qualitätsanalyse unterzogen. Wie oben erwähnt, ist die nominelle Eingliederung der Qualitätsanalyse in die Vorverarbeitung der Daten in der Literatur umstritten. Obgleich die Qualitätsanalyse in diesem Programm auf das Einlesen der Primärdaten folgt, wird sie an dieser Stelle nur kurz erwähnt, da Details zur Qualitätsanalyse bereits beschrieben wurden. Das Resultat der Qualitätsüberprüfung wird in die Ergebnisdatei geschrieben und optional graphisch dargestellt.
Der nächste Vorverarbeitungsschritt besteht aus einer Sortierung der Daten. Zunächst werden die Daten entsprechend ihrer Gen-Bezeichnungen (*GeneID*) sortiert, so dass gegebenenfalls auf dem Microarray vorkommende Gen-Replikate untereinander stehen. Diese Sortierung bietet

außerdem den Vorteil der Auffindung nicht bedruckter Spots auf dem Microarray, Positionen auf dem Microarray also, die beim Spotten nicht mit DNA-Molekülen bedruckt wurden. Solche oftmals als *Blanks* oder *Empties* bezeichneten leeren Spots können teilweise in hoher Zahl auf Microarrays auftauchen und sind durch das vorgegebene Format von GenePix Pro 6.0 für die Gitter- (*gal*)-Dateien bedingt. Dieses Format lässt an einigen Positionen der Blöcke keine leeren Spots zu, so dass diese bei der Erstellung der Gitter (*Grids*) entsprechend benannt werden müssen und später in der Primärdaten-Datei auftauchen. Die Lokalisierung solcher leeren Spots durch die Sortierung dient der Eliminierung aller entsprechenden Daten, so dass teilweise erheblicher Speicherplatz gespart werden kann.
In einer zweiten Sortierung werden die Daten aus *Gpr*-Dateien unterschiedlicher Scan-Einstellungen eines Microarrays der Größe nach eingeordnet. Da jeder Chip bei unterschiedlichen Einstellungen (Laserstärke und PMT-Verstärkung) gescannt worden ist, existieren zu jedem Microarray auch dementsprechend viele Datensätze, die in unterschiedlichen *Arrays* hinterlegt worden sind und bei der späteren *Within-Array*-Normalisierung eine Rolle spielen (siehe Kapitel 6.2.4). Diese Arrays werden in einem ersten Durchlauf nach der Größe der Laserstärke und anschließend gemäß der Höhe der PMT-Verstärkung sortiert.

Entfernen von Flags

Das hier vorgestellte Filterverfahren entfernt nun die Datenpunkte der Spots, die von der Bildverarbeitungssoftware GenePix Pro 6.0 nicht erkannt oder für nicht verwertbar befunden werden. Dabei handelt in der Regel um sich „negative“, seltener um „schlechte“ Flags.
In Kapitel 6.2.2 wurde bereits die Kenntlichmachung mangelhafter Spots in Form sogenannter Flags durch die Bild-Auswertung erwähnt. Da die Qualitätsanalyse jedoch erstens einer weiteren potentiellen Eliminierung einzelner Spots zeitlich vorangeht und in erster Linie ergänzend zu der von GenePix Pro 6.0 vorgenommenen Analyse der Spot-Qualitäten wirkt, wurde die allgemeine Gültigkeit und Praktikabilität des Flag-Vorgangs, vor allem für *low-density*-Microarrays, in diesem Zusammenhang nicht explizit untersucht.
Ein Vergleich der geflaggten Spots mit den ungeflaggten Spots soll daher vorab den notwendigen Ausschluss dieser Flags validieren. Diese Beurteilung basierte auf zwei Vergleichen: Zunächst wurden die Intensitätswerte einzelner Spots aus beiden Zuständen gegenübergestellt. Die Beurteilung der Flags bezieht sich immer auf beide Kanäle, weswegen die Evaluierung auch diesbezüglich erfolgen muss. Darüber hinausgehend wurden das Signal-Rausch-Verhältnis (*SNR*) ermittelt und mit den Signalwerten verglichen.
In Abbildung 6.8 sind beispielhaft diese mittleren Signalsummen der Kanäle (Kanal 1 mit Cy5-Markierung / Kanal 2 mit Cy3-Markierung) auf einem *low-density*-Microarrays illustriert.

In Tabelle 6.7 sind die dazugehörigen Daten so wie das Signal-Rausch-Verhältnis widergegeben.

Bei diesem Microarray handelt es sich um einen ZNS-Microarray der Ratte, der spezifisch für die Erforschung von Entzündungsreaktionen der Haut von Ratten in Zusammenarbeit mit der Klinik für Plastische, Hand- und Wiederherstellungschirurgie der Medizinischen Hochschule Hannover[141] entwickelt wurde. Dieser Microarray wurde aufgrund der hohen Anzahl an

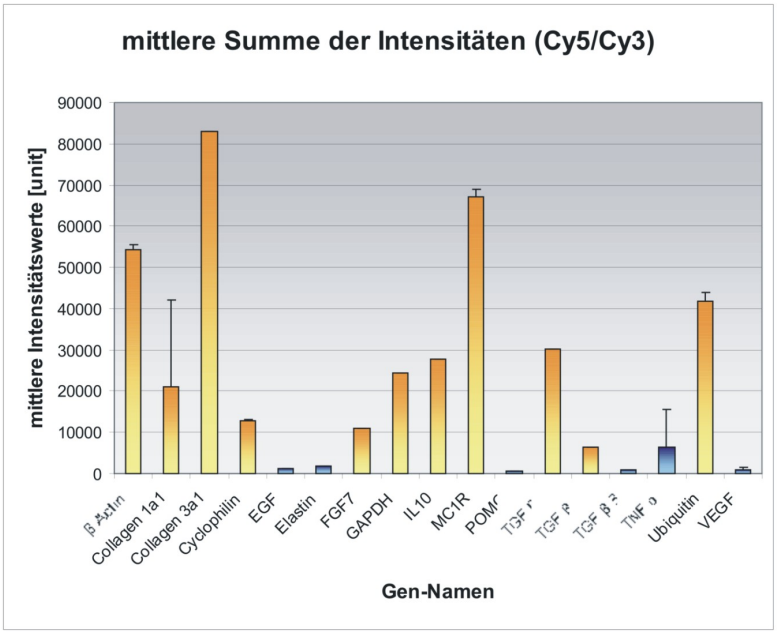


Abbildung 6.8 – *Mittlere Signalintensitäten eines low-density-Microarrays der Ratte. Die Summe der Signalintensitäten beider Kanäle aller Genreplikate wurden gemittelt und der Fehlerindikator als mittlere relative Varianzen dieser Gen-Intensitäten angezeigt. Die zu weniger als 50% (im Vergleich der Genreplikate) geflaggten Gene wurden gelb markiert. Die Gene, die mehr als 50% geflaggte Genreplikate aufweisen, wurden blau markiert.*

Genreplikaten und der übersichtlichen Anzahl an Genen (neun Genreplikate bei 17 verschiedenen Genen) als Darstellungs-Beispiel der Signaldifferenzen und Signal-Rausch-Verhältnis ausgewählt. Diese spezifische Microarray-Familie gewährleistet daher eine signifikante Aussage bezüglich der Genreplikate. Weitere Microarray-Familien (*E.coli*-, *S.cerevisiae*-, *Human Leber Tumor-low-density*-Chips sowie ein *whole-genome*-Chip mit Genen der Tomate) bestätigten das Resultat der Flag-Analyse (siehe unten), sind hier jedoch wegen ihrer Komplexität und ihres höchstens fünf Genreplikate umfassenden Spottingmusters nicht dargestellt. Die Daten stammen von einem Chip, der bei einer Laserstärke 100% und einem PMT-gain von 600 gescannt wurde.

Wie in Abbildung 6.8 und Tabelle 6.7 demonstriert, besitzen die nicht geflaggten Gene erwartungsgemäß höhere mittlere Intensitäten („negative“ Flags weisen sehr geringe Intensitätswerte auf). Das Signal-Rausch-Verhältnis wurde als weiteres Gütekriterium vergleichend herangezogen. Das Signal-Rausch-Verhältnis spiegelt ein Maß für die Qualität eines Signals wider, das von einem Rauschsignal überlagert ist. In diesem Fall stellt es das Verhältnis der mittleren Leistung des Signals zur mittleren Rauschleistung des Störsignals dar. Das Signal-Rausch-Verhältnis wird allgemein vielfach zur Überprüfung der Qualität von Signalen herangezogen und ist auch bereits vielfach in der Literatur als wichtiges Kriterium bei bei der Qualitätsanalyse von Spots auf Microarrays beschrieben.[142,143]

Die Gegenüberstellung der Signal-Rausch-Verhältnisse der geflaggten und der ungeflaggten Gene deuten auf die Relevanz dieser Größe dür die Qualität der Daten hin, da die Gene, die sich aus einem Großteil geflaggter Genreplikate addieren, Signal-Rausch-Verhältnisse unter einem Wert von 3.5 (bis auf TNF α sogar unter eins) aufweisen, was für die größtenteils un-

Tabelle 6.7 *– Flag-Validierung anhand eines low-density-Microarrays mit 17 Genen aus dem ZNS der Ratte. Angegeben sind die mittleren Intensitäts-Summen über alle Genreplikate sowie die entsprechende mittlere relative Varianz (Spalte drei). Die vierte Spalte zeigt das mittlere Signal-Rausch-Verhältnis (SNR) der Gene und die letzte Spalte den Prozentsatz an geflaggten Genreplikaten des Gens.*

Gen-Name	$\overline{Int}$	$\overline{\boldsymbol{SNR}}$	**rel.** $\overline{\sigma^2}$	% Flags
Beta Actin	54343	1654	22.96	0
Collagen 1a1	21152	27649	11.71	0
Collagen 3a1	83120	64	32.81	0
Cyclophilin	12814	641	6.31	0
EGF	1284	1306	0.64	90
Elastin	1764	4403	0.72	90
FGF7	10891	1317	4.55	10
GAPDH	24326	2643	5.46	0
IL10	27636	2181	6.26	0
MC1R	67048	206	17.51	0
POMC	738	195	0.36	100
TGF beta 1	30225	3088	7.05	0
TGF beta 2	6336	3185	3.21	30
TGF beta 3	780	2210	0.29	90
TNF α	6547	53697	3.47	90
Ubiquitin	41887	8179	19.67	0
VEGF	1046	2940	0.49	90

geflaggten Gene nicht gilt.
Eine nähere Untersuchung der Gene mit uneinheitlichem Flagzustand der Genreplikate sollte mögliche fehlerhafte Flagzuweisungen aufspüren bzw. ausschließen.
Die Gene Elastin und TGF β3 sind zu 90% geflaggt und weisen somit einen einzelnen ungeflaggten Spot auf. Wie sich herausstellte, wird dies in beiden Fällen durch ein ungewöhnlich hohes Signal der ungeflaggten Replikate in einem der beiden Kanäle verursacht, während die zu 90% geflaggten Genreplikate zu niederige Signalwerte aufweisen, um ausgewertet werden zu können („negative“ Flags). Bei diesen beiden ungeflaggten Signalen handelt es sich - verglichen mit den Signalwerten (der anderen Genreplikate) dieses Kanals des selben Gens - um einen Ausreißer, was durch den Ausreißertest nach *Nalimov*[144] bestätigt werden konnte. Die hohen Signal in einem der Kanäle führen demzufolge zu der berechtigten Angabe eines verwertbaren Spots. Um dennoch die Ursache dieser hohen Werte zu ergründen, wurde die lokale Verteilung dieser Spots betrachtet. In beiden Fällen befinden sich die Spots direkt neben den Spots von

Replikaten von Genen (Collagen 3a1 und TGF β1), die insgesamt über hohe Signale - teilweise sogar in der Sättigung - in diesen Kanälen verfügen. Die hohen Signale der ungeflaggten Genreplikate in einem der beiden Kanäle der Gene Elastin und TGF β3 könnten demzufolge durch Überlagerung benachbarter Signal zustande gekommen sein, wie sie auch schon in der Literatur beschrieben wurden.[76]
Die Gene TNF α und VEGF zeigen ebenfalls eine deutlich erhöhte, von den anderen Genreplikaten nach *Nalimov* signifikant abweichende Signalwerte für den einen ungeflaggten Spot unter den Genreplikaten. Im Gegensatz zu den oben genannten Genen gilt die Signal-Abweichung jedoch für beide Kanäle. Eine Positionsabhängigkeit konnte nicht ausgemacht werden. Dafür wurde eine sehr geringe Anzahl an Spot-Pixeln für den Signalbereich gefunden. Fehler bei der Zuordnung der Anzahl an Pixeln zum Spot- und Hintergrundbereich könnten durchaus eine Ursache für fälschlich erhöhte Signalwerte sein. Dies würde die in beiden Kanälen erhöhten Werte erklären und ist vermutlich mit minimalen, lokal begrenzten Verunreinigungen verbunden. Es würde sich dabei jedoch eher um einen Fehler in der Histogrammabhängigen Primärdatenauswertung denn um eine falsche Flagzuordnung handeln. Lösungsansätze zu einer verbesserten Primärauswertung auf der Basis von Histogrammen wurden von Saffarian *et al.* beschrieben.[145] Die Entfernung der Ausreißer führt dazu, dass das mittlere Signal-Rausch-Verhältnis der beiden Kanäle von TNF α auf 0.25 sinkt, so dass die Gesamtheit aller Gene mit mehr als 50% geflaggten Genreplikaten ein mittleres Signal-Rausch-Verhältnis unter eins besitzt.
Wie Elastin und TGF β3, so besitzt auch das ungeflaggte Genreplikat der Gensequenzen von EGF nur in einem Kanal signifikant erhöhte Signalwerte (Cy3-Kanal). Dies kann aber weder auf lokale Signalübertragungen durch benachbarte Spots mit hoher Intensität noch durch eine außergewöhnlich kleine Pixelzahl des Signal begründet werden. Viel eher deutet die vergleichsweise sehr hohe interne Varianz des Spots daraufhin, dass es sich bei diesem Spot um einen sehr heterogenen Spot handelt. Ausnahmen besonders ungleichmäßiger Spots wurden bereits von Zhang *et al.* beschrieben und sind vermutlich auf die fehlerhafte Detektion des Cy3-Kanals durch den Laser zurückzuführen.[119] In diesem Fall, ist es jedoch verwunderlich, dass dieser Spot nicht geflaggt wurde, da abgesehen von „negativen“ Spots, also solchen mit geringen Signalintensitäten, dem Handbuch des Bildauswerte-Programms zufolge auch „schlechte“ Spots mit einer hohen internen Standardabweichung geflaggt werden sollten. Es bleibt zu vermuten, dass sich die berechnete Standardabweichung knapp unterhalb des entsprechenden (von GenePix Pro nicht genannten) Grenzwertes befindet.

Bei dem Gen FGF liegt der umgekehrte Fall vor: hier sind 90% der Genreplikate als verwertbar und nur ein Genreplikat als nicht verwertbar (geflaggt) eingestuft worden. Wie auch bei EGF, scheint die Diskrepanz (in diesem Fall andersherum) zwischen den ungeflaggten Signalwerten und dem geflaggten Signalwerte in Kanal Cy3 auf die Heterogenität dieses Spots zurückzuführen zu sein, wobei die erhöhte Varianz in diesem Fall jedoch als erfülltes Flagkriterium detektiert werden konnte.

Die Auswertung weiterer *low-density*- und eines *whole-genome*-Microarrays (siehe oben) unter dem Aspekt der Flag-Validierung ergab ähnliche Ergebnisse. Bei 97% der Flags handelt es sich um „negative“ Flags.
Vor allem der *whole-genome*-Microarrays mit Genen des Tomaten-Genoms wies 85% geflaggte Spots auf. Diese große Zahl erscheint zunächst ungewöhnlich viel, ist jedoch durch das Design des Experiments bedingt ist. Da *whole-genome*-Experimente darauf ausgerichtet sind, wenige, in weiteren Experimenten näher zu untersuchende Gene zu detektieren, ist die Mehrzahl der Gene im Vergleich zweier Zustände nicht reguliert und somit von geringer Intensität (siehe Kapitel 6.2.2). Eine visuelle Analyse dieser Daten mit dem Programm Findspot (siehe Kapitel 6.2.2) ergab, dass keine Spots fälschlicherweise geflaggten worden waren (jeweils sehr schwache ode heterogene Spots). Gleiches gilt für die *low-density*-Microarrays.
Hier wurden zudem - wie auch auf den hier vorgestellten Microarray mit spezifischen Rattengenen - nur wenige Gene gefunden, die im Vergleich der Gen-Replikat-Spots uneinheitlich geflaggt waren. Ausnahmen konnten durch Unreinheiten auf dem Chip und somit lokale Unterschiede erklärt werden, die vermutlich während des Herstellungs- und Hybridisierungsprozess entstanden sind.

Die relative Einheitlichkeit der Flagzustände (geflaggt bzw. ungeflaggt) innerhalb der Replikate (abgesehen von TNF β2 existieren unter zehn Spots pro Gen höchstens einer mit abweichendem Flagstatus auf - siehe Tabelle 6.7), die von jeder genspezifischen Sequenz auf ein Microarray gespottet wurden, verweist auf die Validität der durch das Bildauswerte-Programm vorgenommenen Beurteilung einzelner Spots bezüglich des Flag-Zustands. Die erwähnten Ausnahmen von ungeflaggten bzw. geflaggten Spots (10% bzw. 90% geflaggte Gene) konnten durch Chip-bedingte Variabilitäten erklärt werden. Die Entfernung der geflaggten Spots scheint berechtigt, da es sich bei diesen Spots um erwiesenermaßen mangelhafte Spots handelt. So werden zum einen Werte von Spots, die aufgrund technischer Probleme (z.B. fehlerhafte PCR-Reaktion oder unzureichende Flüssigkeitsübertrag beim Spotten) nicht verwendbar sind, und zum anderen Spots mit hoher Varianz entfernt. Letzteres betrifft vor allem Datenpunkte mit sehr niedriger Intensität, weil bei der Datenauswertung ausschließlich das Verhältnis der Expression eines Gens zwischen zwei Proben betrachtet wird. Da dieses über die Intensität der für das Gen spezifischen Spots berechnet wird, führt das Hintergrundrauschen bei niedrigen Intensitäten zu wesentlich höheren Abweichungen bei diesem Quotienten. Anzumerken bezüglich des automatischen Flag-Vorgangs ist allenfalls eine möglicherweise nicht ausreichende Kennzeichnung mangelhafter Spots (wie an dem Gen EGF gezeigt).

Diese Befunde verdeutlichen neben der erwiesenen Notwendigkeit, die Flags vor der weiteren Auswertung zu entfernen, außerdem die Unumgänglichkeit einer zusätzlich evaluierenden Qualitätsanalyse (siehe Kapitel 6.2.2) sowie eines Vergleichs der Genreplikate (siehe dazu Kapitel 6.2.4).
Auf die Qualitätsanalyse und dem damit potentiell einhergehenden Ausschluss fehlerhafter Microarrays folgt somit zunächst die Eliminierung aller zu den von GenePix Pro 6.0 geflagg-

ten Spots gehörigen Daten.

Subtraktion des Hintergrundes

Eine der gebräuchlichsten Methoden in der Vorverarbeitung ist die Hintergrund-Korrektur.[146,147] Sie berechnet sich wie folgt:

$$Spotwert = Median(Intensität) - Median(Hintergrund) \tag{6.1}$$

Zur Berechnung der einzelnen Hintergrund-bereinigten Intensitätswert aller Spots wird mit den Medianen aller Pixel eines Spots sowie des lokalen Hintergrunds gerechnet. Mediane haben im Vergleich zu Mittelwerten den Vorteil, dass sie Ausreißern gegenüber weniger anfällig und somit robuster sind.[148] Die um den Hintergrund bereinigten Intensitätswerte repräsentieren die Verteilung der Hybridisierung der gelabelten Probenmoleküle (*targets*) an die Fängermoleküle (*probe*) abzüglich nicht-spezifischer Hybridisierung und der natürlichen Fluoreszenz der Fängermoleküle selbst. Die Signalintensitäten sollten daher immer höher sein als die Hintergrundwerte, da negative Intensitätswerte keine verwertbaren Informationen liefern.
Solche negativen Differenzen können lokale Mängel des Chips darstellen, die beispielsweise durch Staub oder Schrammen auf der Oberfläche des Microarrays zustande kommen. Daher werden diese Werte in der Regel als unverlässliche Intensitätswerte verworfen.[149]

Da die Literatur zur Vorverarbeitung der Microarray-Daten inklusive der Korrektur des Hintergrunds weitestgehend auf *whole-genome*-Microarrays ausgerichtet ist, das im Rahmen dieser Doktorarbeit entwickelte Auswerteprogramm aber auch sdie Auswertung von *low-density*-Microarrays vorsieht, werden im Folgenden Intensitätswerte von Microarray-Experimenten, die um den Hintergrundbereich korrigiert worden sind, mit Intensitätswerten, bei denen keine Hintergrund-Korrektur stattgefunden hat, verglichen. Dazu werden erneut Microarrays aus unterschiedlichen Experimenten getestet. Die Ergebnisse dieser Untersuchung werden anhand eines Beispiel dargestellt, dass möglichst gute Bedingungen für einen solchen Vergleich bietet. Bei diesem Beispiel handelt es sich um ein *low-density*-Microarray-Experiment mit 13 Genen aus dem Bakterium *E.coli*, das gezielt für eine absolute Quantifizierung von *E.coli*-Genen entwickelt wurde (nähere Erläuterungen zu dem Versuch siehe Kapitel 6.2.4). Da dieses Microarray-Experiment nicht auf den Vergleich mehrerer Zustände abzielte, befindet sich auf den Microarrays jeweils nur der Cy5-Zustand. Die 20-fache, replikative Ausfertigung jedes Gens auf den Chips dieser Microarray-Familie eignet sich für eine signifikante Aussage über die Notwendigkeit der Hintergrund-Korrektur. Das Experiment besteht aus einer Verdünnungsreihe, die in dreifacher Wiederholung durchgeführt wurde (Verdünnungsreihe 1-3), wobei ein Verdünnungsschritt (1:20480) nur in einfacher Ausführung angefertigt wurde.
Da selbst das hier veranschaulichte *E.coli*-Experiment bereits aus 13 Chips bestand, die jeweils mindestens acht Mal gescannt wurden, wurde das Einlesen und Berechnen der resultierenden umfangreichen Datenmenge automatisiert, wozu das für die Validierung der Hintergrund-Korrektur implementierte Programm *Validate Background-Correction* genutzt wurde. Mithilfe

dieses Programms können alle zu einem Experiment zählenden Primärdaten-Dateien eingelesen und berechnet werden. Für die Validierung der Background-Korrektur wurden zunächst alle geflaggten Daten entfernt, um anschließend die Standardabweichungen der Mediane bzw. der Hintergrund-bereinigten Mediane (siehe Formel 6.1) über alle Genreplikate eines Gens berechnen und miteinander vergleichen zu können. Da sowohl Gene in der Sättigung (unverhältnismäßig kleine Standardabweichung) wie auch Gene, die ausschließlich aus geflaggten Genreplikaten (mittlerer Intensitätswert liegt bei null) bestehen, die Ergebnisse des Vergleichs verfälschen, wurden dazu die entsprechenden Gene durch das Programm *Validate Background-Correction* vorab herausgefiltert.
Jede Verdünnung wurde auf einen separaten Microarray hybridisiert, der wiederum mindestens acht Mal gescannt wurde (unterschiedliche Laserstärke und PMT-Verstärkung). Die Ergebnisse unterschiedlicher Scans wurden gemittelt und die relative Standardabweichung berechnet. Die Daten der ersten Verdünnungsreihe sind in Abbildung 6.9 dargestellt, die entsprechenden Daten befinden sich in Tabelle 6.8, die der zweiten und dritten Verdünnungsreihe in Anhang E.
Jede Säule repräsentiert somit den Mittelwert der Gesamtzahl der verwertbaren Gene über alle Scans eines Microarrays, die sich aus der mittleren Anzahl der Gene zusammensetzen, bei den die Standardabweichung (σ) mit Hintergrund-Bereinigung (gelb) größer ist und den Genen, bei denen die Standardabweichung ohne Hintergrund-Korrektur größer ist (blau).

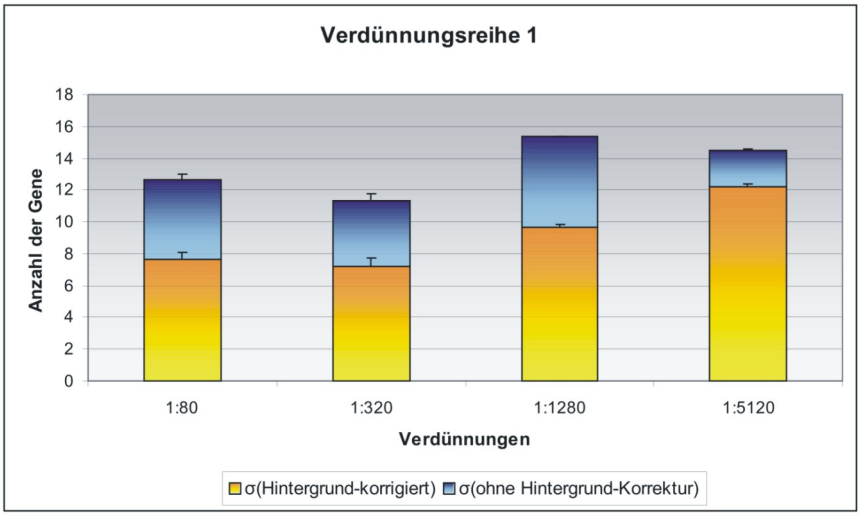


Abbildung 6.9 – *Validierung der Hintergrund-Korrektur. In der Abbildung sind die vier Verdünnungsschritte der ersten Verdünnungsreihe widergeben. Jeder Verdünnungsschritt befindet sich auf einem Chip, der mehrfach gescannt wurde. Die Ergebnisse der unterschiedlichen Scans wurden gemittelt und die relative Standardabweichung der gemittelten Daten berechnet. Die Säulen repräsentieren die unterschiedlichen Standardabweichungen der Gene nmit und ohne Hintergrund-Korrektur an.*

Wie in der Abbildung sowie der oben gezeigten Tabellen 6.8 und den im Anhang E befindlichen Tabellen zu sehen, weisen die Standardabweichungen der Hintergrund-bereinigten Mediane geringere Werte auf als die Mediane ohne Hintergrund-Korrektur. Eine detaillierte Untersuchung der Daten zu den Scan-Einstellungen, bei denen diese Aussage nicht zutrifft, sollte mögliche Regelmäßigkeiten in diesen Abweichungen aufdecken. Dabei konnte ein deutlich

Tabelle 6.8 – *Validierung der Hintergrund-Korrektur. In der Tabelle sind die vier Verdünnungsschritte der ersten Verdünnungsreihe widergeben. Jeder Verdünnungsschritt befindet sich auf einem Chip, der mehrfach gescannt wurde. Die Ergebnisse der unterschiedlichen Scans wurden gemittelt und die relative Standardabweichung der gemittelten Daten berechnet. Die zweite Spalte enthält die mittlere Anzahl der Gene (über alle Scans des entsprechenden Verdünnungsschrittes) mit geringerer Standardabweichung der Mediane aller Genreplikate nach der Hintergrund-Korrektur im Vergleich zu den unkorrigierter Medianwerten an. Spalte drei gibt die entsprechende mittlere Standardabweichung dieser Daten an. Spalte vier enthält die Anzahl der insgesamt auswertbaren Genen. Insgesamt sind 16 Gene auf den Microarrays vorhanden, von denen jedoch gesättigte bzw. vollständig geflaggte Gene eliminiert wurden.*

Verdünnungsreihe 1				
Verdünnung	$\overline{\sigma_{HG} < \sigma_{Median}}$	**rel.** $\overline{\sigma}$	$\overline{N_{auswertbar}}$	**rel.** $\overline{\sigma}$
1:80	8	0.42	13	0.33
1:320	7	0.53	11	0.46
1:1280	10	0.19	15	0.07
1:5120	12	0.21	15	0.10

sichtbarer Trend dahingehend gefunden werden, dass je weniger auswertbare Gene insgesamt unter den Daten einer bestimmten Scan-Einstellung zu finden sind, umso eher ist die mittlere Standardabweichung der Mediane ohne Hintergrund-Bereinigung geringer gegenüber der mittleren Standardabweichung der Mediane mit Hintergrund-Bereinigung. Wenn beispielsweise von 16 Genen nur fünf eine ausreichende Güte für diese Untersuchung aufweisen, also weder gesättigt noch vollständig geflaggt sind, so ist die Wahrscheinlichkeit, dass unter diesen fünf Genen die Mehrzahl der Gene eine geringere Standardabweichung ohne Hintergrund-Korrektur aufweisen größer als wenn alle 16 Gene auswertbar sind (siehe Beispiel Anhang E). Da jedoch davon ausgegangen werden kann, dass eine geringe Anzahl auswertbarer Gene Daten eines Scans von tendenziell minderer Qualität bedeuten und somit Abweichungen innerhalb der Genreplikate ohnehin eher zu erwarten sind, scheint die Hintergrund-Bereinigung insgesamt eine sinnvolle Maßnahme, um lokale Fehler auf den Microarrays zu korrigieren. Dies konnte auch durch Untersuchungen an weiteren Microarrays mit weniger Genreplikaten bestätigt werden

Obgleich eine Hintergrund-Korrektur der Medianwerte der Intensitäten in der Regel empfohlen wird, kann sie auch Nachteile mit sich bringen. Die Eliminierung der aus der Hintergrund-Bereinigung resultierenden negativen Werte führt zu fehlenden Datenpunkten. Dies kann beispielsweise bei einer möglicherweise anschließend durchgeführten Logarithmierung der Quotienten zweier Zustände auf einem Microarray zu Problemen führen.[72] Erste Forschungsansätze zur Lösung dieses Problems durch eine Transformation der negativen Differenzen in einen positiven Wertebereich werden jedoch bereits evaluiert.[150] Haaland *et al.* zweifelt ferner die prinzipielle Befähigung zur korrekten Bestimmung des lokalen Hintergrunds an.[151]
Da sich diese Methode jedoch beim der Untersuchung unterschiedlicher Microarray-Familien

durchaus bewährt hat, werden die Signalwerte in dieser Analyse nach Entfernung aller geflaggten Daten um den lokalen Hintergrund bereinigt.

Darstellung und Transformation

Eine routinemäßig graphisch sondierende Darstellung der rudimentär vorverarbeiteten Daten erleichtert eine erste Einschätzung der Ergebnisse der Microarray-Experimente. Auf diese Weise kann zu einem frühen Zeitpunkt der Analyse eine erste Abschätzung des Erfolgs des Experiments vorgenommen werden und mögliche spezifische Probleme erkundet werden. Zu solchen typischerweise auftretenden Problemen gehören Intensitätsabhängige Farbstoff-bedingte Fehler.
Sowohl bei der graphischen Darstellung wie auch bei der darauf folgenden Transformation sind jedoch bestimmte Grundannahmen zu berücksichtigten, die nicht auf alle Microarray-Experimente auch tatsächlich zutreffen. Diese Unterscheidung betrifft vor allem die Art des Microarrays: während *low-density* aufgrund der gezielten Auswahl bestimmter Gene so konzipiert werden, dass die Mehrzahl der Gene weitgehend differentiell exprimiert sind, gilt das Gegenteil für Experimente, die das ganze Genome eines Organismus untersuchen.
Das differentielle Expressionsmuster von *low-density*-Chips im Allgemeinen ist also in keiner Weise vorhersagbar. Im Gegensatz dazu werden bei *whole-genome*-Chips vorab die folgenden Annahmen getroffen:[81]

- Die Mehrheit der Gene ändert die Expression im Vergleich zweier Zustände nicht, so dass etwa die gleiche Anzahl an Genen hoch- und runterreguliert ist
- Die Gesamt-Intensität über alles Spots auf einem Microarray sollte gleich sein.

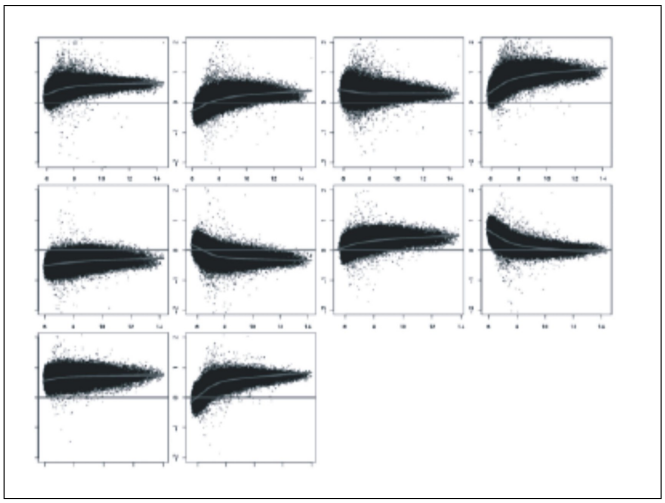
Basierend auf diesen beiden Annahmen können graphische Werkzeuge also durchaus hilfreich sein, um Abweichungen dieser Trends optisch vorab detektieren zu können und gegebenenfalls geeignete statistische Kompensations-Algorithmen anzuwenden.
Die erste und offensichtlichste Form der Darstellung ist sicherlich das während des Scan-Vorgangs generierte *tif*-Bild mit den übereinander gelegten Rohdaten der beiden unterschiedlich markierten Zustände. Solche Bilder ermöglichen einen Eindruck von der Einheitlichkeit der Spotformen und des Hintergrunds, sie deuten auf Artefakte von Staub oder Schrammen auf dem Microarray hin und geben einen ersten groben Überblick über die Stärke und Menge der Regulation der Gene.

Während die *tif*-Bilder einen sinnvollen Überblicke für *low-density* und *whole-genome*-Experimente gleichermaßen bieten, ist die im Folgenden vorgestellte Illustration der vorverarbeiteten Microarray-Daten nur für Microarray-Experimente geeignet, für die die oben definierten Prämissen Gültigkeit haben. Diese Bedingungen treffen auf *whole-genome*-Microarray zu, aber auch auf wenige, spezifische *low-density*-Experimente mit vielen Spots und einem Experiment-Design, das auf das Auffinden einiger wenige Gene ausgerichtet ist.
Für solche Experimente wird vor weiteren Darstellungen und Berechnungen eine Logarithmierung der Werte empfohlen.[152] Eine logarithmische Transformation bietet zweierlei Vor-

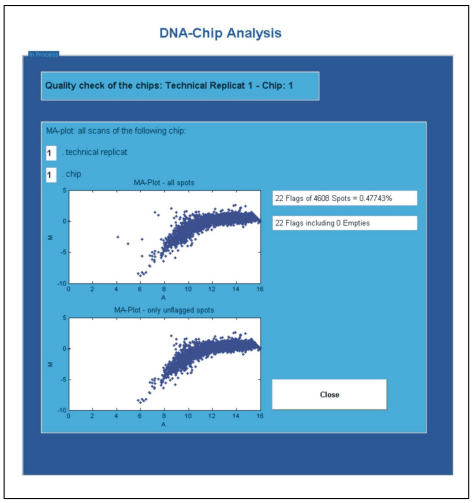
teile: Alle verbleibenden Rohdaten der Hintergrund-bereinigten Intensitätswerte sind positiv und erstrecken sich über einen großen Wertebereich von 0 bis 63535, von denen der überwiegende Teil in der unteren Hälfte dieser Spanne rangiert. Die Berechnung logarithmisch-transformierter Werte erübrigt also eine gesonderte Darstellung der Daten im unteren Wertebereich, da die Daten gleichmäßiger über die Spannweite der gesamten Größenordnung verteilt werden. Außerdem steigt die Höhe der Standardabweichung der Intensitätswerte typischerweise annäherungsweise mit der durchschnittlichen Signalstärke.[153] Die Logarithmierung der Werte tendiert daher dazu, aus den Quotienten zweier Zustände R/G (mit R für Rot und G für Grün) die Differenzen $M = log_2(R)$-$log_2(G)$ zu erzeugen. Wenn die oben definierten Annahmen durch das experimentelle Design des Microarray-Experiments erfüllt sind, können die Daten dieses Experiments also logarithmisch transformatiert werden, um so Daten zu liefern, die leichter interpretiert werden können.

Eine der gebräuchlichsten Darstellungsformen der einzelnen Microarrays eines Experiments ist der sogenannte *Scatterplot* der beiden Kanalintensitäten $log_2(R)$ und $log_2(G)$. Trotz des Vorteils einer einfachen Darstellung hebt diese Abbildung der Daten eher die Korrelation zwischen den Intensitäten der beiden Kanäle hervor anstatt den interessanteren Aspekt der Unterschiede zwischen den beiden Zuständen auf einem Chip zu verdeutlichen.[72] Das Hauptaugenmerk des Experimentators liegt also vielmehr auf den Abweichungen von der Diagonalen einer solchen Darstellungsform. Bei der von Dudoit *et al.* eingeführten Veranschaulichungsform der Daten, dem sogenannte *MA-Plot*, wird der *Scatterplot* daher um 45 Grad gedreht und die Achsen neu skaliert.[81] Demzufolge werden auf der Ordinate die logarithmischen Differenzen M (siehe oben) und auf der Abszisse die logarithmischen mittleren Intensitäten $A = log_2(\sqrt{RG})$ aufgetragen.
Eine solche Darstellung der vorverarbeiteten Daten wird auch von dem hier vorgestellten Microarray-Auswerte-Programm zur Verfügung gestellt. Die Eingaben der in Kapitel 6.2.1 vorgestellten Benutzeroberfläche werden dahingehend untersucht, ob erstens die Voraussetzung für eine solche Transformation und Darstellungsform gegeben sind (siehe oben) und zweitens eine Anzeige der Zwischenergebnisse vom Benutzer gewünscht wird. Abbildung 6.10 zeigt Beispiele für solche *MA-Plots*. In der linken Abbildung (Abbildung 10(a)) wurden *MA-Plots* von unterschiedlichen Verdünnungsreihen aus Leberzellen erstellt. In der rechten Graphik (Abbildung 10(b)) ist ein im Rahmen der Auswertung eines *whole-genome*-Microarray-Experiments erzeugter *MA-Plot* zu sehen. In diesem Experiment wurden zwei Zeitpunkte einer *E.coli*-Kultivierung miteinander verglichen. So wurde eine Probe dieser Kultur vor und eine Probe nach Induktion eines potentiell die Zellmembran perforierenden Proteins unterschiedlich markiert und auf einen Chip hybridisiert.

Abbildung 6.10 zeigt einen nach der Vorverarbeitung der Daten von dem Auswerte-Programm erzeugten *MA-plot*. Die obere Abbildung zeigt vergleichend *MA-Plots*, die von Bolstad *et al.* veröffentlicht wurden und die Variationenvielfalt dieser Darstellungsform verdeutlichen. Abbildung 10(b) hingegen beinhaltet einen *MA-Plot* aus institutsinternen Versuchen. Während im oberen Teil der Abbildung 6.10 alle Daten inklusive der geflaggten Spots gezeigt werden, wurden die Flags in der unteren Graphik vorab entfernt.



(a) Beispiele *MA-Plots* aus der Literatur[154]



(b) Beispiel *MA-Plot* eigene Experimenten

Abbildung 6.10 *– Beispiele für MA-Plots aus der Literatur und aus eigenen Experimenten. In Abbildung 10(a) sind Beispiele für einige in der Literatur veröffentlichte MA-Plots gezeigt, die aus unterschiedlichen Verdünnungsreihen mit Leberzellen entstanden sind.[154] In Abbildung 10(b) ist ein Beispiel eines MA-Plots dargestellt, der aus einem eigenen whole-genome-Experiment mit E.coli-Zellen erstellt wurde. Die obere Graphik dieser Abbildung schließt alle Datenpunkte ein. In der unteren Graphik wurden die geflaggten Spots eliminiert.*

Insgesamt verdeutlichen diese Beispiele die erweiterte Möglichkeit, die Spannbreite an Intensitäten, die diesen Experimenten innewohnt, graphisch zu erfassen und die Vereinfachung zur Erkennung nicht-linearer Zusammenhänge zwischen den logarithmischen Intensitäten. Die Darstellung der Intensitäten in Form solcher *MA-Plots* hebt außerdem den Zusammenhang zwischen dem Grad der differentiellen Expression und der Intensitätsstärke hervor, worauf in späteren Kapitel noch detaillierter eingegangen wird.

Abgesehen von dieser universell genutzten Darstellungsform existieren noch weitere Möglichkeiten der Darstellung, die auf den oben gemachten Annahmen der mehrheitlich unveränderten Expression und der gleichen Gesamtintensität beruhen.[83] Auf diese Darstellungsraten soll an dieser Stelle jedoch nicht weiter eingegangen werden soll, da diese Auswertung auch spezifisch für *low-density* entwickelt wurde, für die diese Annahme in der Regel nicht zutrifft.

GEO-Daten schreiben

Im letzten Schritt der Vorverarbeitung der Daten werden alle für die Veröffentlichung der Experimente in der *GEO*-Datenbank notwendigen Informationen zusammengetragen und in einer Datei gespeichert.[102] Durch diese Zusammenfassung aller nötigen Daten soll eine Veröffentlichung der Daten in der Literatur erleichtert werden.
Zu diesen Daten gehören für jeden Chip:

- die Gennamen für alle auf dem Chip befindlichen Spots,
- eine Zuordnungszahl zu den entsprechenden Positionen auf dem Microarray,

- die Summe der Mediane des Signals und
- die Summe der Mediane des Hintergrunds,
- Angaben zu den Flags sowie der
- Logarithmus zur Basis zwei des Quotienten der beiden Kanäle.

6.2.3.3 Fazit

Vor die eigentliche Auswertung der Microarray-Daten werden üblicherweise einige Vorverarbeitungsschritte vorgeschaltet. Diese Schritte zielen auf eine Bereinigung und erste Übersicht der Daten ab. Aus diesem Grund werden einerseits die von der Primärauswerte-Software für schlecht befundenen Daten (Flags) von der weiteren Auswertung ausgeschlossen. Andererseits wird von den Signalwerten aller Spots der entsprechende lokale Hintergrund subtrahiert, um lokale Unregelmäßigkeit zu kompensieren. Die Validität dieser Schritte wurde anhand einer Vielzahl von Daten aus unterschiedlichen Microarray-Experimenten, die mit verschiedenen Microarray-Familien durchgeführt worden waren, überprüft und bestätigt. Einen ersten optischen Eindruck dieser bereinigten Daten wird über den sogenannten *MA-Plot* vermittelt, der nicht-lineare Tendenzen aufzudecken vermag und einen groben Überblick über das Ausmaß der differentiellen Expression der Experimente verschafft. Darüberhinausgehend wurde eine Zusatzfunktion eingeführt, die die unmittelbare Veröffentlichung der Ergebnisse in *GEO* vereinfacht, indem die dafür erforderlichen Informationen in der Ergebnisdatei in separaten Tabellenblättern zur Verfügung gestellt werden.

6.2.4 *Within-Array*-Normalisierung

6.2.4.1 Hintergrund und Anforderungen

Der Vorgang der Normalisierung ist eine bioinformatische Notwendigkeit bei der Auswertung von Microarray-Daten, um Faktoren wie die unterschiedliche Inkorporation in cDNA und die unterschiedlichen molaren Fluoreszenzeigenschaften der Farbstoffe und andere Experiment-bedingte Variabilitäten zu kompensieren.
Einige Variationsarten, wie Unterschiede in der Expression eines Gens, können zwar höchstgradig informativ sein, da sie biologischen Ursprungs sind.[119] Andere Typen von Variationen sind jedoch unerwünscht und können die nachfolgende Analyse der Daten nachhaltig beeinflussen, da sie zu falschen Schlussfolgerungen bezüglich der Aussage des Experiments verleiten. Als Ursache solcher unerwünschten Variationen kommen insbesondere systematische Quellen in Betracht, die auf Eigenarten dieser Technologie zurückzuführen sind, und vor der Ermittlung der differentiellen Expression korrigiert werden sollten. Der Prozess der Entfernung bzw. Korrektur dieser systematischen Variabilitäten wird *Normalisierung* genannt und spielt eine besonders wichtige Rolle für die Qualität der Datenanalyse.
Aufgrund der vielfältigen Ursachen für systematische Fehler bei Microarray-Experimenten, wird dieser Vorgang üblicherweise entsprechend des kausalen Zusammenhangs dieser Fehler

in die *Within-Array*- und die *Between-Array*-Normalisierung unterteilt. Die im Folgenden vorgestellten Normalisierungs-Algorithmen haben den Ausgleich systematischer Fehler zum Ziel, die auf die Daten einzelner Microarrays (*within-array*) beschränkt und somit Chip-unabhängig sind.

6.2.4.2 Durchführung und Ergebnisse - Lineare Regression

Die Anzahl der Veröffentlichungen zur Microarray-Technologie sind entsprechend der gestiegenen Relevanz dieses Verfahrens in Forschungslaboren enorm. Sie betreffen vor allem das experimentelle Design, die Microarray-Herstellung, die RNA-Isolierung und Amplifikation, das Labeling- und Hybridisierungs-Verfahren und die Analyse der bei diesen Experimenten generierten Daten. Eine vernachlässigbar geringe Menge an Publikationen beschäftigen sich mit dem Post-Hybridisierungsprozess der Daten-Akquirierung. Dementsprechende Veröffentlichung zu Auswerteverfahren, die sich mit der statistischen Bewertung und Bereinigung möglicher Fehler dieses Vorgangs befassen, sind ebenfalls selten.[73]

Die mehrheitlich empfohlene Scan-Prozedur besteht aus dem einmaligen Scannen der Microarrays bei einer möglichst Sättigungseffekte minimierenden Scan-Einstellung.[155,156] Dafür wird die PMT-Stärke so gewählt, dass der komplette dynamische Bereich genutzt wird. Von steigenden PMT-Werte wird abgeraten, da diese das Signal-Rausch-Verhältnis nicht verbessern. Steigende PMT-Stärken verstärken eher das Rauschen als das Signal und führen somit zu unerwünschten Effekten. Andererseits führt auch eine verringerte PMT-Stärke aufgrund der weniger effizienten Photodetektion bei niedrigen Verstärkungen nicht zu einem verbesserten Signal-Rausch-Verhältnis.[119] Diese Tatsachen resultieren in der allgemein angenommenen Konsequenz, dass eine einzige Scan-Einstellung ausreichen sollte, da jede weitere Einstellung keine Verbesserung des Signal-Rausch-Verhältnisses bringt.
Da jedoch bei dieser Annahme immer von mittleren bzw. summativen Signal-Rausch-Verhältnissen ausgegangen wird, hat sie auch nur für das Mittel bzw. die Summe aller Spots Gültigkeit. Auf diese Weise kann daher zwar eine Scan-Einstellung mit möglichst hohem Signal-Rausch-Verhältnis für die Gesamtheit aller Spots eines Microarrays gefunden werden. Dabei werden aber immer auch spezifische Intensitätsbereiche in der Gesamt-Intensitätsspanne weniger gut berücksichtigt, so dass insgesamt ein Gesamtgefüge aus unterschiedlich gut erfassten Expressionswerten entsteht.
Die aus diesen Defiziten resultierenden methodischen Überlegungen orientieren sich daher an einer möglichst optimalen Einzelerfassung der Spots auf den Microarray, die letztlich eine optimale Gesamterfassung aller Spots ergibt. Dazu muss der gesamte Intensitätsbereich auf den Microarrays möglichst detailliert erfasst werden, was vor allem im unteren und oberen Bereich spezifische Schwierigkeiten mit sich bringt. Da das Signal-Rausch-Verhältnis von Spots mit niedrigen Intensitäten verhältnismäßig gering ist,[157] werden diese Spots oftmals in der Vorverarbeitung der Daten durch die Definition eines unteren Grenzwertes entfernt. Dadurch gehen wichtige Informationen zu der Expression dieser Gene verloren. Im oberen Intensitätsbereich tritt ab einem Signalwert von 65535 ($2^{16} - 1$) Sättigung ein, so dass keine differenzierten

Expressions-Daten zu den entsprechenden Genen vorliegen. Ein umfassender Algorithmus zur Erfassung Spot-spezifischer Signale sollte diese Einschränkungen berücksichtigten und korrigieren.

Statistische Auswertung

Der methodische Ansatz zur Lösung dieses Problems besteht darin, alternativ zu einer einfachen Scan-Einstellung (mit möglichst hohem allgemeinen Signal-Rausch-Verhältnis) möglichst viele unterschiedliche Scan-Einstellungen zu wählen. Dieses multiple Scan-Verfahren hat den Vorteil einer umfassenden Intensitätspalette für jeden Spots, aus der die verwertbaren Scans für die Ermittlung des Expressionswertes dieses Spots genutzt werden können.
Um jedoch eine Vorhersage über die Expressionsgrade der Gene unter Verwendung multipler Scans nutzen zu können, muss der Zusammenhang zwischen gemessenem Intensitätswert und dem wahren Wert der Expression bekannt sein. Laut Kerr *et al.* folgt der logarithmierte Intensitätswert einem linearen Trend.[158] Es kann daher ein universelles lineares Modell für den globalen Ansatz eines jeden Microarray-Experiments definiert werden, das alle verwendeten Microarrays beinhaltet. Unter Vernachlässigung aller Microarray-spezifischen Aspekte kann von diesem globalen Ansatz der bei der *Within-Array*-Normalisierung gültige lokale Ansatz abgeleitet werden.
Dem globalen *ANOVA*-Ansatz zufolge wird y_{ijkg} als die logarithmisch transformierte gemessene Intensität des Zustands k auf dem Microarray i für das Gen g definiert, das mit dem Farbstoff j gelabelt wurde. Kerr *et al.* zufolge stellen also der Chip, der Farbstoff, die Art der Behandlung (z.B. vor Induktion versus nach Induktion) und das jeweilige Gene die Komponenten dar, die dem größten Einfluss an Variationen ausgesetzt sind. Das auf dieser Annahme basierende *ANOVA*-Modell für die gemessenen, logarithmierten Intensitätswerte ist in der folgenden Gleichung widergegeben:

$$y_{ijkg} = \mu + \alpha_i + \theta_j + \beta_k + \gamma_g + (\alpha\gamma)_{ig} + (\theta\gamma)_{jg} + (\beta\gamma)_{kg} + \epsilon'_{ijkg} \tag{6.2}$$

μ steht dabei für den wahren Wert, α, θ, β und γ je für den Einfluss des Microarrays, des Farbstoffs, der Behandlung und des Gens. $(\alpha\gamma)$, $(\theta\gamma)$ und $(\beta\gamma)$ entsprechen der Interaktion zwischen Gen und jeweils Chip, Farbstoff und Behandlung. Die differentielle Genexpression wird demnach durch die Interaktion zwischen Gen und Behandlung ausgedrückt. Von diesen Haupteffekten werden also Unveränderliche definiert. Für den zufälligen Fehler ϵ'_{ijkg} wird eine unabhängige Verteilung angenommen. Die Reduktion dieser Gleichung auf die lokale Ebene des einzelnen Spots auf einem Microarray führt zu der stark vereinfachten Gleichung:

$$y_{ijkg} = \mu + \epsilon'_{ijkg} \tag{6.3}$$

Diese Gleichung verdeutlicht den linearen Zusammenhang zwischen Expressionswert und logarithmierten Signalwert. Diese lineare Abhängigkeit der Werte beruht auf den spezifischen Eigenschaften von Laser-Geräten.[159] Bei dem für die Quantifizierung der Fluoreszenzintensi-

täten hier verwendeten Laser handelt es sich um den Axon 4000B Scanner der Firma Axon Instruments (Axon Instruments, Foster City, CA), der zur Anregung der beiden Farbstoffe Cy3 und Cy5 zwei Laser benutzt sowie zwei Photoverstärker-Röhren zur Detektion des emittierten Lichtes. Die Verstärkung der Signale hängt von der Spannung zwischen Anode und Photokathode ab und soll laut Herstellerangaben optimalerweise in einem Bereich zwischen 500 und 900 Volt liegen, um ein gutes Verhältnis zwischen Signalintensität und Rauschen zu gewährleisten.
Der in Gleichung 6.3 widergegebene lineare Trend gilt demnach nur für einen eingeschränkten Bereich an PMT-Stärken und lässt sich aus eben dieser Abhängigkeit herleiten. So ergaben Untersuchung von Lyng *et al.* bezüglich des Einflusses unterschiedlicher Scanner auf die gemessenen Intensitäten eine logarithmisch-lineare Abhängigkeit zwischen der gemessenen Signalstärke und der PMT-Verstärkung,[160] die sich auf einen definierten dynamischen Bereich beschränkt. Einschränkungen dieser Abhängigkeit ergeben sich aber nicht nur durch die Gültigkeit für nur einen eingeschränkten (mittleren) Intensitätsbereich. Darüberhinausgehend variiert der Verlauf dieser Kurve bei gleicher PMT-Stärke für unterschiedliche Farbstoffe (Cy5 und Cy3).[125]
Eine Optimierung der Berechnung von Genexpressionswerten einzelner Spots muss demzufolge nicht nur einen eingeschränkten dynamischen Bereich beim Scannen der Microarrays berücksichtigen, sondern auch differierende Intensitätswerte in Abhängigkeit der Farbstoffe bei gleicher PMT-Einstellung.
Für die Implementierung eines entsprechenden Algorithmus wurden zunächst Chips aus unterschiedlichen Microarray-Familien mehrfach gescannt. Von den drei möglichen Laserstärken (10%, 33% und 100%) wurden nach ersten Vorversuchen die beiden höchsten Stärken gewählt, da mit der 33%-igen Einstellungen in Kombination mit niedrigen PMT-Verstärkungen alle relevanten Scan-Einstellungen im unteren Intensitätsbereich erfasst werden können und höhere Laserstärken zu einem besseren Signal-Rausch-Verhältnis beitragen.[160] Die PMT-Verstärkungen wurden so variiert, dass sie möglichst breite Intensitätsbereiche erfassen konnten. Insgesamt wurden also zwei unterschiedliche Laserstärken verwendet, die mit jeweils bis zu acht verschiedene PMT-Verstärkungen kombiniert wurden, so dass insgesamt 16 verschiedene Einstellungen für jeden Microarray vorlagen. Ein Beispiel für mögliche Scan-Einstellungen eines Microarrays ist in Anhang F gegeben.
Zur Wiederholung sei kurz erwähnt, dass es sich bei den für die *Within-Array*-Normalisierung vorliegenden Daten um die in der Präprozessierung um den Hintergrund bereinigten Microarray-Daten handelt. Für jeden Spot auf den individuellen Microarrays des entsprechenden Experiments existieren nun komplette Datensätze für unterschiedliche Scan-Einstellungen in beiden Farbstoffen. Die Informationen zu den mit den Datensätzen verbundenen Scan-Einstellungen wurden - wie in Kapitel 6.2.1 erwähnt - in eine Abfragemaske der Datenblätter eingegeben, so dass zu jedem Datensatz der unterschiedlichen Scans gleichzeitig die Kombination aus Laserstärke und PMT-Verstärkung vorliegt.

Für die Berechnung eines Spotwerts aus dieser Vielfalt an Daten für jeden Datenpunkt beginnt das Programm mit einer Sortierung der Datensätze nach den Scannereinstellungen. In einer

Vorsortierung werden die Datensätze nach der Laserstärke sortiert, so das für jeden Zustand eine Datenmenge aus zwei Gruppen generiert wird. Anschließend wird innerhalb dieser beiden Gruppen nach der PMT-Verstärkung sortiert - dieser Vorgang ist in Anhang F exemplarisch gezeigt.
Nach der Sortierung sollten also für jeden Spot in beiden Zuständen zwei Geraden der Laserstärken 33% und 100% mit jeweils bis zu acht Datenpunkten vorhanden sein. Da jedoch gerade Datenpunkte von Scans bei niedrigen PMT-Einstellungen geflaggt und somit in der Daten-Vorverarbeitung (Kapitel 6.2.3) entfernt worden sein können, existiert nicht für alle Spots die vollständige Zahl an Datenpunkten. Um jedoch eine sinnvolle Gerade durch eine Wertemenge legen zu können, ist eine minimal Anzahl von drei Punkten notwendig. Aus diesem Grund werden die entsprechenden Geraden inklusive aller dazugehörigen Datenpunkte gelöscht, wenn diese Voraussetzung nicht gegeben ist.
Im Gegensatz zur der in Standardmethoden vorgenommenen Eliminierung von Spots mit niedrigen Intensitätswerten, die aus einem einzigen Scan resultieren, werden bei dem hier vorgestellten multiplen Scan-Verfahren jedoch weniger Spots entfernt, da jeder Spot auch mehrfach im höheren PMT-Bereich gescannt wird. Während bei dem üblicherweise angewandten Scan-Prozess nur einfach im mittleren PMT-Bereich gemessen wird, gewährleistet das multiple Scannen auch ein mehrfaches Messen bei höheren PMT-Einstellungen, da auf ein optimales Gesamt-Signal-Rausch-Verhältnis in einem Scan über alle Spots verzichtet werden kann. Auf diese Weise können auch Spots im niederen Intensitätsbereich quantitativ erfasst werden.
Die validen Datenpunkte werden daraufhin auf den absoluten Sättigungswert von 65535 normiert, um den Wertebereich so auf Daten zwischen 0 und 1 zu beschränken.
Im Gegensatz zu dem von Dudley *et al.* vorgestellten Verfahren, welches multiple Scans zur *Within-Array*-Korrektur von Ratios der beiden Kanälen vorsieht,[161] arbeitet der hier vorgestellte Algorithmus mit den Einzelwerten der beiden Kanäle. Die separate Berechnung aller Spots der beiden Zustände kalkuliert somit die von Shi *et al.* beschriebenen Abweichungen des linearen Verlaufs bei unterschiedlichen Farbstoffe ein.[125]
Zusammenfassend kann also jeder Spot

- einem Microarray, und auf dem Microarray
- einem Zustand (Cy5 bzw. Cy3)

zugeordnet werden.

Eine Zusammenfassung der Daten - wie Dudley *et al.* sie durch die Quotientenbildung beider Zustände vornimmt[161] - findet an dieser Stelle also noch nicht statt.

Dementsprechend liegen für jeden Spot die folgenden Daten vor:

- Zwei Gerade, die je bei einer konstanten Laserstärke (33% oder 100%) gescannt wurden, mit
- je vier bis acht Datenpunkten (PMT-Verstärkung = 400 bis maximal 800 für Cy3 bzw.

maximal 1048 für Cy5) - je nach Anzahl unterschiedlicher PMT-Verstärkungen bei einer konstanten Laserstärke.

Basierend auf der Annahme der linearen Abhängigkeit zwischen logarithmierten Intensitätswerten und PMT-Verstärkung, die in Gleichung 6.3 resultiert, werden die vorliegenden normierten Geraden eines jeden Spots in beiden Kanälen nacheinander auf Sättigungseffekte kontrolliert.

Detektion von Sättigungseffekten

Die einfachste Methode, um den Sättigungseffekte zu detektieren, ist die Festlegung eines Grenzwertes, oberhalb dessen alle Werte als Sättigungswerte gekennzeichnet werden.[162] Dieser Annahme liegt ein einfaches Sättigungsmodell zugrunde, das beispielsweise für Analog-zu-Digital-Konvertier-Geräten zutrifft, *Clipping* genannt wird und ab einem fixen Grenzwert alle Daten auf diesen Grenzwert herabsetzt.[163] Insbesondere für Photodetektoren und andere optoelektrische Gerätschaften können Sättigungseffekte treffender durch graduellere Modelle wie die sogenannte *gamma*-Sättigung beschrieben werden. Aus diesem Grund wurde eine erweitere Methode entwickelt, die in einem variablen Datenbereich die Detektion von Sättigungseffekten zulässt. Die in Abbildung 6.11 und 6.12 gezeigten Graphen sind Beispiele der graphischen Darstellungen der Zwischenergebnisse bei der Auswertung eines *whole-genome*-Microarrays mit Genen des Bakteriums *E.coli* entnommen.

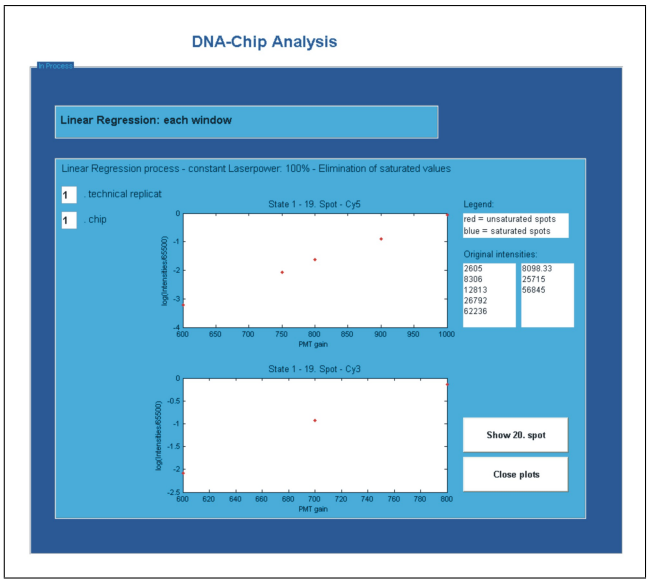


Abbildung 6.11 – *Auffinden von Sättigungseffekten am Beispiel eines whole-genome-Microarrays mit Genen des Bakteriums E.coli. Dargestellt sind je die logarithmierten normierten Intensitätswerte gegen die PMT-Verstärkungen. Die obere Gerade spiegelt die Datenpunkte eines Gens bei einer konstanten Laserstärke von 100% im Cy5-Kanal wider. Die untere Graphik zeigt den entsprechenden Cy3-Kanal bei konstanter Laserpower von 100%. Je nach Anzahl geflaggter und somit eliminierte Datenpunkte, kann eine Gerade aus mehr oder weniger Wertepaare bestehen. Rote Datenpunkte stehen für ungesättigte Intensitätswerte, blaue dagegen für gesättigte. Bei diesem Spot wurde bei konstanter Laserpower von 100% in beiden Kanälen keine Sättigung gefunden.*

Während die Geraden des Spots in Abbildung 6.11 keine Sättigungseffekte (bei einer Laserstärke von 100%) aufweisen, wurden in Abbildung 6.12 im Cy5-Kanal zwei Datenpunkte bei PMT-Verstärkungen 900 von 1000 und detektiert, die Sättigungseffekte bei einer konstanten Laserstärke von 100% zeigen.

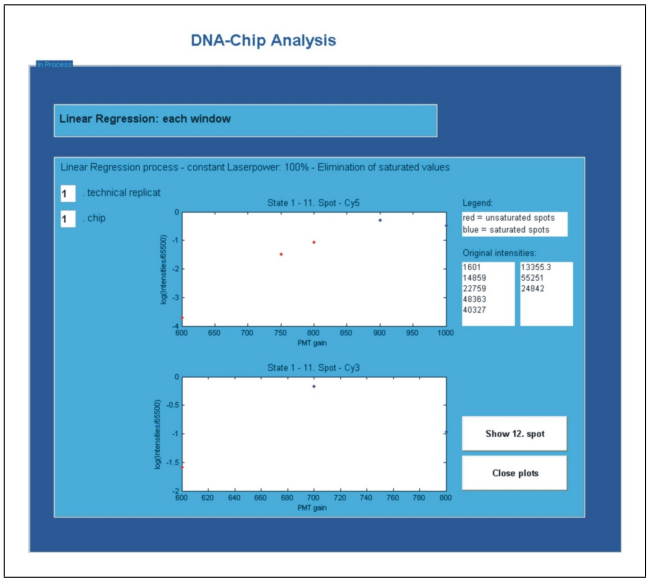


Abbildung 6.12 – *Auffinden von Sättigungseffekten am Beispiel eines whole-genome-Microarrays mit Genen des Bakteriums E.coli. Dargestellt sind je die logarithmierten normierten Intensitätswerte gegen die PMT-Verstärkungen. Es gelten die gleichen Angaben wie in Abbildung 6.11. Bei diesem Spot wurden jedoch bei einer Laserstärke von 100% Sättigungseffekte im Cy5-Kanal detektiert.*

Bei diesem Versuch handelt es sich um das bereits in Kapitel 6.2.3 beschriebene Microarray-Experiment. Der in den Abbildungen 6.11 und 6.12 gezeigte Spot befindet sich somit auf dem ebenfalls in Abbildung 6.10 als *MA-Plot* dargestellten Microarray.

Die Funktion zum Finden von Sättigungseffekten beruht - wie die Auftragung zeigt - auf der Bildung von Geraden aus den logarithmisch transformierten normiertem Intensitätswert gegen die entsprechende PMT-Verstärkungen bei konstanter Laserstärke. Die Detektion von Sättigungseffekten basiert auf der Annahme eines nicht-linearen Verlaufs der Kurve in diesem Bereich. Da diese Effekte im oberen Bereich der Kurve auftreten, setzt die Suchmaske im oberen Bereich der PMT-Verstärkung ein und setzt sich punktuell fort: zunächst werden die letzten beiden Datenpunkte der Gesamt-Geraden herangezogen und eine Gerade durch diese Punkte gelegt, wenn die Steigung dieser „Unter“-Geraden nicht positiv ist, werden die Datenpunkte markiert. Anschließend wird ein weiterer Datenpunkte (der drittletzte) in die Gerade mit aufgenommen und erneut auf das Vorzeichen der Steigung dieser um einen Datenpunkt erweiterten „Unter“-Geraden getestet. Dieser Vorgang wird so lange durchlaufen, bis alle Datenpunkte nach und nach in die Anfangsgerade mit aufgenommen wurden. Daraufhin werden diejenigen Datenpunkte, die vom letzten Datenpunkt fortlaufend eine nicht-positive Steigung der entsprechenden „Unter“-Geraden aufweisen, als Sättigungswerte markiert und aus der Gesamt-Geraden ausgeschnitten. Die Festlegung eines zusätzlichen Mindestwertes für Sättigungspunkte erübrigt sich, da mit dieser Methode ohnehin nur Werte über 40000 gefunden wurden, einem Bereich also, in dem der Literatur zufolge bereits Sättigungseffekte zu

vermerken sind.

Ermittlung der Werte im niedrigen Intensitätsbereich

Im unteren Intensitätsbereich weisen die markierten Spots zwar keine Sättigungseffekte auf, zeigen jedoch ebenfalls nicht-lineares Verhalten. Die markierten Datenpunkten dieser Spots liegen unterhalb der Bestimmungsgrenze, so dass Scans unterschiedlicher PMT-Einstellungen bei konstanter Laserstärke aufgrund der hohen Varianz stark schwanken. Bei diesen Spots wurden mit der oben beschriebenen Selektionsmethode, die lineare Verläufe untersucht, jeweils immer mehr als die Hälfte der Datenpunkte markiert. Aufgrund dieses insgesamt nicht-linearen Verhaltens, ergeben diese Spot keine verwertbaren Expressionsinformation und können somit als nicht exprimiert von der weiteren Auswertung ausgeschlossen werden. Durch dieses Verfahren fallen also nur solche Spots mit niedrigen Intensitätswerten raus, die erwiesenermaßen nicht verwertbar sind, da eine Linearität des Kurvenverlaufs (auch bei hohen PMT-Verstärkungen) nicht gegeben ist. Im Gegensatz dazu werden bei der standardmäßigen Festlegung eines fixen Grenzwertes pauschal alle Spots unterhalb dieses Grenzwertes eliminiert, so dass durchaus verwertbare Spots nicht berücksichtigt werden können.

Im weiteren Verlauf des Programms werden die Geraden mit Sättigungseffekten extrapoliert, so dass die eliminierten Datenpunkte ersetzt werden. Zu diesem Zweck werden die validen Daten linear regressiert, in dem eine Gerade durch die validen Datenpunkte gelegt wird. Unter Verwendung der Steigung und des Achsenabschnitt dieser Geraden können neue Modellwerte für die Wertepaare aus x- (PMT-Verstärkungen) und y-Werten (logarithmisch transformierte normierte Intensitätswerte) berechnet werden.

Ausgehend von der in der Literatur beschriebenen Beobachtung, dass die Linearität der Geraden bei unterschiedlichen PMT-Verstärkungen eines Scanners leichte Abweichungen zeigt,[125] werden die PMT-Werte darauf folgend mithilfe eines Simplex-Algorithmus über alle validen Geraden so optimiert, dass der Gesamt-Korrelationskoeffizient aller Geraden maximal war. Da es dabei nur zu leichten Korrekturen der x-Werte kommt, kann ein Zeit-sparender, lokal arbeitender Optimierungsalgorithmus wie der Simplex-Algorithmus verwendet werden. Die Korrektur der Daten erfolgt über die Minimierung der in der folgenden Gleichung gezeigten Fehlerquadratsumme:[164]

$$F := \sum_{i=1}^{n} (f(x_i) - y_i)^2 \tag{6.4}$$

mit $f(x_i) = ax_i + b$ gilt demzufolge:

$$F(a, b) := \sum_{i=1}^{n} (ax_i - b - y_i)^2 \tag{6.5}$$

In den auf diese Weise optimierten Geraden können sich nach wie vor Datenpunkte befinden, die statistisch signifikante Abweichungen der restlichen Geraden darstellen und somit deutlich außerhalb der Geraden platziert sind. Die Detektion solcher Ausreißer kann mithilfe des Mandel-Tests erfolgen.
Der Mandeltest ist ein statistischer Linearitätstest, dessen Nullhypothese H_0 annimmt, dass keine Varianzinhomogenitäten beim Vergleich der Reststandardabweichungen zweier Regression zu erkennen (P=99%) sind. Die erste der beiden zu vergleichenden Regressionen stellt immer die Gerade inklusive aller Datenpunkten dar. Bei der zweiten Gleichung handelt es sich um die selbe Gerade, aus der jedoch vorab der Reihe nach ein Datenpunkt entfernt wurde. Für den Anpassungstest nach Mandel wird die Varianzdifferenz Δs^2 der beiden Gerade berechnet, um so eine Aussage darüber treffen zu können, ob Varianzinhomogenitäten vorliegen.

$$\Delta s^2 = [(N_A - 2) \cdot s_A^2] - [(N_{NA} - 2) \cdot s_{NA}^2] \tag{6.6}$$

Mit dem Index A als linearer Anpassung *aller* Datenpunkte und dem Index NA als der linearer Anpassung der um einen Datenpunkt reduzierten Geraden (*nicht alle* Datenpunkte). Die Anzahl verwendeter Datenpunkte wird mit N angegeben.
Die berechnete Varianzendifferenz Δs^2 wird über einen F-Test abgeglichen. Dazu wird die Prüfgröße PG berechnet nach:

$$PG = \frac{\Delta s^2}{s_{NA}^2} \tag{6.7}$$

Auf den Ausreißertest folgt ein zweiter Optimierungsdurchlauf mit dem Simplex-Algorithmus, um möglichst optimale Parameter für die Geraden der Spots zu gewährleisten. Die daraus resultierenden Geraden werden letztlich zur Bestimmung eines neuen Spotwertes herangezogen.
Da für jeden Spot jedoch potentiell zwei Geraden vorliegen (eine bei einer konstanten Laserstärke von 33% und eine bei einer konstanten Laserstärke von 100%), die sich wiederum aus einer unterschiedlichen Anzahl an Datenpunkten (je nach Anzahl ungeflaggter Werte, die bei den verschiedenen PMT-Verstärkungen aufgenommen wurden), müssen diese Daten miteinander zu einem Wert verrechnet werden. Demnach müssen zunächst die beiden Geraden miteinander verrechnet werden bevor Einzelwerte für die Spots ermittelt werden können. Für die Mittelung der Geraden wurden zwei Verfahren konstruiert, die in unterschiedlichen Varianten umgesetzt wurden, um die bestmögliche Berechnungsmethode eines mittleren Geradenwertes zu ergründen:

1. Das erste Verfahren sieht die **Projektion** einer der beiden Geraden auf die andere vor. Bei dieser Projektionsmethode werden also beispielsweise die Datenpunkte der Geraden, die bei einer konstanten Laserstärke von 33% entstanden ist, auf die Gerade mit einer unveränderlichen Laserpower von 100% verschoben. An die Verwirklichung dieses Verfahrens wurde auf unterschiedliche Weise herangegangen:

    - Die Bestimmung des Abstandes für die Projektion konnte einerseits über den mittleren Abstand aller bei gleicher PMT-Verstärkung vorliegenden Datenpunkte erfolgen. Diese Abständen wurden unter Anwendung des Tests nach *Nalimov* auf Aus-

reißer untersucht, die vor der Mittelwertbildung für einen einheitlichen Projektions-Abstand entfernt wurden.

- Eine andere Methode der Abstandsberechnung bestand Ermittlung einer mittleren Differenz unter Verwendung der Achsenabschnitte der Regressionsgeraden aller Spots. Auch hier wurden vorab Ausreißer mit dem *Nalimov*-Test gefunden und eliminiert.

Mit diesen jeweils unterschiedlich berechneten, fixen Abstand wurden anschließend alle Datenpunkte auf die jeweils andere Gerade projiziert.

2. Das zweite Verfahren beruht auf der Bildung des **Mittelwerts aus den Steigungen der Regressionsgeraden** der beiden Kurven bei unterschiedlichen Laserstärken.

   - Die Ermittlung des Mittelwerts erfolgte einerseits unter Verwendung aller validen (ungeflaggten, ungesättigten und von Ausreißern befreiten) Datenpunkte beider Geraden. Die Steigungen der sich ergebenden Regressionsgeraden werden zu einer mittleren Steigung zusammengefasst.
   - Ein anderer Ansatz zur Ermittlung einer mittleren Steigung aus den beiden Einzelgeraden der unterschiedlichen Laserstärken sah die Mittelwertbildung von gewichteten Steigungen vor. Die Güte der Geraden wurde über die Korrelationskoeffizienten bestimmt, in dem diese anteilig in die Wichtung der Geraden eingingen.

Die Zwischenergebnisse dieser Methoden können nach der Berechnung optional graphisch verfolgt und beurteilt werden.
Das Projektions-Verfahren stellte sich als weniger robust heraus gegenüber dem Verfahren, bei dem die mittlere Steigung der beiden Geraden berechnet wurde. Die größere Anfälligkeit gegenüber Geraden geringer Güte führte zu einer ungleich gewichteten Verschiebung der Datenpunkte, so dass die Werte der schlechteren Gerade die Linearität der neuen Gesamt-Gerade stark beeinträchtigten. Außerdem verkompliziert dieses Verfahren die Bildung einer neuen Gesamtgeraden, wenn eine der beiden Geraden aufgrund mangelnder Güte nicht existent ist (alle Datenpunkte geflaggt oder während der Sättigungsfindung eliminiert). In diesem Fall muss unter Umständen trotz einer fehlenden Geraden eine Geradenprojektion erfolgen, um die Einheit bezüglich der anderen Spots beizubehalten. Die Verwendung der Regressionssteigungen führt insgesamt zu weniger Fehlern. Besonders die Wichtung der Regressionsgeraden bezüglich ihres Korrelationskoeffizienten führte zu guten Ergebnissen.
Die beide Einzelgeraden umfassende neue Gesamt-Gerade kann nun herangezogen werden, um aus der Vielfalt an Datenpunkten für einen Spot jeden Zustands einen einzelnen Spotwert zu berechnen. Zu diesem Zweck wird eine durch die neue Gerade verlaufende Regressionsgerade berechnet. Um eine vergleichbare Einheit der Spot-Intensitäten zu gewährleisten, wird anschließend mit der Steigung und dem Achsenabschnitt der zusammengefassten Geraden ein neuer y-Wert (logarithmisch transformierter normierter Intensitätswert) bei einer mittleren PMT-Verstärkung von 700 berechnet.

Eine sinnvolle Validierung dieser Methode zur Ermittlung der Spotwerte im Vergleich zu Standardmethoden setzt einen möglichst verlässlichen Kenntnisstand der wahren Intensitätswerte voraus. Da die Ergebnisse von Microarray-Experimenten typischerweise nur in Form relativer Vergleiche zweier Zustände vorliegen und somit keine absolute Aussage bezüglich der Expressionsgrade der Zustände zulassen, wurde ein experimentelles Verfahren entwickelt, welches eine absolute Quantifizierung der Expressionswerte der Gene ermöglichte. Die Kenntnis über die absolute Menge an RNA auf dem Microarray ermöglicht einen direkten Abgleich der in den Microarray-Experimenten ermittelten Expression mit der bekannten RNA-Menge in der auf den Microarray aufgetragenen Probe. Diese Methode stellte also insofern eine Neuheit dar, als dass der mit zusätzlichen Variabilitäten behaftete Vergleich zweier Zustände auf einem Microarray durch die einfache Hybridisierung eines Farbstoffes (Cy5) ersetzt werden kann.[72] Ferner kann das Ausmaß der verbleibenden Variabilitäten durch einen Vergleich der erwarteten Werte mit den Ergebnissen abgeschätzt werden. Insgesamt kann das Verfahren der absoluten Quantifizierung also nicht nur zur Validierung des implementierten Auswerteverfahrens multipler Scans herangezogen werden, sondern bietet zugleich vergleichsweise zu relativen Methoden einen realistischen Spiegel der Varianzen im Zusammenhang mit Microarray-Experimenten.
Um die Ergebnisse der Auswertemethode vergleichend beurteilen zu können, wurde parallel zu der oben beschriebenen *Within-Array*-Normalisierungmethode ein Standardauswerteverfahren angewandt. Diese standardmäßig angewandte Methode beinhaltet die Division der Einzelspots einer Scan-Einstellung durch die Gesamt-Summe über alle Datenpunkte eines Scans. Auf diese Weise sollen ungleiche Expressionswerte, die auf verschiedenen Scan-Einstellungen beruhen, korrigiert werden.

Material und Methoden zur Validierung der Auswertemethode

Für die Validierung der Methode wurden eigens Microarrays hergestellt, die eine absolute Quantifizierung der RNA-Spezies in der Probe ermöglichten. Diese spezifischen *low-density*-Chips tragen 13 unterschiedliche Oligonukleotide der Firma Eurofins MWG Operon (Ebersberg) für die *E.coli*-Gene LacY, LacA, DnaK, TatD, SecD, FtsH, ProA, Signalpeptidase2 und GlyA so wie vier künstliche Gene (RFP und drei unterschiedliche GFP-Arten, die komplementär an das gleiche Probenmolekül binden). Die Oligonukleotide wurden gemäß der Vorgaben von Kane et al. designt[165] und in 20facher Ausfertigung je auf die Microarrays aufgetragen.
Die Präparation der komplementären Probenmoleküle bekannter Konzentrationen erfolgte durch eine *in vitro*-Transkription. Die Längen der dabei verwendeten PCR-Produkte reichten von 457 bis 960 Basenpaaren. Die erzeugten RNA-Moleküle wurden in einem Reversen Transkriptionsschritt in cDNA umgeschrieben und gleichzeitig mit Cy5 markiert. Die Einbaurate sowie die Menge an cDNA wurde mit dem Nanodrop Spektrometer aufgenommen. Die Proben wurden auf die Microarray hybridisiert und bei unterschiedlichen Scan-Einstellungen gemessen.
Für detaillierte Versuchsangaben zur Herstellung und Hybridisierung der Microarrays sei auf

Anhang G hingewiesen.
In Tabelle 6.9 sind die durchgeführten Ergebnisse zusammengefasst.

Tabelle 6.9 – *Zusammenfassung des Microarray-Experiments zur absoluten Quantifizierung*

Verdünnung	**unabhängige Verdünnungsreihe (VR)**		
	VR 1	**VR 2**	**VR 3**
1:80	X	X	X
1:320	X	X	X
1:1280	X	X	X
1:5120	X	X	X
1:20480			X

Wie in Tabelle 6.9 dargestellt, wurden drei unabhängige Verdünnungsreihen durchgeführt, wobei die dritte Verdünnungsreihe einen zusätzlichen Verdünnungsschritt umfasste.

Neben diesen dreifach durchgeführten Experimenten mit unterschiedlichen Konzentrationen, die jeweils mit dem Faktor vier weiterverdünnt wurden, und bei denen alle auf den Microarrays immobilisierten Gene auch in den hybridisierten Lösungen vorlagen, wurden weitere Versuche durchgeführt, bei denen lediglich ein Gen, die Signalpeptidase2, untersucht wurde. Diese Versuche wurden in den Verdünnungen 1:10, 1:100, 1:1000 und 1:10000 ausgeführt und dienten der Ermittlung der Konzentrationsabhängigkeit der Signalintensitäten (im Gegensatz zur Abhängigkeit von der Basenpaarlänge bzw. dem dCTP-Gehalt).

Die von Shi *et al.* genannten Einschränkungen multipler Scans umfassen neben den bereits erwähnten divergierenden Resultaten unterschiedlicher Farbstoffe bei gleichen PMT-Verstärkungen des Weiteren vermutete Bleichungseffekte (*Photobleaching*) durch mehrfaches Scannen und einen potentiellen hohen experimentellen Zeitaufwand.[125] Wie bereits angeführt, kann der durch unterschiedliche Farbstoffe erzeugte differierende Einfluss der PMT-Verstärkungen durch Vermeidung einer vorangegangenen Quotientenbildung umgangen werden, in dem eine separate Berechnung der beiden Zustände eines Chips vorgenommen wird.
Um darüberhinausgehend die weiteren, erwähnten potentiell negativen Einflussfaktoren mehrfachen Scannens ausschließen zu können, wurde die Microarrays zunächst auf *Bleaching*-Effekte getestet.[166] Zu diesem Zweck wurde ein Microarray einer Versuchsreihe mit dualer Markierung (Apoptose-Microarray) zwölf Mal bei den höchsten Laser-Einstellungen (Laserstärke = 100%, PMT-Verstärkung = 1043 im Cy5-Kanal und 800 im Cy3-Kanal) gescannt. Die Medianwerte der beiden Zustände wurden in der Reihenfolge der vorgenommenen Scans aufgezeichnet und auf durch Bleicheffekte verursachte abnehmende Intensitätswerte hin untersucht. Diese Analyse ergab zwar eine mittlere relative Standardabweichung der Scanwerte von 0.05 über alle Gene, es konnte jedoch kein einheitlich abnehmender Trend der Signalwerte festgestellt werden. Dieser Befund der allenfalls sehr mäßigen *Bleaching*-Effekte wurde eben-

falls durch Skibbe *et al.*, Romualdi *et al.* und Hsieh *et al.* bestätigt.[73,167,168]

Der Aspekt des Zeitaufwands multipler Scans wurde getestet, in dem die Zeit für den Scanprozess eines *whole-genome*-Microarrays des Bakteriums *E.coli* mit 4608 Genen gestoppt wurde. Diese Messung ergab eine maximale Dauer von zwei Minuten bei einer Auflösung von 10 μm für beliebige Scan-Einstellungen. Unter Berücksichtigung der Vorgabe, dass für jede der beiden Laserstärken (33% und 100%) jeweils mindestens vier unterschiedliche PMT-Verstärkungen erforderlich sind, beträgt die Dauer für das Scannen von *whole-genome*-Microarrays mit dem Genom des Organismus *E.coli* insgesamt also mindestens 16 Minuten. Entsprechend vergrößert sich der Zeitaufwand für Chips mit einer größeren Anzahl an Genen und verringert sich für *low-density*-Microarrays. Verglichen mit dem gesamten experimentellen Aufwand sowie unter Beachtung der Tatsache, dass das Auswerten multipler Scans eine signifikante Verbesserung der Auswertung hervorbringt (siehe unten), stellt dieser vergleichsweise einen vernachlässigbare Einschränkung dar.

Im vorgegangenen Abschnitt wurde gezeigt, dass die in der Literatur erwogenen Einwände bezüglich einer multiplen Scan-Prozedur keine hinderlichen Einschränkungen. Im folgenden Abschnitt soll belegt werden, dass die hier vorgestellte, auf mehrfachen Scans beruhende Auswertemethode auch eine signifikante Verbesserungen der Expressionsdaten mit sich bringt.

Statistische Berechnungen zur absoluten Quantifikation

Da die Neuheit dieser Methode sowohl die Nutzbarmachung multipler Scans sowie das Herausfiltern und Korrigieren von Sättigungseffekten enthält, wurde zum Vergleich eine Standardmethode herangezogen, die keine der genannten Implementierungen beinhaltet (siehe oben). Durch die quantitative Festlegung absoluter Werte kann eine direkte Korrelation zwischen Erwartungs- und Ergebniswerten (der beiden zu vergleichenden Auswerteverfahren) vollzogen werden. Zu diesem Zweck wurde einerseits die Anzahl der mRNA-Moleküle aller auf den Microarrays befindlichen Gene ermittelt und andererseits die Anzahl der Desoxycytidintriphosphate (dCTPs) jener Gene. Die Anzahl der Moleküle lässt sich mit der folgenden Formel berechnen:[169]

$$x_{RNA} = \frac{m_{RNA}/V_F}{M_{RNA}} \cdot N_A \tag{6.8}$$

wobei x_{RNA} für die Anzahl der RNA-Moleküle steht, m_{RNA} für die Nukleotidmasse, V_F für den Verdünnungsfaktor, M_{RNA} für die Molekularmasse und N_A für die Avogadrokonstante. Der dCTP-Gehalt der Gene spielt insofern eine wichtige Rolle, als dass die Fluoreszenz-Markierung der auf die Microarrays hybridisierten cDNA-Moleküle über den Einbau gelabelter dCTP-Moleküle erfolgt. Da die Schnittstellen der gesamten ursprünglichen Sequenz bekannt sind, kann über die eingebaute Sequenz auch die Anzahl an dCTPs berechnet werden. Prinzipiell wird bei ausreichender Länge der Probenmoleküle zwar eine Gleichverteilung

der Nukleotide (dATP, dCTP, dGTP, dTTP) in den unterschiedlichen cDNA-Sequenzen vorausgesetzt, so dass die Molekülzahl bzw. die Länge der RNA direkt mit der Anzahl eingebauter dCTPs korrelieren sollte.[170] Da in diesen Versuchen jedoch designte Nukleinsäure-Fragmente eingesetzt wurden, sollte vorab getestet werden, ob diese Annahme in den hier vorgestellten Experimenten Gültigkeit besitzt. Die Berechnung der Verhältnisse aus der Gesamtzahl der Nukleotide und dem dCTP-Gehalt der Gensequenzen ergab einen mittleren Faktor von 4.14 mit einer relativen Standardabweichung von 0.13 über alle Gene. Aufgrund der relativ hohen Abweichungen der dCTP-Gehalte der Gensequenzen, sollte zwischen der Basenpaarlänge und dem dCTP-Gehalt unterschieden werden. Daher werden im Folgenden die Berechnungen zunächst parallel mit beiden Größen durchgeführt.

Ermittlung der Güte der Auswertemethoden mit Hilfe von Genreplikaten

Eine weitere Aussage über die Güte der unterschiedlichen Verfahren kann anhand der Genreplikate getroffen werden. Da die Intensitätswerte der Genreplikate möglichst gering voneinander abweichen sollten, ermöglicht der Vergleich der relativen Standardabweichungen der Genreplikate der Gene eine Aussage über die Qualität der Auswertung. Diese Bewertung der unterschiedlichen Auswerteverfahren beruht dabei jedoch immer auf der Annahme, dass keine während des Experiments zustande gekommenen Unreinheiten zu technischen Fehlern und somit Intensitätsabhängigen Abweichungen der Spots der jeweiligen Genreplikate führen.
In den Tabellen H.1 bis H.3 im Anhang I sind die relativen Standardabweichungen der Gene für alle drei Verdünnungsreihen dargestellt. Die folgenden Tabellen 6.10 bis 6.12 zeigen den Vergleich der beiden Auswertemethoden, in dem jene Gene, bei denen die relative Standardabweichung zwischen den Genreplikaten in der Standardauswertung höher ist als in der neuartigen Auswertemethode, mit einer **1** gekennzeichnet sind.

Wie in allen drei Verdünnungsreihen sichtbar, weichen die Genreplikate aller Gene, die mit der Linearen Regression berechnet wurden, bei kleinen Konzentrationen (stärkerer Verdünnung) geringer voneinander ab als die Genreplikate, die mit der Standardauswertung berechnet wurden. Es kann demzufolge angenommen werden, dass das Signal-Rauschen bei niedrigen Konzentrationen durch die Extrapolation der Geraden, die mittels einer Linearen Regression ermittelt wurde, besser kompensiert werden kann als durch einfache Mittelwertbildung der normierten Intensitätswerte. Diese Beobachtung entspricht den Erwartungen, da Scans mit höheren Scaneinstellungen, die hohe Signal-Rausch-Verhältnisse aufweisen, in die Gerade integriert werden können und so eine verbesserte Diskriminierung niedriger Konzentrationswerte vermuten lassen. Dahingegegn ermöglicht die Standardauswertung keinerlei Ausgleich der im niedrigen Intensitätsbereich stärker schwankenden Intensitätswerte.

Bei höheren Konzentrationen hingegen weichen die Genreplikate, die mit der Linearen Regression berechnet wurden, teilweise stärker voneinander ab als die der Standardauswertemethode. Diese Ergebnisse lassen zunächst eine höhere Güte der Standardauswertemethode

Tabelle 6.10 *– Vergleich der relativen Standardabweichungen der Gene zwischen den beiden Auswertemethoden - 1. Verdünnungsreihe. Eine 1 steht für eine höhere relative Standardabweichung der Gene bei der Standardauswertung im Vergleich zur neuen Auswertemethode.*

Relative Standardabweichung der Gene - 1. VR				
Gene	1:80	1:320	1:1280	1:5120
DnaK		1	1	1
GFP (EGFP)	1	1	1	1
FtsH			1	1
GFP (GFP)		1	1	1
GFP (GFPuv)	1	1		1
GlyA				1
GlyA				1
LacA				1
LacY				1
RFP			1	1
RpoA			1	
SecD		1		1
Sig2		1	1	1
TatD			1	

für höhere Konzentrationen vermuten. Eine detaillierte Analyse der in die Auswertung eingeflossenen Originaldaten zeigt aber, dass die Originalwerte der Genreplikate hoher Scans bei den kleineren Verdünnungen 1:80 und 1:320 für alle Gene außer RpoA und SecD im Sättigungsbereich liegen. Auch bei der 1:1250-Verdünnung liegen teilweise noch drei der acht Scans der Gene GFPuv, LacA und RFP im Sättigungsbereich. Aufgrund der in diesem Bereich beschränkten Sensitivität der Intensitätsmessung kann hier also technisch bedingt keine signifikante quantitative Messung der Signale erfolgen. Aus diesem Grund liegen die Originaldaten unterschiedlicher Scans bei diesen Verdünnungen unverhältnismäßig nah beieinander (im methodisch nicht signifikant differenzierbaren Intensitätsbereich zwischen 50.000 und 65.000).
Da die Standardmethode keine Korrektur von Sättigungseffekten beinhaltet, können Sättigungseffekte verfälschend auf die Berechnung der wahren Intensität wirken. Aufgrund der - teilweise identischen - Originalwerte von 65.000 im Sättigungsbereich weisen die Genreplikate also sehr viel geringere relative Standardabweichungen auf (siehe Tabelle H.3 bis H.3 im Anhang).
Dahigegen sieht die neuartige Auswertemethode eine Kompensation der Sättigungseffekte durch die Detektion und anschliessende Extrapolation dieser Werte vor. Aus diesem Grunde

Tabelle 6.11 *– Vergleich der relativen Standardabweichungen der Gene zwischen den beiden Auswertemethoden - 2. Verdünnungsreihe. Eine 1 steht für eine höhere relative Standardabweichung der Gene bei der Standardauswertung im Vergleich zur neuen Auswertemethode.*

Relative Standardabweichung der Gene - 2. VR					
Gene	1:80	1:320	1:1280	1:5120	1:20480
DnaK	1			1	1
GFP (EGFP)	1	1	1	1	1
FtsH			1	1	1
GFP (GFP)	1		1	1	1
GFP (GFPuv)				1	1
GlyA				1	1
GlyA				1	1
LacA				1	1
LacY		1			1
RFP				1	1
RpoA	1	1			1
SecD					1
Sig2				1	1
TatD			1	1	1

liegen die Intensitätswerte der Genreplikate bei geringer Verdünnung nach der Auswertung nicht mehr so dicht beieinander wie die entsprechenden Originaldaten und ergeben somit höhere relative Standardabweichungen. Entsprechend vermag diese Methode auch besser zwischen Genen hoher Konzentrationen zu diskriminieren.
Die mittleren relativen Standardabweichungen der Originaldaten der Genreplikate über alle Scans eines Gens nehmen entsprechend bei höheren Verdünnungen (geringeren Konzentrationen) zu, da hier einerseits das Signal-Rausch-Verhältnis abnimmt und andererseits konsistente, undifferenzierte Werte im Sättigungsbereich keinen verfälschenden Einfluss mehr auf die relative Standardanweichung ausüben. Diese mittleren relativen Standardabweichungen der Rohdaten können zur Interpretation der Güte herangezogen werden, da stark divergierende Originalwerte der Genreplikate auf eine Fehlerquelle für technische Mängel auf dem Chip hinweisen könnten. Aus diesem Grunde wurde der im vorangegangen Abschnitt wenig untersuchte mittlere Konzentrationsbereich (1:5120-Verdünnungen der drei Verdünnungsreihen) auf die mittleren relativen Standardabweichungen der Originalwerte der Genreplikate hin untersucht, um mögliche Erklärungen für abweichende Intensitäten der Genreplikate (relative Standardabweichungen, siehe Tabelle H.1 bis H.3) nach Berechnung mit der Linearen

Tabelle 6.12 *– Vergleich der relativen Standardabweichungen der Gene zwischen den beiden Auswertemethoden - 3. Verdünnungsreihe. Eine 1 steht für eine höhere relative Standardabweichung der Gene bei der Standardauswertung im Vergleich zur neuen Auswertemethode.*

Relative Standardabweichung der Gene - 3. VR				
Gene	1:80	1:320	1:1280	1:5120
DnaK	1	1		1
GFP (EGFP)	1	1	1	1
FtsH	1	1	1	1
GFP (GFP)	1		1	1
GFP (GFPuv)	1			1
GlyA	1			1
GlyA	1			1
LacA				1
LacY	1	1		1
RFP				
RpoA		1		1
SecD				1
Sig2	1			1
TatD		1		1

Regression zu finden. Die stärksten Schwankungen (mittlere relative Standardabweichungen) innerhalb der Genreplikate sind demnach in allen Verdünnungen für die Gene TatD ($\overline{\sigma_{rel}}$ von 0.66 für die erste Verdünnungsreihe, 0.68 für die zweite und 0.53 für die dritte Verdünnungsreihe) und RpoA ($\overline{\sigma_{rel}}$ von 0.47 für die erste Verdünnungsreihe, 0.40 für die zweite und 0.28 für die dritte Verdünnungsreihe) zu finden. Die stärkste mittlere relative Standardabweichung der Rohdaten der Genreplikate weist das Gen LacY ($\overline{\sigma_{rel}}$ von 0.9) in der zweiten Verdünnungsreihe auf. Diese starken Schwankungen der Rohdaten der Genreplikate könnten somit eine Erklärung für die relativen Standardabweichungen der mit der Linearen Regression ermittelten Intenstitätswerte liefern, da diese Intensitätswerte bei den genannten Gene teilweise auch geringere relative Standardabweichungen im Vergleich zu den mit der Standardmethode berechneten Intensitätswerte aufweisen.
Insgesamt können also die im Vergleich der beiden Auswertemethoden beobachteten Tendenzen der relativen Standardabweichungen der Genreplikate durch die genannten Effekte bei Chipexperimenten zustande gekommen sein und somit möglicherweise teilweise stärkere relative Standardabweichungen der Genreplikate bei den Werten der Linearen Regression erklären.

Ermittlung der Abhängigkeit der Signalintensitäten vom dCTP-Gehalt

Die ermittelten Erwartungsgrößen (Molekülzahl und Anzahl an dCTPs) werden zunächst über Quotientenbildung mit den Signalwerten der beiden Auswertemethoden korreliert. Die Ergebnisse dieser Korrelation sind in Tabelle I.1 und Tabelle I.2 im Anhang I widergegeben. Unter der Annahme eines linearen Zusammenhangs zwischen der Basenpaarlänge (bzw. dem dCTP-Gehalt) und der Signalintensität sollten die resultierenden Quotienten erwartungsgemäß gleich groß sein. Demnach gibt weniger die Größenordnung der Quotienten als vielmehr das Ausmaß der Abweichung im Vergleich unterschiedlicher Quotienten Auskunft über die Validität der Methode. Je geringer also die relative Standardabweichung der Quotienten aller Gene eines Microarrays, als umso höher kann die Korrelation zwischen Erwartungswert und Signalintensität eingestuft werden. Die berechneten mittleren relativen Standardabweichungen sind in Abbildung 6.13 und 6.14 gezeigt. Die dazugehörigen Werte sind in Tabelle 6.13 aufgelistet.

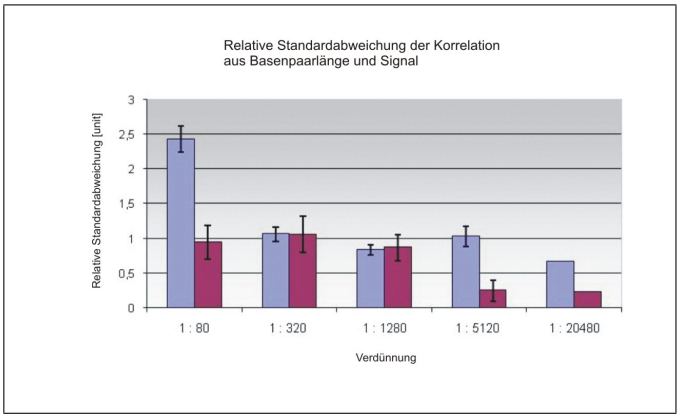


Abbildung 6.13 – *Relative Standardabweichungen der Quotienten aus Basenpaarlänge und den Signalintensitäten über alle Gene auf den Microarrays. Die Säulen repräsentieren die unterschiedlichen Verdünnungen sowie deren relative Standardabweichung über die Wiederholungsversuche. Alle Versuche mit Ausnahme der 1:20480-Verdünnung wurden dreifach durchgeführt. Die entsprechenden Fehlerbalken der unterschiedlichen Verdünnungsreihen sind in diesen Verdünnungen angegeben.*

Wie in den Abbildungen 6.13 und 6.14 ersichtlich, weichen die Quotienten, die die Korrelation zwischen Signalintensitäten und Basenpaarlänge bzw. dCTP-Gehalt dieser Sequenzen widerspiegeln, in geringerem Maße voneinander ab, wenn die Signalintensitäten mittels der neuartigen *Within-Array*-Methode berechnet wurden. Darüberhinausgehend unterscheiden sich die Variationsstärken der relativen Standardabweichungen im Vergleich unterschiedlicher Verdünnungen untereinander: während die relativen Standardabweichungen insbesondere der Quotienten aus der Anzahl dCTPs und den mittels Linearer Regression berechneten Signalwerten eine größere Ähnlichkeit aufweisen, sind die entsprechenden Quotienten, die mit der Standardmethode ermittelt wurden, in Bezug auf ihre Größenordnung hochgradig variabel. Insgesamt korrelieren die Signalintensitäten, die mit der Linearen Regression berechnet wurden, besser mit den Basenpaarlängen wie auch den dCTP-Gehalten der Gensequenzen. Eine höhere Stabilität der Korrelation im Vergleich unterschiedlicher Verdünnungsstufen konnte

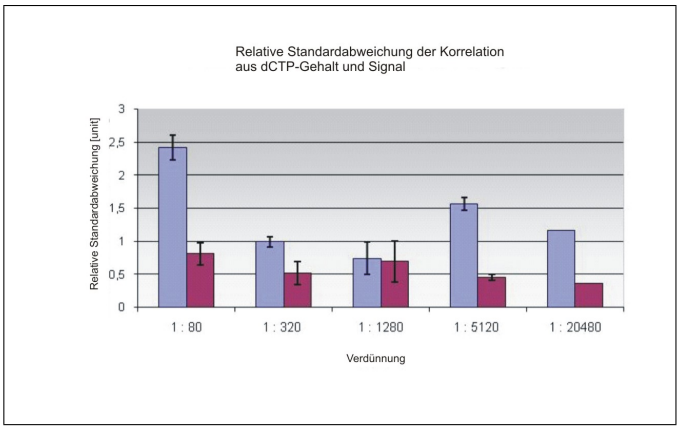


Abbildung 6.14 *– Relative Standardabweichungen der Quotienten aus der Anzahl dCTPs und den Signalintensitäten über alle Gene auf den Microarrays. Die Säulen repräsentieren die unterschiedlichen Verdünnungen sowie deren relative Standardabweichung über die Wiederholungsversuche. Alle Versuche mit Ausnahme der 1:20480-Verdünnung wurden dreifach durchgeführt. Die entsprechenden Fehlerbalken der unterschiedlichen Verdünnungsreihen sind in diesen Verdünnungen angegeben.*

beim dCTP-Gehalt beobachtet werden. Die vorgestellte Methode ermöglicht somit eine höhere Flexibilität gegenüber unterschiedlichen Konzentrationsmengen der Proben und bringt insgesamt eine Verbesserung bei der Ermittlung wahrer Werte gegenüber der Standardmethode.

Im weiteren Verlauf der Untersuchung der weiterentwickelten Methode wird ein Vergleich mit der Basenpaarlänge vernachlässigt. Die Beschränkung der Berechnungen auf die dCTP-Gehalte beruht auf der Annahme, dass die wahren Fluoreszenz-Intensitäten durch diese Größe besser widergegeben werden. Diese Annahme wird durch die oben gezeigte höheren Korrelation mit den Signalwerten der neuartigen Auswertemethode bestätigt.

Über die Ermittlung der oben vorgestellten Quotienten hinausgehend, kann die Detektion der Ausreißer unter den Quotienten zusätzliche interessante Informationen bezüglich der Auswertemethode liefern. Unter Anwendung des Ausreißertests nach *Nalimov* wurden dementsprechend die Ausreißer unter den Quotienten einer Verdünnung für alle drei Verdünnungsreihen im Einzelnen errechnet.[144] Die Ergebnisse sind in Tabelle 6.14 dargestellt.

Statistische Analysen der Ausreißer (siehe Tabelle 6.14) unter den Quotienten (Anzahl der dCTP-Moleküle der Sequenzen / neu berechnete Signalintensitäten) aller Gene der unterschiedlichen Konzentrationen (Verdünnungen) aller drei Verdünnungsreihen zeigten, dass vor allem die Signalintensitäten des Gens EGFP (eins der drei GFP-Varianten) mehrfach von den erwarteten Werten abwichen. In jeweils einer der Verdünnungen wurden aber auch LacY und GlyA als Ausreißer detektiert.
Während die Gene EGFP und GlyA vergleichsweise niedrige Signalintensitäten (hohe Quotienten) besitzen, ist der Expressionswert von LacY in der vierten Verdünnung der zweiten Verdünnungsreihe relativ hoch ím Vergleich zum dCTP-Gehalt dieser Gensequenz (niedriger Quotient).

Tabelle 6.13 *– Relative Standardabweichungen der Korrelation zwischen Erwartungswerten und Ergebnissen.*

rel. σ der Quotienten: $\frac{Basenpaarlänge}{Signalwerte}$					
	1:80	1:320	1:1280	1:5120	1:20480
Standardmethode	2.43	1.06	0.84	1.03	0.67
Lineare Regression	0.94	1.06	0.87	0.25	0.23

rel. σ der Quotienten: $\frac{dCTP-Gehalt}{Signalwerte}$					
	1:80	1:320	1:1280	1:5120	1:20480
Standardmethode	2.41	0.99	0.74	1.57	1.16
Lineare Regression	0.81	0.52	0.70	0.45	0.36

Um die Ursachen für diese Abweichungen zu ergründen, wurden zunächst die Einbauraten des Fluoreszenzfarbstoffes nach der folgenden Formel berechnet:[171]

$$Cy5\ Inkorporation = 35.1 \cdot \frac{A_{650}}{A_{260}} \tag{6.9}$$

Die Berechnung der Inkoporationsrate des Farbstoffes Cy5 (*Frequency of Incorporation* = *FOI*) gemäß der Formel 6.9 erfolgt durch Multiplikation eines sich aus dem Lambert-Beerschen Gesetz ergebenden Faktors mit der Adsorption bei der Wellenlänge von Cy5 geteilt durch die Wellenlänge der cDNA.[171] Die Messung der Adsorption bei unterschiedlichen Wellenlängen erfolgte während des experimentellen Teils mit dem NanoDrop Spectrophotometer (siehe Anhang G). Die Ergebnisse liegen als Anzahl Cy-dCTP eingebauter Moleküle pro 1000 Nukleotide der cDNA-Sequenzen vor. Diesen Kalkulationen zufolge beträgt die mittlere *FOI* über alle Gene 8.16. Demgegenüber liegt die *FOI* von GlyA mit 4.55 am unteren Rand der Wertemenge. Dieser Befund erklärt den Ausreißer von GlyA in der ersten Verdünnungsreihe bei einer Verdünnung von 1:5120 sowie die allgemein vergleichsweise schwächeren Signalintensitäten (hohen Korrelations-Quotienten) bei einem relativ hohen dCTP-Gehalt dieses Gens. Bezüglich des Ausreißers LacY in der zweiten Verdünnungsreihe konnte kein analoger Trend in anderen Konzentrationen und Verdünnungsreihen beobachtet werden: im Gegensatz zu den Genen EGFP und GlyA zeigte LacY eine einmalige Signal-Abweichung. Aus diesem Grund liegt die Vermutung nahe, dass diesem Ausreißer kein systematischer Fehler (wie bei GlyA gefunden) zugrunde liegt. Daher wurden die Einzelintensitäten aller Genreplikate dieses Gens bei der entsprechenden Verdünnung untersucht. Die entsprechenden Original-Genwerte

Tabelle 6.14 *– Ausreißertest nach Nalimov für die Quotienten aus $\frac{dCTP-Gehalt}{Intensitätswerte}$. Die Intensitätswerte wurden mittels der oben beschriebenen Within-Array-Methode unter Anwendung einer Linearen Regression berechnet.[144] Gezeigt sind die signifikanten Ausreißer („X“) für alle Verdünnungen in den replikativen Verdünnungsreihen (1. - 3.V)*

Ausreißertest nach *Nalimov* für die Quotienten aus $\frac{dCTP-Gehalt}{Intensitätswerte}$													
Gene	1:80			1:320			1:1280			1:5120			1:20480
	1.V	2.V	3.V	1.V	2.V	3.V	1.V	2.V	3.V	1.V	2.V	3.V	3.V
DnaK													
EGFP	**X**	**X**		**X**				**X**					
FtSH													
GFP													
GFPuv													
GlyA													
GlyA										**X**			
LacA													
LacY											**X**		
RFP													
RpoA													
SecD													
Sig2													
TatD													

können bei *GEO* (http://www.ncbi.nlm.nih.gov/geo/) eingesehen werden.[102] Die Untersuchung der Genreplikate von LacY der entsprechenden Verdünnung führte zur Offenlegung einer bemerkenswert hohen Standardabweichung (0.15 im Vergleich zu einem Mittelwert über alle Gene dieses Microarrays von 0.08) unter den Genreplikaten, die sowohl die Standardabweichung aller anderen Gene dieses Microarrays als auch die Standardabweichung über die Genreplikate von LacY auf anderen Chips übertrifft. Die Ungleichheit der Genreplikate auf diesem Microarray im Vergleich zu den entsprechenden Genreplikaten dieses Gens auf anderen Chips deutet auf Fehler dieses spezifischen Microarrays hin, die beispielsweise durch lokale Unebenheiten an Spot-Positionen dieses Gens zum Tragen kommen können. Eine derart hohe Standardabweichung resultiert darin, dass relativ stark abweichende Genreplikate nicht als Ausreißer detektiert werden können, was auf die Berücksichtigung der Standardabweichung durch den Ausreißertest zurückzuführen ist. Aus diesem Grund werden zum Beispiel hohe

Intensitätswerte, die beispielweise durch Verunreinigungen auf dem Chip zustande kommen, in die Berechnung eines Genwertes aus den Genreplikaten einbezogen und verfälschen somit den Gesamtintensitätswert von LacY. Demzufolge könnte die Ausreißerdetektion von LacY in der vierten Verdünnung der zweiten Verdünnungsreihe nicht aufgrund systematischer, sondern viel mehr durch Microarray-bedingte Fehler zustande gekommen zu sein.
Das dreifache Ausreißer-Gen EGFP stellt - wie in Anhang G erläutert - eines von drei Oligonukleotid-Molekülen auf der Microarray-Familie dar, die an das selbe Fängermoleküle, GFP mut2, binden. Um den Einfluss der unterschiedlichen Sequenzzusammensetzungen der drei Fängermoleküle GFPuv, wildtype GFP und EGFP zu untersuchen, wurde ein DNA-Alignment der Fängermoleküle mit der Probensequenz durchgeführt. Demzufolge stimmt die Sequenzabfolge der Genvariationen GFPuv und wildtype GFP zu 100% mit der Probensequenz auf den Microarrays überein. Das EFGP - GFP mut2 Alignment hingegen wies einen Ähnlichkeitsindex von nur 83% auf, was die verringerte Signalintensität der EGFP-gebundenen GFP mut2-Sequenz erklärt. Um eine verfälschende Gesamtaussage über die Qualität der Auswertemethode aufgrund der technisch nachweisbaren unvollständigen Hybridisierungskapazität von EGFP an GFP mut2 zu vermeiden, wurden die Signalintensitäten dieses Gens von weiteren Berechnungen ausgeschlossen.

Die Korrelation zwischen Erwartungswerten und Ergebnis kann wie beschrieben mittels der relativen Standardabweichung der Quotienten aus den entsprechenden Größen überprüft werden. Ein weiteres aufschlussreiches Qualitätskriterium stellt der Verlauf der Kurven aus den Signalintensitäten in Abhängigkeit der Anzahl an dCTP-Molekülen in den Gensequenzen dar. Die technischen Replikate (Verdünnungsreihen) aller vorkommenden Verdünnungen wurden zu diesem Zweck gemittelt und normalisiert, in dem die jeweiligen Signalintensitäten (aus beiden Auswertemethoden) durch das entsprechende Maximum der Verdünnungen dividiert wurden. Die jeweiligen Daten der beiden Auswertemethoden wurden in einer visuellen Gegenüberstellung miteinander verglichen, in dem die resultierenden Genwerte aller Verdünnungen gegen den dCTP-Gehalt aufgetragen wurden. Ein Beispiel einer solchen Auftragung ist in Abbildung 6.15 dargestellt. Die lineare Regressionsgerade veranschaulicht die Steigung der Kurve und spiegelt die Verlaufstendenz der Abhängigkeit der beiden aufgetragenen Größen wider.

Bei diesem in Abbildung 6.15 gezeigten Beispiel handelt es sich um die Gegenüberstellung der beiden Daten beider Auswertemethoden aus der vierten Verdünnung (1:1280). Die Regressionsgeraden deuten die lineare Abhängigkeit der Signalintensitäten vom dCTP-Gehalt an. Eine quantitative Aussage über das Ausmaß der Linearität der Abhängigkeit der beiden aufgetragenen Größen kann über die Reststandardabweichungen getroffen werden. Sie dient als Maßzahl für die Abweichung der Wertepaare von den Modellwerten der Geraden und kann somit als Qualitätsangabe der beiden verglichenen Methoden angeführt werden. Tabelle 6.15 enthält die Reststandardabweichungen aller Geraden der verschiedenen Verdünnungen.

Wie in Tabelle 6.15 anhand der Reststandardabweichungen zu sehen, weichen die Wertepaare aus den Signalintensitäten, die mit der Standardmethode berechnet wurden, und der

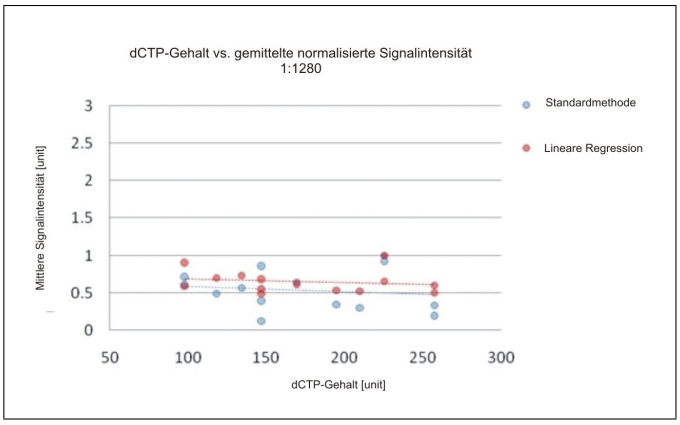


Abbildung 6.15 – *Abhängigkeitsverhältnis der Signalintensitäten vom dCTP-Gehalt der Gene mit Ausnahme von EGFP (Abszisse) in beiden Auswertemethoden in einer 1:1280-Verdünnung. Die Regressionsgerade ist als gepunktete Linie dargestellt.*

Tabelle 6.15 – *Reststandardabweichungen der Wertepaare aus normalisierten Signalintensitäten und dCTP-Gehalten*

Reststandardabweichungen					
Methode	1:80	1:320	1:1280	1:5120	1:20480
Standard method	0.26	0.26	0.27	0.25	0.25
Linear Regression	0.23	0.23	0.26	0.17	0.08

Anzahl jeweiliger dCTP-Moleküle in stärkerem Maße von der Geraden ab als entsprechende Werte der neu eingeführten Methode.
Bezüglich der Parameter der Regressionsgeraden kann die Annahme getroffen werden, dass sich die Achsenabschnitte in Abhängigkeit der Verdünnung unterscheiden. Die Steigungen der Geraden hingegen sollten idealerweise um den Nullpunkt liegen, wenn eine lineare Abhängigkeit der Signale vom dCTP-Gehalt vorliegt. Tatsächlich liegt die mittlere Steigung aller Verdünnungen der Standardauswertemethode bei $2.91 \cdot 10^{-4}$ bei einer relativen Standardabweichung von 0.85 und bei $-1.43 \cdot 10^{-4}$ bei einer relativen Standardabweichung von 0.63, wenn mit der Linearen Regression ausgewertet wurde.

Die Regressionsgeraden verdeutlichen den linearen Trend der normalisierten Signalintensitäten in Abhängigkeit vom dCTP-Gehalt über die Gesamtheit der Gene. Die Linearität trifft vor allem unter Anwendung der Linearen Regression zu, da ein breiterer Intensitätsbereich erfasst werden kann, was durch die Steigung nahe dem Nullpunkt und einer geringeren relativen Standardabweichung unterstrichen wird. Die signifikante Abdeckung eines breiteren Intensitätsbereich kann vermutlich durch die implementierte Korrektur gesättigter Werte und den Ausgleich stärkerer Schwankungen im unteren Intensitätsbereich sowie die im Rahmen der Mandeltests ermöglichten Entfernung von Spots aus fehlerhaften Scans erklärt werden.

Ermittlung der Abhängigkeit der Signalintensitäten von der Konzentration

Nach dem im ersten Validierungspart der beiden Auswertemethoden die Abhängigkeit der Signalintensitäten der beiden Methoden vom dCTP-Gehalt (bzw. der Basenpaarlänge) ermittelt wurde, hat der zweite Teil die Ermittlung des Abhängigkeitsverhältnisses der Signalwerte von der Konzentration zum Ziel.
Zu diesem Zweck wurden die Signalintensitäten der Genreplikate des in diesem Experiment untersuchten Gens Sig2 gemittelt und gegen den Verdünnungsfaktor aufgetragen (siehe Abbildung 6.16).

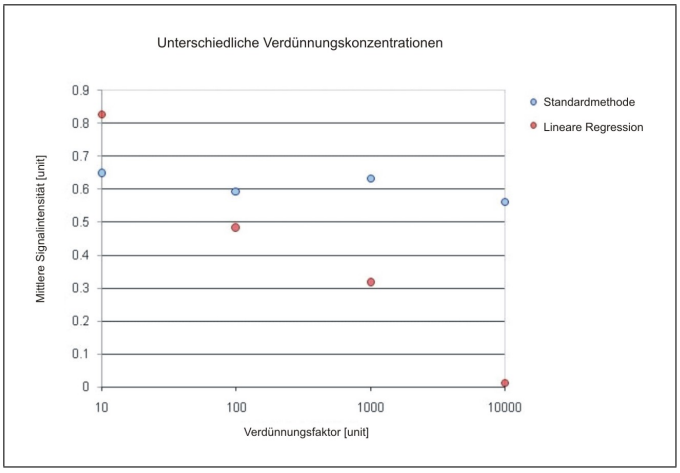


Abbildung 6.16 – *Signalintensitäten von Sig2 gegen die Verdünnungsfaktoren in logarithmischer Auftragung.*

Abbildung 6.16 verdeutlicht die lineare Abhängigkeit der Signalintensitäten von der Konzentration der Lösungen in beiden Methoden. Wie in der Abbildung zu sehen, steigen die Signalintensitäten mit steigenden Verdünnungskonzentrationen signifikant stärker an, wenn erweiterte *Within-Array*-Auswertemethode angewandt wurde. Folglich unterscheidet sich die Steigungen der Regressionsgeraden, die den Intensitätsverlauf gegen die logarithmierten Verdünnungsfaktoren repräsentieren, zwischen den beiden Methoden um eine Größenordnung ($-0.181 \cdot \ln(x)$ mit einem Korrelationsfaktor von 0.98 für die fortgeschrittene Auswertemethode und $-0.012 \cdot \ln(x)$ mit einem Korrelationsfaktor von 0.52 für die Standardmethode).
Wie bereits beim Vergleich der Signalintensitäten mit den dCTP-Gehalten erwähnt, beruht dieser Unterschied vermutlich auf dem Ausgleich von Sättigungseffekten bei der erweiterten Auswertemethode: Während die Standardmethode die vor allem bei geringen Verdünnungen vielfach auftretenden Sättigungswerte nicht auszugleichen vermag, können durch die fortgeschrittene Methode mittels Extrapolation der linearen Scan-Verläufe neue Spotwerte ermittelt werden, die realistischerweise näher an wahren Werten liegen. Aus diesem Grund scheint die erweiterte Methode Konzentrationsunterschiede effektiver erfassen zu können.

Im Anschluss an die Lineare Regression werden die Replikate eines Gens gegebenenfalls einem Ausreißertest nach Nalimov unterzogen.[144] Die Ausreißer werden eliminiert und alle

verbleibenden Gen-Informationen gemittelt, so dass es an dieser Stelle unter Umständen zu einer ersten umfangreichen Reduktion der Datenvielfalt kommen kann. Dieser Schritt erfolgt im Anschluss an die Korrektur der Einzelspots, um gegebenenfalls statistisch nicht sinnvoll verwertbare Spots (in der Sättigung bzw. im Niedrigintensitätsbereich) korrigieren und als Genreplikate wieder nutzbar machen zu können. Aufgrund teilweise der in folgenden Normalisierungsschritten zugrunde liegenden Annahme einer gleichartigen Expression aller Gene über die Gesamtheit aller Gene auf einem Chip, werden die Genreplikate vorab zusammengeführt, um durch eine Eliminierung fehlerhafter Genreplikate eine Überbewertung solcher Spots vermeiden zu können.

6.2.4.3 Fazit - Lineare Regression

Die Registrierung einer erweiterten Anzahl an Daten durch multiple Scans in Microarray-Experimenten wird in der Literatur oftmals nicht berücksichtigt.[155] Wie anhand der oben aufgeführten Beispiele gezeigt, resultiert das mehrfache Scannen der Microarrays unter geringen Zeitaufwand und vernachlässigbaren Ausbleicheffekten in einer ausgedehnten Datenpluralität, die für jeden einzelnen Spot wichtige Informationen beinhaltet. Eine signifikante Verbesserung der Signalintensitäten kann so erreicht werden, in dem die lineare logarithmische Abhängigkeit der Signalintensitäten von den PMT-Verstärkungen bei konstanten Laserstärken ausgenutzt wird. Auf diese Weise fließt zusätzlich verwertbares *A-priori*-Wissen in die Auswertung der Daten ein. Im Gegensatz zur Auswertung einfacher Scans kann durch dieses integrierte Wissen eine Untermenge der zahlreichen Variabilitäten, die Microarray-Experimenten innewohnen, nämlich die durch den Scanner verursachte Variationsquelle, eliminiert oder zumindest auf ein Mindestmaß reduziert werden.
Während die von Dudley *et al.* vorgestellte Methode zur Berücksichtigung multipler Scans auf die Korrektur der logarithmierten Ratios beider Zustände auf einem Microarray abzielt,[161] werden in der hier vorgestellten Methode die Signalintensitäten der Einzelzustände berechnet. Diese Herangehensweise ermöglicht so einerseits differierende PMT-Einstellungen für beide auf einem Microarray befindlichen Farbstoffe während eines Scans und berücksichtigt andererseits durch die Implementierung eines Optimierungs-Algorithmus die den unterschiedlichen Farbstoffen immanenten Variabilitäten bei gleichen PMT-Verstärkungen.[161]
Obgleich die Anwendung der vorgestellten *Within-Array*-Methode signifikant zur verbesserten Identifikation differentiell exprimierter Gene beiträgt, bleibt die Entwicklung einer optimalen Scan- und Datenanalysestrategie ein Bereich, der aufgrund des bis dato vergleichsweise geringeren Wissenschatzes weiterer Forschung bedarf.

6.2.4.4 Durchführung und Ergebnisse - Lowess-Regression

Im ersten Abschnitt der *Within-Array*-Normalisierung wurden unter Anwendung der Linearen Regression und von Optimierungs-Algorithmen Scanabhängige Variabilitäten entfernt und Sättigungseffekt ausgeglichen. Auf der Basis dieser Spot-spezifisch korrigierten Signalin-

tensitätswerte können weitere Normalisierungsschritte der Daten einzelner Microarrays durchgeführt werden.
Da die beiden auf einen Microarray hybridisierten Zustände mit unterschiedlichen Farbstoffen (in der Regel Cy5 und Cy3) markiert werden, besteht die Notwendigkeit, die auftretenden physikalischen Unterschiede zu kompensieren.[8,78,172] Farbstoff-bedingte Unterschiede können beispielweise durch die unterschiedliche Inkorporation in cDNA-Moleküle sowie die verschiedenen molaren Fluoreszenzeigenschaften der beiden Farbstoffe zustande kommen. Darüberhinausgehend kommen jedoch noch vielfältige weitere Quellen infrage, die zu Farbstoffabhängigen Differenzen führen können.[173–178]
Mithilfe der in Kapitel 6.2.3 vorgestellten und in Abbildung 10(a) und 10(b) bereits einführend dargestellten *MA-Plots* können solche Variabilitäten, die durch die unterschiedliche Markierung zweier Zustände auf einem Microarray verursacht werden, auf *whole-genome*-Microarrays visualisiert werden. Der in Abbildung 6.17 gezeigte *MA-Plot* verdeutlicht die Notwendigkeit einer Kompensation der Farbstoffabhängigen Variabilitäten. Aus diesem Grund wird bei den im Folgenden vorgestellten Normalisierungs-Algorithmen mit den im *MA-Plot* verwendeten logarithmierten Quotienten (*log ratios*) aus den Signalwerten beider Kanäle eines Zustand gerechnet.
Um diese Methode zu testen wurden unterschiedliche *whole-genome*-Microarray-Familien (jeweils mit dem Genom des Bakteriums *E.coli*, der Tomate und der Hefe *Saccharomyces cerevisiae*) getestet. Bei den Versuchen handelte es sich um ein *Dye-swap*-Design, bestehend aus zwei Chips (Tomate), einem *Loop*-Design mit drei Microarrays (*Saccharomyces cerevisiae*) sowie um ein *Loop*-Design mit jeweils acht Microarrays mit Zuständen, die unterschiedliche Kultivierungsbedingungen darstellten (*E.coli*). In diesem Abschnitt werden repräsentativ die Ergebnisse der *E.coli*-Kultivierung gezeigt.

Wie in Abbildung 6.17 zu sehen, verläuft die Datenwolke - bestehend aus beiden Zuständen des gezeigten Chips - nicht entlang der Null-Linie und erfüllt somit nicht die ebenfalls in Kapitel 6.2.3 für *whole-genome*-Microarrays definierte Annahme, dass die Mehrheit aller Gene eines Genoms nicht differentiell exprimiert sind. Diese Darstellung erweckt den Anschein, dass die meisten Gene als reguliert identifiziert werden können, obwohl dieses Ergebnis lediglich auf der fehlenden Normalisierung beruht. Im Gegensatz zur Linearen Regression, basiert diese Methode also auf der in Kapitel 6.2.3 für *whole-genome*-Microarrays bereits erwähnten Voraussetzung, dass die Mehrheit der Gene nicht differentiell exprimiert sind. Dementsprechend kann der im Folgenden eingeführte Algorithmus auch nur auf solche Microarrays appliziert werden.

Für die Kompensation dieser Variabilitäten sind in der Literatur zahlreiche Ansätze beschrieben,[1,52,58,155,178] die prinzipiell drei Hauptkategorien zugeordnet werden können: einer Normalisierung über die Gesamtintensität, einer Ratios-basierten Normalisierung sowie einer Normalisierung, die mit Regressionstechniken arbeitet.
Die hierbei angewandten Verfahren basieren jedoch alle auf verschiedenen Grundannahmen, beispielsweise dass die Expression aller Gene im Mittel gleich bleibt oder dass eine Normal-

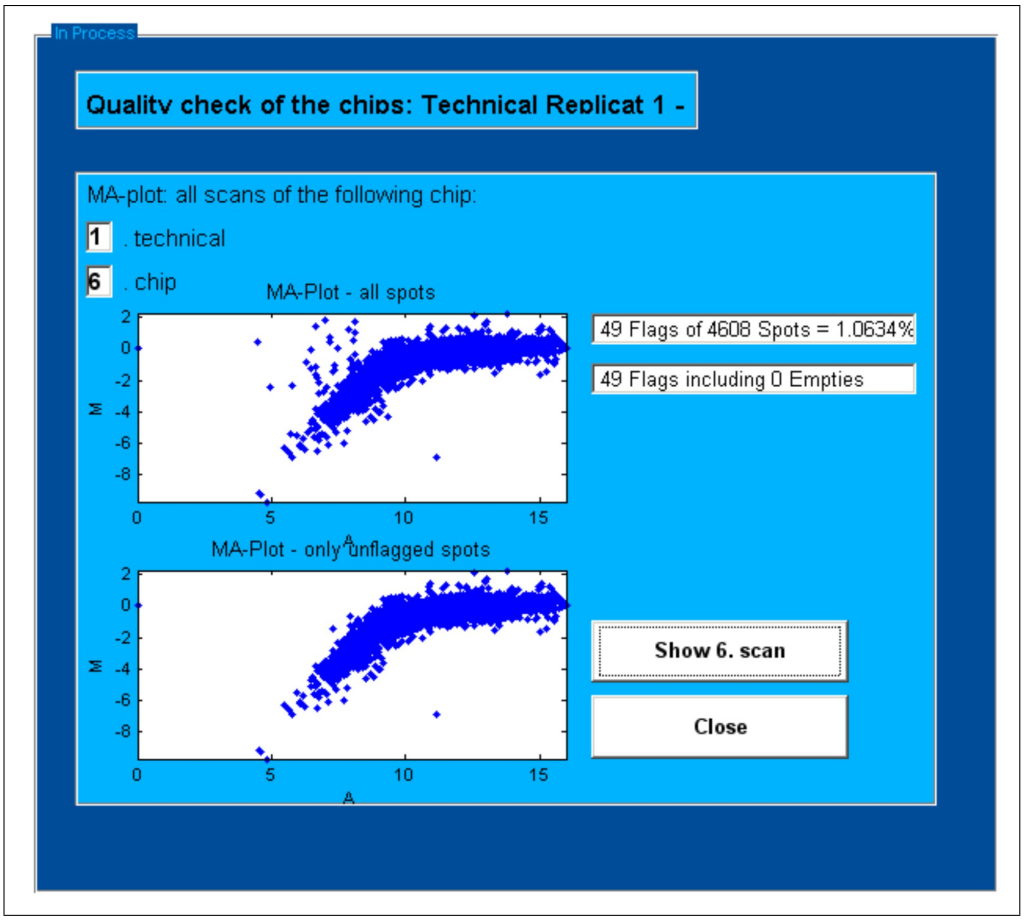


Abbildung 6.17 – *MA-Plot eines E.coli whole-genome-Microarrays in der graphischen Zwischenergebnisausgabe des Programms. Die obere Graphik zeigt die Gesamtheit aller Daten inklusive der geflaggten Spots, welche in der unteren Graphik entfernt wurden. Der Verlauf der Kurve zeigt deutlich einen Intensitätsabhängigen Farbstoffunterschied.*

verteilung über die resultierenden, logarithmisch transformierten Quotienten existieren muss. Abweichungen von dieser Norm werden als technisches Artefakt bewertet und in der Normalisierung durch geeignete Transformation entfernt bzw. reduziert.
Aufgrund der von Stekel *et al.* beschriebenen Einschränkungen der beiden zuerst genannten Methoden,[127] wurde eine erweiterte Lowess-Regression implementiert, die auf Regressions-Routinen basiert.
Die Lowess-Transformation, auch als Loess-Transformation bekannt, steht für die sogenannte *Locally Weighted Polynomical Regression*.[86,154,179] Die Funktion, die durch die Lowess-Regression angeglichen wird, entstammt einem Polynom der Form:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... \tag{6.10}$$

Unter Anwendung dieses Polynoms werden die in den *MA-Plots* dargestellten Abhängigkeiten der beiden Kanäle

$$M = log_2 \frac{Cy5}{Cy3} \tag{6.11}$$

und

$$A = log_2\sqrt{Cy5 \cdot Cy3} \tag{6.12}$$

genutzt, um eine Korrektur der systematischen Intensitätsabhängigkeit der Farbstoffunterschiede vorzunehmen. Eine solche Intensitätsabhängigkeit wurde ebenfalls von Shi *et al.* beschrieben,[125] die vor allem als Abweichungen der log_2-Werte (Ratios) von null im niederen Intensitätsbereich auftreten.
Demgegenüber sollte der Gesamtquotient $log_2\frac{Cy5}{Cy3}$ unabhängig vom Intensitätslevel nach dem Ausschluss solcher systematischer Fehler bei null liegen, da auf *whole-genome*-Microarrays im Mittel keine differentielle Expression zu erwarten ist. Mithilfe der Lowess-Regression werden die Abweichungen vom erwarteten Verhalten detektiert und korrigiert, indem für jeden validen Datenpunkt der *MA-Plots* eine lokal gewichtete lineare Regression durchgeführt wird. Die Lowess-basierte Intensitätsabhängige Normalisierung resultiert somit letztlich in einer Glättung (*Smoothing*) der *MA-Plots* von Microarrays,[119] um so das den Experimenten zugrunde liegende Expressionsmuster offenzulegen und nicht-lineare Abhängigkeiten, in diesem Fall zwischen M (Formel 6.11) und A (Formel 6.12), zu identifizieren.
Zu diesem Zweck wird jeweils eine Untermenge (sogenannte *Fenster* oder *windows*) der Daten selektiert und nacheinander geglättet. Der Benutzer des hier vorgestellten Auswerteprogramm kann optional die vordefinierte Fensterbreite beliebig variieren, umso unterschiedlich ausgeprägte Glättungseffekte zu erzielen. Zu kleine Fenster reagieren empfindlicher auf lokale Veränderungen und erzeugen daher Schlangenlinien anstelle durchgehender Linien, während zu große Fenster möglicherweise nicht sensitiv genug sind, um die Intensitätsabhängigen systematischen Fehler tatsächlich zu erfassen.[180]
Innerhalb dieser Spot-weise bearbeiteten Fenster wird jeweils der Wert in der Mitte mittels einer Regression über alle Daten dieses Fensters bereinigt. Um den Einflu einzelner fehlerhafter Spots zu vermeiden, wurde zusätzlich zu den routinemäßig angewandten Algorithmen ein Linearitätstest nach Mandel implementiert,[181] der zur Detektion von Ausreißern verwendet wurde. In dem in Abbildung 6.18 dargestellten Datenausschnitt wurde ein Ausreißer in dem untersuchten Fenster gefunden.

Auf die Eliminierung der Ausreißer eines Fensters folgt die Glättung der Daten über die separate Korrektur des in dem jeweiligen Fenster untersuchten Spots. Letztlich handelt es sich bei dieser Korrektur gemäß der in der folgenden Gleichung widergegebenen Formeln um eine einfache Subtraktion der Glättungskurve von den logarithmierten Originalwerten.

$$M_i' = log_2\frac{Cy5_i'}{Cy3_i'} \;=\; log2(Cy5_i) \;-\; log2(Cy3_i) \;-\; l(A_i) \tag{6.13}$$

$$= log_2\frac{Cy5_i}{Cy3_i} \;-\; l(A_i) \tag{6.14}$$

$$= M_i - l(A_i) \tag{6.15}$$

Wenn $l(A_i)$ den Wert der Lowess-Kurve als eine Funktion von A am i-ten Punkt des Microarrays (Spot) darstellt, so gilt die oben dargestellte Gleichung unter der Annahme, dass

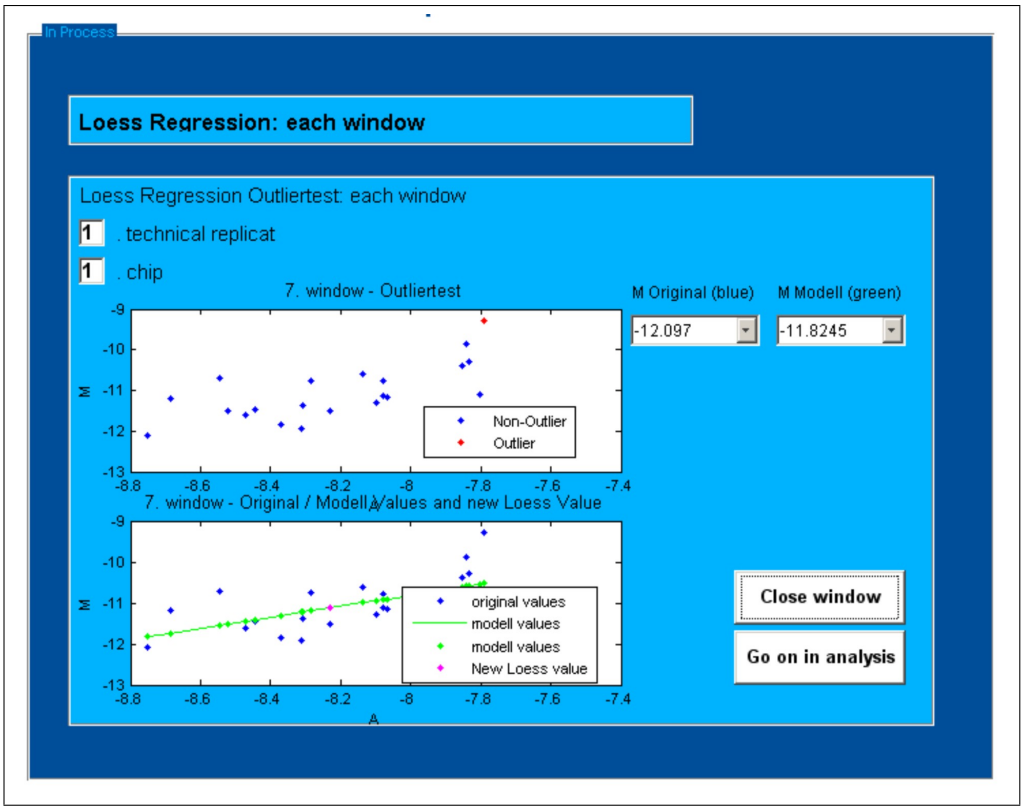


Abbildung 6.18 – *Ausreißertest nach Mandel in einem Fenster der Lowess-Regression.*

$Cy3'_i = Cy3_i \cdot 2^{l(A_i)}$ und $Cy5'_i = Cy5_i$ ist.
Die Anwendung einer Intensitätsabhängigen Normalisierung im M vs. A-Raum ist einem direkten Einsatz im $log_2(Cy5_i)$ vs. $log_2(Cy3_i)$-Raum vorzuziehen, da die erstgenannte Umgebung die Variabilitäten in beiden Kanälen zugleich anspricht, in dem der geometrischen Mittelwert beider Zustände adressiert wird.[119]

Von denen, im Rahmen der Validierung des Programms getesteten unterschiedlichen Fensterbreiten, wurde ein 31 Spots umfassendes Fenster als für diese Microarray-Familie optimale Breite gewählt. Ferner wurden auf die Anzahl zu eliminierender Ausreißer in einem Fenstern getestet. Zu diesem Zweck wurden null bis fünf Ausreißer überprüft. Dabei stellte sich heraus, dass eine Anzahl von zwei Ausreißern pro Fenster einen Kompromiss aus einer sinnvollen Eliminierung signifikanter Ausreißer und einer ausreichenden Datenmenge zur Bestimmung des Lowess-Glättung ausmacht.
In Abbildung 6.19 werden die bezüglich der Intensitäten unnormierten Daten eines Microarrays denen in einer Lowess-Regression normierten Signalwerten in Form von *MA-Plots* gegenübergestellt. Beim Vergleich der beiden Graphen ist zu berücksichtigen, dass die während der Lowess-Regression bereinigten Intensitätswerte in einem vergleichsweise schmaleren Wertebereich dargestellt sind, so dass die optisch breiter wirkende Streuung der niedrigen Werte keiner quantitiver Grundlage entspricht.

Bei der in Abbildung 6.19 gezeigten Graphik handelt es sich um Daten des bereits in Abbildung 6.17 und Abbildung 6.18 untersuchten *whole-genome*-Microarrays. Die Graphiken

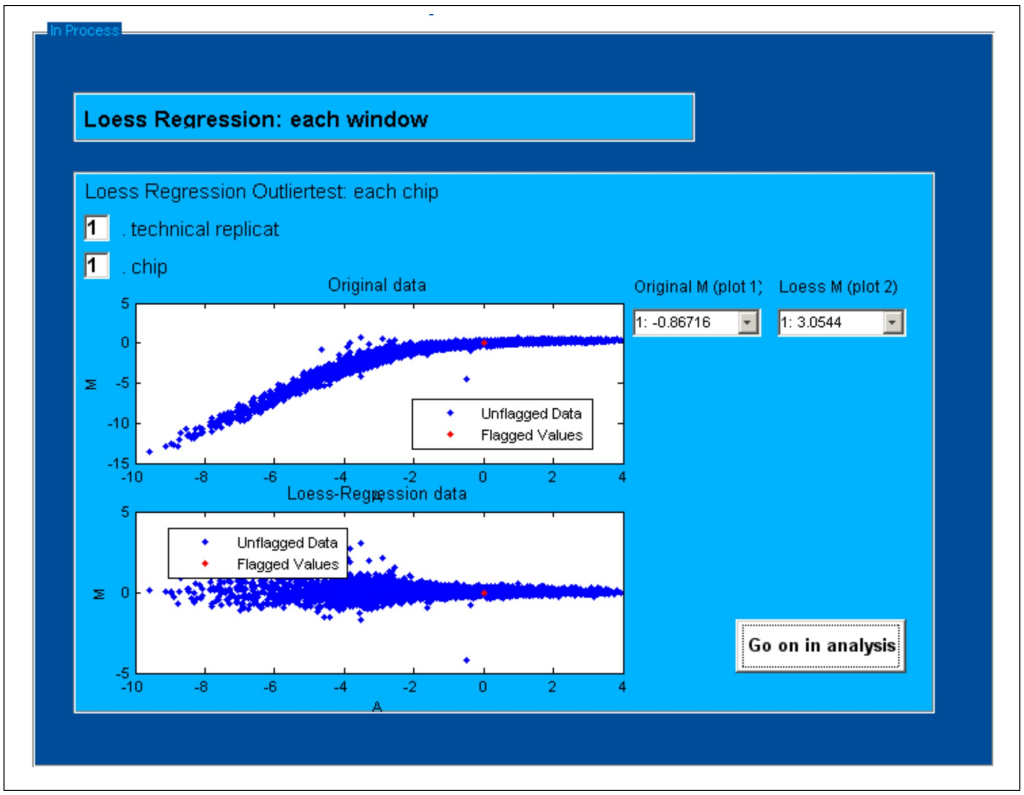


Abbildung 6.19 – *Vergleich eines MA-Plots vor und nach der Applikation einer Lowess-Regression.*

stammen aus der visuellen Widergabe der Ergebnisse, die vom Benutzer optional erfragt werden können.
Wie in der Abbildung deutlich zu sehen, konnte der abfallende Trend im geringen Intensitätsbereich unter Anwendung der Statistik-Routinen vollständig ausgeglichen werden.

6.2.4.5 Fazit - Lowess-Regression

In dem vorangegangenen Abschnitt wurden systematische Variationen, die Zwei-Kanal-Experimenten innewohnen, anhand der Fluorophore Cy5 und Cy3 diskutiert. Es wurde die Vorgehensweise einer zur Bereinigung solcher Variabilitäten implementierte Regressions-Methode (Lowess-Regression) vorgestellt. Mithilfe dieser bereits in der Literatur beschriebenen Methode werden die Einzelquotienten aus den Intensitätswerten beider Kanäle neu skaliert.[86,154,189] Auf diese Weise konnten Intensitätsbedingte Fehler minimiert werden. In dem hier vorgestellten Programm wurde die auch in anderen Programmen angewandte Lowess-Regression um einen Ausreißertest erweitert, so dass die Berechnung neuer Lowess-Werte auf validen Datenpunkten beruht.
Neben der hier vorgestellten globalen Lowess-Regression existieren auch lokale Ansätze, die statt der Korrektur des gesamten Microarrays in Einem, gesonderte Abschnitte auf dem Chip fokussieren.[8,72] Eine lokal begrenzte Bereinigung der Daten kann beispielsweise Pin-abhängige systematische Fehler adressieren. Da auf den im Rahmen dieser Arbeit untersuchten Microarrays jedoch keine Pin-abhängigen Intensitätsunterschiede ausgemacht werden konnten, wur-

de auf eine derartige Korrektur verzichtet.
Eine weitere Möglichkeit, lokale Variabilitäten erfassen und korrigieren zu können, besteht in der Verwendung exogener Nukleinsäuren als Kontrollsequenzen. Auch auf diese Herangehensweise wurde in der hier vorgestellten Auswerte-Software verzichtet, da diese Methode die Einkalkulierung solcher Sequenzen während des experimentellen Designs voraussetzen würde. Die vorgestellte Methode wurde jedoch weitestgehend anhand bereits gedruckter Microarrays entwickelt.
Die Güte der Lowess-Regression selber wird durch ihre Parameter bestimmt. Dabei spielt vor allem die Fensterbreite, aber auch die Anzahl an Ausreißern, die in einem Fenster zugelassen werden, eine entscheidende Rolle.[182] Weitergehende Untersuchung möglicher Parameter könnten zu einer zusätzlichen Verfeinerung der Auswertemethode führen.

6.2.5 *Between-Array*-Normalisierung

6.2.5.1 Hintergrund und Anforderungen

Wie bereits in Kapitel 6.2.4 erwähnt, treten bei Microarray-Experimenten zahlreiche Variabilitäten unterschiedlichsten Ursprungs auf. Die den Microarray-Versuchen meistens zugrunde liegende Fragstellung nach der differentiellen Expression der Gene unter verschiedenen Bedingungen kann daher nur unzureichend beantwortet werden, wenn eine Korrektur dieser Variabilitäten ausbleibt und es somit zur Verfälschung der Ergebnisse kommt. Anstelle rein biologischer Unterschiede werden dann Differenzen zwischen den Zuständen detektiert, die biologischen Ursprungs sind, aber auch durch zufällige und systematische Fehler zustande kommen. Die Bereinigung und Eliminierung einiger dieser systematischen Fehler wurde bereits in vorangegangenen Abschnitten (Kapitel 6.2.2, 6.2.3 und 6.2.4) vorgenommen. Diese Kapitel beschränkten sich auf die separate Normalisierung der Daten auf einzelnen Chips (*Within-Array*- oder *within-slide*-Korrektur). Im folgenden Abschnitt werden Normalisierungsmethoden vorgestellt, die auf den Vergleich der Daten von unterschiedlichen Microarrays abzielen (*Between-Array* oder *slide-to-slide*-Normalisierung).

6.2.5.2 Durchführung und Ergebnisse

Die Notwendigkeit der Anwendung von *Between-Array*-Normalisierung-Methoden kann durch die simultane Hybridisierung von technischen Replikaten überprüft werden. Zu diesem Zweck wurden zehn Hybridisierungen einer identischen *E.coli*-Probe (Kultivierung bei 37 °C und 5% CO_2-Zufuhr) auf *whole-genome*-Chips der selben Microarray-Familie hybridisiert. Der Variationskoeffizient der 4608 Spots zwischen den Chips reichte von 0 bis 22.4%.
Detaillierte Untersuchungen auf Microarrays unterschiedlicher Hersteller zu Positions- wie Chip-bedingten Varianzen wurden auch von Balázsi *et al.* vorgenommen und bestätigen die Notwendigkeit der Nachbearbeitung der Daten.[183]
Solche Chip-bedingten Varianzen führen zu der Verfälschung von Expressionsvergleichen der Zustände auf unterschiedlichen Microarrays durch die künstliche Anhebung des kompletten Gensatzes eines Zustands gegenüber dem anderen und sind somit nicht vernachlässigbar. Eine nachträgliche Korrektur durch statistische Methoden ist somit erforderlich, da auch bei einer

sorgfältiger Überwachung des Druckvorgangs die Erzeugung von Fehlspots nicht ausgeschlossen werden kann.[55]

Die Variabiliäten, die aufgrund des Herstellungsprozesses und während der Hybridisierungsreaktionen zwischen unterschiedlichen Microarrays auftreten, können also differierende Gesamtintensitäten im Vergleich unterschiedlicher Microarrays bedingen. Um die Zustände, die auf unterschiedliche Microarrays hybridisiert wurden, auf gleicher Basis miteinander vergleichen zu können, werden demnach die Variabilitäten auf jedem Microarray so korrigiert, dass die Spots unterschiedlicher Chips anschließend direkt miteinander verglichen werden können.
Wie bereits in Kapitel 6.2.4 bei der Loess-Regression beruhen auch die im Folgenden eingeführten korrigierenden statistischen Methoden wiederum auf der Annahme, dass die Mehrheit der Gene nicht reguliert ist. Es wird also vorausgesetzt, dass die Variationen in den Verteilungen zwischen den Microarrays Resultate experimenteller Bedingungen sind und somit keine biologischen Variabilitäten repräsentieren.[53] Aus diesem Grund wird wie bei der Loess-Regression mit den logarithmierten Ratios (*log ratios*) aus den Intensitätswerten beider Zuständen eines Microarrays gerechnet.

Eine sinnvolle simultane Darstellung der Signale unterschiedlicher Microarrays kann durch *MA-plots* nur bedingt vorgenommen werden, da statistische Größen durch die Diagramme nur unzureichend wiedergegeben werden.
Stattdessen eignen sich so genannte *Box-Plots* oder *Box-Whisker-Plots* als vergleichende Methode zur parallelen Visualisierung unterschiedlicher Verteilungen, indem sie die gleichzeitige Darstellung von statistischen Basisinformationen von logarithmierten Intensitäten (*log ints*) oder Quotienten (*log ratios*) auf verschiedenen Microarrays gewährleisten.[184,185] Als graphische Darstellung der wichtigsten Verteilungsmerkmale von Wertausprägungen einer Variablen, werden in *Box-Plots* das Zentrum, die Streuung, die Schiefe und die Spannweite der Verteilung teilweise inklusive möglicher Ausreißer zusammengefasst. Als Box wird das durch die Quartile bestimmte Rechteckt bezeichnet, das 50% der Daten umfasst, das teilweise auch durch sogenannten *„hinges“* (der obere *„hinges“* entspricht dem 25. Perzentilwert, der untere dem 75.) festgelegt wird. Als Maß für die Streuung der Daten wird die die Länge der Box als Angabe des Interquartilsabstands angegeben. Als zusätzliches Quantil dient der Median, der als Kreuz angeben ist und durch seine Lage innerhalb der Box einen Eindruck von der Schiefe der den Daten zugrunde liegenden Verteilung vermittelt.
Die als vertikale Linien gekennzeichneten, sogenannten *„Whisker“* oder *„Antennen“* besitzen eine Länge von maximal dem 1.5-fache des Interquartilsabstands und werden durch einen Wert aus den Daten bestimmt. Dieser Grenzwert wird zu Ermittlung von Ausreißern herangezogen (Werte, die über dieser Grenze liegen), die separat in das Diagramm eingetragen werden können. Ausreißer, die sich zwischen einem 1.5-fachen und dreifachen Interquartilsabstands befinden, werden als „milde“ Ausreißer bezeichnet. Werte, die über einem dreifachen Interquartilsabstands hingegen als „extreme“ Ausreißer. Diese Bezeichnungen gehen auf Tukey

et al. zurück, wurden inzwischen jedoch teilweise abgewandelt und um zusätzliche Angaben erweitert.[184,186] In Abbildung 6.20 ist ein der Literatur entnommener typischer vertikaler *Box-Plots* mit den Kennzeichnungen der entsprechenden statistischen Größen gezeigt.

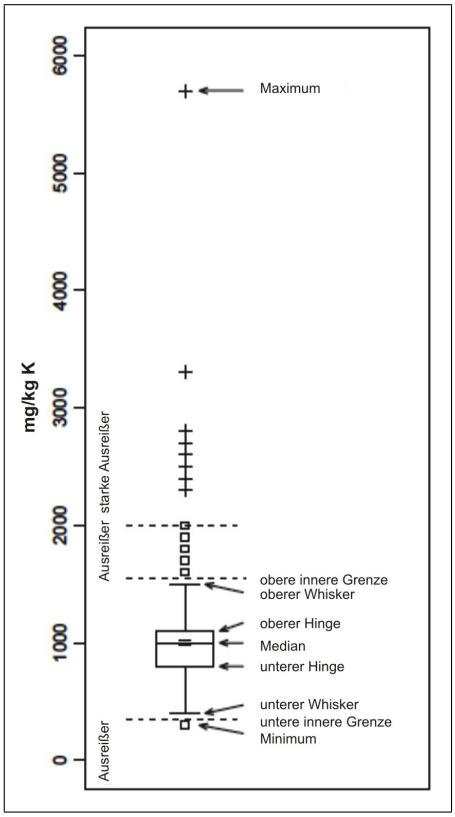


Abbildung 6.20 – *Tukey-Box-Plot für Kalium (K%)-Konzentrationen im O-Horizont von Podsole im Gebiet der Halbinsel Kola (Reimann* et al.*)[187] Die Abbildung zeigt die in einem Boxplot enthaltenen statistischen Größen anhand eines Beispiels aus der Literatur.*

Mit Hilfe solcher *Box-Plots* können die Ergebnisse erweiterter Normalisierungsmethoden anschaulich dargestellt und bewertet werden. Zu diesen *Between-Array*-Normalisierungs-Schritten zählen unter anderem:

- das Zentrieren (*Centering*)[87]
- das Skalieren (*Scaling*) und[72]
- die Normalisierung der Verteilung (*Distribution normalization*)[154]

der in den *Within-Array*-Normalisierungen vorverarbeiteten normalisierten Daten unterschiedlicher Microarrays. Da diesen Auswertemethoden die Annahme einer in der Summe aller Spots nicht vorhandenen Regulation der Gene zugrunde liegt, werden diese Methoden - wie eingangs erwähnt - mit den *log ratios* der beiden jeweiligen Zustände auf den Chips durchgeführt.
Die Methodenauswahl erfolgt unter Verwendung von fünf verschiedenen *whole-genome*-Microarray-Experimenten mit den Genen des Bakteriums *E.coli*. Bei diesen Experimenten handelt es sich um unterschiedlich lange inkubierte K12-MG1655-Stämme, die mit unterschiedlichen

Dosen Arabinose induziert wurden.
Um eine Verfälschung der Normalisierungsparameter einzelner Microarrays zu vermeiden, werden zunächst alle Spots, deren logarithmierten Quotienten aufgrund nicht auswertbarer Daten eines einzelnen oder beider Kanäle dieses Spots den Wert 0 besitzen, markiert und vor der weiteren Berechnung ausgeschlossen. Diese notwendige Vernachlässigung einzelner Datenpunkte stellt eine bereits in Kapitel 6.2.3 beschriebene Einschränkung bei mathematischen Operationen mit logarithmierten Werten dar, die bei der Gesamtaussage über die Ergebnisse berücksichtigt werden sollte. Die derart transformierten Daten können anschließend einer der genannten Normalisierungsmethoden unterzogen werden.
Die einfachste Methode stellt die Zentrierung der Daten dar, die wiederum Bestandteil der Skalierung ist, die ihrerseits einen Schritt während der Normalisierung der Verteilung der Microarrays darstellt. Die am wenigsten manipulative Normalisierungsmethode, das Zentrieren, dient einem Angleich der Mittelwerte der unterschiedlichen Microarrays eines Experiments. Durch Subtraktion des Mittelwertes über alle *log ratios* eines Microarrays von den jeweiligen Einzelwerten (*log ratios* der Spots) dieses Chips befindet sich der gesamte Mittelwerte eines jeden Microarrays nach der Operation auf der Nulllinie.[53,87] Da die Bildung von Mittelwerten äußerst anfällig für auf dem Microarray befindliche Ausreißer ist, bietet sich alternativ zur Subtraktion der Chip-Mittelwertes die Subtraktion der Chip-Mediane an.[186] Da es sich bei diesen Experimenten ausschließlich um Datensätze handelte, die in der Summe keine nennenswerte differentielle Expression aufweisen sollten (siehe Annahme), sollten Ausreißer vermieden werden. Aus diesem Grunde werden die Zentrierungen mit den Medianen aller *log ratios* durchgeführt. In den Abbildungen 6.21 und 6.22 sind die Daten vor den Werten nach der Zentrierung der Microarrays gegenübergestellt. Bei dem in diesen Darstellungen gezeigten Experiment wurden sieben Zustände in einem *Loop*-Design auf sieben unterschiedliche Microarrays hybridisiert. Die einzelnen *Box-Plots* stellen jeweils die einzelnen Microarrays dar.

Wie in den Abbildungen 6.21und 6.22 im Vergleich zu sehen, weisen die Microarrays bereits vor der Zentrierung fast alle einen Median (gekennzeichnet durch grüne, gestrichelte Linien innerhalb der Boxen) nahe dem Null-Punkt auf. Diese Beobachtung unterstreicht die Gültigkeit der getroffenen Annahme einer im Mittel nicht auftretenden differentiellen Expression, da es sich bei den dargestellten M-Werten um die logarithmierten Quotienten aus den beiden Zuständen handelt, die bei gleicher Expression einen Wert von eins und somit nach Logarithmierung zur Basis zwei einen Wert von null erhalten. Während die optische Analyse der Daten über die Diagramme einen ersten Eindruck der Daten vermittelt, werden durch das Programm auch die numerischen Ergebnisse der Zentrierung angegeben. Die Mediane und Mittelwerte (gekennzeichnet durch blaue Kreuze) der Microarrays der Microarrays liegen sowohl vorher als auch nachher alle unterhalb eines Wertes von 0.01. Eine Ausnahme bildet der sechste Microarray, bei dem der Median vor der Zentrierung einen Wert von 0.034 und der Mittelwert von -0.032 besitzt. Damit weicht dieser Microarray von den anderen Chips durch einen vergleichsweise großen Abstand zum Nullpunkt ab. Außerdem fällt bei diesem Microarray die starke Konvergenz zwischen Median und Mittelwert auf, was auf einen qualitativ vergleichsweise geringer wertigen Microarray hindeutet (hohe Anzahl an Ausreißern). Da

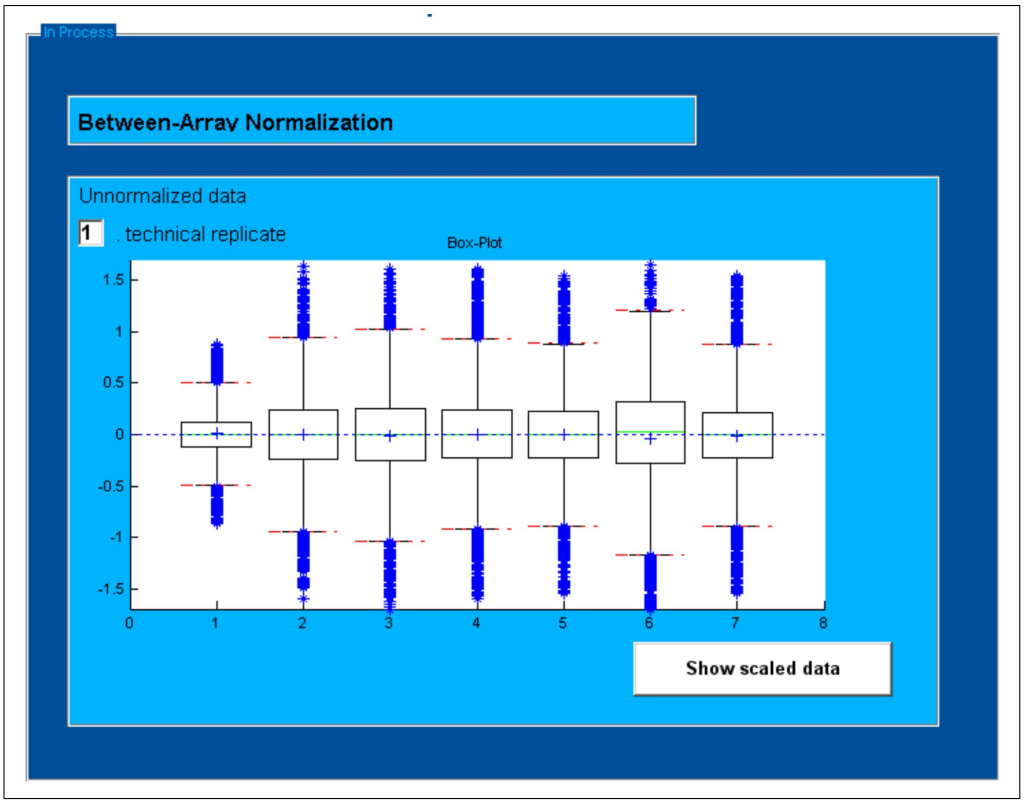


Abbildung 6.21 – *Box-Plots der Chips eines Microarray-Experiments (whole-genome, E.coli) vor einer Between-Array-Normalisierung.*

dieser Microarray jedoch während der Qualitätsanalyse (siehe Kapitel 6.2.2) keine auffälligen Abweichungen zeigte und somit zu einer zur Datenanalyse ausreichenden Güte befunden wurde, soll die *Between-Array*-Normalisierung zur Bereinigung der verbleibenden Chip-bedingten Variabilitäten dienen.
Wie anhand des *Box-Plots* dieses Microarrays nach der Zentrierung zu sehen, konnte mit Hilfe der Zentrierung der Gesamtmedian über alle validen *log ratios* dieses Microarrays der Nulllinie angenähert werden (Medianwert nach der Zentrierung $< -10^{-4}$). Die Zentrierung der Daten führt also - wie anhand dieses Beispiels gezeigt - zu einer Vereinheitlichung der Gesamtmediane aller Microarrays und somit letztlich zu einer optimierten Erfüllung der vordefinierten Bedingung gleicher Gesamtexpressionen über alle Gene von *whole-genome*-Microarrays.

Wie in Abbildung 6.22 deutlich zu sehen, weichen nach einer Zentrierung der Daten die Standardabweichungen der unterschiedlichen Microarrays nach wie vor stark voneinander ab. Um die Standardabweichungen der Microarrays anzugleichen, wird daher eine Skalierung der Daten vorgenommen. Auf diese Weise soll also sichergestellt werden, dass die Mittelwerte bzw. Mediane der Chips als auch deren Standardabweichungen möglichst gleich sind. Die neuen Spot werden dementsprechend nach der folgenden Formel berechnet:

$$I_i' = \frac{I_i - \mu_i}{\sigma_i} \tag{6.16}$$

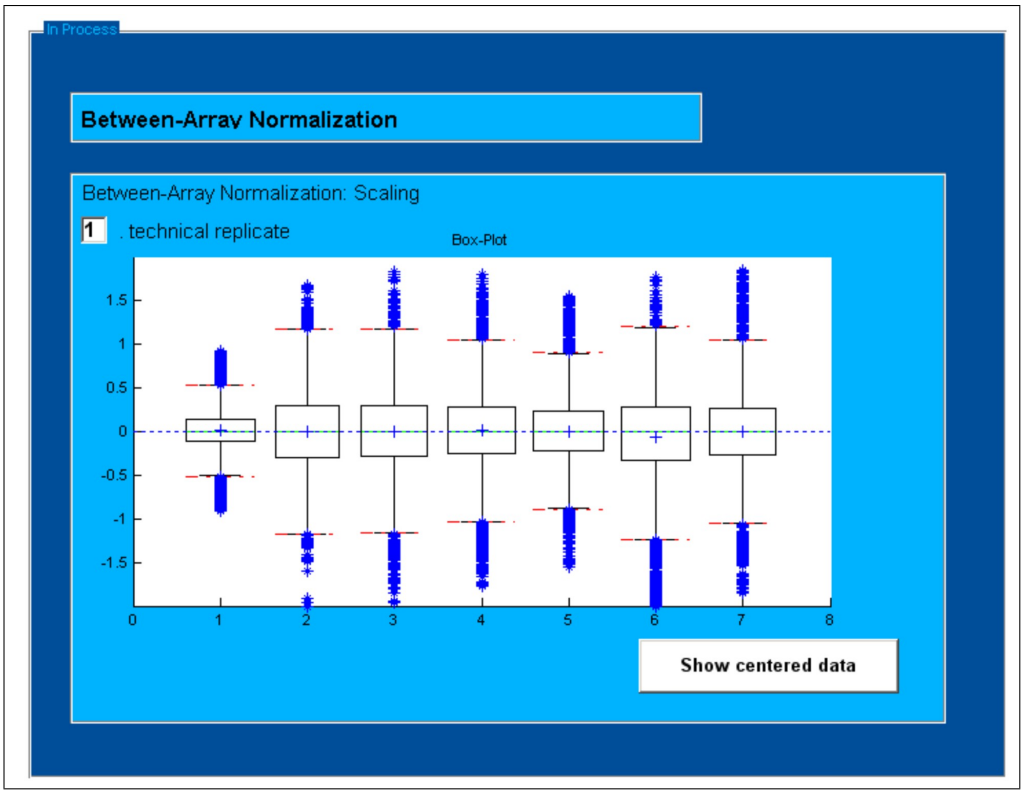


Abbildung 6.22 – *Box-Plots der Chips eines Microarray-Experiments nach der Zentrierung der Datensätze. Die Bedeutung der einzelnen Elemente der Box-Plots sind in Abbildung 6.20 erläutert.*

mit i als Index des Microarrays 1 bis 7, μ dem Mittelwert bzw. Median über alle *log ratios* und σ als der Standardabweichung aller Werte eines Chips.
Die Berechnung normalisierter Werte mit dieser Formel sollte demgemäß in Mittelwerten bzw. Medianen der Microarrays um den Nullpunkt resultieren, während sich die Standardabweichungen eins annähern sollten.
Das Skalieren der Expressionswerte wird standardmäßig in erweiterten Auswertemethoden zur Korrektur von *slide-to-slide*-Effekten angewandt und kann besonders dann an Bedeutung gewinnen, wenn anschließend an die Auswertung *Pearson-Korrelationskoeffizienten* in Clusteranalysen berechnet werden sollen, da mittels der Korrelationskoeffizienten Distanzmatrices berechnet werden können, die wiederum die Basis für Verwandtschaftsbeziehungen in Clusteranalysen stellen können.
Wie bereits bezüglich der Zentrierung erwähnt, können auch bei der Skalierung wieder die Mittelwerte durch Mediane ersetzt werden, was einerseits im Zähler bei der Differenzbildung eine Rolle spielt, andererseits aber auch für die Berechnung der sogenannten „,mittlere absoluten Abweichung vom Median“ (MAD, engl.: median absolute deviation, auch: MedMed) als Pendant zur regulären Standardabweichung gilt.[53]
Die mittlere absolute Abweichung MAD ist definiert durch:[188]

$$P(|X - \tilde{x}| \leq MAD) = 0.5 \tag{6.17}$$

Im Falle einer konkreten Stichprobe wird sie errechnet durch:[188]

$$MAD = median\,|x_i - \tilde{x}| \tag{6.18}$$

Durch die Definition ergibt sich im Falle von normalverteilten Daten folgender Zusammenhang zur Standardabweichung:[188]

$$MAD = z_{0.75} \cdot \sigma \tag{6.19}$$

$z_{0.75}$ ist das 0.75-Quantil der Standardnormalverteilung. Die Berechnung normalisierter Expressionswerte über die Mediane der Einzelwerte hat wiederum den Vorteil einer erhöhten Robustheit gegenüber Ausreißern.
Um einen direkten Vergleich zwischen den Berechnungen mit den Mittelwerten bzw. den Medianen anstellen zu können, wurden die Ergebnisse beider Berechnungen gegenübergestellt (Ergebnisse nicht gezeigt). Dabei stellte sich eindeutig heraus, dass das Skalieren der Daten über die Mediane und MAD der Spots zu wesentlich einheitlicheren Ergebnissen führte (geringen Abweichungen der Gesamtmediane vom Nullpunkt und geringere Schwankungen innerhalb der Standardabweichungen der Microarrays). Aus diesem Grund wurde auch hier weitergehend mit den Medianen gerechnet.
Die Ergebnisse der Skalierung aller Microarrays dieses Experiments sind in Abbildung 6.23 dargestellt.

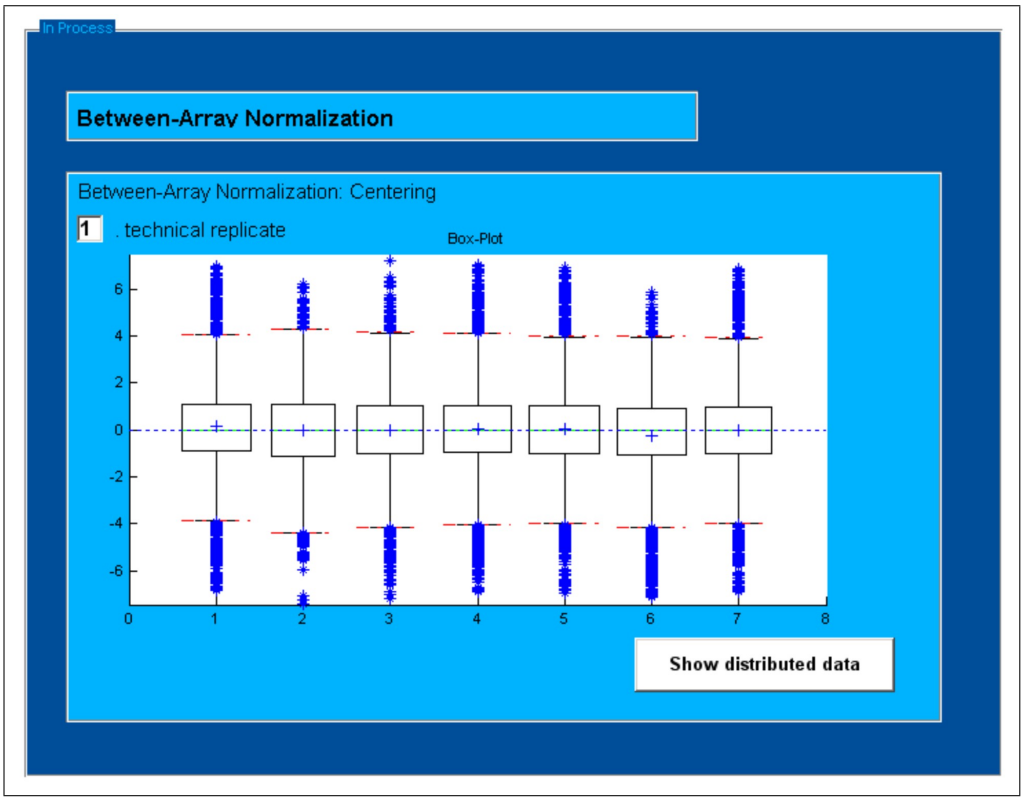


Abbildung 6.23 – *Box-Plots der Chips eines Microarray-Experiments nach der Skalierung der Datensätze.*

Die Division der zentrierten Daten durch die mittlere absolute Abweichung vom Medi-

an zeigt eindeutig die Annäherung der Standardabweichungen der Microarrays an eins. Die entsprechenden Daten sind in Tabelle 6.16 angegeben.

Tabelle 6.16 – *Standardabweichungen vor und nach der in der Between-Array-Normalisierung skalierten Microarrays.*

	Standardabweichungen der Microarrays						
	Chip 1	Chip 2	Chip 3	Chip 4	Chip 5	Chip 6	Chip 7
Vor BA	0.12	0.24	0.26	0.24	0.22	0.32	0.22
Nach BA	1.06	1.07	1.05	1.06	1.02	0.95	0.97

Wie in der Tabelle ersichtlich, streuen die normalisierten Daten um den Wert eins. Die Nähe zu einem festen Wert ist jedoch weniger bedeutsam für die Vergleichbarkeit unterschiedlicher Microarrays als die relative Ähnlichkeit der Standardabweichungen untereinander. Als vergleichendes Maß für die Streuung der Daten kann die relative Standardabweichung der Werte über alle Microarrays herangezogen werden. Diese Berechnung ergab eine relative Standardabweichung über alle Werte von 0.25 vor der Skalierung der Daten und entsprechend von 0.05 über alle Einzelstandardabweichungen der Microarrays nach dieser *Between-Array*-Normalisierung.
Mittels dieser *Between-Array*-Normalisierung konnten also deutliche Verbesserungen bezüglich der Vergleichbarkeit unterschiedlicher Microarrays eines Experiments erreicht werden, und zwar sowohl hinsichtlich der angenäherten Mediane wie auch in Bezug auf deren mittlere absolute Abweichung.

Aufgrund der erwiesenen Validität der Skalierung, wurde diese Methode ebenfalls in die dritte aufgeführte *slide-to-slide*-Normalisierungsmethode, die *Distribution normalization*, integriert. Die Entwicklung dieser Quantil-Methode durch Bolstad *et al.* beruht auf der Prämisse, dass bei einem Microarray-Experiment bestehend aus unterschiedlichen Chips eine geringe Anzahl an Genen zwar differentiell exprimiert sein mag, die Gesamtverteilung der Spot-Intensitäten jedoch nicht zu stark variieren sollte.[154] Die Normalisierung der Daten über die Verteilung resultiert in gleichen Streuungen der Intensitätswerte der Daten. Die Programmiermethodik dieses Normalisierungsschrittes übertrifft die bisher vorgestellten Methoden an Komplexität, da sie mehrere Schritte beinhaltet:

- Zunächst werden die Daten skaliert.
- Anschließend werden die skalierten Daten eines jeden Microarrays D_i entsprechend ihrer Größe vom niedrigsten zum größten Wert sortiert, so dass D_{i1} der niedgrigste Werte des Arrays D_i wäre und D_{in} der größte bei n Messwerten insgesamt.
- Die sortierten Werte aller Microarrays werden in einer Matrix angeordnet, so dass in einer Spalte die Werte mit gleichen Größenindices zu finden sind (die jedoch für Intensitätswerte unterschiedlicher Spots auf den Chips stehen können). Diese Matrix wird

zur Berechnung einer neuen Verteilung *D'* durch Mittelung der sortierten Werte über alle Microarrays herangezogen, so dass für die kleinste normalisierten Werte auf allen Chips gilt, so dass $D'_1 = avg\{D_{11}, ...D_{1m}\}$ mit m als der Anzahl der Arrays. Die zweitkleinsten Werte aller Chips wären dementsprechend $D'_2 = avg\{D_{21}, ...D_{2m}\}$. Auf diese Weise wird für jeden Intensitätswert aller Chips eine neuer Wert bis D'_n berechnet.

- Im letzten Schritt dieser *Between-Array*-Normalisierung wird jeder Messwert auf jedem Microarray durch den entsprechenden Mittelwert in der neuen Verteilung entsprechend der jeweiligen Position ersetzt (Rücksortierung der Daten). Wenn beispielsweise ein bestimmter Intensitätswert des Microarrays D'_i der 100 kleinste Wert dieses Microarrays ist, so wird dieser Wert durch den 100 kleinsten Messwert D'_{100} der neuen Verteilung ersetzt.

Die Auswirkung einer Verteilungsnormalisierung auf die Daten ist in Abbildung 6.24 zu sehen.

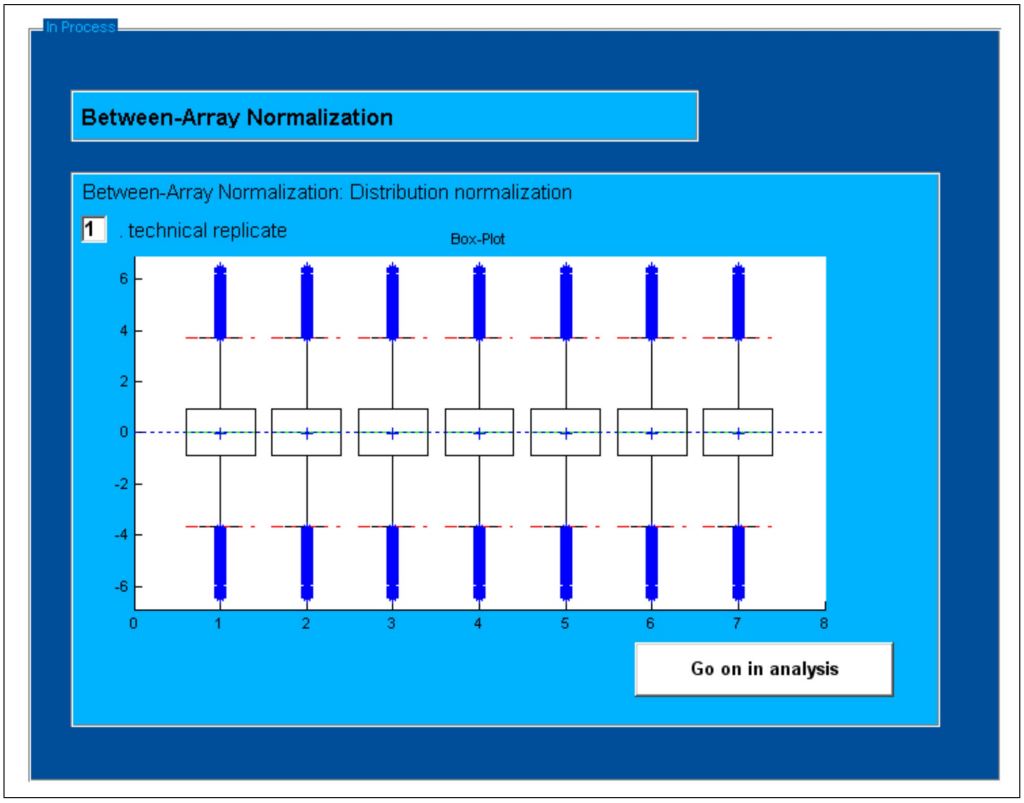


Abbildung 6.24 – *Box-Plots der Chips eines Microarray-Experiments nach der Normalisierung der Verteilung der Datensätze.*

Bei dieser Normalisierungsmethode werden ähnlich wie bei der in Kapitel 6.2.4 vorgestellten Loess-Regression Intensitäten aneinander angeglichen. Während die Loess-Regression jedoch durch Veränderung des Mittelwertes des Datensatzes für *Within-Array*-Normalisierungen herangezogen wird, wird die *Distribution normalization* auf die Gesamtheit der Microarrays angewandt, um die Verteilung der Spotintensitäten über die Microarrays aufeinander abzustimmen.

Wie anhand der Beschreibung der Methodik sowie in Abbildung 6.24 sichtbar, handelt es sich bei dieser Normalisierung um eine äußerst manipulative Angleichung der Daten. Auf der Annahme der gleichen Verteilung der Daten beruhend, werden die Intensitätswerte unterschiedlicher Microarrays vereinheitlicht, so dass durch Mittelwertbildung letztlich auf allen Microarrays insgesamt ausschließlich identische Messwerte vorliegen. Eine Gleichverteilung der Intensitätswerte ist zwar theoretisch anzunehmen, entspricht jedoch in der Praxis nicht der Realität. Die starken Schwankungen der Verteilung der unterschiedlichen Microarrays sind in Abbildung 6.23 deutlich zu sehen. Sie können durch vielfache Variabilitäten auf den Microarrays (schwankender Hintergrund, Unreinheiten auf der Oberfläche, Scan-Unterschiede, usw.), aber auch durch biologische Unterschiede (starke Expression selektierter Gene in definierten Zuständen) zustande kommen. Bei der stark invasiven Normalisierung der Verteilung können somit wichtige Informationen verloren gehen, weshalb diese Methode in der Literatur selten empfohlen wird. Zu den seltenen Anwendungsgebieten dieses Verfahrens zählen vor allem Affymetrix-Microarrays mit ihrem von den hier verwendeten Microarrays abweichenden Chip-Design.[154]

6.2.5.3 Fazit

In diesem Kapitel wurden unterschiedliche *Between-Array*-Normalisierungs-Verfahren vorgestellt, die jedoch alle auf der Annahme einer im Mittel nicht vorhandenen differentiellen Expression basieren. Verfahren, die auch auf *low-density*-Microarrays anwendbar sind, bei denen solche Annahmen also nicht zutreffen, sind in der Literatur bisher nicht beschrieben. Diesem Mangel an *Between-Array*-Normalisierungen für *low-density*-Microarrays liegen die biologische bedingten starken Unterschiede (starke differentielle Expression im Vergleich unterschiedlicher Zustände) zugrunde, so dass bisherige *Between-Array*-Methoden immer auch die biologische Aussage der Experimente verfälschen würden. Aus diesem Grund beschränken sich die hier vorgestellten Methoden auf Microarrays, bei denen die Voraussetzung erfüllt ist, dass über die Gesamtheit aller Gene keine Regulation erfolgt.
Wie in den Abbildungen 6.22, 6.23 und 6.24 zu sehen, führten alle drei der im Rahmen dieser Arbeit getesteten Methoden (*Centering*, *Scaling* und *Distribution normalization*) zu den erwarteten Ergebnissen (angeglichener Median, Streuung, bzw. Verteilung). Dennoch erwies sich das Skalieren der Daten als die sinnvollste Methode, da sie einerseits zur sinnvollen Angleichung der Daten führt, andererseits aber nicht zu stark in die ursprüngliche Datenstruktur eingreift - wie dies beispielsweise bei der *Distribution normalization* der Fall sein kann. Die Integration der Skalierung als Standard-*Between-Array*-Normalisierung in dem implementierten Auswerteprogramm entspricht dem in der Literatur beschriebenen gängigen Vorgehen.[8,72] Eine Erweiterung der gemeinhin angewandten Zentrierung konnte durch die Verwendung von Medianen bzw. MAD in dieser Methode erreicht werden. Damit bestätigt sich der bereits in vorangegangenen Abschnitten (Kapitel 6.2.2 und 6.2.3) beschriebene Vorteil einer bei ausreichender Datenmenge vorgenommenen Verwendung des Medians anstelle des Mittelwerts, da so eine erhöhte Stabilität gegenüber Ausreißern gewährleistet werden kann.

Eine weitere vielversprechende Methode zur Garantie einer erhöhten Robustheit, die auf

grundsätzlich unterschiedlichen Voraussetzungen beruht, ist das von Khan *et al.* eingeführte sogenannte „*Mutliplexing*“,[189] das die Bündelung mehrerer Gene innerhalb eines Spots vorsieht. Diese Methode ermöglicht zwar einerseits eine signifikante Korrektur der *slide-to-slide*-Variationen setzt jedoch ein grundsätzlich neues Microarray-Design voraus, was in diesem Fall nicht gegeben war.

6.2.6 Endauswertung

6.2.6.1 Hintergrund und Anforderungen

In den Kapiteln 6.2.2 bis 6.2.5 wurden die während der Daten-Prozessierung implementierten Algorithmen eingehend beschrieben und analysiert. Obgleich dieser Prozess der Bearbeitung und Normalisierung der Daten auf die Korrektur zahlreicher den Microarray-Experimenten innewohnenden Variabilitäten abzielt und somit der mathematisch aufwendigste Schritt eines Auswerteprogramms für Microarrays darstellt, dient er letztlich nur der Vorbereitung der Werte auf die Datenanalyse. Die Datenanalyse beschränkt sich zwar in der Regel auf wenige statistische Methoden, diese sehen jedoch die Beantwortung der Fragestellung des Experiments vor und sind somit für den Experimentator von entscheidender Bedeutung.
Die Untersuchung des Genexpressionsprofils auf die differentielle Expression einzelner Gene wird „*Multiple Testing*“ genannt und stellt die ursprüngliche und wichtigste Aufgabe der Datenanalyse dar. Diese Definition der Datenanalyse liegt diesem Kapitel zugrunde, weswegen der folgende Abschnitt ausschließlich der Untersuchung relativer Unterschiede zwischen denen im Microarray-Experiment vorliegenden Zuständen gewidmet ist.
Darüber hinausgehend können die vom Experimentator gestellten Fragestellungen jedoch weitergehend sein und beispielsweise die Klassifizierung bzw. das Clustern der Gene umfassen. Die Beantwortung derartiger Fragestellungen wird in Kapitel 6.2.7 vorgenommen.

6.2.6.2 Durchführung und Ergebnisse

Für die Selektion differentiell exprimierter Gene aus Microarray-Daten werden in der Literatur eine Vielzahl unterschiedlicher Methoden vorgeschlagen.[72,83,87,90,91,190] In diesem Kapitel werden einige dieser beschriebenen Methoden vorgestellt und auf die Eignung für die hier verwendeten Microarray-Experimente gestestet. Die Methoden umfassen sowohl eine klassische Implikation als auch neuartigere Ansätze, unter deren Anwendung signifikante Aussagen über die Hoch- bzw. Runterregulation der Gene getroffen werden sollen.

Die während der Datenanalyse auf die differentielle Expression der Gene zu vergleichenden Zustände können durch eine oder mehrere Proben pro Gen repräsentiert werden. Die Probenmenge der Zustände wird in der Regel durch die Anzahl der Replikate bestimmt: bei *whole-genome*-Microarrays liegen zwar zumeist keine Genreplikate vor, in Abhängigkeit vom Experiment-Design können technische Replikate aber die Anzahl der Proben erhöhen. Die Anzahl der Proben der Zustände auf *low-density*-Chips hingegen kann sowohl durch Genreplikate wie auch durch technische Replikate gekennzeichnet werden.

Dabei kann auf die Nullhypothese H_0, dass keine Unterschiede in der Expression zwischen den Proben existieren, gestestet werden. Wird ein Signifikanztest durchgeführt, so wird im Stichprobenraum eine so genannte „kritische Region“ K_α derart festgelegt, dass bei zutreffender Nullhypothese H_0 das Ergebnis einer Zufallsstichprobe vom Umfang n höchstens mit der bestimmbar kleinen Wahrscheinlichkeit α in diese kritische Region fällt. Wenn das Ergebnis der Zufallsstichprobe in diese kritische Region fällt, so wird H_0 bei Zugrundelegung der Irrtumswahrscheinlichkeit α abgelehnt. Der so genannte „p-Wert“ beschreibt hierbei den kleinsten Wert α, für den H_0 abgelehnt wird.

Die einfachste Methode, differentiell exprimierte Gene herauszufiltern, besteht in der klassischen Herangehensweise, für jedes Gen das Expressionsverhältnis zwischen zwei Bedingungen zu untersuchen, und alle Gene, die sich um mehr als einen festgelegten Wert unterscheiden, als differentiell exprimiert zu betrachten. Dieser Test, meist als „*Fold-Change*“ bezeichnet, findet oftmals Anwendung bei der Auswertung von Microarray-Daten,[190] da die Anzahl der Proben für die jeweiligen Zustände variieren kann. Mit dieser Methode können also auch *whole-genome*-Microarray-Experimente ausgewertet werden, bei denen experimentell bedingt keine technischen Replikate verwendet werden konnten (z.B. wegen nicht ausreichend verfügbaren biologischen Materials oder aus Kostengründen), so dass für jeden Zustand höchstens ein Wert vorliegt.

In Abbildung 6.25 ist *MA-Plot* eines *whole-genome*-Microarrays dargestellt, der die Gene des Bakteriums *E.coli* trägt und dem bereits in Kapitel 6.2.5 diskutierten Experiment entnommen wurde. Die beiden gebräuchlichsten Grenzwerte sind durch Linien gekennzeichnet. Aus Symmetriegründen werden die Ergebnisse des Tests meist logarithmisch transformiert angegeben.

Wie in Abbildung 6.25 verdeutlicht, findet sich in der Literatur kein einheitlicher Konsens bezüglich des Grenzwertes, ober- bzw. unterhalb dessen alle Gene als reguliert betrachtet werden. Üblicherweise wird ein Grenzwert von 2 / -2 (der Logarithmus zur Basis 2 von 2 entspricht einem Wert von 1 oder -1 für den negativen Bereich - siehe Abbildung 6.25) festgelegt.[72] Schena *et al.* hingegen definieren die fünffache Expression des einen Zustands im Vergleich zum anderen als sinnvollen Schwellenbereich, und DeRisi *et al.* identifizieren differentiell exprimierte Gene über den 3-*“Fold“* der logarithmierten Quotienten.[1,191]
Diese Uneinigkeit in Bezug auf die eindeutige Determinierung einer differentiellen Expression beim Vergleich unterschiedlicher Zustände basiert auf der statistischen Unsicherheit dieser Methode. So dass die Festlegung eines Grenzwertes letztlich auf empirischen Studien mit einer Vielzahl unterschiedlicher Microarrays basieren muss. Aus diesem Grunde wurde sowohl die in Kapitel 6.2.4 erwähnten, für die Untersuchung der Lowess-Regression verwendeten *whole-genome*-Microarrays verwendet, um einen sinnvollen Grenzwert abzuschätzen. Diese Analyse ergab, dass selbst bei einem Grenzwert von zwei ($log_2(2) = 1$) die Mehrzahl der Gene nicht differentielle exprimiert ist (eine maximale Regulation über alle Experimente bei 5% der Gene im Gegensatz zu einer maximalen Regulation von 1% der Gene bei einem Grenzwert von

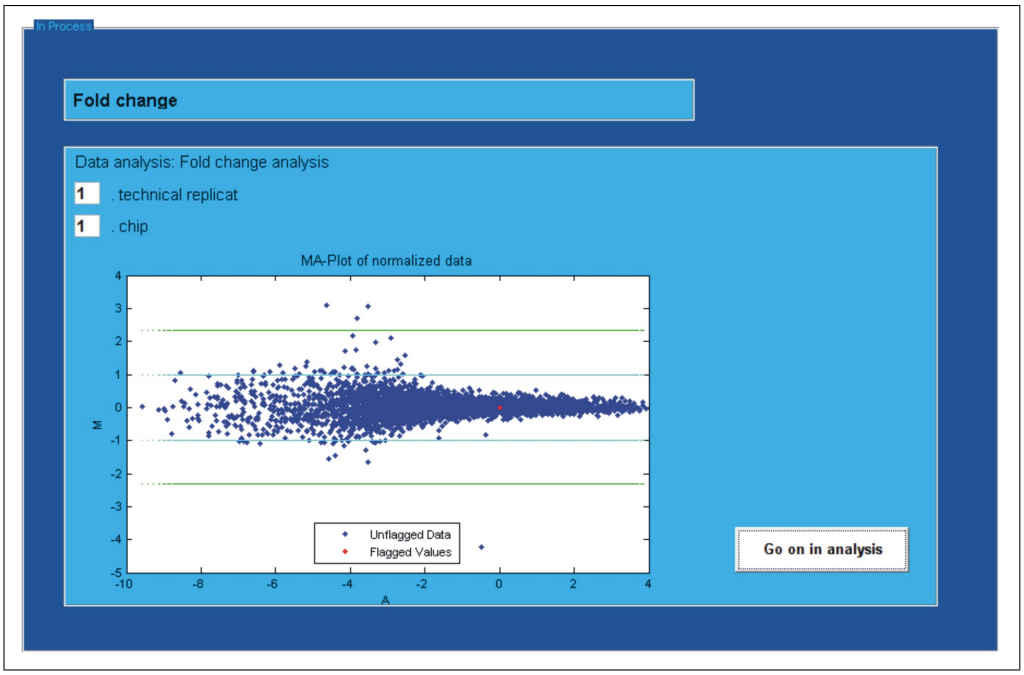


Abbildung 6.25 – *MA-Plot eines whole-genome-Microarrays mit Kennzeichnung möglicher „Fold-Change"-Grenzwerte. Die cyanfarbene sowie die grüne Linien zeigen unterschiedliche in der Literatur erwähnte Grenzwerte zur Auffindung differentiell exprimierter Gene.*[190]

fünf) und somit die Annahme einer nicht vorliegenden differentiellen Gesamtexpression für *whole-genome*-Microarrays erfüllt ist. Um den in der Literatur beschriebenen Problemen bei der exakten Festlegung eines Grenzwertes zur Bestimmung einer differentiellen Expression, die reale biologische Unterschiede beschreibt, gerecht zu werden, werden die Quotienten oberhalb der Grenzwerte zwei und fünf jedoch gekennzeichnet und unterschiedlich markiert.

Die Berechnung der Regulation mittels der *„Fold-Change"*-Methode erfolgt in der Regel über die logarithmisch transformierten Quotienten der beiden Zustände. Um die auf diese Weise erreichte Symmetrie jedoch auch bei unlogarithmierten Werten zu erhalten, können alternativ alle Quotienten zwischen null und eins gemäß der folgenden Gleichung in den negativen Bereich transformiert werden: $x' = -\frac{1}{x}$. Auf diese Weise liegen die Quotienten der beiden Zustände vergleichbar gespiegelt um die Nulllinie vor. Die Verwendung der Quotienten hat den Vorteil, dass der Experimentator den realen Unterschied der beiden Zustände abschätzen kann.

Wie bereits erwähnt, ermöglicht diese Methode die Berechnung der differentiellen Expression von *whole-genome*-Microarrays, die weder Genreplikate tragen noch als technische Replikate vorliegen. Die Fähigkeit der Auswertung auch nicht-replikativer Datensätze durch die *„Fold-Change"*-Methode bedeutet jedoch gleichzeitig, dass es sich bei dieser Anwendung nicht um einen statistischen Test handelt. Somit können auch keine Fehlerwahrscheinlichkeiten für die Zuweisung der Genregulation existieren. Die Aussage über die differentielle Expression der Gene kann daher im Vergleich zu statistischen Tests nicht gleichwertig verlässlich

sein. Diese Einschränkung sollte bei Ergebnissen, die in Ermangelung von Replikaten mit der „*Fold-Change*“-Methode errechnet wurden, berücksichtigt werden.

Wenn für die Gene der zu vergleichenden Zustände jedoch mehr als eine Probe vorliegt, so kann auf weitergehende Tests zurückgegriffen werden.
Einige dieser Tests erwiesen sich für die Anwendung im realen Forschungsbetrieb als impraktikabel (z.B. „*Noise sampling*“, „*Model based maximum likelihood estimation*“), da sie die separate detaillierte Analyse einer Vielzahl von Chips für jede Microarray-Familie voraussetzen.[192]
Der so genannte t-Test, der unter Anwendung statistischer Implikationen die Selektion differentiell exprimierter Gene vorsieht, hingegen erfordert für Einzelexperimente keine weiteren Vorversuche und kann somit zur Identifizierung regulierter Gene herangezogen werden. Dieses Verfahren liefert eine Entscheidungshilfe darüber, ob ein gefundener Mittelwertsunterschied zufällig entstanden ist oder ob bedeutsame Unterschiede zwischen den zwei untersuchten Gruppen (aus Proben bestehende Zuständen) existieren.[193]
Dazu wird für jedes untersuchte Gen auf der Grundlage der Expressionswerte die Wahrscheinlichkeit (p-Wert) berechnet, dass die beobachteten Mittelwertsunterschiede zwischen unterschiedlichen Gruppen durch einen Stichprobenfehler erklärbar sind und es daher keinen echten Unterschied zwischen den beiden Gruppen gibt. Je kleiner der p-Wert ist, umso geringer ist demnach die Wahrscheinlichkeit, dass die beobachteten Ergebnisse zufällig entstanden sind und umso signifikanter sind folglich die Ergebnisse. Im Gegensatz zur „*Fold-Change*“-Methode handelt es sich beim t-Test also eindeutig um einen Signifikanztest.
Der t-Test basiert auf der t-Verteilung, die der Normalverteilung ähnelt, jedoch auf eine geringere Stichprobengröße angewandt werden kann und sich mit zunehmender Größe dieser Stichprobe der Normalverteilung annähert. Die sogenannte „Student's t-Verteilung“ ist somit ein Spezialfall der allgemeinen hyperbolischen Verteilung.[194] Eine Voraussetzung zur Berechnung differentieller Expressionsgrade mit Hilfe des t-Tests sind folglich identische Varianzen der beiden zu vergleichenden Gruppen. Zu diesem Zweck wird mit Hilfe eines weiteren statistischen Tests, dem F-Test, mit einer bestimmten Konfidenz entschieden, ob sich die beiden Stichproben (Gruppen) hinsichtlich ihrer Varianz wesentlich unterscheiden. Er dient damit unter anderem der generellen Überprüfung von Unterschieden zwischen zwei statistischen Populationen. Der F-Test berechnet sich aus dem Quotienten der beiden Varianzen nach folgender Formel:

$$F_{Stichprobe} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \tag{6.20}$$

Wenn die Nullhypothese des F-Tests (gleiche Varianzen) also Gültigkeit hat und die Stichprobenvarianzen Varianz σ_1^2 und σ_2^2 der Gruppen beider Zustände berechnet werden können (die Gruppen einen Umfang n_1 und n_1 von mindestens eins haben), so kann der t-Wert für zwei unabhängige Proben gemäß der folgenden Formel berechnet werden:

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sigma} \cdot \sqrt{\frac{N_1 \cdot N_2}{N_1 + N_2}} \tag{6.21}$$

mit der sogenannten gewichteten Varianz σ^2:

$$\sigma^2 = \frac{(x_1 - 1) \cdot \sigma_1^2 + (x_2 - 1) \cdot \sigma_2^2}{n_1 + n_2 - 2} \tag{6.22}$$

wobei $n_1 + n_2 - 2$ der Anzahl der Freiheitsgrade entspricht. Die in Gleichung 6.21 gezeigte Formel zeigt die Berechnung des t-Wertes für die unlogarithmierten Mittelwerte $\bar{x_i}$ beider Zustände. In der Literatur sind jedoch auch Beispiele zu finden, bei denen statt der absoluten Differenz beider Zuständen der Mittelwert der *log ratios* gebildet wird.[53,72] Der Einfluss einer vorab durchgeführten logarithmischen Transformation auf die Ergebnisse der Datenanalyse wurden von Li *et al.* ausführlich diskutiert. Da es bei den im Rahmen dieser Arbeit analysierten Microarrays oftmals *low-density*-Chips untersucht wurden und diese nur eine begrenzte Anzahl an Genen tragen, wurde auf die Transformation in den logarithmischen Wertebereich verzichtet, um einen unnötigen Datenverlust zu vermeiden (zu Transformationsverfahren siehe Kapitel 6.2.3).

Wenn die Nullhypothese des F-Tests hingegen nicht zutrifft und die Varianzen der beiden Stichproben voneinander abweichen, kann eine Berechnung der differentiellen Expression nicht über die in Gleichung 6.21 dargestellte Formel erfolgen. In diesem Fall gilt für die Berechnung des t-Wertes die in Gleichung 6.23 vorgestellte Formel:

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sigma} \tag{6.23}$$

mit einer gewichteten Varianz σ^2, die wie folgt berechnet wird:

$$\sigma^2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.24}$$

Die Untersuchung der Validität des t-Tests zur Detektion differentiell exprimierter Gene wird anhand eines Experiments mit Schwannzellen gezeigt, das zugleich den möglichen Einsatz von Microarrays bei der Beantwortung aktueller Forschungsfragen demonstriert.

Da das periphere Nervensystem selbst im späten Erwachsenenstadium ein bemerkenswertes Regenerationspotential besitzt, werden derzeit zahlreiche therapeutische Strategien untersucht, die auf die Regeneration verkürzter Nerven abzielen. Besonders stärkere nervliche Verletzungen können mit bisherigen Regenerationsansätzen nicht geheilt werden, so dass eine Reinnervierung der betreffenden Nerven ausbleibt.[196,197]
Neurobiologische Ansätze versuchen bisherige transplantationsbedingte Probleme zu umgehen, indem sie die Einführung synthetischer Nervenkanäle vorsehen.[198,199] Die Entwicklung von synthetischem peripheren Nervengewebe basiert also auf der Konstruktion geeigneter biodegradabler Matrices, die eine Fibrose minimieren und die Regeneration der Heilung fördern

bevor sie im Körper abgebaut werden. Um die Regeneration zu beschleunigen, werden gewöhnlich Schwannzellen in Nervenkanäle injiziert, die auf molekularer Ebene zur Regeneration beitragen, in dem sie zur Expression von Zelladhäsionsproteinen, löslichen Wachstums- und Differenzierungsfaktoren sowie Komponenten der Extrazellulärmatrix (*ECM*) beitragen.[200,201]
Eine Vielzahl dieser Bestandteile der Extrazellulärmatrix des peripheren Gewebes spielen eine wichtige Rolle während der Entwicklung des Peripheren Nervensystems (*PNS*), weswegen einige dieser Komponenten bereits auf ihre Tauglichkeit als Materialen für potentielle Matrices zur Regeneration von Nervenverletzungen getestet wurden. Diese meisten dieser Materialen erwiesen sich aus unterschiedlichen Gründen als unbrauchbar zum Einsatz im Gewebe. Jüngere Ansätze zielen daher auf die Verwendung neuartiger Moleküle, die an der adhäsiven Interaktion in der synaptischen Zone beteiligt sind.[202] Zu diesen Molekülen zählen neben Integrinen auch neurale Zelladhäsionsmoleküle (*NCAMs*), die ebenfalls von Schwannzellen exprimiert werden.[202]
Den Isoformen von NCAM, die ein Carbohydrat-Polymer der Polysia-Säure (*PSA*) tragen, werden in der Literatur aufgrund der negativen Ladung von PSA besondere Eigenschaften zugeschrieben, die mit der Fähigkeit von PSA, als Spacer zu fungieren, in Verbindung gebracht werden.[203]

In diesem Experiment wird daher der postulierte positive Einfluss von PSA auf die Synaptogenese und das axonales Wachstum während der Nervenregeneration bei Schwannzellen *in vitro* untersucht. Zu diesem Zweck wurde ein kommerziell erhältliches Homopolymer der Polysia-Säure, die Colomin-Säure (*CA*), als neues Beschichtungsmaterial für Matrices im *Tissue Engineering* getestet. Das Polysaccharid CA ist immunologisch identisch mit dem biochemisch gewonnenen PSA aus Säugetieren, besitzt diesem gegenüber jedoch den Vorteil, dass es aus dem *E.coli*-Stamm K1 isoliert werden kann.[204]
Im Gegensatz zu vorangegangenen Studien, die vor allem den Einfluss von CA auf neurale Zelllinien und primäre Neuronen untersucht haben,[205] liegt der Fokus der hier vorgestellten Studien auf dem Effekt der Kultivierung immortalisierter Schwannzellen auf CA. Da die Angabe der Expressionsprofile aus Microarray-Experimenten in Form relativer Werte der Transkription aus dem Vergleich zweier Zustände erfolgt, wurden die auf CA kultivierten immortalisierten Schwannzellen (*iSZ*) einer Negativkontrolle (unbeschichtete Spinnerflaschen-Oberfläche) sowie einer Positivkontrolle (Kultivierung auf einer Laminin-beschichteten Oberfläche) in einem *Loop*-Design gegenübergestellt.
Auf die für dieses Experiment hergestellten *low-density* Microarrays wurden 352 neurospezifische Rattengene gespottet. Eine detaillierte Versuchsbeschreibung ist im Anhang J zu finden.

Laminin (*Lam*) versus unbeschichtete Oberfläche (*uo*)

Im ersten Vergleich des *Loop*-Designs wurde die Kultivierung der iSZ auf Laminin (Lam) mit der entsprechenden Kultivierung auf der unbeschichteten Oberfläche (uo) verglichen. Da

es sich bei Laminin um ein bereits etabliertes Oberflächenbeschichtungsmaterial handelt, kann dieser Vergleich einen Überblick über mögliche Regenerationsprozesse von Schwannzellen liefern.

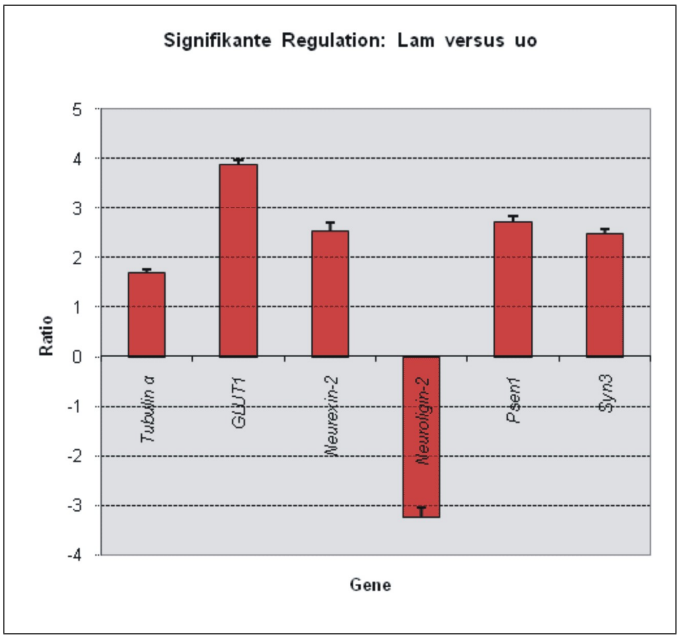


Abbildung 6.26 – *Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (Lam gegen uo) bei einer Irrtumswahrscheinlichkeit $p < 0.001$. Gezeigt ist die Höhe der differentiellen Expression als Quotient.*

In Abbildung 6.26 sind die Gene dargestellt, die bei einer Wahrscheinlichkeit von $p < 0.01$ signifikant um den auf der Ordinate angegebenen Faktor hoch- bzw. runterreguliert sind. Eine positive Regulation bedeutet in diesem Fall, dass die entsprechenden Gene in iSZ bei der Kultivierung auf Lam um den gezeigten Faktor stärker exprimiert sind als in iSZ, die auf einer unbeschichteten Oberfläche gezüchtet wurden.
Die Tauglichkeit von Lam im Einsatz beim *Tissue Engineering* beruht auf der erwiesenermaßen regenerationsfördernden Interaktion dieses ECM-Moleküls mit Integrinen.[206–208] Während dieses Prozesses wird das Laminin aus der Basallamina an das Cytoskelett gebunden. Diese Bindung initiiert schließlich die Ummantelung der Axone durch Schwannzellen.[209]
An diesem Vorgang ist auch das auf Laminin stärker exprimierte Tubulin-α beteiligt, das in Form des acetylierten Tubulin-α beim intrazellulären Transport im mikrotubulären Netzwerk nicht-neuronaler Zellen beteiligt ist.[210]
Auch das ebenfalls hochregulierte Syndecan-3 ist als Bestandteil von perinodalen Prozessen in Rattennerven beteiligt,[211] die wiederum während der Myelinbildung von Schwannzellen eine Rolle spielen.
Dasselbe gilt für die β-Untereinheit von $GABA_A$-Rezeptoren, die bei einer Irrtumswahrscheinlichkeit von 5% (Daten nicht angebenen) auf Laminin hochreguliert ist. Die nachgewiesenermaßen von Schwannzellen exprimierten $GABA_A$-Rezeptoren spielen eine entscheidende Rolle während der Entwicklung und Differenzierung glialer Vorläuferzellen, da selektive Liganden

von $GABA_A$- und $GABA_B$-Rezeptoren die Schwannzellen-Proliferation kontrollieren.[212]
Dong *et al.* zufolge wirkt Neuroligin-2 in Zusammenwirken mit Neurexin-2 und postsynaptischen $GABA_A$-Rezeptoren zentral auf die Zusammensetzung von GABA-Synapsen.[213]
Obwohl dieser Mechanismus in iSZ bisher nicht nachgewiesen werden konnte, wurde eine Expression von Neuroliginen und Neurexinen durch Schwannoma-Zellen von Olivera *et al.* bestätigt.[214] Wie in Abbildung 6.26 zu sehen, zeigt Neurexin-2 eine signifikant höhere Expression auf Lam im Vergleich zur unbeschichteten Oberfläche, wohingegen Neuroligin-2 auf Laminin runterreguliert ist.
Das ebenfalls von Schwannzellen induzierte Glucose-Transporter Protein (*GLUT1*) wirkt durch eine starke perineurale Expression auf die Regeneration von Axonen und ist vermutlich an der Wallerschen Degeneration beteiligt, in dem phänotypische Veränderungen der perineuralen Zellen und Schwannzellen eingeleitet werden.[215,216]
Wie bei den bereits beschriebenen Genen, konnte auch die Expression von Presenilin-1 in Schwannzellen bereits bestätigt werden.[217] Aussagen über den Einfluss von Presenilin-1 auf das periphere Nervensystem hingegen bleiben spekulativ.

Colominsäure (*CA*) versus unbeschichtete Oberfläche (*uo*)

Um den Einfluss der Kultivierung von iSZ auf CA zu untersuchen, wurde die RNA kultivierter iSZ auf einer CA-Beschichtung auf *low-density*-Chips mit der RNA von iSZ in Kultivierung auf unbeschichteten Oberflächen verglichen. Dieser direkte Vergleich zeigt eine signifikant höhere Expression für 13 Gene bei der CA-Kultivierung. Wie bereits bei Lam gezeigt, wirkt sich die Kultivierung der iSZ auf der unbeschichtete Oberfläche nicht signifikant auf die Expression der iSZ aus.

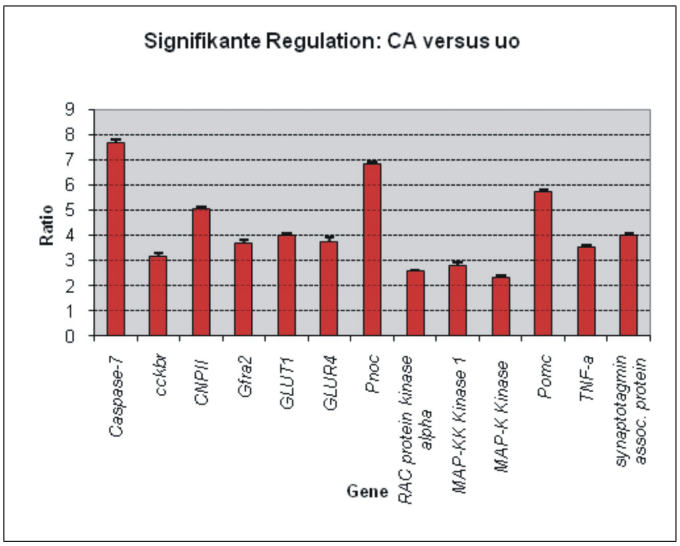


Abbildung 6.27 – *Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (CA gegen uo) bei einer Irrtumswahrscheinlichkeit $p < 0.001$. Gezeigt ist die Höhe der differentiellen Expression als Quotient.*

In der Literatur vielfach beschriebene Charakteristiken der PSA wurden herangezogen, um den erwarteten Einfluss des spezifischen PSA-Derivats CA auf kultivierte iSZ herleiten zu können. Bezüglich der Schwannzellen ist vor allem die Anbindung der großen negativen Ladung von PSA an das NCAM von besonderer Bedeutung. Die in der Literatur beschriebene signifikante Expression von NCAM auf axonalen und Schwannzell-Membranen konnte auch nach Kultivierung auf allen drei untersuchten Oberflächen (Lam, CA, uo) beobachtet werden (eine absolute Expression auf hohem Niveau konnte zwar beobachtet werden, da es sich dabei jedoch um keine relative differentielle Expression zwischen den Zuständen handelt, ist dieses Gen nicht in den vergleichenden Abbildungen 6.26 bis 6.28 wiederzufinden).[218] Das integrale Membranprotein NCAM wirkt als temporärer Modulator bei einer Vielzahl von Zell-Zell-Interaktionen durch homophile und heterophile Bindemechanismen.[219] Isoformen von NCAM, werden wegen des starken Einflusses von PSA auf NCAM besondere Eigenschaften im Regenerationsprozess von Nerven zugeschrieben.[220] Diese äußern sich zum Teil in Form dynamischer Veränderungen in den Membrankontakten, wodurch die zelluläre und Gewebeplastizität erhöht wird. Außerdem wird eine PSA-vermittelte Regulation der Zell-Zell-Interaktion vermittelt, die sich vor allem auf die Zellmigration und den axonale Auswuchs auswirkt.[221] In Bezug auf therapeutische Maßnahmen ist ein hohes PSA-NCAM-induziertes migratorisches Potential wünschenswert, da die Fähigkeit glialer Zellen in verletztes Gewebe zu migrieren hauptsächlich vom Potential auswachsender Axone abhängt, in die entsprechenden Regionen zu wandern.[222] Diese Migrationsfähigkeit ist vor allem für Schwannzellen von Bedeutung, da diese die Ausbildung regenerierender Axone durch leitende Myelinierung dirigieren. Der proliferative Status der peripheren Myelinbildung durch Schwannzellen kann also durch die Proliferation und Migration prämyelinierender Schwannzellen charakterisiert werden. Diese Begebenheit unterstreicht die Bedeutung der Migration vor allem durch Schwannzellen während der peripheren Regeneration.
Der G-Protein gekoppelte cholecystokinin-B-Rezeptor (*cckbr*), der erwiesenermaßen durch Schwannoma exprimiert wird und in dem hier vorgestellten Vergleich auf CA hochreguliert ist, weist ein in mehreren Experimenten nachgewiesenes Migrationspotential durch unterschiedliche parakrine regulatorische Pfade auf.[223,224] Innerhalb dieser Pfade stimuliert cckbr die Aktivierung des EGF-Rezeptors und der erB2-Rezeptortyrosin-Kinase einerseits und transaktiviert den MAP-Kinase-Pfad andererseits. In diesem Zusammenhang ist die Expression spezifischer Gene des zuletzt genannten MAP-Kinase-Pfads erwähnenswert. Zu diesen Genen, deren Expression durch Induktion von CA verändert wird, zählt die MAP-Kinase-Kinase-Kinase (MAP-KK Kinase), die MAP-Kinase-Kinase (MAP-K Kinase), die Rac Protein-Kinase (Rac-PKα) und der Tumor-Nekorse-Faktor α (TNF α). Die Aktivierung der MAP-Kinase-Kaskade kann demnach einerseits durch unterschiedliche Stimuli hervorgerufen werden und so zur migratorischen Antwort von cckbr beitragen oder andererseits TNF α ausgelöst werden.
Die MAP-Kinase JNK, eine der drei Klassen der MAP-Kinasen, die direkt durch die MAP-KK Kinase und die MAP-K Kinase durch eine oder mehrere der erwähnten Stimuli aktiviert wird, aktiviert sowohl die Vermehrung und Differenzierung der Zellen als auch deren Onkogenese. Jüngere Studien unterstreichen jedoch die ebenfalls ermittelte Beteiligung des JNK-Pfads an regulatorischen Prozessen der Zellmigration.[225] Diese Verbindung wurde auch bei Schwann-

zellen beobachtet.[226]
Obgleich die Aktivierung des JNK-Pfades die Expression der Caspase-7 eingeleitet und somit ein in apoptotische Vorgänge involviertes Gen aktiviert haben könnte, könnte diese Diskrepanz zu dem oben erwähnten migratorischen Effekt auf regulatorische Effekte zurückgeführt werden. Solche regulatorischen Effekte werden in der Literatur oftmals als notwendig für den Erhalt der Zellzahl, insbesondere bei immortalisierten Zelllinien und Tumorzellen, beschrieben.[227] Darüberhinausgehend wird neben der Expression proaptoptischer Gene wie der Caspase-7, aber auch von TNF α und unterschiedlichen MAP Kinasen, auch das gegenteilig wirkende Gen Rac-PK durch die Kultivierung auf CA induziert. Bei Rac-PK handelt es sich um ein die Apoptose inhibierendes Protein.[228] Dementsprechend wurden neben bekannten Transkriptionsfaktoren, die als Substrate des JNK-Pfades fungieren, jüngst auch einige Cytoskelett-assoziierte Proteine als Substrate des JNK-Pfades identifiziert. Zu diesen Proteinen gehören vor allem Proteine, die mit Mikrotubuli zusammenwirken und bei denen die Beteiligung an der Reorganisation des Cytoskeletts nachgewiesen werden konnte.

Wie bereits bei der Kultivierung auf Lam beobachtet, ist GLUT1 auch auf CA hochreguliert. Wie bereits erwähnt ist GLUT1 unter Einwirkung der Schwannzellen an Regenerationsvorgängen der Axone durch die Wallersche Degeneration und Regeneration beteiligt. Interessanterweise wird für dieses Gen auch eine Induktion der JNK-Aktivierung in retinalen Endothelzellen bei Ratten beschrieben.[229]
Dem Einfluss von PSA auf die Zelladhäsion werden in unterschiedlichen Veröffentlichungen gegensätzliche Beobachtungen zugeschrieben.[203,219] In den im Rahmen dieser Arbeit durchgeführten Studien, konnte ein durch Gfra2 induzierter, potentiell stimulierender Effekt auf die Zellausbreitung beobachtet werden. Die besondere Gewebe-Lokalisation der RNA von Gfra2 sowie die diesem Molekül zugeordnete Fähigkeit, GDNF-Signale zu vermitteln, lässt außerdem eine bedeutende Rolle von Gfra2 in Transduktionsprozessen des neurotrophilen Faktors GDNF sowie positive Effekte auf die Schwannzell-Kultivierung vermuten.
Entsprechend ist auch die hochregulierte 2',3'-cyclische Nukleotid 3'-Phosphodiesterase (*CNPII*) bei der Ausbildung von Mikrotubuli beteiligt und könnte somit die zelluläre Verteilung der Mikrotubuli regulieren. Als Myelinprotein der Cytoskelett-Maschinerie dirigiert CNPII die Ausbreitung von Gliazellen und wird daher oft als Differenzierungsmarker von Schwannzellen benutzt.[230]
Der detaillierte Mechanismus, der den durch PSA induzierten Veränderungen der Zellmorphologie zugrunde liegt, konnte zwar bisher nicht geklärt werden. PSA besitzt jedoch erwiesenermaßen ein erhöhtes Potential zur Förderung des Heranwachsens von Schwannzellen.

Neben den bereits untersuchten Genen, wurde auch eine erhöhte Expression der Rezeptor-Untereinheit GLUR-4 auf CA-behandelten Kulturplatten gefunden. Eine Expression dieses Proteins wurde sowohl in Oligodendrozyten wie auch in Schwannzellen beobachtet. Obgleich eine Funktion dieses Proteins in dieses Zellen bisher nicht beschrieben wurde, ist die Aktivierung der Glutamat-Antwort durch PSA in AMPA-Rezeptoren des Hippocampus bereits

bestätigt worden.[231]

Die Ausbildung von Synapsen wird vor allem durch eine Gruppe von Genen gefördert, zu denen unter anderem das Neuroligin, Neurexin und Synaptotagmin zählen.[230] Zwei dieser Gene, das Neuroligin und das Neurexin sind mit 5%iger Fehlerwahrscheinlichkeit (Daten nicht angegeben) reguliert, während das Synaptotagmin mit nur 1%iger Irrtumswahrscheinlichkeit signifikant höher reguliert ist nach der Kultivierung der iSZ auf CA. Im Gegensatz zu Synaptotagmin, das bisher in Schwannzellen nicht nachgewiesen werden konnte (obgleich in Gliazellen), wurden Neurexine und Neuroligine bereits in Schwannzellen detektiert.[214,232]
Die Hochregulation von Proopiomelanocortin (*POMC*) könnte - wie in Abbildung 6.27 zu sehen - ein Indikator für den Regenerations-beeinflussenden Effekt von PSA auf Schwannzellen sein. Es wurde bereits dokumentiert, dass POMC-Peptide die Nervenregeneration im peripheren (PNS) und zentralen Nervensystem (ZNS) *in vivo* beschleunigen können.[233]
Letztlich konnte auch die Beteiligung von PSA-NCAM auf die Zell-Zell-Interaktion durch Glass *et al.* bestätigt werden, was durch die starke Expression von Prepronociceptin (*Pnoc*) unterstrichen wird, da dieses Gen in Astrocyten an der Zell-Zell-Signalübertragung mitwirkt.[234,235]

Zusammenfassend steht das vergleichende Expressionsprofil von iSC, die auf CA kultiviert wurden, im Vergleich zur Kultivierung auf einer unbehandelten Oberfläche weitestgehend in Einklang mit den in der Literatur beschriebenen, durch PSA hervorgerufenen Veränderungen. PSA bzw. CA beeinflusst demzufolge nachweislich das Wachstum, die Differenzierung sowie die Aptoptose und die Regeneration von Schwannzellen.

Colominsäure (*CA*) versus Laminin (*Lam*)

Um die unterschiedliche Wirkung der beiden Oberfächenbeschichtungen abschätzen zu können, wurden die Genexpressionsprofile der iSZ auf CA und Lam verglichen.

Wie bereits im Vergleich Lam versus uo gesehen, ist das Tubulin-α auch auf Lam im Vergleich zu CA stärker exprimiert.
Bezüglich aller verbleibenden regulierter Gene wurde eine erhöhte Expression ausschließlich auf CA im Vergleich zu Lam detektiert (siehe Abbildung 6.28). Diese Aussage trifft insbesondere für die Gene des JNK-Pfades zu, die für eine erhöhte Migration der Schwannzellen verantwortlich sein könnte (siehe Vergleich CA versus uo). Die Induktion dieser Gene könnte das Gleichgewicht der Rho-Rac-Signalkaskade zugunsten des axonalen Wachstums beeinflussen.[236] Obwohl Caspasen einerseits mit der Induktion der Apoptose und somit mit dem programmierten Zelltod Verbindung gebracht werden, können sie andererseits auch regulatorische Funktionen übernehmen. Demzufolge könnten die Caspasen-3 und -7 das Chaperon Calnexin spalten, dessen Spaltprodukt zur Abschwächung der Apoptose in Mauszellen führt.[237] Caspasen könnten also durch Inhibierung unterschiedlicher in Schwannzellen bereits detektierter

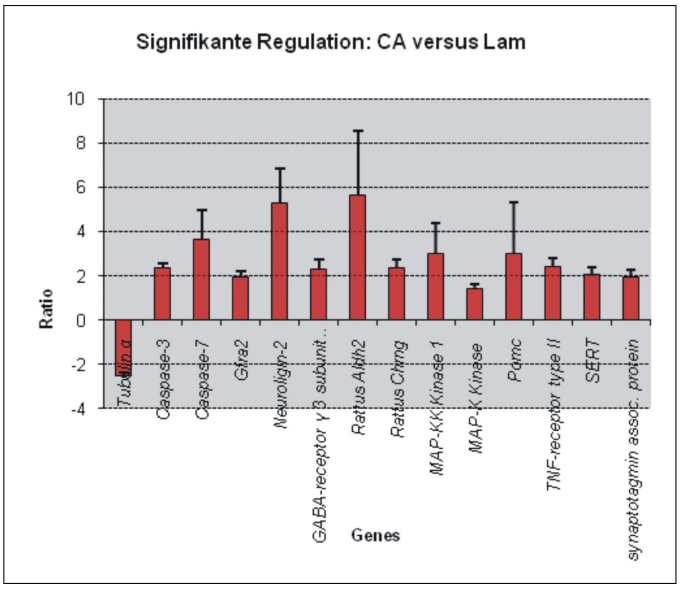


Abbildung 6.28 – *Signifikant regulierte Gene nach der Kultivierung auf verschiedenen Matrices (CA gegen Lam) bei einer Irrtumswahrscheinlichkeit $p < 0.001$. Gezeigt ist die Höhe der differentiellen Expression als Quotient.*

Chaperone zur Regulation der Zellzahl beitragen.
Die durch CA potentiell ausgelösten migratorischen Effekte wurden bereits anhand des (ebenfalls im Vergleich CA versus uo) regulierten Gens GDNFR-β (*Gfra2*) diskutiert. Gleiches gilt für POMC.
Neuroligin-2 wurde jüngst in Co-Expression mit $GABA_A$-Rezeptoren in nicht-neuronalen Zellen gefunden, was den potentiellen Einfluss dieses Gens auf die entsprechenden Rezeptoren hervorhebt.
Ein weiteres Gen, das möglicherweise in die Synapotgenese involviert ist, ist das Chrng, welches die γ-Untereinheit des nicotinoiden Acetylcholin-Rezeptors AChR kodiert, und welches auch bereits in Schwannzellen nachgewiesen werden konnte.[238] AChR wird auch durch den fibroiden Wachstumsfaktor bFGF aktiviert, welches wiederum bekanntermaßen durch PSA-NCAM stimuliert wird und somit die erhöhte Expression auf CA erklären könnte.[232] Neben AChR aktiviert das bFGF aber auch den Serotonintransporter SERT, welcher in Gliazellen exprimiert wird und sowohl im ZNS als auch im PNS bereits gefunden wurde.[200]
Die Funktion des verbleibenden regulierten Gens, der Alkoholdehydrogennase-2 (*Aldh2*) ist unklar. Die Aldh2 könnte einen Beitrag zur Abschwächung des Zelltods in Schwannzellen leisten, da Xu *et al.* der mitochondrialen Aldh2 in Epithelzellen der Lunge eine abschwächende Funktion während der Hyperoxia-induzierten Apoptose durch Aktivierung des MAP-K-Pfades zuschreiben konnten.[239] Da dieses Gen bisher weder in glialen noch in Schwannzellen detektiert worden ist, könnte es für weitere Untersuchungen von besonderem Interesse sein.
Wie auch beim Vergleich zwischen CA und uo, so sind auch im Vergleich von CA zu Lam Gene auf CA hochreguliert, die in den gleichen regulatorischen Pfaden eine große Rolle spielen. Diese Analogie unterstreicht die diskutierte Wirkung von PSA / CA auf die Kultivierung von Schwannzellen.

6.2.6.3 Fazit

Die Validität des *t*-Tests als geeignete Methode zur Ermittlung differentiell exprimierter Daten wurde anhand eines Experiments mit immortalisierten Schwannzellen diskutiert. Das Microarray-Experiment umfasste drei Chips und wurde im *Loop*-Design durchgeführt, so dass ein Vergleich der drei Zustände miteinander möglich war. Die Ergebnisse dieser Versuche wurden anhand der in der Literatur verfügbaren Informationen zu den regulierten Genen detailliert diskutiert. Dabei stellte sich heraus, dass einerseits der aus der Literatur bekannte positive Einfluss von PSA auf die Regeneration von Nervenzellen auch für Schwannzellen bestätigt werden konnten. In diesem Zusammenhang wurde einerseits eine auf dem PSA-Derivat CA erhöhte Expression von solchen Genen detektiert, die bekanntermaßen an dem Regenerationsprozess neuronaler Verletzungen beteiligt sind. Andererseits konnten in diesem Kontext zusätzlich neue Gene gefunden werden, die auf der Basis bekannter Merkmale zu PSA sinnvoll eingeordnet werden konnten.
Anhand dieses Beispiels konnte die Praktikabilität dieses Auswerte-Algorithmus für *low-density*-Microarrays aufgezeigt werden.

Neben dem hier aufgeführten Beispiel eines Microarray-Experiments, das für die aktuelle Forschung auf dem Gebiet des *Tissue Engineerings* große Relevanz besitzt, konnte die Anwendbarkeit des *t*-Tests bei der Detektion regulierter Gene auch anhand eines weiteren selbstgespotteten *low-density*-Microarray-Experiments mit nur 17 Genen gezeigt werden. Die im Rahmen dieser Experimente in Zusammenarbeit mit dem Oststadkrankenhaus entstandenen Veröffentlichungen beschreiben die Regeneration von Rattenzellen nach einem sogenannten Neeling-Verfahren.[239,240]

Der *t*-Tests, der zur Analyse von Microarraydaten bereits als weit verbreitete und effiziente Vorgehensweise etabliert wurde, wurde daher in das Auswerteprogramm integriert und kann bei einer ausreichenden Gruppengröße die Signifikanz der Aussage zu den Ergebnissen erheblich erhören.

6.2.7 Clusteranalyse und regulatorische Pfade

6.2.7.1 Hintergrund und Anforderungen

Die Detektion differentiell exprimierter Gene dient als Grundlage zur Beantwortung der Fragestellung eines Microarray-Experiments. Die Datenanalyse stellt somit vor allem in Bezug auf *low-density*-Microarrays den für den Experimentator entscheidenden Schritt der Datenauswertung dar.
Da *low-density*-Microarrays nur eine begrenzte Anzahl an Genen tragen, kann die Analyse der Ergebnisse der Auswertung ohne weitere Schritte erfolgen. Wenn es sich bei den Microarrays jedoch um *whole-genome*-Chips bzw. Arrays mit größeren Genverbänden handelt, kann eine Interpretation der Ergebnisse alleine auf der Basis der differentiellen Expression sehr um-

fangreich werden. Aus diesem Grunde und um komplexere Zusammenhänge zwischen den Genen aufdecken zu können, werden oftmals Cluster-Techniken eingeführt. Mit Hilfe dieser Cluster-Algorithmen können Experimente auf bestimmte Signaturen oder Muster in der Genexpression untersucht werden. Der Begriff Clustering steht dabei für das Partitionieren der Daten in Teilmengen, den sogenannten Clustern, in denen die Elemente möglichst ähnliche Eigenschaften besitzen.
Im Gegensatz zu Hypothesentests liefern die Ergebnisse der Clusteranalysen jedoch keine Informationen über die Signifikanz der Ergebnisse, sondern werden vielmehr angewandt, um die Daten auf biologische sinnvolle Muster hin zu untersuchen und daraus gegebenenfalls neue Hypothesen zu entwickeln. Biologische Muster können jedoch nur detektiert werden, wenn die Dimensionalität der Microarray-Experimente ausreichend hoch ist, was vor allem für *whole-genome*-Microarrays gilt. Aus diesem Grunde werden Clusteranalysen vorwiegend bei *whole-genome*-Experimenten eingesetzt, um beispielsweise Gene zur anschließenden Entwicklung neuer *low-density*-Microarrays oder anderer biologischer Tests für spezifische Fragestellungen selektieren zu können.

Neben den Clusteranalysen kann auch eine funktionelle Zuordnung der entsprechenden Proteine zu den untersuchten Genen zusätzliche Auskunft über deren Aktivität innerhalb bestimmter regulatorischer Pfade geben. Entsprechende Datenbanken, die funktionelle Informationen zu Genen unterschiedlicher Organismen liefern und mit den eindeutig identifizierbaren *Accession*-Nummern der Gene arbeiten, sind im Internet verfügbar.

Die Implementierung entsprechender Datenbank-Zuweisungen in die Microarray-Auswertung sowie die Korrelation dieser Angaben mit den Ergebnisse aus den Clusteranalysen kann weitereichende neue Informationen bezüglich der funktionellen Position definierter Gene im Genom eines Organismus liefern.

6.2.7.2 Durchführung und Ergebnisse

Eins der Hauptmerkmale von Microarray-Experimenten ist die parallele Messung tausender Genexpressionswerte und die damit einhergehende hohe Dimensionalität. So kann bei beispielsweise 10000 Messwerten für eine Probe jeder Genexpressionswert als Punkt in einem 10000-dimensionalen Raum aufgefasst werden. Clusteranalysen erlauben in der Regel einen entsprechenden visuellen Überblick über die Daten und können dementsprechend zur Klassenerkennung beitragen. So konnten Clusteralgorithmen beispielweise bereits zur Detektion unterschiedlicher Krankheitstypen eingesetzt werden.[241]

Wie bereits erwähnt (siehe Kapitel 5.5.3), existieren zwei grundlegende Typen von Clusterverfahren: die *unsupervised* und die *supervised* Clusteranalyse. Während die *supervised* Clusteranalyse die Verwendung zusätzlicher Informationen für ein Trainings-basiertes Lernen

voraussetzt, ermöglicht das *unsupervised* Clustern die Identifikation und Visualisierung von Gruppenstrukturen im Datensatz ohne vorher verfügbares Wissen von existierenden Gruppen. Aus diesem Grund liegt der Fokus der im Rahmen dieser Arbeit untersuchten Methoden auf den *unsupervised* Clusterverfahren.

Charakteristisch für Genexpressionsdaten ist die Bedeutsamkeit der Sortierkriterien, die es ermöglichen, die Gene und Proben zugleich zu clustern.[242] Gene werden dabei als Elemente behandelt, Proben stellen Zustände oder Eigenschaften dar. Andererseits können Proben in homogene Gruppen unterteilt werden, die bestimmten makroskopischen Phänotypen entsprechen. Die Unterscheidung des Gen-basierten und Proben-basierten Clusterings basiert auf unterschiedlichen Charakteristika der Aufgabenstellung beim Clustern von Genexpressionsdaten. Einige Clusteralgorithmen wie K-means- oder hierarchische Ansätze können sowohl zur Gruppierung von Genen als auch zur Partitionierung von Proben verwendet werden.
Aufgrund des Vorteils der beidseitigen Betrachtung der Daten durch diese methodischen Ansätze werden häufig K-means- oder hierarchische Methoden zur Analyse von Genexpressionswerten verwendet.
Hierarchische Clustermethoden liefern eine Anzahl ineinander verschachtelter Klassen zurück, deren Zusammengehörigkeit in einem Baum, dem sogenannten Dendogramm, dargestellt ist. Nichthierarchische Clustermethoden wie das K-means-Clustering teilen Objekte in verschiedene Klassen ein, ohne vorab Beziehungen zwischen den Klassen anzunehmen.
Da hierarchische Methoden erstens längere Laufzeiten aufweisen und dies bei *whole-genome*-Daten ins Gewicht fällt und da die Objektzugehörigkeit bei dieser Methode zweitens irreversibel ist, was zu einer Fehlerfortpflanzung führt, wurde lediglich das K-means-Clustern als Clusteralgorithmus in das Programm aufgenommen.[243–245]

Wie bei allen anderen Clustermethoden beruht auch das K-means-Clustern auf der Berechnung von Distanzen zwischen jeweils zwei Objekten. Für Microarray-Daten wird hierzu die Distanz zwischen den Genexpressionsvektoren berechnet. In Abhängigkeit davon, ob Gene oder Proben geclustert werden sollen, werden Gen- oder Probenvektoren zur Distanzbestimmung herangezogen. Zu den metrischen paarweisen Distanzen zählen z.B. die

- Euklidische Metrik

$$d_{euc}(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{6.25}$$

- Manhatten-Metrik

$$d_{man}(\vec{x},\vec{y}) = \sum_{i=1}^{m}|x_i - y_i| \tag{6.26}$$

Daneben existieren eine Reihe weiterer korrelationsbasierter Distanz-Methoden, die ebenfalls auf Microarray-Daten angewendet werden können, u.a. die:

- Pearson-Korrelation

$$d_{pear}(\vec{x},\vec{y}) = 1 - \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2 \sum_{i=1}^{m}(y_i-\bar{y})^2}} \tag{6.27}$$

- Spearman-Korrelation

$$d_{spear}(\vec{x},\vec{y}) = 1 - \frac{\sum_{i=1}^{m}(x'_i-\bar{x}')(y'_i-\bar{y}')}{\sqrt{\sum_{i=1}^{m}(x'_i-\bar{x}')^2 \sum_{i=1}^{m}(y'_i-\bar{y}')^2}} \tag{6.28}$$

mit $x'_i = rank(x_i)$ als Beschreibung des Spearman's rank Korrelationskoeffizienten.

Das in das Auswerteprogramm integrierte K-means-Clustering kann mit unterschiedlichen Distanz-Maßen berechnet werden. Der Experimentator erhält über eine Benutzer-Oberfläche die Option zur freien Auswahl des Distanz-Maßes sowie zur Veränderung weiterer Parameter (siehe Abbildung 6.29 und 6.30). Eine kurze Beschreibung der jeweiligen Einzeloptionen kann über den linken Mausklick angewählt werden (siehe Abbildung 6.30).

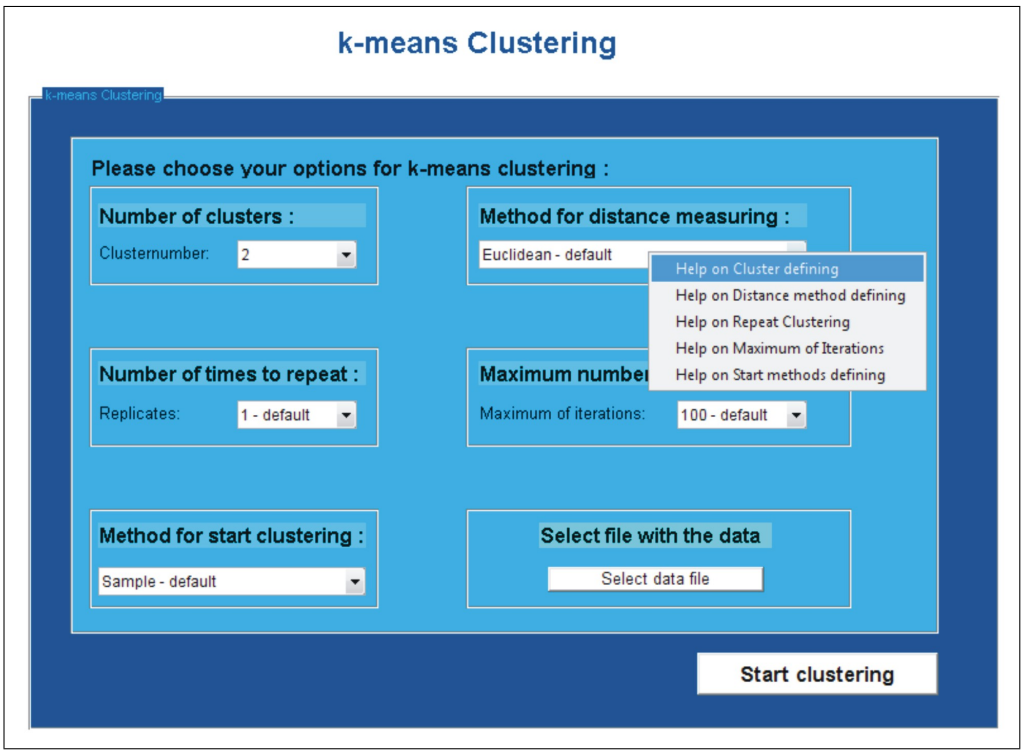


Abbildung 6.29 – *Benutzeroberfläche zum Cluster der Microarray-Ergebnisse. Eine Hilfe zu den einzelnen Optionen ist über die linke Maustaste erhältlich (siehe Abbildung 6.30)*

Unter Anwendung eines der optionalen Distanzmaße kann dann das K-mean-Clustering gestartet werden. Dabei handelt es sich - wie bereits erwähnt - um ein Partitions-basierte

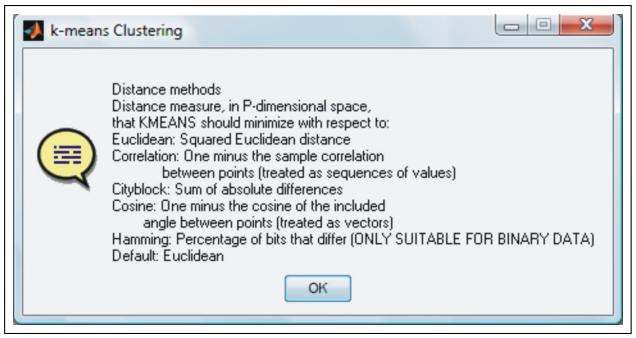


Abbildung 6.30 *– Hilfe zu einer Distanz-Maß-Option beim Clustern. Mit der linken Maustaste kann zunächst die entsprechende Hilfe angewählt werden (siehe Abbildung 6.29). Entsprechend der Auswahl öffnet sich dann ein erklärendes Hilfefenster.*

Clusterverfahren, das die Daten zunächst unabhängigen, untereinander nicht verzweigten Clustern zuteilt. Vor Abbruch des Clustering werden iterativ hunderte Cyclen durchlaufen. Die Anzahl der Cyclen wiederum können optional variiert werden (siehe Abbildung 6.29). Je geringer die Anzahl der Cyclen, umso schneller stehen die Ergebnisse zur Verfügung. Da das K-means-Clustern wie andere iterative Verfahren jedoch auch die Gefahr birgt, in lokalen Minima zu enden, bietet sich eine möglichst hohe Anzahl an Wiederholungen an.

In Abhängigkeit der vom Benutzer gewählten Anzahl Cluster (k), beginnt der Cluster-Prozess mit dem Anlegen k zufälliger Vektoren in sogenannte Cluster-Centroiden c_i. Die Objekte werden dem am nächsten gelegenen Cluster-Centroid zugeordnet und liegen somit letztlich partitioniert in k separate Cluster vor.[246] Anschließend werden die Cluster-Centroiden neu berechnet, in dem der Mean oder Median aller zu einem Cluster gehörigen Objekte ermittelt wird. Diesen Cluster-Centroiden werden nun erneut die Objekte entsprechend ihrer räumlichen Nähe zugewiesen. Dieser Vorgang wiederholt sich so lange bis der vorgegebene Grenzwert von Iterationen erreicht ist.[247] Insgesamt handelt es sich bei der K-means-Clustering also um ein iteratives Minimierungsverfahren gemäß der folgenden Gleichung:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} d(x_j c_i) \tag{6.29}$$

mit der Matrix u_{ij}, bestehend aus Werten zwischen 0 und 1 in Abhängigkeit davon, ob es sich bei dem Objekt j um ein Mitglied des Clusters i handelt. Wie oben beschrieben minimiert diese Technik die Distanz der zu ihren Centroiden gehörigen Objekte.

Trotz dem es sich beim K-means-Clustern um einen mathematisch einfachen Ansatz handelt, vermag er viele mit dem hierarchischen Clustern verbundenen Probleme zu lösen.[244] Die Einschränkung des K-means-Clusterns hingegen, dass die Anzahl der Cluster vorab durch den Benutzer festgelegt werden muss, kann durch wiederholtes Durchlaufen des Prozesses mit unterschiedlichen Clustervorgaben zur einer Optimierung der Parameter führen und somit aufgehoben werden.

Ein Beispiel einer Clusteranalyse anhand eines *whole-genome*-Microarrays des Bakteriums

E.coli soll die Verwendbarkeit des *K*-means-Clusterns zur sinnvollen Gruppierung des Datensatzes verdeutlichen. Um die Validität dieser Methode überprüfen zu können, können die Gene ebenfalls ihren Funktionen in regulatorischen Pfaden gemäß den Angaben aus der *KEGG*- und gegebenenfalls weiteren Datenbanken zugeordnet werden.[101] Während die *KEGG*-Datenbank universelle Informationen zu einer Vielzahl an Organismen enthält und daher auch für die Analyse von Human- und Rattenchips herangezogen wurde, enthalten andere Datenbanken wie beispielsweise *Multifun* oder *SNAP* rein *E.coli*-spezifische Informationen. Für eine detaillierte Untersuchung der Microarrays dieses Bakteriums wurden daher die Informationen aus diesen drei Datenbanken (*KEGG*, *Multifun* und *SNAP*) ausgewertet. Eine solche Datenbank-basierte funktionelle Analyse der Daten kann unabhängig von der Dimension des Arrays (*whole-genome* bzw. *low-density*) mit all den Microarrays realisiert werden, für die entsprechende Informationen in anerkannten Datenbanken (w.z.B. *KEGG*) vorhanden sind.

Um dem Benutzer die Interpretation der Ergebnisse zu vereinfachen, werden in die erste Spalte die Gen-Namen entsprechend der auf dem Microarray verwendeten GeneID geschrieben, die darauf folgenden Spalten enthalten die Unterebenen der Regulationspfade der entsprechenden Datenbanken.

Ein zusätzliches Programm ermöglicht das separate Einlesen aller regulierten Gene, so dass optional automatisch zusätzliche Dateien angelegt werden können, in denen einerseits nur die Datenbankinformationen aller regulierten bzw. aller hoch- oder runterregulierten Gene gespeichert werden. Diese Zusatzdateien ermöglichen eine nachträgliche Sortierung entsprechend der Pfade, so dass ein möglicher Zusammenhang zwischen regulierten Genen und deren gemeinsamen Auftreten innerhalb bestimmter Pfade einfacher detektiert werden kann.

In dem hier gezeigten Beispiel wurden Kultivierungen mit einem Arabinose-defizienten *E.coli*-Stamm durchgeführt, denen ab einem bestimmten Zeitpunkt *L*-Arabinose appliziert wurde. In dem als *Loop*-Design angelegten Microarray-Experiment wurden das Genom des Bakteriums vor der Induktion mit unterschiedlichen Zeitpunkten nach der Induktion verglichen.

Die Ergebnisse der Auswertung wurden für weitere Clusteranalysen verwendet. Zu diesem Zweck wurde eine in der Ergebnisdatei mitgelieferte Zusammenfassung die in der folgenden Liste aufgeführten zehn Vergleiche zwischen den Zeitpunkten herangezogen, um eine Untermenge an Genen für die Clusteranalyse zu selektieren:

- 10 Minuten (-10) vor Induktion gegen Zeitpunkt der Induktion (0 Minuten)
- 2 Minuten nach Induktion gegen 0 Minuten
- 5 Minuten nach Induktion gegen 0 Minuten
- 10 Minuten nach Induktion gegen 0 Minuten

- 30 Minuten nach Induktion gegen 0 Minuten
- 45 Minuten nach Induktion gegen 0 Minuten
- 60 Minuten nach Induktion gegen 0 Minuten
- 2 Minuten nach Induktion gegen 10 Minuten vor Induktion
- 30 Minuten nach Induktion gegen 10 Minuten nach Induktion
- 45 Minuten nach Induktion gegen 10 Minuten nach Induktion

Da diese Microarray-Versuche für die gezielte Selektion von Genen vorgesehen sind, die eine bedeutende Rolle im Verlauf des Experiments spielen, wurden anhand der Tabelle eine Auswahl an weiter Genen getroffen, die weiter untersucht werden sollte. Zu diesem Zweck wurden die Genexpressionswerte derjenigen Gene, die in mindestens vier der zehn Vergleiche reguliert waren, an die Clusteranalyse übergeben. Dabei handelte es sich um 207 (4.49%) der insgesamt 4608 Gene des *E.coli*-Genoms.
Das Clustering wurde mit den folgenden Parameter durchgeführt:

- $N_{Cluster}$: **5**
- $Distanzmethode$: **Korrelation**
- $N_{Cluster-Wiederholungen}$: **50**
- $N_{Iterationsschritte}$: **1500**

Mit diesen benutzerabhängigen Eingaben wurden fünf Cluster unterschiedlicher Größe generiert. Die entsprechenden Gene wurden einer Datenbank-Analyse unterzogen, so dass letztlich für jedes der 209 Gene einerseits eine Cluster- und andererseits eine Multifun-Zuordnung vorlag.
Anhand dieser Informationen sowie unter Einbeziehung der KEGG-Datenbank und einer umfassender Literaturrecherche werden im Folgenden selektiv Gengruppen aus diesen Cluster analysiert.

In Anhang K sind die Gene sowie die ersten drei der fünf Multifun-Ebenen des ersten der fünf Cluster aufgelistet (Eine komplette Liste inklusive aller Multifun-Level sowie der Nähe der einzelnen Gene zum Centroiden werden in jeder Ergebnisausgabe mitgeliefert. Der Übersicht halber wurde hier jedoch auf die Anzeige der Ergebnisse in vollem Umfang verzichtet).

Cluster 1

Infolge der Induktion des Arabinospromotors *L*-Arabinose varriert der *E.coli*-Stamm die Metabolisierung dieser alternativen Kohlenstoffquelle. Wie zu erwarten kann eine erhöhte Expression aller direkt an diesem Vorgang beteiligten Gene beobachtet werden. Dabei handelt

sich um die Gene araE, araF, araG, araH und araJ, die alle innerhalb eines Clusters wiederzufinden sind. Die dahingegen nicht signifikante Expression der ebenfalls auf dem Microarray befindlichen Gene araA, araB und araD, die ebenfalls in den primären Arabinose-Stoffwechsel involviert sind, kann durch die spezifische Stamm-Entwicklung erklärt werden: Aufgrund der Arabinosedegradation-fördernden Eigenschaften des araBAD-Promotors wurden die entsprechenden Gene während der Entwicklung des *E.coli*-Stammes deletiert.[248] Im Gegensatz zu der degradierenden Wirkung dieses Promotors konnten substantielle Unterschiede in der Funktionen zu den Gene des araFGH-Operon nachgewiesen werden.[249] So kodiert der araFGH-Promotor für Proteine, die in den hoch-affinen *L*-Arabinose-Transport involviert sind und somit die Verstoffwechslung des Induktors fördern. Wie aus den Multifun-Daten ersichtlich gehört dieses Transportersystem zur ATP-binding Cassette (ABC) Superfamilie (ABC-Transporter), deren Elemente eine hohe Substrat-Spezifität aufweisen und mit entsprechend hoher Affinität transportieren. Ursache hierfür ist der Transport unter ATP-Hydrolyse und weiterhin die Expression von spezifischen Bindeproteinen im Periplasma. Diese Proteine binden das Substrat und befördern es zum Transporter, so dass Substrate auch gegen den Konzentrationsgradienten transportiert werden können, beispielsweise unter Kohlenhydratmangel.[250] Für die Gene araE und araJ wird ebenfalls eine Beteiligung am Arabinose-Transport angenommen.[251,252]
Während sich der araFGH-Promotor strukturell stark vom araBAD abhebt, konnten Hendrickson *et al.* eine große strukturelle Ähnlichkeit zum galP1-Promotor nachweisen.[249] Dieser Promotor umfasst die Gene galT, galE und galK,[253] von denen die ersten beiden ebenfalls innerhalb dieses Clusters gefunden wurden. Zusätzlich wurde das Gen galU gefunden, welches innerhalb des Galactose-Pfades benachbart zu galT zu finden ist.

Ein weiteres ABC-Transporter-System, das bei Glucose-Mangel verstärkt exprimiert wird, ist der Galactose-ABC-Transport mglABC, kodiert durch die Gene mglA, mglB und mglC.[254] Dieses Transportsystem befindet sich also in einem Cluster mit denen den galp1-Promotor kodierenden Gene galT und galE, das wiederum eine strukturelle Ähnlichkeit zum hochaffinen ABC-Transporter aus den araFGH-Genen aufweist.
Eine strukturelle Ähnlichkeit zu den Proteinen des Arabinose-Transporter-Systems konnte auch für das Protein eines weiteren ABC-Transporters gefunden werden, das glpT. Die strukturelle Ähnlichkeit entsteht durch ein gemeinsames 12-Membran-umspannende α-Helix-Segment.[255]

Zu der Superfamilie der ABC-Transporter zählen auch die uncharakterisierten Mitglieder ytfR und ytfQ.[256] ytfR gehört vermutlich zu den ATP-bindenden Komponenten, während ytfQ voraussichtlich das zu transportierende Protein bindet. Basierend auf Sequenzähnlichkeiten wurde eine Funktion als ATP-abhängiger Zuckertransporter vorgeschlagen. Eine Expression dieser Gene wird ebenfalls unter Glucoselimitierung beobachtet.

Das Gen nuoB kodiert für eine Untereinheit der NADH-Dehydrogenase I (*NDHI*) und

spielt eine Rolle bei der Protonentranslokation der NDHI. Die detaillierte Untersuchung durch Lemming *et al.* ergab eine erhöhte Expression dieses Gens in *E.coli* nach der Zugabe von Arabinose,[257] was ebenfalls für das Gen fkpA gilt.[258] Wie die Mehrzahl der Proteine zu den bereits untersuchten Gene, ist das Genprodukt von nuoB also auch Membran-ständig lokalisiert und spielt - wie aus den Multifundaten ersichtlich - eine Rolle beim Transport über die Membran.

Das erste Cluster umfasst also vor allem Gene, die für Proteine des membran-ständigen Transportmechanismus codieren. Dazu zählen auch eine Reihe von Proteinen, die am primären Arabinose-Stoffwechsel beteiligt sind. Eine verstärkte Expression der entsprechenden Gene wird vermutlich durch die Induktion dieses Zuckers eingeleitet.

Cluster 2

Die Aminosäureverfügbarkeit spielt für die Zelle eine entscheidende Rolle. Der Aminosäuremangel gilt als einer der entscheidenden Auslöser der stringenten Kontrolle.[259] Eine der größten Herausforderungen der Zellbiologie in diesem Zusammenhang stellt das Verständnis der Regulation der intern verbundenen metabolischen Pfade dar. Die Kenntnis solcher Interaktionen dient dazu, einen Überblick über die dabei entstandenen zellulären Produkte zu liefern, die in Folge einer veränderten Expression von Metaboliten oder Enzymen entstehen.
In *E.coli* sind die Arginin- und Pyrimidin-Biosynthese-Pfade miteinander eng durch einen Metaboliten verbunden, das Carbamoylphosphats, das die Synthese der entsprechenden Pfade reguliert.[260] Das carAB-Operon wirkt dabei regulierend auf die Synthese des Carbamoylphosphats, weshalb das hier untersuchte carA als Teil dieses Operons ebenso wie das argI aktiv an der Arginin-Biosynthese beteiligt ist.[261,262]
Darüberhinausgehend spielt das carA eine Rolle bei der Pyrimidin-Synthese, die ebenfalls in die Aminosäure-Synthese involviert ist.[263] Auch das in diesem Cluster befindliche Gen codA kodiert für ein Protein der Pyrimidin-Synthese. Die Expression dieses Gens wiederum wird durch das Genprodukt von purR reguliert, das ebenfalls nahe des Centroiden dieses Clusters liegt.[264] Der Regulator purR reguliert auch ein weiteres Gen der Pyrimidin-Synthese dieses Clusters, das pyrC,[265] das wiederum über ein Zwischenprodukt im Biosynthesepfad mit carA verbunden ist. Die Beteiligung dieser Gene an der Pyrimidinsynthese wird auch durch die Multifun-Angaben dokumentiert. Über die vierte Ebene diese Multifun-Angaben wird auch die Funktion von purR in der Histidin- und Glycin-Synthese deutlich.
Innerhalb des zweiten Clusters befinden sich weitere Gene der Biosynthese von Aminosäuren, die teilweise in gleichen Pfaden synthetisiert werden. So läuft die Synthese von Arginin (hier durch das Gen argI repräsentiert) beispielsweise zunächst parallel zur Synthese Histidine, in welche das hier gefundene Gen hidF erwiesenermaßen involviert ist.[266] Ähnlich verzweigt sich auch der Cystein-Pfad in zwei Teile, die einerseits in der Cystein- und andererseits über das Gen metB in der Methionin-Synthese resultieren.[266] In diesem Kontext spielen vermutlich auch die Gene cysP und CysU eine Rolle, die Cystein-auxotrophe *E.coli*-Stämme komplementieren und als Transporter Aminosäuren über die Zellmembran befördern.[267] Auch die

Gene livF und lifH sind aktiv in den membranständigen Transport von Aminosäuren eingegliedert.[268]
Entscheidende Gene für die Biosynthese von Tryptophan und Phenylalanin wiederum sind das trpA, das trpB, das trpD und das trpE, die ebenfalls in diesem Cluster zu finden sind.[269]

Parallel zur Biosynthese von Aminosäuren konnten auch Gene der Fettsäure-Synthese (aceE, fabA und fadR) gefunden werden.[270]

Insgesamt umfasst das zweite Cluster weitestgehend die Gene der Biosynthese von Aminosäuren und zeigt somit eine große funktionelle Ähnlichkeit zwischen den untersuchten Genen dieser Einheit.

Cluster 3

Das dritte Cluster ist das kleinste der fünf Cluster. Innerhalb dieses Clusters sind vor allem Gene zu finden, die im zeitlichen Verlauf nach Induktion der Arabinose zunächst eine verminderte Expression und schließlich wieder eine leicht erhöhte Expression erfahren. Die Expressionswerte des Clusters sind in Abbildung 6.31 beispielhaft gezeigt.

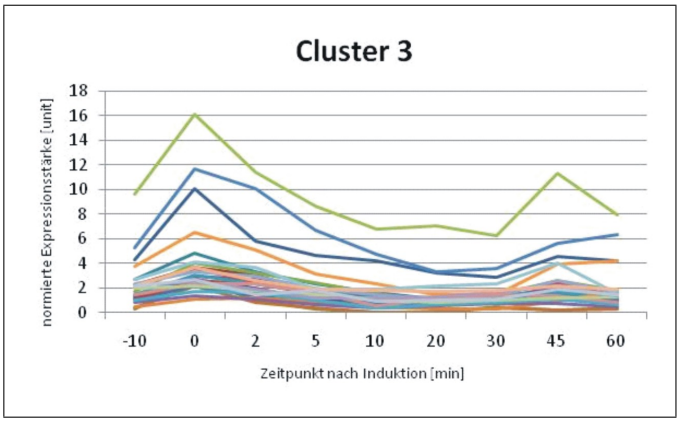


Abbildung 6.31 – *Clusterbeispiel anhand des dritten Clusters. Die Abbildung zeigt den Expressionsverlauf der regulierten Gene dieses Clusters im zeitlichen Verlauf relativ zum Induktionszeitpunkt.*

Eins der dargestellten Gene ist das crp, dessen Genprodukt, CRP, als Regulator Katabolit-sensitiver Operone wirkt, die unter anderem eine entscheidende Rolle bei der Regulation des *L*-Arabinose Operons araBAD spielen.[271] Dieses Operon wurde - wie oben erwähnt - in dem hier vorliegenden *E.coli*-Stamm deletiert, weswegen eine Induktion mit Arabinose keine erhöhten Expression dieser Gene einzuleiten vermag, so dass es zu keiner enzymatischen Degradation der Arabinose durch diese Gene kommt.
Die Induktion mit Arabinose als alternative Zuckerquelle initiiert also eine Runterregulation des Regulons CRP, das an der Bildung hochaffiner Glucosetransportsysteme und der Verstoffwechslung von Acetyl-CoA beteiligt ist. CRP reguliert auch das Gen yhfA dieses Clusters, dessen Funktion noch nicht geklärt werden konnte. Außerdem hat CRP Einfluss auf das ubiG, das sich in einem Pfad mit den Genen ubiB und ubiC dieses Clusters befindet. Diese Gene

wiederum sind Bausteine der Ubichinon-Synthese und wirken somit als Elektronenüberträger der Atmungskette. In einem benachbarten, alternativen Pfad der Biosynthese von Ubichonon befindet sich auch das Genprodukt des nuoH.[272]

Der Rückgang im Glucose-Abbau, der sich durch CRP zeigt, wird auch anhand des Gens mhpR deutlich, dessen Genprodukt den Abbau von Fettsäuren zur Bildung von Acetyl-CoA vorsieht.[273] Gleiches gilt für pfjK und mdh, beides Gene, die unmittelbar die Glucose-Verstoffwechslung beeinflussen. Das Gen mdh wird dabei durch acrA reguliert, welches ebenfalls innerhalb dieses Clusters gefunden wurde.

Im Gegensatz zum ersten Cluster, in dem vermutlich Gene enthalten sind, die durch *L*-Arabinose induziert werden und dadurch die Energieversorgung der Zellen einleiten, beinhaltet das dritte Cluster die Gene, die durch die *L*-Arabinose-Zugabe reprimiert werden, da *L*-Arabinose als alternative Energiequelle vorliegt. Die leicht erhöhte Expression gegen Ende der Kultivierung könnte mit der finalen Verstoffwechslung des Induktors zusammenhängen, dessen Metaboliten schließlich als Substrat für die entsprechenden Cyclen vorliegen

Cluster 4

In dem vierten der fünf Cluster befinden sich unter anderem Gene, die als essentielle Bestandteile der Biosynthese aller Aminosäuren fungieren, in dem sie an Translationsprozessen beteiligt sind. Dazu zählen vor allem Gene, die ribosomale Untereinheiten kodieren. So wirken die Genprodukte von rplA, rplD und rplQ als 50S ribosomale Untereinheiten und die Genprodukte von rpsD und rpsN als 30S ribosomale Untereinheiten am Aufbau der Ribosome mit.[274–278]
Als Gene, die für Ribosomen-assoziierte Elongationsfaktoren kodieren, sind in diesem Zusammenhang auch tsf und fusA zu nennen.[279,280] Vermutlich ist auch das Protein von ibpA an diesem Prozess beteiligt, da es während der Biosynthese als kleines Hitzeschockprotein an aggregierte Proteine bindet.[281]

Neben dieser Gruppe an Genen konnten auch Gene der Leucin- und Isoleucin-Synthese in diesem Cluster detektiert werden. Als globaler Regulator dieser Gene wirkt dabei das Genprodukt von lrp (LRP). Die Expression von LRP selber wird teilweise durch die Nährstoffe reguliert und ist in Minimalmedien mit alternativen Carbon-Quellen (wie in diesem Versuch) erwiesenermaßen erhöht.[282]
Im Allgemeinen wird durch LRP die Expression von Genprodukten, die in biosynthetischen Wegen wirksam sind, stimuliert. Genprodukte, die katabolisch wirksam sind, werden in der Regel entsprechend negativ reguliert. Durch LRP werden in *E. coli* also eine große Zahl von Genen und Operons reguliert, die eine zentrale Rolle in der Aminosäurebiosynthese und dem Aminosäurekatabolismus spielen. In einigen Fällen potenziert sich die Wirkung von LRP durch

Zugabe von *L*-Leucin.
Die Regulation des Operons leuPABCD durch LRP, das bei Synthese von Leucin beteiligt ist, beispielweise verläuft dabei nach einem transkriptionellen Attenuationsmechanismus.[283] Aber auch die Expression von ilvM, welches letztlich in der Synthese von Isoleucin resultiert, wird durch LRP beeinflusst.[284] Spezifischen LRP-Homologen wurde ebenfalls eine bedeutende Aufgabe während der Aminosäure-Synthese von unter anderem Prolin nachgewiesen, was die erhöhte Expression von prlA erklären könnte, welches wiederum an der Prolinsynthese beteiligt ist.[285]

Innerhalb dieses Clusters befinden sich auch die Gene atpA und atpG, die auf einem Operon für eine Membran-gebundene ATP-Synthase lokalisiert und die teilweise aneinander gekoppelt sind.[286] Eine Wachstums-bedingte parallele Expression dieser Gene im *Corynebacterium glutamicum* mit Genen für ribosomale Untereinheiten und Elongationsfaktoren wurde auch von Bendt *et al.* gefunden.[287] Die Kopplung von Untereinheiten der ATP-Synthase scheint insgesamt die Anbindung neu-synthetisierter Ribosomen an die atpA translationale Initiationsregion zu erleichtern.[288]
Denkbar wäre auch eine Initiierung der Expression von atpA und atpG in Zusammenhang mit einer verstärkten Expression von hslV.[289]

Insgesamt umfasst dieses Cluster eine Reihe von Genen, die in Wachstumsprozesse involviert sind. Dieser Aspekt wird durch die gemeinsame Gruppierung ribosomaler Gene sowie einiger damit in Zusammenhang stehender Elongationsfaktoren unterstrichen. Wie bereits in anderen Clustern gesehen, wurden auch nahe dieses Centroiden einige Gengruppen gefunden, die einem gemeinsamen Expressionspfad angehören. Auch in diesem Cluster konnte demnach eine sehr sinnvolle Anordnung von Genen nachgewiesen werden.

Cluster 5

Das letzte Cluster stellt die größte der fünf Gruppen dar. Dieses Cluster weisst gleichzeitig die größte Heterogenität auf. Mit 13 unbekannten Genen beinhaltet es die meisten mit der Datenbank *Multifun* nicht zuzuordnenden Gene. Der zeitliche Verlauf der Expression der Gene dieses Clusters ist in Abbildung 6.32 gezeigt.

Die in diesem Cluster befindlichen Gene kodieren unter anderem für Proteine, die an einer Stress-abhängigen Antwort der Zellen teilhaben. Dazu gehört beispielsweise das Gen yfhJ, welches am Aufbau von Chaperonen beteiligt ist, sowie das sulA, das in der SOS-Antwort der Zellen eine Rolle spielt.[290,291] Auch sspB wird stärker exprimiert, wenn die Zellen erhöhten Stress ausgesetzt sind, da das entsprechende Genprodukt falsch gefaltete Proteine für den anschließenden Abbau durch Proteasen markiert.[292] Auf der Membranebene hingegen vermag das exprimierte Gen proX ungünstige Bedingungen in Form einer hohen Osmolarität auszugleichen.[293]
Das Gen nfnB wird durch den Regulator marA aktiviert, welches als globaler Regulator auf

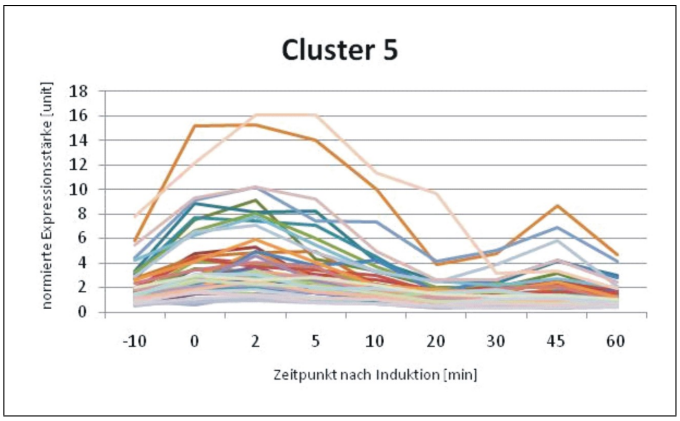


Abbildung 6.32 – *Clusterbeispiel anhand des fünften Clusters. Die Abbildung zeigt den Expressionsverlauf der regulierten Gene dieses Clusters im zeitlichen Verlauf relativ zum Induktionszeitpunkt.*

unterschiedliche Stresssituationen reagiert.[294] Zu solchen Stresssituationen zählen auch beispielsweise erhöhte Temperaturen, auf die auch das Gen htrA, das für eine Serin-Protease kodiert, mit verstärkter Expression reagiert.[295]

Eine weitere ubiquitäre Antwort auf Stress besteht in der vermehrten Produktion von Guanosine Tetraphosphat (ppGpp), einem niedermolekularen Effektormolekül, das in hohem Masse akkumuliert. Das für die Synthese dieses Moleküls zustände Protein wird durch das Gen relA kodiert, welches in diesem Cluster vorzufinden ist.[296] Tedin *et al.* konnten nachweisen, das RelA wiederum Genprodukt von ilvI aktiviert, welches sich ebenfalls innerhalb dieses Clusters wiederzufinden ist.[296]

Andere Gene innerhalb dieses Clusters wiederum sind an Glycosilierungsprozessen beteiligt. So bewirken die Genprodukte von galfF, galP und gatR die Synthese von Galactose bzw. dem reduzierten Produkt von Galactose: Galactitose.[297,298]
Die Gene evgA und ftsZ spielen eine Rolle bei der Signaltransduktion, die auch im Rahmen einer Stressantwort eine Rolle spielen könnte. Sie gehören zu den so genannten Zwei-Komponenten-Systemen und befinden sich vermutlich im selben regulatorischen Pfad wie das bereits erwähnte htrA.[299]

Eine weitere Gruppe an Genen ist an der Synthese von Coenzymen, prosthetischen Gruppen und Cofaktoren beteiligt. Dabei handelt es sich um die Gene dfp, metc, mfd und mog.[300–303]

Die Gene innerhalb dieses letzten Clusters spiegeln die größte funktionelle Heterogenität im Vergleich aller Cluster wider. Eine große funktionelle Übereinstimmung wurde bezüglich der Reaktion einzelner Gene auf Stress gefunden.

6.2.7.3 Fazit

Clusteranalysen spielen eine wichtige Rolle bei der Interpretation größerer Genverbände, in dem funktionelle Analysen die Detektion von Genen mit ähnlichen regulatorischen Aufgaben ermöglichen. Anhand des hier gezeigten Beispiels einer *E.coli*-Kultivierung konnte die funktionelle Einteilung der Gene in ähnliche funktionelle Untereinheiten unter Zuhilfenahme von Literaturnachweisen bestätigt werden.
Clusteranalysetechniken stellen also sehr wertvole Werkzeuge bei der Interpretation von Microarray-Experimenten dar. Dennoch müssen dioe Ergebnisse solcher Clusteralgorithmen immer unter Vorbehalt betrachtet werden, denn obwohl die Methoden reproduzierbar sind, weisen sie noch immer subjektive Charakteristiken auf, da aus der Vielzahl unterschiedlicher Algorithmen mit teilweise sehr unterschiedlichen Ergebnissen ausgewählt werden muss, denen wiederum unterschiedliche Normalisierungen oder Differenzmetriken zugrunde liegen. Darüberhinausgehend entstehen durch das Cluster Gruppen von Genen, die biologische nicht verwandt sind, da immer eine Einteilung aller Gene ohne Ausnahme in unterschiedliche Cluster vorgenommen wird. Die Herausforderung von Clusteranalysen besteht also in der gezielten Interpretation selektiver Gene, deren Clusterergebnisse aufgrund intensiver Literaturrecherche plausibel erscheint. Erst der detaillierte Nachweis einzelner Cluster mit Hilfe selektiver Folgeversuche kann Gewissheit über funktionelle Zusammenhänge von gruppierten Genen bringen. Clusteranalysen stellen somit sehr hilfreiche Instrumente zur Auffindung funktionell und regulatorisch ähnlicher Gene dar, die aber immer mit Vorsicht behandelt werden müssen.

7 Zusammenfassung

Ziel dieser Arbeit war es, eine optimierte Methode zur Analyse von Daten aus Microarray-Experimenten zu entwickeln. Das in dieser Arbeit entwickelte Softwaretool wurde in ein umfassendes Auswerteprogramm zur fehlerkompensierten Analyse von sowohl *low-density*- wie auch von *whole-genome*-Chipdaten integriert und stellt somit eine Neuheit auf dem Gebiet der Microarray-Analyse dar.
Ein grundlegenes Problem bei der Analyse von Microarray-Daten ist die Vielzahl an Variabilitäten, die aufgrund der zahlreichen experimentellen Einzelschritte in die Daten einfliessen. Aus diesem Grunde sind umfassende statistische Berechnungen erforderlich, die eine Minimierung der durch die verschiedenen Einflussgrößen entstandenen Fehler vorsehen. Umfangreiche Studien haben sich in diesem Kontext mit der statistischen Kompensation der durch die Verwendung mehrerer Chips und Farbstoffe in einem Experiment auftretenden Fehlerquellen befasst. Verhältnismäßig wenig Aufmerksamkeit wurde bisher jedoch einer weiteren, ebenfalls technische bedingten Einflussgröße geschenkt, die durch die den kompletten Intensitätsbereich abdeckenden Daten der Microarrayexperimente zustande kommt: Die unzureichende Differenzierung des gesamten Intensitätsspektrums durch den Scanner führt in diesem Zusammenhang zu spezifischen Fehlern bei niedrigen wie auch bei hohen Intensitäten. Da das Signal-Rausch-Verhältnis von Spots mit niedrigen Intensitäten verhältnismäßig gering ist werden diese Spots oftmals in der Vorverarbeitung der Daten durch die Definition eines unteren Grenzwertes entfernt. Dadurch gehen wichtige Informationen zu der Expression dieser Gene verloren. Im oberen Intensitätsbereich tritt ab einem Signalwert von 65535 Sättigung ein, so

dass keine differenzierten Expressions-Daten zu den entsprechenden Genen vorliegen. Das im Rahmen dieser Arbeit eingeführte Analyse-Instrument nutzt als methodischen Ansatz zur Bereinigung dieser Effekte die bislang wenig diskutierte Möglichkeit der Verwendung multipler Scans. Dieses multiple Scan-Verfahren bietet den Vorteil einer differenzierten umfassenden Intensitätspalette für jeden Spot, aus der die entsprechenden verwertbaren Informationen der Scans für die Ermittlung des Expressionswertes dieses Spots genutzt werden können. Es gestattet somit die valide Erfassung der Genexpression im gesamten Intensitätsspektrum.
Um für die Validierung des Analyse-Tools eine möglichst optimale Datenquelle zur Verfügung zu stellen, wurde ein neuartiges Experiment-Design gewählt. Im Gegensatz zu bisherigen Microarray-Experimenten, die relative Aussagen über die Expression eines Genes in zwei Zuständen treffen, erlaubt dieses *low-density*-Microarray-Experiment, bestehend aus 13 Genen (in zwanzigfacher replikativer Anordnung) aus dem Bakterium *E.coli* eine gezielte absolute Quantifizierung der *E.coli*-Gene. Um mögliche Optimierungen des neu entwickelten Analyse-Instruments gegenüber der bisher standardmäßig angewandten Auswertemethode von Scans (Mittelwertbildung über alle normierten Scanintensitäten) beurteilen zu können, wurden diese beiden Methoden bei den entsprechenden Validierungsschritte jeweils mit absoluten Expressionswerten aus der Quantifizierung ins Verhältnis gesetzt und die Ergebnisse anschliessend vergleichend gegenüber gestellt. Unter Verwendung unterschiedlicher Verdünnungsreihen der *E.coli*-Gene wurde so die Güte der Erfassung des gesamten Intensitätsbereichs getestet.
Unter der in der Literatur beschriebenen Annahme eines linearen Trends der logarithmierten Intensitätswerte können mittels einer Linearen Regression funktionelle Geraden für jedes einzelne Gen ermittelt werden. Basierend auf diesen Geraden wurden zunächst Sättigungseffekte detektiert. Die als Unterfunktion der Auswerte-Methode implementierte Sättigungsdetektion erweist sich insofern als vorteilhaft gegenüber bisherigen Standardmethoden als dass sie einfachen *Cut-off*-Verfahren überlegen ist. Den Verlauf der Geraden berücksichtigend können hier in einem flexiblen Hochintensitätsbereich Sättigungswerte gefunden und durch Extrapolation der Geraden kompensiert werden. Auch im niedrigen Intensitätssektor kann die Linearität des Verlaufs der Intensitätskurve eines Gens genutzt werden, um stärkeren Schwankungen der Werte in diesem Bereich korrigierend entgegenzuwirken. Entsprechend geringer fallen die relativen Standardabweichungen der Intensitäten der Genreplikate bei niedrigen Konzentrationen der Gene aus (eine mittlere relative Standardabweichung über alle Verdünnungsreihen von 0.17 im Vergleich zu 0.27 bei der Standardauswertemethode für eine 1:5210-Verdünnung). Da im mittleren Intensitätsbereich nur geringfügige qualitative Unterschiede zwischen den Auswertemethoden zu verzeichnen sind, scheint das neue Analyse-Tool vor allem die in der Literatur beschriebenen Fehlerquellen im Randbereich der Intensitätswerte auszugleichen zu können. Auf diese Weise kann das Spektrum der siginifkant detektierbaren Intensitäten durch die Methode erheblich erweitert werden, was eine sehr viel sensitivere Auswertung der Expressionswerte zur Folge hat. Auch kann ein als Unterfunktion integrierter implementierter Mandel-Test zur nachhaltigen Verbesserung auch mittlerer Intensitätsbereiche beitragen, da dieser die erfolgreiche Detektion von Ausreissern unter den Scans vorsieht.

Um eine ganzheitliche und benutzerfreundliche Analyse von Microarray-Daten zu gewährleisten, wurde das neu erstellte Analyse-Tools in ein umfassendes Programm zur Auswertung von Microarray-Daten aufgenommen. Die Entwicklung dieses Programms basierte auf der Integration validierter Auswertestatisken für alle notwendigen Teilschritte einer Microarray-Auswertung. Das Programm sieht eine detaillierte Informationseingabe bezüglich der Microarray-Experimente vor, die anschliessend spezifisch für unterschiedliche Auswerteanforderungen genutzt wird. So können individuell verschiedene Experiment-Designs (beispielsweise mit verschieden vielen Chips und Genreplikaten oder aus unterschiedlichem Zellmaterial, usw.) mittels eines Programms ausgewertet werden.
Mit Hilfe neu entwickelter Algorithmen werden die unterschiedlichen Scans zunächst automatisch eingelesen. Die Daten werden entsprechend der Informationen aus der Eingabemaske sortiert und prozessiert. Erste Datenbereinigungsschritte zielen auf die valide Eliminierung unzulässiger Daten ab, was unter anderem durch die Implementierung eines Ausreissertests nach *Nalimov* realisiert wurde. Die oben beschriebene Methode zur signifikanten Analyse der Intensitätswerte auf den Microarrays wurde als *within-array*-Verfahren in die Normalisierungsverfahren integriert, zu denen auch eine *between-array*-Analyse zählt. Diese erfolgt entsprechend der gängigen, in der Literatur beschriebenen Vorgehensweise. Die normalisierten Intensitätswerte können daraufhin unter Anwendung eines *t*-Tests einer Analyse der Expression der Gene unterworfen werden.
Die Validität und Güte des Auswerteprogramms konnte anhand von zahlreicher institutsinternen Microarray-Experimenten sowie weiteren Forschungskooperationen mit der medizinischen Hochschule Hannover, der Universität Bielefeld, dem Oststadtkrankenhaus Hannover, der Universität Leibzig und der University of Haifa bestätigt werden.

Für die aus diesen Experimenten hervorgehenden Veröffentlichungen wurde ein zusätzliches Tool programmiert und in die Auswertesoftware eingefügt, welches die Daten bei jeder Auswertung automatisch in dem für die Microarray-Datenbank *GEO* kompatiblen Ausgabeformat speichert. Entsprechend wurden bereits Datensätze aus drei Microarray-Experimenten für Veröffentlichungen genutzt. Zusätzlich dazu wurde ein weiteres Tool für die Datenbank-Recherche von *E.coli*-Genen implementiert, mit dessen Hilfe die Datenbanken *KEGG* und *Multifun* nach allen aktuell vorhandenen regulatorischen Pfade zu den Genen des Bakterienorganismus durchsucht werden. So konnten basierend auf *whole-genome*-Microarray-Experimenten des Bakteriums *E.coli* umfangreiche funktionelle Studien durchgeführt werden. Für diese *E.coli*-Experimente ebenso wie für Microarray-Versuche an Schwannzellen wurde auch eine weitere, über die Standardanalyse der Genregulation hinausgehende Funktion genutzt, die unter Anwendung eines *K*-means-Clusterverfahren die Sortierung ähnlich exprimierter Gene zu Genclustern ermöglicht. Diese Zuordnung ähnlicher Gene zu Genverbänden eröffnet die Möglichkeit zu weitreichenden Forschungsansätzen wie beispielsweise in Zusammenarbeit mit der Universität Bielefeld die Entdeckung neuer Genpfade des Bakteriums *E.coli*, die in darauf folgenden *low-density*-Experimenten näher untersucht werden können.
Die im Rahmen dieser Dissertation entwickelte Auswertesoftware zur Berechnung von Microarray-Experimenten stellt somit ein innovatives Analyse-Werkzeug zur umfassenden und er-

leichterten Interpretation von Chip-Experimenten unterschiedlicher Herkunft dar.

Literatur

[1] Schena M., Shalon D., *Science* **1995**, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*, S.467–470

[2] Haab B.B., Dunham M.J., *Genome Biol.* **2001**, *Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions*, 2.

[3] Iyer V.R., Horak C.E., *Nature* **2001**, *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF*, S.533–538

[4] Pollack J.R., Perou C.M., *Nat. Genet.* **1999**, *Genome-wide analysis of DNA copy-number changes using cDNA microarrays*, S.41-46

[5] Yan H., Park S.H., *Science* **2003**, *DNA-templated self-assembly of protein arrays and highly conductive nanowires*, S.1882–1884

[6] Imbeaud S., Auffray C., *Drug Discovery Today* **2005**, *'The 39 steps' in gene expression profiling: critical issues and proposed best practices for microarray experiments*, S.1175-1182

[7] Bajcsy P., *CVPR* **2005**, *An overview of DNA microarray image requirements for automated processing*, S.1-6

[8] Quackenbush J., *Nature Gen.* **2002**, *Microarray data normalization and transformation*, S.496-501

[9] Stears R.L., Martinsky T., Schena M., *Nature Medicine* **2003**, *Trends in microarray analysis*, S.140-144

[10] Southern E.M., *Blotting at 25.* **2000**, *Trends Biochem. Sci.*, S.585-588

[11] Halbeisen R.E., Galgano A., *Cell. Mol. Life Sci.* **2007**, *Post-transcriptional gene regulation: From genome-wide studies to principles*, S.1-14

[12] Harbison C.T., Gordon D.B., *Nature* **2004**, *Transcriptional regulatory code of a eukaryotic genome*, S.99-104

[13] Ihmels J., Bergmann S., *Science* **2005**, *Rewiring of the yeast transcriptional network through the evolution of motif usage*, S.938-940

[14] Schulze A., Downward J., *Nat. Cell. Biol.* **2001**, *Navigating gene expression using microarrays–a technology review*, S.E190-E195

[15] Schulze A., Downward J., *technology review* **2001**, *Navigating gene expression using microarrays - a technology review*, S.E190-E195

[16] Zellweger T., Ninck C., *The prostate* **2003**, *Tissue microarray analysis reveals prognostic significance of syndecan-1 expression in prostate cancer*, S.20-29

[17] Petricoin E.F., Hackett J.L. *Nat. Genet.* **2002**, *Maedical applications of microarray technologies: a regulatory science perspective*, S.474-479

[18] Marx J., *Science* **2000**, *Medicine. DNA arrays reveal cancer in its many forms*, S. 1670-1672

[19] Sinclair A., *CMAJ* **2002**, *Genetics 101: detecting mutations in human genes*, S.275-279

[20] Nazmul-Hossain A.N.M., Patel K.J., *Oral Diseasis* **2008**, *Microarrays: applications in dental research*, S.25-29

[21] Horn L.C., Purz S., *Ann. Diagn. Pathol.* **2008**, *Application of tissue microarrays in placental research*, S.48-49

[22] Karssen A.M., Li J.Z., *Models for Primate Behavior* **2006**, *Application of microarray technology in primate behavioural neuroscience research*, S.227-234

[23] Buchholz M.,Boeck W., *Pancreatology* **2001**, *Use of DNA Arrays/Microarrays in Pancreatic Research*, S.581-586

[24] Baron J.K.,Merk H.F., *Hautarzt* **2003**, *Array technology in skin pharmacology and allergology*, S.315-320

[25] Schena M., Heller R.A., *Trends Biotechnol.* **1998**, *Microarrays: biotechnology's discovery platform for functional genomics*, S.301-306

[26] Southern E.M., *J. Mol. Biol.* **1975**, *Detection of specific sequences among DNA fragments separated by gel electrophoresis*, S.503-517

[27] Name, *Journal* **Jahr**, *Veröffentlichung*, 27.

[28] Blohm D.H., Guiseppi-Elie A. *Curr. Opinion in Biotechnol.* **2001**, *New developments in microarray technology*, S.41-47

[29] Fodor S.P., Read J.L. *Science* **1991**, *Light-directed, spatially addressable parallel chemical synthesis*, S.767-773

[30] Pease A.C., Solas D., *Proc. Natl. Acad. Sci USA* **1994**, *Light-generated oligonucleotide arrays for rapid DNA sequence analysis*, S.5022-5026

[31] Phillips M.F., Lockett M.R., *Nucleic Acids Res.* **2007**, *In situ oligonucleotide synthesis on carbon materials: stable substrates for microarray fabrication*, S.1-9

[32] Singh-Gasson S., Green R.D., *Nat. Biotechnol.* **1999**, *Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array*, S.974-978

[33] Heller M.J., *Annu. Rev. Biomed. Eng.* **2002**, *DNA microarray technology: devices, systems and applications*, S.129-143

[34] Lorkowski S., Lorkowski G., *Chemie in unserer Zeit* **2000**, *Biochemische Methoden*, S.356-372

[35] Schena M., *Bioessays* **1996**, *Genome analysis with gene expression microarrays*, S.427-431

[36] Shalon D., Smith S.J., *Genome Res.* **1996**, *A DNA microarray system for analyzing complex DNA samples using two-color fluorenscent probe hybridization*, S.639-645

[37] Ermantraut E., Wohlfahrt K., *Ultramicroscopy* **1998**, *Perforated support foils with predefined hole size, shape and arrangement*, S.75-81

[38] Golze S., *Dissertation* **2001**, *Oberflächengebundene Polymermonolagen für die Herstellung von DNA-Chips*, S.4-6

[39] Wrobel G., *Dissertation* **2004**, *Aufbau, Validierung und Verwendung eines cDNA-Microarray-Systems für die Expressionsanalyse humaner Meningiome*, S.1-103

[40] Brown P.O., Botstein D., *Nat. genet.* **1999**, *Exploring the new world of the genome with DNA microarrays*, S.31-37

[41] Jarrett R.G., Ruggiero K., *Biometrics* **2007**, *Design and Analysis of Two-Phase Experiments for Gene Expression Microarrays—Part I*, S.208-216

[42] Kerr, M. K. *Biometrics* **2003**, *Design considerations for efficient and effective microarraystudies*, S.822–828

[43] Nadon R., Shoemaker J., *Trends Genet.* **2002**, *Statistical issues with microarrays: processing and analysis*, S.265-271

[44] Stalteri M.A., Harrison A.P., *BMC Bioinformatics* **2007**, *Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips*, S.13

[45] Weniger M., *Dissertation* **2007**, *Genome Expression Pathway Analysis Tool Analyse und Visualisierung von Microarray Genexpressionsdaten unter genomischen, proteomischen und metabolischen Gesichtspunkten*, S.24-26

[46] Kothapalli R., Yoder S.J., *BMC Bioinformatics* **2002**, *Microarray results: how accurate are they?*, e22

[47] Harbig J., Sprinkle R., *Nucleic Acids Res* **2005**, *A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.*, e31

[48] Hamadeh H.K., Bushel P., *Biotechniques* **2002**, *Detection of diluted gene expression alterations using cDNA microarrays.*, S.322-329

[49] Pietu G., Alibert O., *Genome Res.* **1996**, *Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array*, S.492–503

[50] Schuchhardt J.,Beule D., *Nucleic Acids Res.* **2000**, *Normalization strategies for cDNA microarrays*, e47

[51] Yang Y.H., *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE* **2001**, *Normalization for cDNA microarray data*, S.141–152

[52] Lander E.S., *Nat. Genet.* **1999**, *Array of hope*, S.3-4

[53] Stekel D., *Cambridge University Press* **2003**, *Microarray Bioinformatics*, S.100-110

[54] Velculescu V.E., Zhang L., *Science* **1995**, *Serial analysis of gene expression*, S.484-487

[55] Kerr M.K., *J. Comput. Biol.* **2000**, *Analysis of variance for gene expression microarray data*, S.819-837

[56] Brazma A., Hingamp P., *Briefings in Bioinformatics* **2001**, *Minimum information about a microarray experiment (MIAME)—toward standards for microarray data*, S.365-371

[57] Tomiuk S. Hofmann K., *Briefings in Bioinformatics* **2001**, *Microarray probe selection strategies*, S.329-340

[58] Emmert-Streib F. , Seidel C., *Analysis of Microarray Data: A Network-Based Approach, WILEY-VCH* **2008**, *Introduction to DNA Microarrays*, S.1-19

[59] Chen M., ten Bosch J., *Abstract TGR Conference* **1999**, *Covalent attachment of sequence optimized PCR products and oligos for DNA microarrays*

[60] Sinibaldi R., O'Connell C., *Methods in molecular biology* **2001**, *Gene expression analysis on medium-density oligonucleotide arrays*, S.211-222

[61] Bozdech Z., Zhu J., *Genome Biology* **2003**, *Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray*, R9

[62] Yang Y.H., Speed T., *Nature* **2002**, *Design issues for cDNA microarray experiments.*, S.579–588

[63] Naidooa S., Denby K.J., *South African J. of Sci.* **2005**, *Microarray experiments: considerations for experimental design*, S.347-454

[64] Livesey F.J., *Brief Funct. Genomic Proteomic* **2003**, *Strategies for microarray analysis of limiting amounts of RNA*, S.31–36

[65] Dobbin K., Simon R.M., *Biotechniques* **2003**, *Experimental design of DNA microarray experiments*, S.16–21

[66] Shih J.H., Michalowska A.M., *Bioinformatics* **2005**, *Effects of pooling mRNA in microarray class comparisons*, S.3318–3325

[67] Dobbin K., Shih J.H., *Bioinformatics* **2003**, *Statistical design of reverse dye microarrays*, S.803–810

[68] Huitema E., Vleeshouwers G.A.A., *Mol. Plant Path.* **2003**, *Active defence responses associated with non-host resistance of Arabidopsis thaliana to the oomycete pathogen Phytophthora infestans*, S.487–500

[69] Oleksiak M.F., Churchill G.A., *Nature Gen.* **2002**, *Variation in gene expression within and among natural populations*, S.261–266

[70] Moser J.M., Freitas T., *Molecular Biochem. Parasitol.* **2005**, *Gene expression profiles associated with the transition to parasitism in Ancylostoma caninum larva* S.39-48

[71] Chowers I., Gunatilaka T.L., *Invest. Ophthalmol. Vis. Sci.* **2003**, *Identification of novel genes preferentially expressed in the retina using a custom human retina cDNA microarray*, S.3732–3741

[72] Smyth G.K., Yang Y.H. *Methods Mol. Biol.* **2003**, *Statistical Issues in cDNA Microarray Data Analysis*, S.111-136

[73] Skibbe D.S., Wang X., *Bioinformatics* **2006**, *Scanning microarrays at multiple intensities enhances discovery of differentially expressed genes*, S.1863-1870

[74] Schena M., Basarsky T., *Microarray Biochip Technology* **2000**, *Overview of a Microarray Scanner: Design Essentials for an Integrated Acquisition and Analysis Platform*, S.265-284

[75] Buhler J., Ideker T., *University of Washington CSE Technical Report UWTR 2000-08-05* **2000**, *Dapple: improved techniques for finding spots on DNA microarrays*

[76] Brown C.S., Goodwin P.C., *Proceedings of the National Academy of Science* **2000**, *Image metrices in the statistical analysis of DNA microarray data*, S.8944-8949

[77] Bowtell D.D., *Nat. Genet.* **1999**, *Options available-from start to finish-for obtaining expression data by microarray*, S.25-32

[78] Quackenbush J., *Nat. Genet.* **2002**, *Microarray data normalization and transformation*, S.496-501

[79] Reimers M. *Addiction Biology* **2005**, *Statistical Analysis of Microarray Data*, S.23-35

[80] Kim J.H., Shin D.M., *Exp. Mol. Med.* **2002**, *Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles*, S.224-232

[81] Dudoit S., Yang Y.H., *Statistica Sinica* **2002**, *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*, S.111-139

[82] Huber W., Von Heydebreck A., *Bioinformatics* **2002**, *Variance stabilization applied to microarray data calibration and to the quantification of differential expression*, S.96–104

[83] Quackenbush J., *Nat. Rev. Genet.* **2001**, *Computational analysis of microarray data*, S.418-427

[84] Yang Y.H., Dudoit S.D., *SPIE BioE* **2001**, *Normalization for cDNA microarray data*

[85] Park T., Sung-Gon Y., *BMC Bioinformatics* **2003**, *Evaluation of normalization methods for microarray data*, S.1-13

[86] Cleveland W.S., *J. Amer. Stat. Assoc.* **1979**, *Robust locally weighted regression and smoothing scatterplots*, S.829–836

[87] Yang Y.H., *Nucleic Acids Res.* **2002**, *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*, e15

[88] Tusher V., Tibshirani R., *PNAS* **2001**, *Significance analysis of microarrays applied to the ionizing radiation response*, S.5116-5124

[89] Chen Y., Dougherty E.R., *J. Biomed. Optics* **1997**, *Ratio-based decisions and the quantitative analysis of cDNA microarray images*, S.364-374

[90] Lönnstedt I., Speed T.P., *Statistica Sinica* **2002**, *Replicated microarray data*, S.31-46

[91] Efron B., Tibshirani R., *Journal of the American Statistical Association* **2001**, *Empirical Bayes analysis of a microarray experiment*, S.1151-1160

[92] Shaffer J.P., *Annals of Review Psychology* **1995**, *Multiple hypothesis testing*, S.561-576

[93] Benjamini Y., Hochberg Y., *J.R.Statist.Soc.* **1995**, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, S.289-200

[94] Wolfinger R.D., Gibson G., *J. Comput. Biol.* **2001**, *Assessing gene significance from cDNA microarray expression data via mixed models*, S.625-637

[95] Ideker T., Thorsson V., *J. Comput. Biol.* **2000**, *Testing for differentially-expressed genes by maximum-likelihood analysis for microarray data*, S.805-817

[96] Golub T.R., Slonim D.K., *Science* **1999**, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, S.531-537

[97] Mardia K.V., Kent J.T., *Academic Press, London* **1979**, *Multivariate Analysis*

[98] Breiman L., Friedman J.H., *Wadswoth, Monterey, CA* **1984**, *Classification and Regression Tress*

[99] Breiman L., *Machine Learning* **1996**, *Bagging predictors*, S.123-140

[100] Riply B.D., *Cambridge University Press, Cambridge* **1996**, *Pattern Recognition and Neural Networks*

[101] Kanehisa M., Goto S., *Nucleic Acids Res.* **2006**, *From genomics to chemical genomics: new developments in KEGG*,S.354-357

[102] Barrett T., Suzek T.O., *Nucleic Acids Res* **2005**, *NCBI GEO: mining millions of expression profiles-database and tools*, S.562-566

[103] Kothapalli R., Yoder S.J., *BMC Bioinformatics* **2002**, *Microarray results: how accurate are they?*, e22

[104] Kuo W.P., Jenssen T., *Bioinformatics* **2002**, *Analysis of matched mRNA measurements from two different microarray technologies*, S.405-412

[105] de Longueville F., *Int. J. Oncology* **2005**, *Molecular characterization of breast cancer cell lines by a low density microarray*, S.881–892

[106] Rangel Lopez A., *BMC Biotechnol.* **2005**, *ow-density DNA microarray for detection of most frequent missense point mutations*, e5

[107] de Longueville F., *Toxicolog. Sci.* **2003**, *Use of a low-density microarray for studying gene expression patterns induced by hepatoxicants on primary cultures of rat hepatocytes*, S.378–392

[108] Bouchie A., *Nature Biotechnol.* **2008**, *Shift anticipated in DNA microarray market*, e8

[109] Tejedor D.,Castillo S., *Clinical Chemistry* **2005**, *Reliable Low-Density DNA Array Based on Allele-Specific Probes for Detection of 118 Mutations Causing Familial Hypercholesterolemia*, S.1137–1144

[110] Naderi A., Ahmed A.A., *BMC Genomics* **2004**, *Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling.*, e9

[111] *http://www.ncbi.nlm.nih.gov/geo/*

[112] Coombes K.R., Highsmith W.E., *J. Comp. Biol.* **2002**, *Identifying and Quantifying Sources of Variation in Microarray Data Using High-Density cDNA Membrane Arrays*, S.655-669

[113] Novak J.P., Sladek R., *Genomics* **2002**, *Characterization of variability in large-scale gene expression data: implications for study design*, S.104-113

[114] Han E.S., Wu. Y. *J. Geron. Biol. Sci.* **2004**, *Reproducibility, Sources of Variability, Pooling, and Sample Size: Important Considerations for the Design of High-Density Oligonucleotide Array Experiments*, S.306-315

[115] Lee M.L., Kuo F.C., *Proc. Natl. Acad. Sci. USA* **2004**, *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations*, S.9834-9839

[116] van den Doel L.R., *Dissertation* **2002**, *Quantitative Microscopic Techniques for Monitoring Dynamic Processes in Microarrays*, S.5

[117] Grainger D., Gong P., *European Cells and Materials* **2005**, *Fundamental DNA-surface analysis to understand microarray assay limitations*, BS10

[118] Diehl F., Grahlmann S., *Nucleic Acids Res.* **2002**, *Manufacturing DNA microarrays of high spot homogeneity and reduced background signal*, e79

[119] Zhang W., Shmulevich I., Astola J., *Wiley-Liss, New Jersey, USA* **2004**, *Microarray quality control*

[120] Kim P.-G., Park K. *International Society for Computational Biology* **2005**, *A quality measure model for microarray images*

[121] Räke C., *Bachelorarbeit Studiengang Life-Science, TCI Universität Hannover* **2008**, *Untersuchung charakteristischer Spotformen von Microarrays*

[122] Bylesjö M., Eriksson D., *BMC Bioinformatics* **2005**, *MASQOT: a method fo cDNA microarray spot quality control*, e250

[123] Draghici S., Kuklin A. , *Curr. Opin. Drug Discov. Devel.* **2001**, *Experimental design, analysis of variance and slide quality assessment in gene expression arrays*, S.332-337

[124] Schena M., *John Wiley & Sons, New York* **2003**, *Microarray analysis*

[125] Shi L., Tong W., *BMC Bioinformatics* **2005**, *Microarray scanner calibration curves: characteristics and implications*, S11

[126] Carter B., Wu G., *BMC Genomics* **2008**, *A process for analysis of microarray comparative genomics hybridisation studies for bacterial genomes*, e53

[127] Stekel D., *Cambridge University Press, Cambridge USA* **2003**, *Microarray Bioninformatics*

[128] Kreil D.P., Auburn R.P., *GCB FlyChip* **2003**, *Quantitative microarray spot profile optimization: A systematic evaluation of buffer/slide combinations*, S.77-81

[129] Ekstrom C.L., Bak S., *Bioinformatics* **2004**, *Spot shape modelling and data transformations for microarrays*, S.2270–2278

[130] Kim H.E., Lee S.E., *BMC Bioinformatics* **2007**, *Characterization and simulation of cDNA microarray spots using a novel mathematical model*, e485

[131] Li Q., Fraley C., *Bioinformatics* **2005**, *Donuts, scratches and blanks: robust model-based segmentation of microarray images*, S.2875-2882

[132] Petrov A., Shams S., *Journal of VLSI Signal Processing Systems* **2004**, *Microarray Image Processing and Quality Control*, S.211-226

[133] Ruosaari S., Hollmen, *Lecture notes in computer science* **2002**, *Image analysis for detecting faulty spots from microarray images*, S.259-266

[134] Cutler D.J., Zwick M.E., *Genome Res.* **2001**, *High-Throughput Variation Detection and Genotyping Using Microarrays*, S.1913-1925

[135] Novikov E., Barillot E., *Bioinformatics* **2007**, *Software package for automatic microarray image analysis (MAIA)*, S.639-640

[136] Brody J.P., Williams B.A., *Proc Natl Acad Sci USA* **2002**, *Significance and statistical errors in the analysis of DNA microarray data*, S.12975–12978

[137] Novikov E., Barillot E., *BMC Bioinformatics* **2003**, *An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments*, S

[138] Novikov E., Barillot E., *Vision Systems: Segmentation and Pattern Recognition* **2007**, *Robust Microarray Image Processing*, S.546

[139] Lundgren M. Andersson A., *Proc Natl Acad Sci USA* **2004**, *Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination*, S.7046-7051

[140] Dumur C.I., Nasim S. *Clinical Chemistry* **2004**, *Evaluation of quality-control criteria for microarray gene expression analysis*, e11

[141] Aust M., Fernandes D., *Plast. Reconstr. Surg.* **2008**, *Percutaneous collagen induction therapy: an alternative treatment for scars, wrinkles, and skin laxity*, S.1421-1429

[142] Johansson P., Häkkinen J., *BMC Bioinformatics* **2006**, *Improving missing value imputation of microarray data by using spot quality weights*, e306

[143] Wang L., Gaigalas A.K., *Methods in Molecular Biology* **2008**, *Evaluating the quality of data from DNA microarray Measurements*, S.121-131

[144] Iglewicz B., Hoaglin D., *ASQ Quality Press, Milwaukee* **1993**, *How to Detect and Handle Outliers*

[145] Saffarian N., Zou J.J., *IEEE International Conference on Video and Signal Based Surveillance* **2006**, *DNA Microarray Image Enhancement Using Conditional Sub-Block Bi-Histogram Equalization*, S.86

[146] Nagarajan R., Peterson C.A., *IEEE Trans. Med. Imaging* **2002**, *Identifying spots in microarray images*, S.78-84

[147] Yang Y.H., *J. Comp. Graph. Stat.* **2002**, *Comparison of methods for image analysis on cDNA microarray data*, S.108–136

[148] Gottardo R., Raftery A.E., *J. Am. Stat. Assoc.* **2006**, *Quality control and robust estimation for cDNA microarrays with replicates*, S.30-40

[149] Blekas K., Galatsanos N.P. *Image Processing* **2003**, *An unsupervised artifact correction approach for the analysis of DNA microarray images*, S.165-168

[150] Kooperberg C., Fazzio T. G., *Journal of Computational Biology* **2002**, *Improved background correction for spotted cDNA microarrays*, S.55-66

[151] Haaland D.M., Timlin J.A., *Proc. SPIE* **2003**, *Multivariate curve resolution for hyperspectral image analysis: applications to microarray technology*, S.55

[152] Smyth G.K., Speed T., *Methods* **2003**, *Normalization of cDNA Microarray Data*, S.265-273

[153] Long A.D., Mangala H.J., *J. Biol. Chem.* **2001**, *Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and A Bayesian Statistical Framework*, S.19937-19944

[154] Bolstad B.M., Irizarry R.A., *Bioinformatics* **2003**, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, S.185.193

[155] Hedge P., *Biotechniques* **2000**, *A concise guide to cDNA microarray analysis*, S.548-556

[156] Leung Y.F., Cavalieri D. *Trend Genet.* **2003**, *Fundamentals of cDNA microarray data analysis*, S.649-659

[157] Dodd L.E., Korn E.L., *Bioinformatics* **2004**, *Correcting log ratios for signal saturation in cDNA microarrays*, S.2685-2693

[158] Kerr M.K., Churchill G.A., *Biostatistics* **2001**, *Experimental design for gene expression microarray*, S.183-201

[159] Sokal R.R., Rohlf F.J., *Freeman, New York* **1995**, *Biometry: The Principles and Practices of Statistics in Biological Research*

[160] Lyng H., Badiee A., *BMC Genomics* **2004**, *Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction*, e10

[161] Dudley A.M., Aach J. *PNAS* **2002**, *Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range*, S.7554–7559

[162] Gupta R. Auvinen P., *Stat. Appl. Genet. Mol. Biol.* **2006**, *Bayesian hierarchical model for correcting signal saturation in microarrays using pixel intensities*, e20

[163] de la Nava J.G., Hijum S., *Stat. Appl. Genet. Mol. Biol.* **2004**, *Saturation and quantization reduction in microarray experiments using two scans at different sensitivities*, e11

[164] Lewis-Beck M., Bryman A., *Thousand Oaks (CA): Sage* **2003**, *Encyclopedia of Social Sciences Research Methods*

[165] Kane M. D., Jatkoe T. A., *Nucleic Acids Res.* **2000**, *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays*, S.4552-4557

[166] Nagl S., *Microchimica Acta* **2005**, *Fluorescence analysis in microarray technology*, S.1-22

[167] Romualdi C., *Nucleic Acids Res* **2003**, *Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration*, e149

[168] Hsieh H.B., *J. Histochem. Cytochem.* **2001**, *Monitoring signal transduction in cancer: cDNA microarray for semiquantitative analysis*, S.1057-1058

[169] Chen T.-C., *J. Clin. Microbiol.* **2006**, *Combining Multiplex Reverse Transcription-PCR and a Diagnostic Microarray To Detect and Differentiate Enterovirus 71 and Coxsackievirus A16*, S.2212-2219

[170] Zhongming G.E., *Proc. Japan Acad.* **1993**, *The Infectious Transcripts o f Sweet Clover Necrotic Mosaic Virus Bipartite Genome Constructed by the Polymerase Chain Reaction*, S.113-118

[171] Briggs M., *Promega Notes* **2004**, *The Pronto!TM Plus System: Increasing Microarray Reproducibility, Reliability and Ease of Use*, S.86

[172] Bilban M., Buehler L.K., *Curr. Issues Mol. Biol.* **2002**, *Normalizing DNA Microarray Data*, S.57-64

[173] Bartosiewicz M., Trounstine M., *Arch. Biochem. Biophys.* **2000**, *Development of a toxicological gene array and quantitative assessment of this technology*, S.66-73

[174] Wurmbach E., Yuen T., *J. Biol. Chem.* **2001**, *Gonadotropin releasing hormone receptor-coupled gene gene network organization*, S.47195-47201

[175] van Hal N.L., Vorst O., *J. Biotechnol.* **2000**, *The application of DNA microarrays in gene expression analysis*, S.271-280

[176] Tseng G.C., Oh M.K., *Nucl. Acids Res.* **2001**, *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects*, S.2549-2557

[177] Taniguchi M., Miura K., *Genomics* **2001**, *Quantitative assessment of DNA microarrays—comparison with Northern blot analyses*, S.34-39

[178] Yang M.C., Ruan Q.G., *Physiol. Genomics* **2001**, *A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays*, S.45-53

[179] Edwards D., *Bioinformatics* **2003**, *Non-linear normalization and background correction in one-channel cDNA microarray studies*, S.825-33

[180] Lee J.W., Lhung M., *OR Spektrum* **2008**, *An optimal choice of window width for LOWESS normalization of microarray data*, S.235-248

[181] Mandel J., *Dekker, New York* **1991**, *Evaluation and Control of Measurements*

[182] Berger J.A., *BMC Bioinformatics* **2004**, *Optimized LOWESS normalization parameter selection for DNA microarray data*, e194

[183] Balázsi G., Oltavi Z.N., *Methods in Molecular Biology, Humana Press Inc., Totowa, NJ* **2007**, *A pitfall in series of microarrays*, S.153-161

[184] Tukey J., *Addison-Wesley* **1977**, *Exploratory Data Analysis*

[185] Mosteller F., Tukey J., *Addison-Wesley* **1977**, *Data Analysis and Regression*

[186] Reimann C., *Sci. Total Environ.* **2005**, *Background and threshold: critical comparison of methods of determination*, S.1-16

[187] Reimann C., *NGU-GTK-CKE Special Publication. Trondheim* **1998**, *Environmental geochemical atlas of the central barents region*

[188] Hoaglin D.C, Mosteller F., Tukey J., *John Wiley & Sons* **1983**, *Understanding Robust and Exploratory Data Analysis*, S.404-414

[189] Khan A.H., Ossadtchi A., *Genomics* **2003**, *Error-correcting microarray design*, S.157-165

[190] Kim S.Y., Lee J.W., *Stat. Methods Med. Res.* **2006**, *Comparison of various statistical methods for identifying differential gene expression in replicated microarray data*, S.3-20

[191] DeRisi L., Penaland L., *Nature Genetics* **1996**, *Use of a cDNA microarray to analyse gene expression patterns in human cancer*, S.457-460

[192] Draghici S., Kulaeva O., *Bioinformatics* **2003**, *Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays*, S.1348-1359

[193] Rasch B., Friese M., *Heidelberg: Springer* **2006**, *Quantitative Methoden*

[194] Lange K.L., Little R.J.A., *J. Amer. Stat. Assoc.* **1989**, *Robust Statistical Modeling Using the t Distribution*, S.881-896

[195] Li W., Suh Y.J., *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2006**, *Does logarithm transformation of microarray data affect ranking order of differentially expressed genes?*, S.6593-6596

[196] Keilhoff G., *Tissue Eng.* **2006**, *Peripheral Nerve Tissue Engineering: Autologous Schwann Cells vs. Transdifferentiated Mesenchymal Stem Cells*, S.1451-1465

[197] Brockes J., *Brain Res.* **1979**, *Studies on cultured Schwann cells. I. Establishment of purified populations from cultures of peripheral nerve*, S.105-118

[198] Lee CH., *Biomaterials* **2005**, *Nanofiber alignment and direction of mechanical strain affect the ECM production of human ACL fibroblast*, S.1261-1270

[199] Thompson D.M., *Tissue Eng.* **2001**, *Schwann cell response to micropatterned laminin surfaces*, S.247-265

[200] Schlosshauer B., *Brain Res.* **2002**, *Rat Schwann cells in bioresorbable nerve guides to promote and accelerate axonal regeneration*, S.321-326

[201] Bunge R.P., *J.Neurol.* **1994**, *The role of the Schwann cell in trophic support and regeneration*, S19-S21

[202] Hoffmann K.B., *Cell Mol Neurobiol.* **1998**, *The relationship Between Adhesion Molecules and Neuronal Plasticity*, S.461-475

[203] Gascon E., *Brain Res Rev.* **2007**, *Polysialic acid-neural cell adhesion molecule in brain plasticity: From synapse to integration of neurons*, S.101-118

[204] Finne J., *J Immunol.* **1987**, *An IgG monoclonal antibody to group B meningococci cross-reacts with developmentally regulated polysialic acid units of glycoproteins in neural and extraneural tissues*, S.4402-4407

[205] Haile Y., *Biomaterials* **2007**, *Culturing of glial and neuronal cells on polysialic acid*, S.1163-1173

[206] Doyu M., *J Neurochem.* **1993**, *Laminin A, B1, and B2 chain gene expression in transected and regenerating nerves: regulation by axonal signals*, S.543-551

[207] Podratz P.L., *Glia.* **2001**, *Role of the extracellular matrix in myelination of peripheral nerve*, S.35-40

[208] Tsiper M.V., *J Cell Sci.* **2002**, *Laminin assembles into separate basement membrane and fibrillar matrices in Schwann cells*, S.1005-1115

[209] Trapp B.D., *J Neurosci.* **1994**, *Polarization of myelinating Schwann cell surface membranes: role of microtubules and the trans-Golgi network*, S.1797-1807

[210] Cambray-Deakin M.A., *Cell Motil Cytoskeleton.* **1988**, *Colocalisation of acetylated microtubules, glial filaments, and mitochondria in astrocytes in vitro*, S.438-449

[211] Oguievetskaia K., *J Sec. Bio.* **2005**, *Cellular contacts in myelinated fibers of the peripheral nervous system*, S.281-292

[212] Magnaghi V., *JPNS* **2006**, *GABA-A and GABA-B receptors participate in the axon-Schwann cell interactions*, S.179-208

[213] Dong N., *Mol Cell Neurosci.* **2007**, *Molecular reconstitution of functional GABAergic synapses with expression of neuroligin-2 and $GABA_A$ receptors*, S.14-23

[214] Oliveira A.M., *Adv Anat Pathol.* **2003**, *CHIPing Soft Tissue Tumors: Will the Paradigms Be Changed?*, S.1-7

[215] Muona P., *Diabetes* **1992**, *Glucose transporters of rat peripheral nerve. Differential expression of GLUT1 gene by Schwann cells and perineurial cells in vivo and in vitro*, S.1587-1597

[216] Stark B., *Exp Brain Res.* **2000**, *Developmental and lesion-induced changes in the distribution of the glucose transporter Glut-1 in the central and peripheral nervous system*, S.74-84

[217] Kasa P., *Brain Res.* **2001**, *Presenilin-1 and its N-terminal and C-terminal fragments are transported in the sciatic nerve of rat*, S.159-169

[218] Papastefanaki F., *Brain.* **2007**, *Grafts of Schwann cells engineered to express PSA-NCAM promote functional recovery after spinal cord injury*, S.2159-2174

[219] Acheson A., *J Cell Biol.* **1991**, *NCAM Polysialic Acid Can Regulate both Cell-Cell and Cell-Substrate Interactions*, S.143-153

[220] Abderrahman E.M., *Neurochem Res.* **2008**, *Use of PSA-NCAM in Repair of the Central Nervous System*

[221] Grawanis A.I., *Microsurgery* **2004**, *Effect of genetically modified Schwann cells with increased motility in end-to-side nerve grafting*, S.423-432

[222] Blaugrund E., *J Comp Neurol.* **1992**, *Axonal regeneration is associated with glial migration: Comparison between the injured optic nerves of fish and rats*, S.105-112

[223] Lefranc F., *Int J Oncol.* **2003**, *Determination of RNA expression for cholecystokinin/gastrin receptors (CCKA, CCKB and CCKC) in human tumors of the central and peripheral nervous system*, S.213-219

[224] Noble P.J., *Am J Physiol Gastrointest Liver Physiol.* **2003**, *Stimulation of gastrin-CCKB receptor promotes migration of gastric AGS cells via multiple paracrine pathways*, G75-G84

[225] Huang C., Jacobsen K., *J Cell Sci.* **2004**, *MAP kinases and cell migration*, S.4619-4628

[226] Yamauchi J., Chan R., *Proc Natl Acad Sci U S A.* **2003**, *Neurotrophin 3 activation of TrkC induces Schwann cell migration through the c-Jun N-terminal kinase pathways*, S.144421-144426

[227] Cornely C., *Dissertation* **2006**, *Untersuchung des Transkriptionsfaktors Mash2 und der von ihm regulierten Zielgene bezüglich ihres Einflusses auf Proliferation und Apoptose von Schwannzellen*

[228] NCBI Gene; GeneID: 24185

[229] Zhou J., Deo B.K., *Invest Ophthalmol Vis Sci.* **2005**, *Increased JNK Phsosphoylation and Oxidative Stress in Response to Increased Glucose Flux through Increased GLUT1 Expression in Rat Retinal Endothelial Cells*, S.3403-3410

[230] Lee J., *J Cell Biol.* **2005**, *Process outgrowth in oligodendrocytes is mediated by CNP, a novel microtubule assembly myelin protein*, S.661-673

[231] Vaithianathan T., *J Biol Chem.* **2004**, *Neural Cell Adhesion Molecule-associated Polysialic Acid Potentiates α-Amino-3-hydroxy-5-methylisoxazole-4-propionic Acid Receptor Currents*, S.47975-47984

[232] Zhang Q., Fukuda M., *Proc Natl Acad Sci U S A.* **2004**, *Synaptotagmin IV regulates glial glutamate release*, S.9441-9446

[233] Teare K.A., Pearson R.S., *Neuroreport* **2004**, *α-MSH inhibits inflammatory signalling in Schwann cells*, S.493-398

[234] Glass J.D., Shen H. *Neurosci Lett.* **2004**, *Polysialylated neural cell adhesion molecule modulates photic signaling in the mouse suprachiasmatic nucleus*, S.207-219

[235] Dickerson D.S., *Peptides* **1994**, *POMC mRNA levels in individual melanotropes and GFAP in glial-like cells in rat pituitary*, S.247-256

[236] Schworer C.M., Kasker K.K., *J Neurosci Res.* **2003**, *Microarray Analysis of Gene Expression in Proliferating Schwann Cells: Synergetic Response of a Specific Subset of Genes to the Mitogenic Action of Heregulin Plus Forsolin*, S.456-464

[237] Takizawa T., Tatematsu C., *J Biochem.* **2004**, *Cleavage of Calnexin Caused by Apoptotic Stimuli: Implication for the Regulation of Apoptosis*, S.399-405

[238] Ponomareva O.N., *Glial.* **2006**, *Schwann cell-derived neuregulin-2 can function as a cell-attached activator of muscle acetylcholine receptor expression*, S.630-637

[239] Aust M.C., Reimers K., Repenning C., *Plast Reconstr Surg.* **2008**, *Percutaneous Collagen Induction: Minimally Invasive Skin Rejuvenation without Risk of Hyperpigmentation-Fact or Fiction?*, S.1553-1563

[240] AUSTPAPER 2 Namen noch einfügen!, *JOURNAL* **JAHR**, *titel*, S.

[241] Smolkin M., Ghosh D., *BMC Bioinformatics* **2003**, *Cluster stability scores for microarray data in cancer studies*, e36

[242] Ben-Dor A., Shamir R., *J. Comput. Biol.* **1999**, *Clustering gene expression patterns*, S.281-297

[243] Gesu V.D., *Lecture Notes in Computer Science, Verlag Springer Berlin, Heidelberg* **2007**, *Data Analysis and Bioinformatics*, S.373-388

[244] Sebastiani P., Ramoni M.F., *Technical report. Children's Hospital Informatics Program, Harvard Medical School and the Department of Mathematics and Statistics University Massachussetts at Amherst.* **2002**, *Statistical challenges in functional genomics*

[245] Sherlock G., *Curr. Opin. Immunol.* **2000**, *Analysis of large-scale gene expression data*, S.201-205

[246] MacQueen, *University of California Press* **1967**, *Some Methods for classification and analysis of multivariate observations*, S.281-297

[247] Tavazoie S., Hughes J.D., *Nature Genet.* **1999**, *Systematic determination of genetic network architecture*, S.281-285

[248] Morgan-Kiss R.M., Wadler C., *Proc Natl Acad Sci U S A* **2002**, *Long-term and homogeneous regulation of the Escherichia coli araBAD promoter by use of a lactose transporter of relaxed specificity*, S.7373-7377

[249] Hendrickson W., Flaherty C., Molz L., *J Bacteriol.* **1992**, *Sequence elements in the Escherichia coli araFGH promoter*, S.6862–6871

[250] Lengeler J.W., *Biomedizin & Life Sciences* **2004**, *Carbohydrate transport in bacteria under environmental conditions, a black box?*, S.275-288

[251] Stoner C., Schleif R., *J Mol Biol.* **1983**, *The araE low affinity L-arabinose transport promoter. Cloning, sequence, transcription start site and DNA binding sites of regulatory proteins.*, S.369-81

[252] Hendrickson W., Stoner C., Schleif R., *J Mol Biol.* **1990**, *Characterization of the Escherichia coli araFGH and araJ promoters*, S.497-510

[253] Fornwald J.A., Schmidt F.J., Adams C.W., *Proc Natl Acad Sci U S A* **1987**, *Two promoters, one inducible and one constitutive, control transcription of the Streptomyces lividans galactose operon*, S.2130-2134

[254] Death A., Ferenci T., *J Mol Biol.* **JAHR**, *Between feast and famine: endogenous inducer synthesis in the adaptation of Escherichia coli to growth with limiting carbohydrates*, S.5101-5107

[255] Eiglmeier K., Boos W. *Mol Microbiol.* **1987**, *Nucleotide sequence and transcriptional startpoint of the glpT gene of Escherichia coli: extensive sequence homology of the glycerol-3-phosphate transport protein with components of the hexose-6-phosphate transport system*, S.251-258

[256] Saurin W., Hofnung M., Dassa E., *J Mol Evol.* **1999**, *Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters*, S.22-41

[257] Flemming D., *J. Biol. Chem.* **2006**, *Catalytic Importance of Acidic Amino Acids on Subunit NuoB of the Escherichia coli NADH:Ubiquinone*, S.24781-24789

[258] Narayanana N., Chou C.P., *Enzyme and Microbial Technology* **2008**, *Periplasmic chaperone FkpA reduces extracytoplasmic stress response and improves cell-surface display on Escherichia coli*, S.506-513

[259] Cashel M., Gentry D.R., *ASM Press, Washington, D.C.* **1996**, *The Stringent Response. Escherichia coli and Salmonella typhimurium*

[260] Caldara M., Dupont G., *J. Biol. Chem.* **2008**, *Arginine Biosynthesis in Escherichia coli*, S.6347-6358

[261] Piette J., Nyunoya H., *Proc Natl Acad Sci U S A* **1984**, *DNA sequence of the carA gene and the control region of carAB: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoylphosphate*, S.4134-4138

[262] Sens D., Natter W., *Mol Gen Genet.* **1977**, *Transcription of the argF and argI genes of the arginine biosynthetic regulon of Escherichia coli K12, performed in vitro*, S.7-18

[263] Turnbough C.L., Switzer R.L., *Microbiol Mol Biol Rev.* **2008**, *Regulation of Pyrimidine Biosynthetic Gene Expression in Bacteria: Repression without Repressors*, S.266-300

[264] Andersen L., Kilstrup M., *Arch Microbiol.* **1989**, *Pyrimidine, purine and nitrogen control of cytosine deaminase synthesis in Escherichia coli K 12. Involvement of the glnLG and purR genes in the regulation of codA expression*, S.115-118

[265] He B., Choi K. Y., Zalkin H., *J Bacteriol.* **1993**, *Regulation of Escherichia coli glnB, prsA, and speA by the purine repressor*, S.3598–3606

[266] Spirin V., Gelfand M.S., *Proc Natl Acad Sci U S A* **2006**, *A metabolic network in the evolutionary context: Multiscale structure and modularity*, S.8774–8779

[267] Mansilla M.C., de Mendoza D., *Microbiology* **2000**, *The Bacillus subtilis cysP gene encodes a novel sulphate permease related to the inorganic phosphate transporter (Pit) family*, S.815-21

[268] Adams M.D., Wagner L.M., *J Biol Chem.* **1999**, *Nucleotide Sequence and Genetic Characterization Reveal Six Essential Genes for the LIV-I and LS Transport Systems of Escherichia coli*, S.11436-43

[269] Knaggs A.R., *Nat. Prod. Rep.* **2003**, *The biosynthesis of shikimate metabolites*, S.119-136

[270] Cronan J.E., Subrahmanyam S., *Molecular Biology* **2002**, *FadR, transcriptional coordination of metabolic expediency*, S.937-943

[271] Heffernan L., Bass R., Englesberg E., *J Bacteriol.* **1976**, *Mutations affecting catabolite repression of the L-arabinose regulon in Escherichia coli B/r*,S.1119–1131

[272] Spehr V., Splitt A. *Biochemistry* **1999**, *Overexpression of the Escherichia coli nuo-operon and isolation of the overproduced NADH:ubiquinone oxidoreductase (complex I)*, S.16261-16267

[273] Ferrandez A., Garcia J.L., *J. Bacteriol.* **1997**, *Genetic characterization and expression in heterologous hosts of the 3- (3-hydroxyphenyl)propionate catabolic pathway of Escherichia coli K-12*, S.2573-2581

[274] Egebjerg J., Christiansen J., *J. Mol. Biol.* **1991**, *Attachment sites of primary binding proteins L1, L2 and L23 on 23 S ribosomal RNA of Escherichia coli*, S.251-64

[275] Freedman L.P., Zengel J.M., *Proc Natl Acad Sci U S A* **1987**, *Autogenous control of the S10 ribosomal protein operon of Escherichia coli: genetic dissection of transcriptional and posttranscriptional regulation*, S.6516-6520

[276] Metspalu E., Maimets T., Ustav M., *FEBS Lett.* **1980**, *A quaternary complex consisting of two molecules of tRNA and ribosomal proteins L2 and L17*, S.105-108

[277] Vartikar J.V., Draper D.E., *J. Mol. Biol.* **1989**, *S4-16 S ribosomal RNA complex. Binding constant measurements and specific recognition of a 460-nucleotide region*, S.221-234

[278] Graifer D.M., Babkina G.T., *Biochim. Biophys. Acta.* **1989**, *Structural arrangement of tRNA binding sites on Escherichia coli ribosomes, as revealed from data on affinity labelling with photoactivatable tRNA derivatives*, S.146-156

[279] Wieden H.J., Gromadski K., Rodnin D., *J. Biol. Chem.* **2002**, *Mechanism of elongation factor (EF)-Ts-catalyzed nucleotide exchange in EF-Tu. Contribution of contacts at the guanine base*, S.6032-6036

[280] Wintermeyer W., Savelsbergh A., *Cold Spring Harb Symp Quant Biol.* **2001**, *Mechanism of elongation factor G function in tRNA translocation on the ribosome*, S.449-458

[281] Allen S.P., Polazzi J.O., Gierse J.K., *J. Bacteriol.* **1991**, *Two novel heat shock genes encoding proteins produced in response to heterologous protein expression in Escherichia coli*, S.6938-6947

[282] Chen C.F., Lan J., Korovine M., *Microbiology* **1997**, *Metabolic regulation of lrp gene expression in Escherichia coli K-12*, S.2079-2084

[283] Wessler S. R., Calvo J. M., *J Mol Biol.* **1981**, *Control of leu operon expression in Escherichia coli by a transcription attenuation mechanism*, S.579–597

[284] Park J. H., Lee K. H., *Proc. Natl. Acad. Sci. USA* **2007**, *Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation*, S.7797-7802

[285] Brinkman A.B., Ettema T.J., *Mol Microbiol.* **2003**, *The Lrp family of transcriptional regulators*, S.287-294

[286] Gerstel B.,McCarthy E.G., *Mol Microbiol.* **1989**, *Independent and coupled translational initiation of atp genes in Escherichia coli: experiments using chromosomal and plasmid-borne lacZ fusions*, S.851–859

[287] Bendt A.K., *Dissertation* **2002**, *Globale Analyse von Stickstoff-Metabolismus und Stickstoff-Kontrolle in Corynebacterium glutamicum*

[288] Hellmuth K., Rex G., *Mol Microbiol.* **1991**, *Translational coupling varying in efficiency between different pairs of genes in the central region of the atp operon of Escherichia coli*, S.813-824

[289] Yoo S.J., Seol J.H., *Biochem Biophys Res Commun.* **1996**, *Poly-L-lysine activates both peptide and ATP hydrolysis by the ATP-dependent HslVU protease in Escherichia coli*, S.531-535

[290] Tokumoto U., Takahashi Y., *J. Biochem.* **2001**, *Genetic analysis of the isc operon in Escherichia coli involved in the biogenesis of cellular iron-sulfur proteins*, S.63-71

[291] Freire P., Amaral J.D., *Biochimie* **2006**, *Adaptation to carbon starvation: RNase III ensures normal expression levels of bolA1p mRNA and sigma(S)*, S.341-346

[292] Levchenko I., Seidel M., *Science* **2000**, *A specificity-enhancing factor for the ClpXP degradation machine*, S.2354-2356

[293] Lucht J.M., Bremer E., *FEMS Microbiol. Rev.* **1994**, *Adaptation of Escherichia coli to high osmolarity environments: osmoregulation of the high-affinity glycine betaine transport system proU*, S.3-20

[294] Barbosa T.M., Levy S.B., *Mol Microbiol.* **2002**, *Activation of the Escherichia coli nfnB gene by MarA through a highly divergent marbox in a class II promoter*, S.191-202

[295] Seol J.H., Woo S.K., Jung E.M., Yoo S.J., *Biochem Biophys Res Commun.* **1991**, *Protease Do is essential for survival of Escherichia coli at high temperatures: its identity with the htrA gene product*, S.730-736

[296] Tedin K., Norel F., *J Bacteriology* **2001**, *Comparison of relA strains of Escherichia coli and Salmonella enterica Serovar Typhimurium suggests a role for ppGpp in attenuation regulation of branched-chain amino acid biosynthesis*, S.6184-6196

[297] Schäffer C., Wugeditsch T., *Appl Environ Microbiol.* **2002**, *Functional Expression of Enterobacterial O-Polysaccharide Biosynthesis Enzymes in Bacillus subtilis*, S.4722–4730

[298] Raut I., Bhaduri A., *J. Biosci.* **1985**, *Biochemical analysis of galactose induced bacteriostasis in galT mutants of Escherichia coli K 12*, S.71-81

[299] Timothy M. K., Padmalayam I., *Clin Diagn Lab Immunol.* **1998**, *Use of the Cell Division Protein FtsZ as a Means of Differentiating among Bartonella Species*, S.766-772

[300] Kupke T., Uebele M., *J Biol Chem.* **2000**, *Molecular characterization of lantibiotic-synthesizing enzyme EpiD reveals a function for bacterial Dfp proteins in coenzyme A biosynthesis*, S.31838-31846

[301] Awano N., Wada M., *Appl Microbiol Biotechnol.* **2003**, *Effect of cysteine desulfhydrase gene disruption on L-cysteine overproduction in Escherichia coli*, S.239-243

[302] Selby C.P., Witkin E.M., *Proc Natl Acad Sci U S A.* **1991**, *Escherichia coli mfd mutant deficient in mmutation frequency declinellacks strand-specific repair: in vitro complementation with purified coupling factor*, S.11574-11578

[303] Leimkühler S., Wuebbens M.M., *J Biol Chem.* **2001**, *Characterization of Escherichia coli MoeB and its involvement in the activation of molybdopterin synthase for the biosynthesis of the molybdenum cofactor*, S.34695-34701

Anhang

A Liste der Angaben in der Ergebnisdatei

Tabelle A.1 *– Liste aller der Tabellenblätter, die im Laufe der Auswertung in der Ergebnisdatei erstellt werden. In der ersten und zweiten Spalte ist die Kapitelnummer sowie der Name des Kapitels angegeben, in dem Informationen zum Hintergrund dieser Daten gegeben werden. Spalte drei enthält den Namen des Tabellenblatts in Excel und die letzte Spalte eine kurze Zusammenfassung der angegebenen Daten.*

Inhalt der Ergebnisdatei [.xls]			
Kapitelverweis	**Kapitelname**	**Tabellenblatt**	**Inhalt**
Kapitel 6.2.1	Benutzeroberfläche	*Datenblatt 1*	dies und das
		Datenblatt 2	die und das
Kapitel 6.2.2	Qualitätsanalyse	*Datenblatt 1*	dies und das
		Datenblatt 2	die und das
Kapitel	Vorverarbeitung	*Datenblatt 1*	dies und das
		Datenblatt 2	die und das

B Liste der Experimenten-Informationen in den Datenblättern

1. Beschreibung des Microarrays
   - Benutzername
   - Experimentname
   - Chiptype
   - Pfad zum Ordner mit den *.gal*-Files
   - Art des Spottens der Microarrays (selbst oder kommerziell)
     und wenn selbst-gespottet:
     - Datum des Spotting-Vorgangs
     - Array-File-Name
     - Oberflächenbeschaffenheit
     - Spotting-Puffer

  - Beschreibung des *Postprocessing*-Vorgangs
  - UV *crossing*
  - Art der Reduktion
  - Beschreibung des Blockens

2. Beschreibung der Proben

   - Quelle
   - Zellzahl (Zellen/ml)
   - RNA-Aufreinigungsverfahren
   - RNA-Aufreinigungskit
   - RNA-Aufreinigungsdatum
   - Konzentration (μg/μl)
   - Reinheit (260/280)
   - verwendete Menge (μg)
   - Agilent-Messung (ja/nein)
   - Nanodrop-Messung (ja/nein)
   - DNAseI Verdau (ja/nein)

3. Labeling und Priming

   - Labeling-Enzym
   - Labeling-Prozedur (direkt/indirekt)
   - Labeling-Moleküle (z.B. Cy5/Cy3)
   - Priming-Prozedur
   - cDNA-Prozessierung
   - cDNA-Nanodrop-Messung
   - Hybridisierungsdatum
   - Hybridisierungsdauer
   - Hybridisierungspuffer
   - Hybridisierungsabdeckung
   - Hybridisierungprozedur
   - Hybridisierungstemperatur
   - Hybridisierungsschüttelrate
   - Waschprozedur

4. Scannen

   - Pfad zu den (*gpr*)-Dateien

- Anzahl der Chips
- Anzahl der Scans pro Chip
- Scandetails
- Name der verwendeten (*gal*)-Datei

5. Technische Details für die Auswertung

- Chipdesign (*low-density* / *whole-genome*)
- Anzahl der Chips
- Anzahl der durchzuführenden Vergleiche
- Auflistung der zu berechnenden Vergleiche
- Anzahl der technischen Replikate
- Anzahl der Genreplikate
- Angaben zu den Zustandsnamen

C Tabellen der Microarrays 2-4: Summen der Intensitätswerte aller Spots einen Blocks innerhalb der Quartale auf den Microarrays.

Tabelle C.1 *– Visuelle Auswertung der Quartale dreier Microarrays*

Microarray 2					
Quartal I			**Quartal II**		
178.730	>	149.915	**183.626**	>	178.996
221.889	>	190.162	**193.520**	>	171.354
262.039	>	218.286	**229.969**	>	158.669
Quartal III			**Quartal IV**		
239.145	<	**277.105**	**181.881**	>	138.714
224.467	<	**241.589**	**177.648**	>	135.932
189.855	<	**203.957**	153.351	<	**215.887**

Microarray 3					
Quartal I			**Quartal II**		
186.274	<	**233.129**	**187.820**	>	181.922
211.406	<	**250.270**	189.844	<	**206.920**
236.411	<	**263.503**	205.129	<	**208.031**
Quartal III			**Quartal IV**		
110.391	<	**245.188**	**160.654**	>	157.892
98.782	<	**180.363**	**130.093**	>	110.391
91.837	<	**149.379**	113.380	<	**160.162**

Microarray 4

Quartal I		
176.211	<	**235.862**
140.637	<	**193.132**
189.120	<	**196.665**

Quartal II		
171.119	<	**189.249**
222.178	>	189.249
206.234	>	234.447

Quartal III		
95.519	<	**120.146**
130.849	>	124.441
145.210	>	118.667

Quartal IV		
131.914	>	131.746
102.018	<	**110.806**
102.457	>	101.534

.

D Häufigkeiten der sechs charakteristischen Spotform aus Abbildung 6.5

Tabelle D.1 *– Häufigkeiten der sechs charakteristischen Spotform in % sowie die Anzahl der vorkommenden Spots entsprechender Formen auf den untersuchten Microarrays*

Häufigkeitden der charakteristischen Spotform		
Spotform	Anzahl	Häufigkeit [%]
umrandete Ellipse	998 (1434)	69.6
umrandete Sichel	182 (1434)	12.7
Zylinder	154 (1434)	10.7
Donut	67 (1434)	4.7
Zylinder mit Loch	18 (1434)	1.3
zwei umrandete Ellipsen	15 (1434)	1.0

E Validierung der Hintergrund-Korrektur

Tabelle E.1 *– Validierung der Hintergrund-Korrektur - entsprechend der 1. Verdünnungsreihe in Tabelle 6.8 und der in Abbildung 6.9 gezeigten Graphik*

Verdünnungsreihe 2				
Verdünnung	$\overline{\sigma_{HG}} < \overline{\sigma_{Median}}$	**rel.** $\overline{\sigma}$	$\overline{N_{auswertbar}}$	**rel.** $\overline{\sigma}$
1:80	7	0.32	13	0.35
1:320	7	0.29	14	0.15
1:1280	9	0.34	14	0.26
1:5120	9	0.22	15	0.15

Verdünnungsreihe 3				
Verdünnung	$\overline{\sigma_{HG}} < \overline{\sigma_{Median}}$	**rel.** $\overline{\sigma}$	$\overline{N_{auswertbar}}$	**rel.** $\overline{\sigma}$
1:80	7	0.49	12	0.28
1:320	9	0.15	14	0.13
1:1280	7	0.44	13	0.34
1:5120	11	0.24	15	0.22
1:20480	7	0.21	15	0.05

.

Verdünnungsreihe 2 - Verdünnung 1:1280			
Scan-Einstellung	$\sigma_{HG} < \sigma_{Median}$	$N_{auswertbar}$	HG-Korrektur besser
33% - 400	3	6	
100% - 400	10	16	X
33% - 500	12	15	X
100% - 500	9	16	X
33% - 600	9	16	X
100% - 600	6	16	
33% - 700	8	16	X
100% - 700	7	15	
33% - 800	8	15	X
100% - 8400	7	11	X
33% - 1000	3	6	
100% - 1000	2	6	

F Beispiel möglicher Scan-Einstellungen

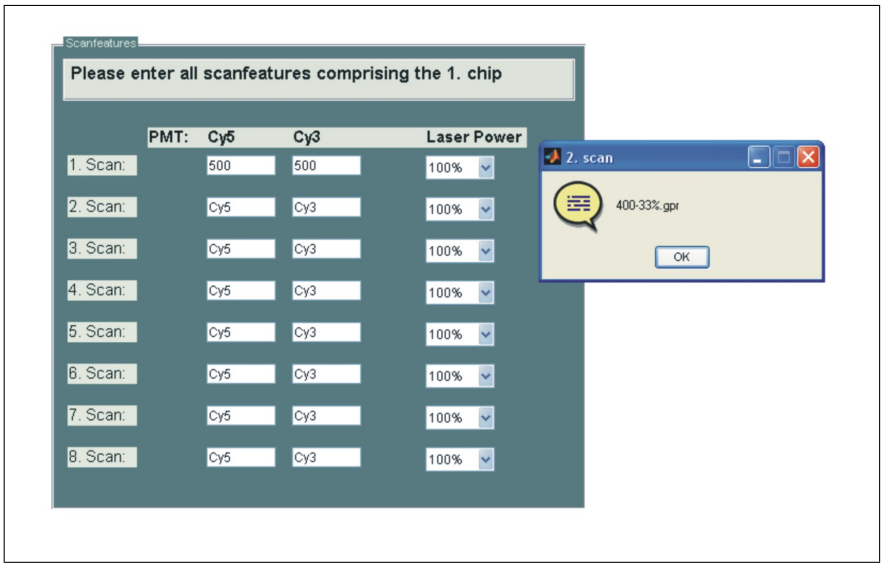


Abbildung F.1 – *Datenblatt zur Abfrage der Scan-Einstellungen eines Microarrays. Die Anzahl der möglichen Einstellungen wird automatisch der Anzahl eingegebener Primärdaten-Dateinamen für den entsprechenden Microarray angepasst. Mit der linken Maustaste kann zu jeder Einstellung der Name der Datei abgefragt werden, der üblicherweise die Scan-Einstellung beinhaltet. Diese Abfragemaske wird automatisch innerhalb der Datenblätter vor dem Experiment aufgerufen.*

Tabelle F.1 *– Beispiel möglicher Scan-Einstellungen für einen Chip eines Microarray-Versuchs*

Scan-Einstellung für einen Microarray		
Laserstärke	**PMT-Stärke Cy5**	**PMT-Stärke Cy3**
33%	500	500
	550	550
	600	600
	650	650
	700	700
	750	750
	800	800
	1048	800
100%	500	500
	550	550
	600	600
	650	650
	700	700
	750	750
	800	800
	1048	800

Tabelle F.2 *– Sortierung der Scan-Einstellung für einen Microarray*

Sortierung der Scan-Einstellung für einen Microarray								
Vor Sortierung			**1. Sortierung**			**2. Sortierung**		
Laser	**PMT Cy5**	**PMT Cy3**	**Laser**	**PMT Cy5**	**PMT Cy3**	**Laser**	**PMT Cy5**	**PMT Cy3**
33%	600	600	33%	600	600	33%	500	500
100%	1048	800		500	500		550	550
100%	500	500		1048	800		600	600
33%	500	500		550	550		650	650
100%	600	600		700	700		700	700
100%	700	700		800	800		750	750
33%	1048	800		750	750		800	800
33%	550	550		650	650		1048	800
33%	700	700	100%	1048	800	100%	500	500
33%	800	800		500	500		550	550
33%	550	550		600	600		600	600
33%	650	650		700	700		650	650
100%	550	550		550	550		700	700
100%	750	750		750	750		750	750
100%	650	650		650	650		800	800
100%	800	800		800	800		1048	800

G Material und Methoden zur Absoluten Quantifizierung

Um die Menge bestimmter RNA-Spezies in einer Probe absolut quantifizieren zu können, wurde ein *low-density*-Microarray bestehend aus 13 unterschiedlichen Oligonukleotiden entwickelt. Zu diesem Zweck wurden *E.coli* 50mer-Oligonukleotide im *sense*-Design von MWG Operon (Ebersberg) erstanden, die für die Gene LacY, LacA, DnaK, TatD, SecD, FtsH, ProA, Signalpeptidase2 und GlyA so wie vier künstliche Gene (RFP und drei unterschiedliche GFP-Arten, die komplementär an das gleiche Probenmolekül binden) kodieren und mit einem C6-Spacer sowie einer 5'-Amino-Modifikation versehen waren. Um lokale Einflüsse evaluieren zu können, wurde ein Gen (GlyA) in zweifacher Ausfertigung (mit jeweils 19 Genreplikaten) auf dem Chip aufgetragen.
Die Oligonukleotide wurden gemäß Kane *et al.* designt[166] und für die weitere Spotting-Prozedur verwendet. Die Oligonukleotide wurden gemäß der Vorgaben im Spotting-Protokoll 1:1 mit dem Spotting-Puffer verdünnt und auf die mit Aldehyd funktionalisierte Oberfläche unter Verwendung des Affymetrix 417 Kontaktprinters gedruckt. Die gespotteten Microarrays wurden für drei Minuten UV-crossverlinkt bei 254nm und anschliessend bei 80 °C zwei Stunden gebacken. Für die folgende Behandlung wurden die Chips mit Sodium-Borhydrid reduziert und mit BSA geblockt.
Auf jedem Microarray befanden sich schliesslich 19 räumlich verteilte Genreplikate für jedes Oligonukleotid (insgesamt also 280 Spots). Die so behandelten Microarrays wurden zur Hybridisierung mt komplementären Targetmolekülen verwendet.
Die Targetmoleküle wurden nach unterschiedlichen experimentellen Verfahren hergestellt. Im ersten experimentellen Teil sollten die Abhängigkeit des Signals von der Basenpaarlänge bzw. des dCTP-Gehalts untersucht werden. Zu diesem Zweck wurden die in Kapitel 6.2.4 in Tabelle 6.9 angegebenen Verdünnungsreihen hergestellt. Die auf die Microarrays hybridiserten Lösungen beinhalteten alle auf den Chips vorhandenen Gene.
Im Gegensatz dazu wurde für den zweiten experimentellen Teil, in dem die Signalabhängigkeit von Konzentration untersucht wurde, ein einzelnes Gen, Sig2 extrahiert und auf Chips in den Verdünnungen 1:10, 1:100, 1:1000 und 1:10000 hybridisiert.
Die Präparation der komplementären Proben bekannter Konzentrationen erfolgte in einer *in vitro*-Transkription. Dazu wurden 10 unterschiedliche *E.coli* Gene bzw. Untereinheiten der Gene von cDNA-Molekülen amplifiziert, in dem eine PCR-Reaktion mit einer Pfu-Polymerase angewandt wurde. Ausserdem wurden zwei künstliche Gene (RFP-Express und GFPmut2-Varianten) ausgewählt. Die PCR-Produkte mit Längen im Bereich zwischen 457 und 960 Basenpaaren wurden mit dem Qiagen-Purification Kit aufgereinigt und einem Blunt-End-Klonierungsschritt unterworfen (in den Promega-Vektor p6EM5Zf(+) zwischen die T7- und SPG-Promotorstelle). Die Vektorkosntrukte sowie die Orientierung der Inserts wurden durch einen restriktiven Verdau und eine Sequenzierung validiert.
In Abhängigkeit der Orientierung wurden unterschiedliche Restriktionsenzyme (Sac I, Nco I) zur Linearisierung der Vektorkontrukte verwendet, um eine *sense*-RNA zu erhalten. Auf diese Weise wurden mit den Restriktionsenzymen überhängende Enden erzeugt, so dass die unerwünschte Entstehung langer Transkripte von falschen Enden vermieden werden konnte.

Die linearisierten Vektore wurden in einer Phenol-Chloroform-Exktraktion gereinigt und die entsprechenden Präzipitate mit 5 M NaCl/Ethanol bei − 80 °C erzeugt.
Anschliessend, wurden die DNA-Pellets in Nuclease-freiem Wasser resuspendiert. 1 μg linearisierten Vektors wurden je während einer *in vitro*-Transkription unter Verwendung des *Ribiprobe in vitro Transcription Systems* von Promega entsprechend der Angaben der Hersteller unterzogen.
Die ungelabelte RNA wurde erneut in einer Phenol-Chloroform-Exktraktion gereinigt und mit 5 M NaCl/Ethanol bei 80 °C ausgefällt. Die in Nuklease-freiem Wasser resuspendierten RNA-Pellets wurden in einer reversen Trankription umgeschrieben. Während der reversen Transkription mit dem Enzym Superscript II (Invitrogen Karlsruhe) wurde Cy5-markierte dCTP-Moleküle in die cDNA eingebaut, so dass schliesslich komplementäre fluoreszierende *anti-sense*-DNA vorlagen.
Die Inkorporationsfrequenz (FOI) sowie die Menge an cDNA wurde mit dem NanoDrop Spectrophotometer (Nanodrop Technologies, Wilmington NC) bestimmt. Die cDNA wurde in einer Speed-Vac pelletiert und in 25 μl Hybridisierungspuffer resuspendiert. Von diesem Bestand wurden unterschiedliche Verdünnungsreihen entnommen und in einem Eppendorfschüttler bei 42 °C und 650 rpm auf die Microarrays hybridisiert.
Die Microarrays wurden dreimal je fünf Minuten mit dem Waschpuffer 1 (2 x SSC, 0.1% SDS), Waschpuffer 2 (1 x SSC) und dem Waschpuffer 3 (0.5 x SSC) gewaschen und mittels Zentrifugation getrocknet.
Die Chips wurden abschliessend bei unterschiedlichen PMT-Verstärkungen und Laserstärken mit dem konfokalen Axon 4000B Scanner (Axon Instruments, Foster City, CA) mit einer Auflösung von 10 μm pro Pixel gescannt. In Übereinstimmung mit vorherigen Versuchen bezüglich der notwendigen minimalen Anzahl an Scans pro Laserstärke, wurden mindestens vier verschiedene PMT-Verstärkungen sowohl bei 33% als auch bei 100% ausgewählt (z.B. 400, 500, 600 und 800 jeweils bei einer Laserstärke von 33% und 100%). Die während des Scanvorgangs entstandenen Bilddateien wurden mit GenePix Pro 6.0 (Axon Instruments, Inc., Union City, CA) in numerische Daten umgewandelt. Die Spots wurden quantifiziert, indem die Mediane der Fluoreszenzintensitäten aller im Spot enthaltenen Pixel berechnet und in einer GPR(GenePix Results)-Datei abgespeichert und der weiteren Auswertung übergeben wurden.

H Relative Standardabweichungen der Gene

Tabelle H.1 *– Relative Standardabweichung der Genreplikate - 1. Verdünnungsreihe. Mit Std = Standardauswerteverfahren und LR = neues Auswerteverfahren mit Linearer Regression.*

Relative Standardabweichung - 1. Verdünnungsreihe								
	1:80		**1:320**		**1:1250**		**1:5120**	
Gene	Std	LR	Std	LR	Std	LR	Std	LR
DnaK	0.081	0.189	0.164	0.062	0.226	0.140	0.369	0.035
GFP (EGFP)	0.173	0.022	0.408	0.070	0.190	0.090	0.357	0.024
FtsH	0.261	0.909	0.122	0.523	0.432	0.302	0.235	0.062
GFP (GFP)	0.124	0.174	0.068	0.032	0.081	0.067	0.356	0.041
GFP (GFPuv)	0.193	0.121	0.292	0.095	0.320	0.485	0.247	0.111
GlyA	0.374	0.564	0.369	0.558	0.420	0.815	0.559	0.235
GlyA	0.474	0.479	0.406	0.435	0.501	0.915	0.406	0.032
LacA	0.313	0.726	0.236	0.596	0.290	0.320	0.345	0.335
LacY	0.099	0.411	0.205	0.386	0.291	0.390	0.390	0.097
RFP	0.171	0.324	0.173	0.706	0.380	0.366	0.468	0.448
RpoA	0.284	0.794	0.100	0.499	0.324	0.141	0.163	0.170
SecD	0.236	0.398	0.410	0.329	0.233	0.640	0.298	0.086
Sig2	0.136	0.260	0.151	0.088	0.0522	0.441	0.233	0.080
TatD	0.161	0.545	0.081	0.588	0.376	0.263	0.421	0.478

Tabelle H.2 *– Relative Standardabweichung der Genreplikate - 2. Verdünnungsreihe. Mit Std = Standardauswerteverfahren und LR = neues Auswerteverfahren mit Linearer Regression.*

Relative Standardabweichung - 2. Verdünnungsreihe										
	1:80		**1:320**		**1:1250**		**1:5120**		**1:20480**	
Gene	Std	LR	Std	LR	Std	LR	Std	LR	Std	LR
DnaK	0.182	0.083	0.237	0.438	0.275	0.102	0.051	0.282	0.024	0.240
GFP(EGFP)	0.187	0.028	0.113	0.049	0.092	0.157	0.019	0.260	0.015	0.308
FtsH	0.204	0.685	0.511	0.856	0.481	0.628	0.162	0.443	0.022	0.283
GFP (GFP)	0.073	0.068	0.101	0.178	0.118	0.165	0.039	0.190	0.033	0.183
GFP (GFPuv)	0.190	0.397	0.292	0.571	0.704	0.150	0.044	0.140	0.056	0.334
GlyA	0.701	2.770	0.369	1.901	0.390	0.171	0.011	0.075	0.018	0.270
GlyA	0.429	1.836	0.247	2.108	0.548	0.481	0.028	0.164	0.032	0.313
LacA	0.526	1.035	0.373	0.921	0.440	0.218	0.160	0.176	0.021	0.152
LacY	0.463	0.739	0.189	0.141	1.221	0.233	2.253	0.363	0.026	0.200
RFP	0.214	0.662	0.274	0.925	0.912	0.516	0.069	0.157	0.142	0.436
RpoA	0.275	0.173	0.147	0.076	2.277	0.436	0.384	0.185	0.056	0.204
SecD	0.359	0.492	0.451	1.405	1.554	0.201	0.160	0.092	0.023	0.235
Sig2	0.203	0.318	0.260	0.586	0.268	0.242	0.122	0.256	0.023	0.158
TatD	0.213	0.494	0.164	0.349	0.246	0.314	0.346	0.428	0.073	0.308

Tabelle H.3 *– Relative Standardabweichung der Genreplikate - 3. Verdünnungsreihe. Mit Std = Standardauswerteverfahren und LR = neues Auswerteverfahren mit Linearer Regression.*

Relative Standardabweichung - 3. Verdünnungsreihe								
	1:80		**1:320**		**1:1250**		**1:5120**	
Gene	Std	LR	Std	LR	Std	LR	Std	LR
DnaK	0.217	0.089	0.199	0.164	0.363	0.657	0.263	0.027
GFP (EGFP)	0.227	0.019	0.175	0.027	0.346	0.063	0.289	0.018
FtsH	0.177	0.163	0.447	0.357	0.618	0.547	0.242	0.058
GFP (GFP)	0.130	0.071	0.068	0.202	0.112	0.100	0.129	0.024
GFP (GFPuv)	0.238	0.087	0.195	0.431	0.324	0.567	0.294	0.066
GlyA	0.430	0.140	0.612	0.678	0.460	0.662	0.283	0.025
GlyA	0.729	0.243	0.172	0.616	0.449	2.391	0.265	0.024
LacA	0.316	0.641	0.320	0.392	0.629	0.708	0.308	0.119
LacY	0.098	0.051	0.360	0.211	0.407	1.396	0.166	0.026
RFP	0.241	0.528	0.230	0.379	0.415	0.838	0.148	0.158
RpoA	0.215	0.267	0.183	0.059	0.230	1.495	0.170	0.066
SecD	0.156	0.189	0.411	0.800	0.448	1.670	0.263	0.051
Sig2	0.149	0.084	0.416	0.595	0.444	0.486	0.342	0.055
TatD	0.332	0.772	0.244	0.124	0.352	0.721	0.287	0.212

I Korrelation zwischen Erwartungswerten und Signalintensitäten

Tabelle I.1 – *Within-Array-Normalisierung: Quotienten aus* $\frac{Molekülzahl}{Signalintensität}$*. Die Quotienten der Verdünnungsreihen wurden gemittelt.*

Mittlere Quotienten aus RNA-Molekülzahl und normalisierter Signalintensität											
	Standardmethode						Implementierte Methode				
Gene	1:80	1:320	1:1280	1:5120	1:20480		1:80	1:320	1:1280	1:5120	1:20480
DnaK	1.28	0.27	0.07	0.04	0.01		3.00	0.76	0.16	0.15	0.32
GFP (EGFP)	20.31	1.89	0.27	0.16	0.03		8.43	1.92	0.45	0.20	0.39
FtsH	2.37	0.39	0.11	0.04	0.02		2.51	0.49	0.23	0.14	0.33
GFP (GFP)	1.32	0.20	0.05	0.03	0.01		3.67	0.92	0.23	0.17	0.34
GFP (GFPuv)	2.42	0.30	0.06	0.02	0.01		3.84	0.66	0.15	0.14	0.34
GlyA	2.45	0.61	0.12	0.05	0.01		1.71	0.36	0.18	0.13	0.29
GlyA	2.19	0.50	0.14	0.06	0.01		1.70	0.44	0.12	0.14	0.28
LacA	2.43	0.60	0.13	0.04	0.02		1.34	0.69	0.25	0.19	0.55
LacY	1.56	0.57	0.19	0.03	0.01		2.30	1.03	0.22	0.11	0.33
RFP	1.48	0.35	0.09	0.01	0.00		1.25	0.31	0.19	0.10	0.28
RpoA	4.92	1.47	0.35	0.04	0.01		4.69	1.76	0.29	0.20	0.47
SecD	2.31	0.63	0.10	0.02	0.01		2.85	0.52	0.11	0.11	0.30
Sig2	1.29	0.36	0.10	0.03	0.01		2.47	0.81	0.12	0.16	0.39
TatD	2.84	0.73	0.23	0.04	0.01		2.41	0.95	0.41	0.16	0.43

Tabelle I.2 – *Within-Array-Normalisierung: Quotienten aus* $\frac{AnzahldCTPs}{Signalintensität}$ *im Vergleich beider Auswertemethoden (siehe Kapitel 6.2.4). Die Quotienten der drei Verdünnungsreihen wurden gemittelt.*

Mittlere Quotienten aus der Anzahl dCTPs und normalisierter Signalintensität											
	Standardmethode						Implementierte Methode				
Gene	1:80	1:320	1:1280	1:5120	1:20480		1:80	1:320	1:1280	1:5120	1:20480
DnaK	0.67	0.57	0.57	1.43	1.21		0.66	0.71	0.49	2.05	5.29
GFP (EGFP)	6.53	2.36	1.34	3.08	1.94		1.04	0.96	0.95	1.55	3.91
FtsH	1.13	0.75	0.84	1.10	1.62		0.58	0.47	0.67	1.80	5.03
GFP (GFP)	0.42	0.26	0.23	0.67	0.40		0.48	0.48	0.45	1.36	3.34
GFP (GFPuv)	0.77	0.38	0.29	0.42	0.47		0.61	0.37	0.30	1.16	3.42
GlyA	1.79	1.74	1.42	2.32	2.29		0.82	0.58	0.83	2.40	6.67
GlyA	1.60	1.43	1.63	2.66	1.42		0.82	0.65	0.40	2.62	6.43
LacA	0.33	0.34	0.29	0.32	0.48		0.10	0.22	0.23	0.67	2.42
LacY	0.35	0.49	0.66	0.44	0.41		0.27	0.37	0.23	0.79	2.38
RFP	0.71	0.67	0.67	0.39	0.33		0.25	0.34	0.50	1.24	4.31
RpoA	0.78	0.94	0.92	0.44	0.22		0.26	0.47	0.30	0.88	2.45
SecD	1.29	1.47	0.97	0.63	1.40		0.59	0.57	0.44	1.82	5.70
Sig2	0.42	0.46	0.55	0.64	0.68		0.40	0.39	0.27	1.44	4.07
TatD	0.62	0.64	0.78	0.61	0.45		0.18	0.38	0.46	0.88	3.06

J Microarray-Experiment mit Schwannzellen

Der experimentelle Teil der Microarray-Versuche wurde von Yvonne Stark durchgeführt.

J.1 Materialien für Zellkulturexperimente

Alle Lösungen wurden mit deionisiertem Wasser (dH_2O) (Arium, Sartorius AG, Göttingen, Deutschland) durchgeführt. Dulbecco's Modified Eagle's medium (DMEM, D7777), nerve growth factor (NGF) and Poly-L-lysin (PLL) wurden von Sigma-Aldrich (Steinheim, Deutschland) erworben. Foetal calf serum (FCS), Newborn calf serum (NCS), L-Glutamin, Natriumpyruvat und die Antibiotika wurden bei PAA Laboratories GmbH (Cölbe, Deutschland) gekauft und das Horse serum (HOS) bei Invitrogen (Karlsruhe, Deutschland). Die Puffer, Salze und andere Reagenzien stammten von Fluka (Buchs, Schweiz) und Sigma-Aldrich (Steinheim, Deutschland) und waren von höchster Reinheit (*per analysis quality*). Die Zellen wurden in Zellkultur-Flaschen von Sarstedt & Co (Nümbrecht, Deutschland) kultiviert.

J.2 Zellkultivierung

J.2.1 iSZ

Die immortalisierten Schwannzellen (iSZ) wurden in DMEM in einer feuchten Atmosphäre bei 37 °C und 5 % CO_2 in Gewebekulturflaschen kultiviert und mit 10 % FCS, Penicillin (100 U/ml), Streptomycin (100 $\mu g/ml$), 2 mM L-Glutamin and 1 mM Natriumpyruvat angereichert.

J.2.2 Beschichtungsprozeß

Um die Unterschiede der Beschichtung der Oberflächen detektieren zu können, wurden Microarray-Experimente durchgeführt, die unter Anwendung des t-Test die differentielle Expression der Gene beschreibt, nach dem die Zellen auf unterschiedlich beschichteten Matrices gewachsen sind.
Hierfür wurde CA mit unterschiedlichen Referenz-Materialien unter etablierten Standardkulturbedingungen verglichen. Laminin wurde wegen seiner nachgewiesenen Fähigkeit, eine erhöhte Viabilität und Anhaftung neuronaler Zellen an die Oberfläche der Zellkulturflaschen einzuleiten, als positive Kontrolle verwendet. Die unbeschichtete Oberfläche der Zellkulturflaschen diente Standardkulturbedingung.
Die Zellkulturflaschen (75 cm^2) wurden also mit jeweils 2 ml Colominsäure (CA) (133 $\mu g/cm^2$), Poly-L-Lysin (0.13 $\mu g/cm^2$) und Laminin (1 $\mu g/cm^2$) bedeckt. Nach einer Stunde Inkubation bei RT wurden die Zellkulturflaschen zweimal mit PBS gewaschen und die Zellen ausgesät.

J.2.3 Zellkultivierung für Microarray-Experimente

Etwa 1.5 - 3 · 10^6 iSZ wurden parallel auf den drei unterschiedlich behandelten Zellkulturoberflächen unter Standardbedingungen kultiviert. Nach vier Tagen wurde die gesmte RNA der iSZ isoliert und für die Microarray-Experimente präpariert.

J.3 Microarray-Experimente

J.3.1 Entwicklung der Ratten-spezifischen neuronalen Microarrays

Ein neuraler Rattenspezifischer Microarray mit 352 Genen (low density Microarray) wurde entwickelt, um regulierte Gene auf dem Genexpressionslevel detektieren zu können. Der Chip ermöglicht so die transkriptionelle Analyse neural-spezifischer Proteine, wie beispielsweise von Neurofilamenten (z.B. NF-F, NF-H, NF-M), Rezeptoren (z.B. GABA-Rezeptoren, Glyzin-Rezeptoren, Dopamin-Rezeptoren) oder Neurotransmitter-ähnlichen Enzymen (z.B. GAD, Chat, Tyrosinhydroxylase) zur Evaluation neuraler und glialer Zellen in unterschiedlichen Wachstums- und Differenzierungsphasen. Die komplette Liste der Gene auf den Arrays ist im Internet erhältlich (http://www.ncbi.nlm.nih.gov/geo/). Unter der Verwendung automatisierter und modifizierter Versionen kommerziell erhältlicher Software-Pakete wurden die Oligonukleotide mit optimalen Sequenzen und Hybridisierungsmustern designt, die innerhalb kodierender Regionen lagen und entsprechend der Kriterien von Kane *et al.* (2000) ermittelt wurden.[165]
Die designten Oligonukleotide wurden über Blast-Analysen mit humanen kodierenden Regionen verglichen. Die Ähnlichkeit der Sequenzen zu den „Nicht-Target"-Sequenzen betrug weniger als 75 %.
Die 50mere wurden bei MWG (ebersberg, Deutschland) gekauft und mit Hilfe eines Affymetrix 417 Arrayer auf die modifizierte Aldehyd-Oberfläche (VSS25, CEL Associates, Inc., Pearland, TX , USA) gespottet.

J.3.2 RNA-Isolation

Nach der Zellkultivierung und den oben erwähnten Bedingungen wurde die RNA mit dem Invitrogen TRIzol Reagent® (Invitrogen, Calsbad, Deutschland) isoltiert. Dazu wurde die konfluente Zellschicht zunächst mit Trypsin/EDTA gelöst. Die Zellen wurden in Kulturmedium resuspendiert und 4 Minuten bei 4000 g zentrifugiert, um Zellpellets zu erhalten. Nach der Entfernung des Medium-Überstandes wurde das Zellpellet in 1 ml Trizol Reagent® resuspendiert und homogenisiert im Ultra-Turrax (IKA-Werke, Staufen, Deutschland). Nach der Entfernung von telltrümmern durch erneute Zentrifugation wurde die RNA in 500 μl Iso-

propanol durch eine Phasentrennung mit Chloroform ausgefällt. Die ausgefällte RNA wurde mit Ethanol (75%) gewaschen, luftgetrocknet und schliesslich in DEPC-Wasser resuspendiert. Die Konzentration der RNA wurden mit Spektralphotometer Nano-Drop-1000® bei 260 nm analysiert.

J.3.3 cDNA-Synthese und Reinigung

Für die cDNA Synthese wurde das NEN® Micromax TSA Labeling and Detection Kit (PerkinElmer, Rodgau - Jügesheim, Deutschland) verwendet. Während der Reversen Transkription wurden 6 μg der gereinigten Gesamt-RNA in Fluorescin- (Fl) und Biotin-gelabelte cDNA umgeschrieben. Die Reinigung der Fluorescin- und Biotin-gelabelten cDNA wurden mit dem QIAquick PCR Purification Kit entsprechend der Angaben der (Qiagen, Hilden, Deutschland) durchgeführt. Ausserdem wurde die cDNA mit 35 %iger Guanidin-Hydrochlorid-Lösung gereinigt und zweimal im EB-Puffer (1:10 Verdünnung, pH 8.5) eluiert.

J.3.4 Hybridisierung

Die Hybridisierung wurde in einer speziellen Form des *Dye-swaps*, dem *Loop*-Design durchgeführt (siehe Kapitel 5.3).
Die gereinigte und gelabelte cDNA zweier Proben wurde also je in einem Experiment über Nacht bei 42 °C auf den gleichen Chip hybridisiert.

J.3.5 Waschen und Detektion

Nach der Hybridisierung, wurde die ungebundene cDNA durch Waschen von den Arrays entfernt. Spezifisch gebundene, gelabelte cDNA-Moleküle wurden anschliessend mit anschliessend durch konjugate Reportermoleküle entsprechend dem TSA-System (Tyramid-Cy3 and Tyramid-Cy5) detektiert. Dazu wurde der DNA-chip zunächst mit Anti-Fl-horseradish Peroxidase (HRP) - Antiköper-Enzym-Konjuagten inkubiert, die spezifisch an hybridisierte Fl-gelabelte cDNA-Moleküle binden. Die Reaktion resultiert in einer sofortigen Anlagerung der Cy3-Labels, die sich in unmittelbarer Nähe zu den immobilisierten HRP-Molekülen befinden. Dadurch wird die Menge an Tyramid (Cy3) relativ zum cDNA-Hapten (Fluorescin) mehrfach amplifiziert. Nach der Inaktivierung der restlichen HRP-Moleküle, bindet das zugefügte Streptavidin-HRP an die Biotin-gelabelte cDNA und katalysiert so die Freisetzung der Cy5-gelabelten Tyramid-Amplifizierungs-Reagenzien.

J.3.6 Scannen

Die hybridisierten Arrays werden anschliessend mit dem Axon 4000B (Axon Instruments, Inc., Union City, CA) Laserscanner bei jeweils acht unterschiedlichen Einstellungen gescannt.

K Clusteranalysen mit *E.coli*-Microarrays

Tabelle K.1 *– Clustergene, Teil 1*

Blattner	Name	MWG
B1901	araf	mwgecov2#1802
B0050	apag	mwgecov2#0046
B0147	yadp	mwgecov2#0140
B0396	araj	mwgecov2#0368
B0758	galt	mwgecov2#0702
B1898	arah	mwgecov2#1799
B1898	arah	mwgecov2#1800
B1900	arag	mwgecov2#1801
B2148	mglc	mwgecov2#2030
B2150	mglb	mwgecov2#2032
B2840	ygea	mwgecov2#2688
B2841	arae	mwgecov2#2689
B2948	yqge	mwgecov2#2789
B3296	rpsd	mwgecov2#3118
B4227	ytfq	mwgecov2#3992
B0170	tsf	mwgecov2#0162
B0753	b0753	mwgecov2#0697
B0757	galk	mwgecov2#0701
B0759	gale	mwgecov2#0703
B0881	ylja	mwgecov2#0822
B0884	infa	mwgecov2#0824
B0924	mukb	mwgecov2#0864
B1652	rnt	mwgecov2#1564
B2149	mgla	mwgecov2#2031
B2424	cysu	mwgecov2#2294
B3357	crp	mwgecov2#3179
B3433	asd	mwgecov2#3253
B3919	tpia	mwgecov2#3708
B0074	leua	mwgecov2#0070
B0430	cyoc	mwgecov2#0402
B0463	acra	mwgecov2#0433
B0517	ylbc	mwgecov2#0487
B1090	plsx	mwgecov2#1021
B1114	mfd	mwgecov2#1044-r
B1676	pykf	mwgecov2#1586
B1825	b1825	mwgecov2#1731
B2350	b2350	mwgecov2#2228
B2379	b2379	mwgecov2#2257
B2477	nlpb	mwgecov2#2347
B2752	cysd	mwgecov2#2605
B2784	rela	mwgecov2#2637
B3307	rpsn	mwgecov2#3129

Tabelle K.2 *– Clustergene, Teil 2*

Blattner	Name	MWG
B3652	recg	mwgecov2#3462
B3686	ibpb	mwgecov2#3495
B3816	cora	mwgecov2#3612
B3941	metf	mwgecov2#3729
B4039	ubic	mwgecov2#3813
B4047	yjbl	mwgecov2#3821
B0030	yaaf	mwgecov2#0027
B0082	yabc	mwgecov2#0078
B0114	acee	mwgecov2#0107
B0346	mhpr	mwgecov2#0324
B0409	secf	mwgecov2#0381
B0413	ybad	mwgecov2#0385
B0438	clpx	mwgecov2#0410
B0854	potf	mwgecov2#0796
B0954	faba	mwgecov2#0893
B0958	sula	mwgecov2#0897
B1033	ycdw	mwgecov2#0966
B1066	rimj	mwgecov2#0997
B1209	hemm	mwgecov2#1138
B1236	galu	mwgecov2#1163
B1263	trpd	mwgecov2#1190
B1479	sfca	mwgecov2#1400
B1582	b1582	mwgecov2#1494
B1609	rstb	mwgecov2#1521
B1723	pfkb	mwgecov2#1633
B2013	yeee	mwgecov2#1900
B2239	glpq	mwgecov2#2118
B2498	upp	mwgecov2#2368
B2607	trmd	mwgecov2#2469-r
B2742	nlpd	mwgecov2#2596
B3008	metc	mwgecov2#2846
B3168	infb	mwgecov2#3000
B3255	accb	mwgecov2#3084
B3294	rplq	mwgecov2#3116
B3559	glys	mwgecov2#3370
B3570	bax	mwgecov2#3381
B3687	ibpa	mwgecov2#3496
B3766	ilvl	mwgecov2#3567
B3946	talc	mwgecov2#3734
B3984	rpla	mwgecov2#3762
B4052	dnab	mwgecov2#3826
B4178	yjeb	mwgecov2#3944

Tabelle K.3 *– Clustergene, Teil 3*

Blattner	Name	MWG
B0134	panb	mwgecov2#0127
B0211	dnir	mwgecov2#0197
B0220	ykfe	mwgecov2#0205
B0337	coda	mwgecov2#0315
B0480	usha	mwgecov2#0450
B0723	sdha	mwgecov2#0674
B0773	ybhb	mwgecov2#0717
B0912	himd	mwgecov2#0852
B0926	ycbk	mwgecov2#0866
B0970	ycca	mwgecov2#0909
B1088	yced	mwgecov2#1019
B1252	tonb	mwgecov2#1179
B1448	b1448	mwgecov2#1371
B1857	yebl	mwgecov2#1762
B2042	galf	mwgecov2#1927
B2087	gatr	mwgecov2#1973
B2240	glpt	mwgecov2#2119
B2369	evga	mwgecov2#2247
B2425	cysp	mwgecov2#2295
B2464	tala	mwgecov2#2334
B2471	yffb	mwgecov2#2341
B2678	prow	mwgecov2#2537
B2943	galp	mwgecov2#2785
B3236	mdh	mwgecov2#3065-r
B3319	rpld	mwgecov2#3141
B3347	fkpa	mwgecov2#3169
B3364	yhfc	mwgecov2#3186
B3611	yibn	mwgecov2#3421
B3623	rfak	mwgecov2#3433
B3638	radc	mwgecov2#3448
B3644	yicc	mwgecov2#3454
B3734	atpa	mwgecov2#3542
B3762	yifda	mwgecov2#3564
B4056	yjbq	mwgecov2#3830
B4069	acs	mwgecov2#3843
B4228	ytfr	mwgecov2#3993
B4252	yjgk	mwgecov2#4015
B3049	glgs	mwgecov2#5125
B0009	mog	mwgecov2#0009
B0025	ribf	mwgecov2#0022
B0032	cara	mwgecov2#0029
B0075	leul	mwgecov2#0071

Tabelle K.4 *– Clustergene, Teil 4*

Blattner	Name	MWG
B0077	ilvi	mwgecov2#0073
B0095	ftsz	mwgecov2#0091
B0102	yacf	mwgecov2#0096
B0143	pcnb	mwgecov2#0136
B0146	sfsa	mwgecov2#0139
B0161	htra	mwgecov2#0153
B0169	rpsb	mwgecov2#0161
B0179	lpxd	mwgecov2#0171
B0280	yagn	mwgecov2#0259
B0281	intf	mwgecov2#0260
B0388	arol	mwgecov2#0362
B0524	ybbf	mwgecov2#0494
B0557	ybcu	mwgecov2#0522
B0578	nfnb	mwgecov2#0543
B0631	ybed	mwgecov2#0595
B0640	hola	mwgecov2#0604
B0641	rlpb	mwgecov2#0605
B0657	lnt	mwgecov2#0620
B0741	pal	mwgecov2#0692
B0760	modf	mwgecov2#0704
B0822	b0822	mwgecov2#0765
B0827	moea	mwgecov2#0769
B0847	b0847	mwgecov2#0789
B0889	lrp	mwgecov2#0829
B0908	aroa	mwgecov2#0848
B0953	rmf	mwgecov2#0892
B0982	yccy	mwgecov2#0920
B1062	pyrc	mwgecov2#0993
B1093	fabg	mwgecov2#1024
B1132	ycfc	mwgecov2#1062
B1152	b1152	mwgecov2#1082
B1187	fadr	mwgecov2#1116
B1216	chaa	mwgecov2#1145-r
B1245	oppc	mwgecov2#1172
B1260	trpa	mwgecov2#1187
B1261	trpb	mwgecov2#1188
B1264	trpe	mwgecov2#1191
B1380	ldha	mwgecov2#1304
B1412	acpd	mwgecov2#1335
B1461	ydce	mwgecov2#1383
B1538	dcp	mwgecov2#1457
B1583	b1583	mwgecov2#1495

Tabelle K.5 – *Clustergene, Teil 5*

Blattner	**Name**	**MWG**
B1584	speg	mwgecov2#1496
B1658	purr	mwgecov2#1570
B1738	cela	mwgecov2#1648
B1852	zwf	mwgecov2#1757
B2023	hish	mwgecov2#1909
B2025	hisf	mwgecov2#1911
B2028	ugd	mwgecov2#1914
B2282	nuoh	mwgecov2#2161
B2287	nuob	mwgecov2#2166
B2413	cysz	mwgecov2#2283
B2524	yfhj	mwgecov2#2393
B2529	b2529	mwgecov2#2398
B2597	yfia	mwgecov2#2460
B2606	rpls	mwgecov2#2468-r
B2679	prox	mwgecov2#2538
B2900	yqfb	mwgecov2#2744
B2977	glcg	mwgecov2#2817
B3036	ygia	mwgecov2#2874
B3228	sspb	mwgecov2#3057
B3300	prla	mwgecov2#3122
B3340	fusa	mwgecov2#3162
B3356	yhfa	mwgecov2#3178
B3454	livf	mwgecov2#3272
B3457	livh	mwgecov2#3275
B3556	cspa	mwgecov2#3367
B3639	dfp	mwgecov2#3449
B3654	yice	mwgecov2#3464
B3672	ivbl	mwgecov2#3481
B3733	atpg	mwgecov2#3541
B3769	ilvm	mwgecov2#3569
B3792	wzxe	mwgecov2#3592
B3844	ubib	mwgecov2#3640
B3860	dsba	mwgecov2#3651
B3932	hslv	mwgecov2#3720
B3939	metb	mwgecov2#3727
B3956	ppc	mwgecov2#3743
B4214	cysq	mwgecov2#3979
B4254	argi	mwgecov2#4017

L Lebenslauf

PERSÖNLICHE INFORMATIONEN

Name	Cornelia Repenning
Geburtsdatum	31.05.1979
Geburtsort	Kiel
Anschrift	Liebigstrasse 10, 30163 Hannover
Telefon	0511/ 1244251 o. 0178/ 5124023
e-mail	Repenning.Cornelia@mh-hannover.de
Familienstand	ledig

SCHULISCHE AUSBILDUNG

August 1985 - Juli 1989	Grundschule Wohltorf
August 1989 - Juli 1998	Sachsenwald Gymnasium Reinbek, Abschluss Abitur

STUDIUM UND BERUFLICHE AUSBILDUNG

Oktober 1999 - Oktober 2005	Studiengang Life-Science an der Leibniz Universität Hannover, Vordiplom Chemie 2001, Bachelor Life-Science 2003
Oktober 2005	Abschluss des Studiums Life-Science als Master of Science, Note: ausgezeichnet (1.0) - mit Auszeichnung
seit Januar 2006	Promotionsarbeit *Statistische Auswertung von Microarrays* zum Dr. rer. nat. unter der Leitung von Prof. Dr. Thomas Scheper

FORTBILDUNGEN

Microarray Technology and Bioinformatics,Summer School, Camerino, Italien

Fortbildungslehrgang mit Abschluss zur *Betriebsbeauftragten für Gewässerschutz*

Fortbildungslehrgang mit Abschluss zur *Qualitätsmanagement/GMP-Beauftragten*

Fortbildungslehrgang GenTSV *Gentechnische Sicherheit*

Ausbildungslehrgang *Marketing-Seminar*

VERÖFFENTLICHUNGEN

Aust MC, Reimers K, Repenning C, Stahl F, Jahn S, Guggenheim M, Schwaiger N, Gohritz A, Vogt PM. Plast Reconstr Surg. **2008**, *Percutaneous Collagen Induction: Minimally Invasive Skin Rejuvenation without Risk of Hyperpigmentation-Fact or Fiction?* 122(5):S. 1553-1563

Heine F, Stahl, F. Sträuber, H., Wiacek C., Benndorf D., Repenning, C., Schmidt, F., Scheper, T., von Bergen, M., Harms H, Mueller. S., Cytometry: PartA **2008**, *Prediction of Flocculation Ability of Brewing Yeast Inoculates by Flow Cytometry, Proteome Analysis and mRNA Profiling*, 75(2), S.140-147

Repenning, C., Reck, M., Stahl, F., Scheper, T., Hitzmann, B., Biospektrum 2009, *Aufgaben und Verfahren zur Mikroarray Analyse*, 6, 646-648

Aust MC, Reimers K, Gohritz A, Jahn S, Stahl F, Repenning C, Scheper T, Altintas MA, Schwaiger N, Redeker J and others. Clin Exp Dermatol **2010**, *Percutaneous collagen induction. Scarless skin rejuvenation: fact or fiction?* 35(4), S.437-439.

A. T. Fleck, T. Nye, C. Repenning, F. Stahl, M. Zahn, and M. K. Schenk, J. Exp. Bot. **2010**, *Silicon enhances suberization and lignification in roots of rice (Oryza sativa)*, S.1-11

Aust MC, Reimers K, Kaplan HM, Stahl F, Repenning C, Scheper T, Jahn S, Schwaiger N, Ipaktchi R, Redeker J and others. J Plast Reconstr Aesthet Surg. **2011**, *Percutaneous collagen induction-Regeneration in place of cicatrisation?* 64(1), S.97-107