

# Microarrayanalyse gynäkologischer Tumorentitäten

## Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen-Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

**Martin Koch**

aus

Ruda-Slaska, Polen

Bonn 2013

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät  
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Michael Wiese
2. Gutachter: Prof. Dr. Martin Hofmann-Apitius

Tag der Promotion: 18.06.2013

Erscheinungsjahr: 2013

Bis zur Unendlichkeit und noch viel weiter!

*Buzz Lightyear*



# Inhaltsverzeichnis

<b>Inhaltsverzeichnis</b>	<b>i</b>
<b>Abkürzungen</b>	<b>v</b>
<b>Abbildungsverzeichnis</b>	<b>viii</b>
<b>Tabellenverzeichnis</b>	<b>x</b>
<b>1 Einleitung</b>	<b>1</b>
1.1 Genexpressionsprofile in der Krebsforschung . . . . .	1
1.2 Microarraytechnologie . . . . .	2
1.2.1 Zweikanal-Microarraytechnologie . . . . .	2
1.2.2 Einkanal-Microarraytechnologie . . . . .	4
1.3 Personalisierte Medizin . . . . .	5
1.3.1 Softwareentwicklung . . . . .	6
1.3.2 Qualitätssicherung von Microarray Daten . . . . .	6
1.3.3 Wirkstoffstudie . . . . .	7
1.3.4 Tumorprogressionsstudie . . . . .	8
1.3.5 Tumor-Subtypisierungsstudie . . . . .	9
<b>2 Zielsetzung dieser Arbeit</b>	<b>11</b>
<b>3 Material und Methoden</b>	<b>15</b>
3.1 Softwareentwicklung für die Analyse von Microarraydaten . . . . .	15
3.1.1 Die Statistiksprache R . . . . .	15
3.1.2 Das Bioconductorprojekt . . . . .	15
3.1.3 Die Software Entwicklungsumgebungen NetBeans und Eclipse .	16
3.2 Statistische Methoden und Maschinelle Lernverfahren . . . . .	18
3.2.1 Statistische Methoden und Tests . . . . .	18
3.2.2 Statistische Klassifikatoren . . . . .	21
3.2.3 Klassifikation über maschinelle Lernverfahren . . . . .	21

3.2.4	Methoden der Microarrayanalyse . . . . .	26
3.3	Microarray-Auswertung und -Visualisierung . . . . .	30
3.3.1	Microarrayarbeitsfluss . . . . .	30
3.3.2	Qualitätssicherung von Microarraydaten . . . . .	32
3.3.3	Das „Microarray Analysis and Reporting Tool: maRt“ . . . . .	33
3.4	Wirkstoffstudie mit liposomalem Cisplatin . . . . .	38
3.4.1	Herstellung der Liposomen . . . . .	38
3.4.2	Zellkultur . . . . .	39
3.4.3	MTT-Zellzytotoxizitäts-Test . . . . .	39
3.4.4	Messung der DNA Platinierung . . . . .	40
3.4.5	RNA Isolierung und Microarraydaten Gewinnung . . . . .	40
3.5	Tumorprogressionsstudie anhand von Zervixkarzinomen . . . . .	42
3.5.1	Microarraydaten . . . . .	42
3.5.2	Zervixkarzinompatienten Charakteristika . . . . .	42
3.5.3	Datenvorverarbeitung . . . . .	43
3.6	Triple-negativ Brustkrebs Subtypisierungsstudie . . . . .	44
3.6.1	Microarraydaten aus öffentlichen Brustkrebs-Studien . . . . .	44
3.6.2	Brustkrebspatienten . . . . .	44
3.6.3	Daten Vorverarbeitung . . . . .	45
<b>4</b>	<b>Ergebnisse und Diskussion</b>	<b>47</b>
4.1	Softwareentwicklung . . . . .	47
4.1.1	Das Microarray-Analysewerkzeug maRt . . . . .	47
4.1.2	Qualitätssicherung von Microarraydaten . . . . .	48
4.1.2.1	Erhebung der Qualitätsparameter diverser in GEO ver- fügbaren Studien . . . . .	48
4.1.2.2	Visualisierung der Qualitätsparameter mit Circos . . . . .	51
4.2	Microarray Analysen . . . . .	55
4.2.1	Liposomales Cisplatin Wirkstoffstudie . . . . .	55
4.2.1.1	DNA Platinierungseffizienz von freiem und liposoma- lem CDDP in Ovarialkarzinom Zellen . . . . .	55
4.2.1.2	Die Aktivität des freien und liposomalem CDDP unter Einfluss des Histon-Deacetylase Inhibitors TSA . . . . .	56
4.2.1.3	Analyse der Microarraydaten von A2780cis Zellen er- laubt Einblick in die Resistenz Mechanismen . . . . .	57
4.2.2	Abschließende Diskussion der Wirkstoffstudie . . . . .	68
4.2.3	Resultate der Zervixkarzinom Tumor Progressionsstudie . . . . .	70

4.2.3.1	Qualitätskontrolle und Normalisierung der Studien aus GEO führte auf eine Patientendatenmatrix . . . . .	70
4.2.3.2	Varianz Komponentenanalyse der Zervixkarzinom Stadien . . . . .	72
4.2.3.3	Konstitutiv über alle Studien exprimierte, wenig heterogenen Proben . . . . .	74
4.2.3.4	GSEA des aus der Varianz-Komponentenanalyse erhaltenen Genexpressions Sets . . . . .	75
4.2.4	Abschließende Diskussion der Zervixkarzinom Progressionsstudie	80
4.2.5	Triple-negativ Brustkrebs Subtypisierungsstudie . . . . .	84
4.2.5.1	Qualitätskontrolle und Normalisierung führen auf eine molekulare subtypspezifische Genexpressionsmatrix . .	84
4.2.5.2	Hauptkomponentenanalyse aller Brustkrebs Studien . .	85
4.2.5.3	Klassifizierung der Expressionsprofile von Referenz Brustkrebs gegen triple-negative Brustkrebs . . . . .	85
4.2.5.4	GSEA der Genexpressionsprofile von triple-negativen Patienten, die über die Klassifikation mit dem Random Forest erhalten wurden . . . . .	87
4.2.5.5	Entwicklung der FECSnV Methode . . . . .	87
4.2.5.6	GSEA der molekularen subtypspezifischen Signaturen, die mit der FECSnV Methode erhalten wurden . . . .	92
4.2.5.7	Klinische Endpunkt-Analyse mit dem Webservice GOBO	96
4.2.5.8	Consensus-Clusteranalyse der triple-negativen Brustkrebs Genexpressionsprofile . . . . .	98
4.2.5.9	Klinische Endpunkt Analyse der potentiellen triple-negativen Subtyp Signaturen mit GOBO . . . . .	99
4.2.5.10	GSEA der potentiellen triple-negativen Subtypen . . .	101
4.2.6	Abschließende Diskussion der Tumor Subtypisierungsstudie . . .	104
<b>5</b>	<b>Zusammenfassung</b>	<b>109</b>
<b>6</b>	<b>Summary</b>	<b>115</b>
<b>7</b>	<b>Appendix</b>	<b>119</b>
7.1	Genexpressions Studien aus “Gene Expression Omnibus” (GEO) . . . .	119
7.2	Wirkstoff Studie . . . . .	119
<b>8</b>	<b>Literaturverzeichnis</b>	<b>131</b>

<b>Danksagung</b>	<b>149</b>
<b>Publikationen</b>	<b>151</b>
<b>Verfassererklärung</b>	<b>153</b>

# Abkürzungen

**aCGH** array Comparative Genomic Hybridisation

**CC** Consensus Clustering

**cDNA:** copyDesoxyriboNucleinAcid

**CDDP** cis-Diamindichloridoplatin(II)

**CRAN** Comprehensive R Archive Network

**CY3** Cyanin-3

**CY5** Cyanin-5

**DNA** DesoxyriboNucleinAcid

**DAVID** Database for Annotated Visualization and Integrated Discovery

**DDA** Diagonale Diskriminanzanalyse

**DLA** Lineare Diskriminanzanalyse

**EMF** Eclipse Modelling Framework

**FDA** Fischers Diskriminanzanalyse

**FDR** False Discovery Rate

**FECSnV** Feature Extraktion via Composite Scoring and Voting

**FIGO** Federation International of Gynaecology and Obstetrics

**FWER** Family Wise Error Rate

**GBM** Gradient Boosting Maschine

**GOBO** Gene expression-based outcome for breast cancer online

**GPL** Gnu Public License

**GSEA** Gene Set Enrichment Analysis

**HER2** Humaner epidermaler Wachstumsfaktor 2

**HPV** Humaner Pappiloma Virus

**HTML** Hyper Text Markup Language

**IVT** In Vitro Transcription

**Limma** Linear Models for Microarray Analysis

**mRNA: messengerRiboNucleinAcid**

**MAQC** Microarray Quality Control Project

**maRt** Microarray Reporting Tool

**MTT** 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazoliumbromid

**OSGi** Open Services Gateway initiative

**PCA** Principal Component Analysis

**PLS-LDA** Partial Least Squares Lineare Diskriminanzanalyse

**RBF** Radiale Basisfunktion

**RCP** Rich Client Plattform

**RF** Random Forest

**RFE** Rekursive Feature Elimination

**SNP** Single Nucleotide Polymorphism

**SVM** Support Vector Machine

**SVN** Subversion

**TDD** Test driven development

# Abbildungsverzeichnis

1.1	Zweikanal-Microarray Datenerhebung . . . . .	3
1.2	Einkanal-Microarraytechnologie . . . . .	4
1.3	Schema der Personalisierten Medizin . . . . .	5
3.1	Schema des FECSnV Algorithmus . . . . .	23
3.2	Resultat der Consensus-Clustering Analyse . . . . .	25
3.3	GO Term Analyse . . . . .	28
3.4	Pathway Analyse mit Genexpressions Daten . . . . .	30
3.5	Die Gewinnung von Microarraydaten . . . . .	31
3.6	HTML-Report einer Microarrayanalyse . . . . .	36
3.7	Eine Heatmap stellt Genexpressions Daten graphisch dar . . . . .	37
4.1	Abbildung der ersten beiden Hauptkomponenten . . . . .	48
4.2	RNA-Degradationskontrolle . . . . .	49
4.3	Darstellung der Qualitätsparameter des Paketes <i>yaqcaffy</i> . . . . .	50
4.4	Legende zur Abbildung der Qualität mit Circos . . . . .	51
4.5	Abbildung der Qualitätskontrollen mit Circos an ausgewählten Datensets	53
4.6	Effizienz der Zytotoxizität und DNA Platinierung des freien und liposomalen CDDP . . . . .	55
4.7	GeneGo Charakterisierung der Prozessnetzwerke der Chemoresistenz .	58
4.8	GeneGo Prozessnetzwerke in A2780cis Zellen nach Behandlung mit freiem CDDP . . . . .	61
4.9	GeneGo Prozessnetzwerke in A2780cis Zellen nach Behandlung mit liposomalem CDDP . . . . .	63
4.10	Induzierte Gene des Wirkmechanismus von freiem CDDP . . . . .	66
4.11	Induzierte Gene des Wirkmechanismus von liposomalem CDDP . . . . .	67
4.12	Boxplot der normalisierten Arrays . . . . .	72
4.13	Resultat Varianz-Komponentenanalyse . . . . .	73
4.14	Heatmap der konstitutiv exprimierten Gene . . . . .	74
4.15	Heatmap der Darapladib Signatur . . . . .	77

4.16	Heatmap der Gaurnier PSMD4 Signatur . . . . .	77
4.17	Heatmap der „KEGG, Graft vs. Host Disease“ Signatur . . . . .	79
4.18	Hauptkomponentenanalyse der Tumor Subtypisierung Studie . . . . .	85
4.19	Ergebnisse der Klassifikation von molekularen Subtypen mit der FECS- nV Methode . . . . .	89
4.20	Visualisierung der Robustheit der FECSnV Methode . . . . .	91
4.21	Visualisierung des NNP Expressionsprofil . . . . .	93
4.22	Visualisierung des PNN Expressionsprofil . . . . .	94
4.23	Visualisierung des PPN Expressionsprofil . . . . .	95
4.24	Visualisierung des PPP Expressionsprofil . . . . .	96
4.25	Ergebnisse der klinischen Validierung molekularen Subtypen mit der FECSnV Methode . . . . .	97
4.26	Ergebnisse der Klassifizierung der Subtypen mit der FECSnV Methode	98
4.27	GOBO Analyse der klinischen Endpunkte der potentiellen Subtypen charakterisierenden Signaturen . . . . .	100
4.28	Clusteranalyse der Subtypen . . . . .	103
7.1	Signifikante Netzwerke der Kategorie GO Lokalisation . . . . .	126
7.2	Signifikante Netzwerke der Kategorie GO Molekulare Funktion . . . . .	127
7.3	Signifikante Netzwerke der Kategorie GO Prozess . . . . .	128
7.4	Legende zu den im Netzwerk verwendeten Symbolen . . . . .	129

# Tabellenverzeichnis

3.1	Die Tabelle fasst die untersuchten Studien aus GEO zusammen. . . . .	32
3.2	Die Tabelle fasst die Anzahl untersuchter Patienten in FIGO Stadien zusammen. . . . .	43
4.1	Die Tabelle listet die $IC_{50}[\mu m]$ der untersuchten Ovarialkarzinomzellen auf, die über die MTT Experimente erhalten wurden. Die Daten repräsentieren die Mittelwerte aus drei bis sechs Experimenten mit jeweils $2e^4$ Zellen für 72 Stunden. . . . .	56
4.2	Die Tabelle listet die Gene auf, die im GeneGo Prozessnetzwerk „Transkription und Chromatin Modifikation“ induziert sind. . . . .	60
4.3	Die Tabelle listet die Histon modifizierenden Gene auf, die in A2780cis Zellen nach Gabe von freiem CDDP induziert sind. . . . .	62
4.4	Die Tabelle fasst die koexprimierten Gene aus dem Prozess „Transkription und Chromatin Modifikation“ zusammen, die nach der Behandlung mit liposomalem CDDP induziert sind. . . . .	64
4.5	Die Tabelle listet alle potentiell verwendbaren Biomarker und deren Annotation. . . . .	75
4.6	Die Tabelle listet alle signifikant angereicherten Gene in Stadium I auf. . . . .	76
4.7	Die Tabelle listet alle signifikant angereicherten Gene in Stadium II und Stadium III auf. . . . .	78
4.8	Die Tabelle 4.8 listet alle klinisch relevanten Merkmale der Brustkrebspatienten sowie der Zellkulturen auf. Insgesamt gibt es Microarray Expressionsdaten von 367 Patienten, davon haben 99 den Status TNBC. Zusätzlich gibt es 20 TNBC Zellkultur Proben, die verschiedene Brustkrebs Zelllinien repräsentieren. . . . .	84

4.9	In Tabelle 4.9 sind die von Markus Hanl erhaltenen Ergebnisse der Klassifizierung mit dem Random Forest Algorithmus aufgelistet. Die Gewichtung der intrinsischen Eigenschaften der Datensets wurde evaluiert durch den Zusammenschluss der Ergebnisse von jeweils 50 RF Modellen a 2000 Entscheidungsbäumen. . . . .	86
4.10	Die Tabelle 4.10 listet 25 signifikant angereicherte Gensignaturen in triple-negativem Brustkrebs auf, die über die Klassifikation mit dem Random Forest Algorithmus erhalten werden. Die ersten zehn Prozesse beschreiben ESR1 beinhaltende Signaturen, die reprimiert sind. Oftmals sind jedoch nur Bruchteile der in der Datenbank für molekulare Signaturen (MSigDB) hinterlegten Signaturen angereichert. . . . .	87
4.11	Die Tabelle 4.11 listet alle signifikant angereicherten Genexpressionsprofile in Brustkrebspatienten auf, die mit der FECSnV Methode erhalten wurden. Die hier dargestellten Prozesse sind signifikant in den molekularen Subtypen angereichert, d.h. der P-Wert ist kleiner 0,01 und die FDR unter 0,05. Die ersten zehn Prozesse sind induzierte, Brustkrebs relevante Signaturen, gefolgt von neun reprimierten Signaturen, die ebenso eine Relevanz in Brustkrebs aufweisen. Fälle in denen eine Signatur in dem betreffenden Subtypen nicht induziert ist, werden mit dem Buchstaben „X“ gekennzeichnet. . . . .	92
4.12	Die Tabelle listet alle über GSEA erhaltenen Prozesse für die untersuchten Subtypen auf, die <b>FDR</b> ist jeweils $< 1e^{-4}$ . . . . .	101
7.1	Die Tabelle fasst die untersuchten Studien aus GEO zusammen. . . . .	119
7.2	Die Tabelle listet die Gene auf, die durch p53 induziert und mit freiem CDDP aktiviert werden. . . . .	122
7.3	Die Tabelle listet die Gene auf, die durch p53 induziert und sowohl mit freiem als auch mit liposomalem CDDP aktiviert werden. . . . .	124
7.4	Die Tabelle listet die Gene auf, die durch p53 induziert und mit liposomalem CDDP aktiviert werden. . . . .	125

# 1 Einleitung

## 1.1 Genexpressionsprofile in der Krebsforschung

Die Krebsforschung erlebt derzeit einen technologischen Wandel. Dieser ist einerseits begründet durch die jüngsten Entwicklungen in der molekularen Biotechnologie und andererseits durch den rapiden Fortschritt der Computertechnologie. Daher ist ein Voranschreiten in der Ergründung der komplexen Wechselwirkungen, die der Krankheit Krebs inne wohnen zu verzeichnen, jedoch wachsen damit einhergehend auch die interdisziplinären Herausforderungen, die deren Erforschung bedingen. Dies führte vor allem zur Ausdifferenzierung der unterschiedlichen naturwissenschaftlichen und medizinischen Bereiche zu der nun als Lebenswissenschaften bezeichneten Forschung. Der Vorteil dieses interdisziplinären Ansatzes wiegt den enormen Aufwand dadurch auf, dass eine Annäherung der Forschung an den Patienten selbst ermöglicht wird und dass der dabei generierte Wissensstand direkt weitergegeben werden kann. So belegen die Erfolge zahlreicher Studien [1, 2], dass ein translativer Ansatz, welcher lebenswissenschaftliche Methoden in der Klinik etabliert, längerfristig die sogenannte personalisierte Medizin ermöglichen kann. Dies ist ein fortschrittliches Konzept, welches auf neuesten molekularbiologischen Methoden gründet und den genetischen Hintergrund des einzelnen Patienten berücksichtigt.

Patienteneigenschaften werden durch genetische Merkmale und in spezifischen Genexpressionsprofilen manifestiert, welche durch die personalisierte Medizin in neu aufkommende Therapiemöglichkeiten umgesetzt werden können. Ein Wegbereiter der personalisierten Medizin ist die Microarraytechnologie. Diese Technologie verbindet die Erkenntnisse und Entwicklungen aus Molekularbiologie und Computertechnologie. Einerseits ist die Erfassung von Genexpressionsprofilen durch die molekularbiologischen Entwicklungen der letzten beiden Dekaden erst ermöglicht worden, andererseits ist die Auswertung der dabei anfallenden Datenmengen überhaupt erst durch die massive Verbreitung und Entwicklung von Computerressourcen gegeben. Welche Möglichkeiten die Microarraytechnologie bietet, Genexpressionsprofile diverser Tumorentitäten zu unter-

suchen und welche Möglichkeiten sich dadurch für den therapeutischen Ansatz bieten soll ein Schwerpunkt dieser Arbeit sein. Des Weiteren liegt ein besonderes Augenmerk auf der softwareseitigen Entwicklung von Analysewerkzeugen, ohne die eine Untersuchung von Microarray Daten nicht möglich wäre.

## 1.2 Microarraytechnologie

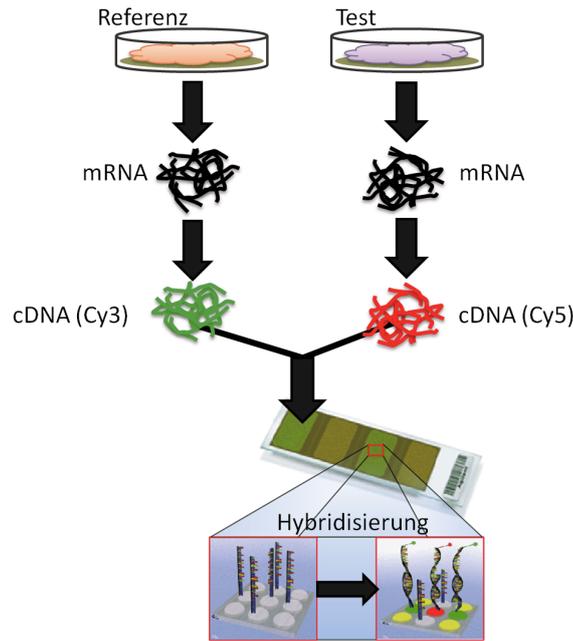
Der Ursprung der modernen Microarraytechnologie geht auf Schena et al. zurück [3]. Die Autoren beschreiben eine Quantifizierung der mRNA Zusammensetzung einer biologischen Probe mit Hilfe sequenzbasierter DNA Sonden. Diese Technik wurde weiterentwickelt und seitdem in fast alle Bereiche der Lebenswissenschaften integriert. Die Microarraytechnologie lässt sich prinzipiell in Einkanal [4] und Zweikanal [5] Systeme unterteilen. Darüber hinaus ist eine formale Untergliederung möglich, welche nach der jeweils zu untersuchenden biologischen Ebene erfolgt.

- DNA Karyotypisierung, mit aCGH Microarray [6]
- DNA Methylierung, mit Methylome Array [7, 8]
- DNA Transkription, mit mRNA Expression Array [3]
- DNA Regulierung, mit mikro RNA Expression Array [9]
- DNA Translation, mit antikörperbasiertem Protein Expression Array [10]
- DNA basierte Gewebetypisierung, mit Tissue Microarray [11]

Die Microarraytechnologie deren Daten in dieser Arbeit bearbeitet werden, ermöglicht eine quantitative Untersuchung der momentanen mRNA Zusammensetzung biologischer Proben. So werden Microarrays im sogenannten Einkanal- und Zweikanal-Format analysiert. Während die Einkanal Arrays aus öffentlichen Datenbanken bezogen wurden, sind die Arrays im Zweikanal Format im Rahmen einer Kooperation mit der Humangenetik der Universität Düsseldorf entstanden.

### 1.2.1 Zweikanal-Microarraytechnologie

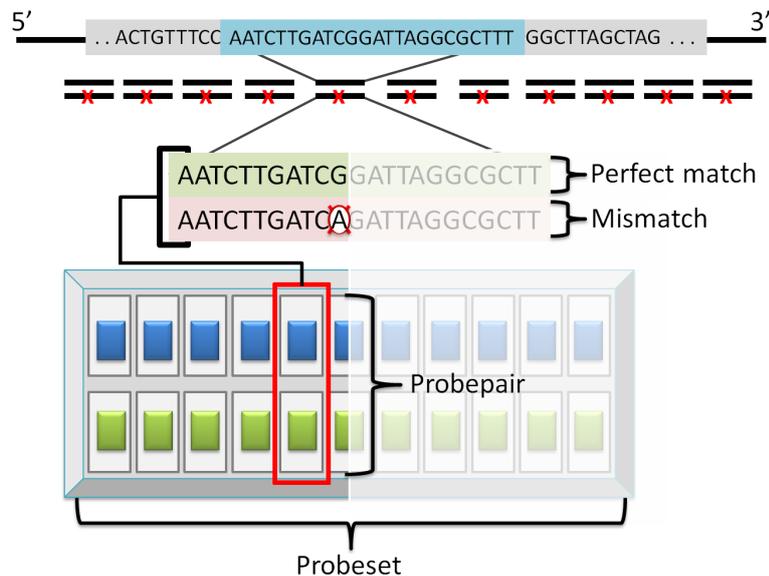
Abbildung 1.1 zeigt ein Schema des experimentellen Ablaufs zur Zweikanal Microarray Datenerhebung. Der erste Schritt beinhaltet die Extrahierung der mRNA. Die Extraktion erfolgt aus der zu testenden Probe und aus einer meist unbehandelten Referenzprobe. Im nächsten Schritt wird die mRNA in die stabilere cDNA umgeschrieben und mit



**Abbildung 1.1:** Die Zweikanal-Microarraytechnologie ermöglicht einen direkten Abgleich der zu testenden Probe mit einer Referenz. Abb. 1.1, modifiziert nach Brown [12].

dem Fluoreszenzmarker Cy3<sup>®</sup> oder Cy5<sup>®</sup> eingefärbt. Beide Proben werden vermischt und anschließend mit den einzelsträngigen DNA Proben, die auf der Array Oberfläche fixiert sind, hybridisiert. Nach einer kurzen Inkubationszeit wird der Array gewaschen, um unspezifische Hybridisierungen zu entfernen. Schließlich werden die Farbstoffe mit Hilfe eines Lasers zu der charakteristischen Emission bei 570nm und 670nm angeregt, die dabei entstandenen Bilder werden mit einem speziellen Microarray-Scanner eingefangen und anschließend überlagert. Die so genannte „Feature Extraction“ Software des jeweiligen Microarray Herstellers ist in der Lage, aus einem solchen Bild Genexpressionsrohdaten zu entwickeln. Dabei gilt ein Gen als exprimiert, wenn sein jeweiliges Reporterkonstrukt erfolgreich mit Fluoreszenzfarbstoff umgesetzt wurde. Der Fall, dass beide Fluoreszenzfarbstoffe gleichermaßen gebunden sind, weist darauf hin, dass auch die Gene in beiden Proben gleich stark exprimiert sind. Eine so gleichmäßige Überlagerung der Farbstoffe wird gemeinhin als gelb wahrgenommen. Neben diesen drei Extremfällen treten auch alle denkbaren Verhältnisse der Farbstoffkombinationen auf. Dies rührt daher, dass auch die Genexpression ausgeprägt differenziert zu sein scheint.

## 1.2.2 Einkanal-Microarraytechnologie

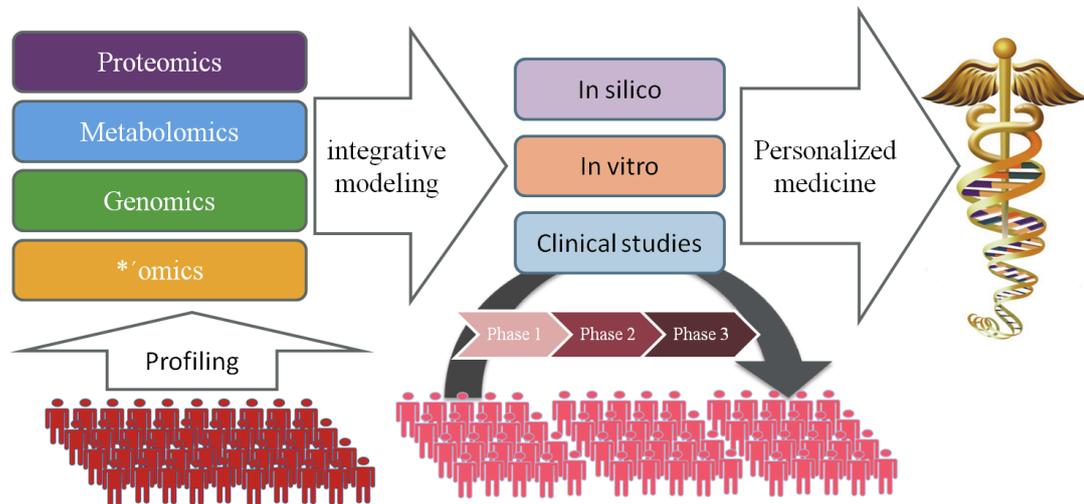


**Abbildung 1.2:** Die Einkanal-Microarraytechnologie erzeugt durch Verwendung einer falsch positiv Kontrolle eine Plattform basierte Referenz. Modifiziert nach Hu [13].

Das sogenannte Einkanal-Microarrayformat gestattet die Erfassung der quantitativen mRNA Zusammensetzung einer biologischen Probe. Der Abgleich mit einer Referenzprobe wird in diesem Fall mittels interner Kontrollproben realisiert. Prinzipiell ist jedoch der experimentelle Ablauf analog zu Zweikanal-Systemen, mit der Ausnahme, dass Biotin als Nucleotidmarker verwendet wird. Zudem gibt es Unterschiede bezüglich der Fertigung. Zweikanal-Arrays werden durch ein Verfahren hergestellt, welches sich aus der Inkjettechnologie ableitet. Einkanal-Arrays hingegen werden in einem, an das Lichtmaskenverfahren aus der Halbleitertechnologie angelehnten Prozess hergestellt. In Abbildung 1.2 ist schematisch der Aufbau eines sogenannten „Probesets“ dargestellt. Dieses ist spezifisch für ein bestimmtes Gen und setzt sich aus elf Proben zusammen, die wiederum aus zwei fast identischen Oligonukleotidsequenzen bestehen. Der Sequenzunterschied ist einzig durch eine polymorphe Base (*i.e.* ein SNP) innerhalb der Oligonukleotidkette gegeben. Diese sogenannte Mismatch Probe, die in der Mitte der Sequenz platziert ist, verhindert die Anlagerung sequenzähnlicher Proben. Auf diese Weise können nur völlig identische, komplementäre Proben angelagert werden, wodurch das Maß an unspezifischer Hybridisierung herabgesetzt wird. Einkanal-Arrays werden bevorzugt zur Detektion von sogenannten Splicevarianten eingesetzt [13]. Inzwischen sind nicht nur 3'-IVT (*i.e.* „in vitro transcription“) Sequenz basierte Arrays,

sondern auch Gen umspannende Exon Arrays und Genome umfassende, sogenannte Tiling Arrays verfügbar [14].

### 1.3 Personalisierte Medizin



**Abbildung 1.3:** Die Etablierung der personalisierten Medizin erfordert die Integration verschiedener biologischer Informationsebenen und ausgiebige Testung [15].

Die personalisierte Medizin trägt der Tatsache Rechnung, dass Patienten mit identischer Diagnose dennoch unterschiedliches Ansprechverhalten auf ein und dasselbe Medikament zeigen können. Dieser Umstand ist begründet durch die unterschiedliche genetische Ausstattung der betroffenen Patienten. Daher ist es unabdingbar, direkten Lesezugriff auf eben diese zu erlangen. Biomedizinisch relevante Informationen umfassen die gesamte genetische Sequenz (DNA), ihre Ausprägung (mRNA), ihre Regulierung (DNA Methylierung, Transkriptionsfaktoren), ihre ausführbaren Transkripte (Proteom) und ihren Stoffwechsel (Metabolom). Bereits heute verfügbare Diagnosesysteme, die für den Einsatz in der personalisierten Medizin geeignet sind, beruhen auf Genexpressionsmessungen mittels dedizierten Microarrays [2, 16, 17]. Künftige Diagnosesysteme könnten jedoch, aus unterschiedlichen biologischen Ebenen stammende Marker etablieren, welche als charakterisierende Merkmale von Patienteneigenschaften fungieren [18, 19]. Die Abbildung 1.3 verdeutlicht die Flussrichtung der Informationen, die zum Etablieren der Personalisierten Medizin führen. Anfangs wird ein Screening von Patientenkohorten durchgeführt, welches die verschiedenen biologischen Informations-

schichten erfasst, deren Resultate in Datenbanken gesammelt werden. Diese Informationen können die integrative Modellierung in verschiedenen Stufen ermöglichen, auch klinische Tests müssen miteinbezogen und gesondert evaluiert werden. Schließlich kann unter Verwendung der erhobenen Expressionsprofile der Patienten ein Abgleich unter Gesichtspunkten erfolgen, die der personalisierten Medizin Rechnung tragen, welche wiederum in diverse Therapieoptionen münden. Somit wird dem Patienten eine maßgeschneiderte Therapie zu Teil, die exakt auf die klinischen Voraussetzungen abgestimmt und idealerweise völlig auf den Bedarf dosiert ist.

### **1.3.1 Softwareentwicklung**

Die Microarray-Genexpressions Technologie wird in Zukunft die personalisierte Medizin in der Klinik etablieren, mit dem Ziel Informationen über Tumorprogression und potentiellen Ausgang der Chemotherapie bereitzustellen [20]. Derzeit stellen zwei öffentliche Datenbanken Array-Express [21] und Gene Expression Omnibus [22] große Microarraydatenbestände zur Verfügung. Die Analyse dieser großen Datenmengen von Patientenkohorten verlangt erhebliche Ressourcen. Die neuesten Entwicklungen im Bereich Computer Hardware haben mehrkernige Hauptprozessoren, sogenannte „multi-core“ Einheiten sogar in gewöhnlichen Arbeitsplatzrechnern eingeführt. Die Nutzbarmachung dieser zusätzlichen Rechenleistung stellt ein attraktives Angebot dar, wurde jedoch bislang wenig umgesetzt. Derzeit sind verschiedene Softwareanwendungen für die Microarrayanalyse frei verfügbar, z.B. lokale Systeme wie TM4 [23] oder webbasierte Systeme wie Babelomics [24] und DAVID [25]. Diese Softwarewerkzeuge sind für ganz bestimmte Aufgaben in der Microarrayanalyse entworfen worden. Das Zusammenführen der Resultate der Microarrayanalysen und das Erstellen spezifischer Berichte erfordert jedoch beides, biologisches Hintergrundwissen und rechentechnische Fachkenntnis.

### **1.3.2 Qualitätssicherung von Microarray Daten**

Die Microarraytechnologie hat den Wechsel von einer spezialisierten Methode zu einer für die lebenswissenschaftliche Forschung alltäglichen Methode bereits vollzogen. Dieser Wechsel wurde jedoch durch geringe Konkordanz in den Resultaten [26] und Diskussionen über die Verwendung von Microarray Analysemethoden begleitet [27–30] und führte schließlich zur Gründung des „Microarray Quality Control Project“ (MAQC). Im MAQC Projekt wurde zunächst die Genauigkeit der Technologie ermittelt [31, 32] und

anschließend die Analysemethoden und die erstellten biologischen Modelle verglichen [33]. Die Ziele des MAQC-II Projektes waren sowohl die Reproduzierbarkeit als auch die Generalisierbarkeit von Microarray Analyseergebnissen zu hinterfragen. Die Bedenken gegenüber der Microarraytechnologie entstammen verschiedenen Quellen. Zum einen waren die Experiment dokumentierende Protokolle unvollständig und zum anderen galt die Methode, welche zum Normalisieren der Daten verwendet wurde als strittig. Letztlich konnte sogar nachgewiesen werden, dass die eingesetzten Analysemethoden fehlerhaft waren. Die MAQC-II Studie brachte zum Vorschein, dass die Reproduzierbarkeit der Ergebnisse hauptsächlich an der Verfügbarkeit einer Dokumentation über den Analyseprozess scheitert; jedoch ist diese ebenso essentiell wie die Analyse auf Basis von qualitativ hochwertigen Daten zu beginnen. Daher sollte zu Beginn der Analyse die Qualität der Rohdaten überprüft werden. Die Kontrolle der Datenqualität soll vor allem potentielle Ausreißer Arrays aufdecken.

Gegenwärtig gibt es keine allgemein akzeptierte Qualitätskennlinie, um die Stichhaltigkeit von Microarraydaten zu bestimmen, obwohl das Bioconductorprojekt [34] eine Vielzahl von Methoden verfügbar macht, mit denen Microarrayqualität bemessen werden kann. Abhilfe wäre zum einen durch Qualitätskontrollen nach der Normalisierung der Daten zu schaffen, um somit auch systematisch fehlerhafte Normalisierung zu erkennen. Zum anderen wäre, wie bereits erwähnt, eine Qualitätskontrolle auch vor der eigentlichen Analyse sinnvoll, um die Normalisierung fehlerhafter Daten zu verhindern [35].

### 1.3.3 Wirkstoffstudie

Die Chemotherapie ovarieller Karzinome wird hauptsächlich mit dem DNA schädigenden Wirkstoff Cisplatin (CDDP, nach IUPAC: (SP-4-2)-Diamindichloridoplatin(II)) eingeleitet. Der Wirkmechanismus beruht auf den DNA interkalierenden Eigenschaften von Cisplatin und den dadurch induzierten intra- und intersträngigen DNA Quervernetzungen. Eine Akkumulierung von Platin-DNA Addukten stört somit essentielle zelluläre Prozesse, insbesondere jene die an der biologischen Replikation partizipieren, da sie direkt auf der Unversehrtheit der DNA basieren. Sobald der Cisplatin verursachte Schaden registriert ist, wird ein komplexer Reparaturmechanismus eingeleitet, der einen vorübergehenden Stillstand des Zellwachstums und gegebenenfalls sogar einen Abbruch der Zellteilung zur Folge hat. Sind jedoch die zellulären DNA Reparaturkapazitäten erschöpft, wird der programmierte Zelltod, die sogenannte Apoptose eingeleitet. Cisplatin ist nicht nur für Tumorzellen, sondern auch für gesunde Zellen äußerst toxisch, daher

ist der Erfolg der Chemotherapie von Anfang an durch gravierende Nebenwirkungen limitiert. Andererseits wird häufig, insbesondere nach anfänglichem Ansprechen auf die Therapie, ein Rückfall beobachtet, der von einem nunmehr Cisplatinresistenten Tumor begleitet ist [36]. Die molekularen Mechanismen, die eine CDDP Resistenz etablieren sind bis heute nicht vollständig aufgeklärt, jedoch sind einige Aspekte in den aktuellen Hypothesen häufiger vertreten. Diese sollen im Folgenden kurz genannt werden. Die Cisplatinresistenz beruht auf einer breiten Grundlage biologischer Prozesse. Die Etablierung der Cisplatinresistenz beinhaltet eine langfristige Partizipation verschiedener Resistenzmechanismen [37].

- Eine verminderte Aufnahme des Wirkstoffes [38]
- Eine reduzierte Wahrnehmung des DNA Schadens [39]
- Die Inaktivierung der platinieren DNA [40]
- Die Toleranz gegenüber platinierter DNA [41]
- Der aktive Export des Wirkstoffes [42]
- Die Expression von Proliferation stimulierenden Genen [43]
- Die Inhibition der Apoptose [44]

Die Cisplatinresistenz stellt derzeit eine Herausforderung für die Krebsforschung dar und die Entwicklung von Therapiealternativen ist nun in den Vordergrund gerückt. Eine neue Möglichkeit bietet in Liposomen eingeschlossenes Cisplatin. Wie bereits in einer vorhergehenden Studie postuliert, werden durch die alternative Darreichungsform des Wirkstoffes effektiv Cisplatin resistente ovarielle Karzinomzellen (A2780cis) eliminiert [45]. Die liposomale Applikation von CDDP hat positive Auswirkungen auf die systemische Toxizität, daher können voraussichtlich klinisch relevante Konzentrationen ohne therapielimitierende Nebenwirkungen erreicht werden. Liposomales CDDP wird derzeit als Lipoplatin in verschiedenen klinischen Studien angewendet [46]. Dieser Umstand stellt die Ergebnisse der vorangegangenen Studie [45] als besonders aussichtsreich dar, jedoch ist der molekulare Wirkmechanismus des liposomalen Cisplatins ungeklärt.

### 1.3.4 Tumorprogressionsstudie

In den USA ist das Zervixkarzinom als gynäkologische Tumorentität die dritthäufigste Todesursache von Frauen [47]. Weltweit ist die Krankheit in der Gynäkologie als

die zweithäufigste durch Krebs bedingte Todesursache anzusehen [48]. Gebärmutterhalskrebs wird fast ausschließlich durch Infektion mit dem Humanen Papilloma Virus (HPV) ausgelöst (zu 99%). Am häufigsten wird Gebärmutterhalskrebs bei Patienten im Alter von 40 Jahren diagnostiziert. Der Zeitraum zwischen HPV Infektion und maligner Tumorprogression beträgt durchschnittlich zehn Jahre [48], Dieser Umstand eröffnet daher einen relativ langen Zeitraum für Detektion und medizinische Intervention. Gegenwärtig sind jedoch keine allgemein akzeptierten onkologischen Marker verfügbar [26]. Obwohl bereits einige Studien Microarray basierte Genexpressionsprofile erstellt haben, gibt es kaum Übereinstimmung in den resultierenden Gensignaturen. Diese geringe Überlappung in den Ergebnissen ist zurückzuführen auf verschiedene Parameter im Studienentwurf, einschließlich der verwendeten Microarray Plattform und Protokolle. Eine weitere Quelle für Variation in den Ergebnissen, die bislang nur selten in Betracht gezogen wurde, ist durch die Intratumorheterogenität gegeben. Bachtiry *et al.* entwickelten die Varianz-Komponenten Analyse zu Untersuchung der genetischen Eigenschaften von Tumorbiopsie-Replikaten, um die Generalisierbarkeit der Ergebnisse zu verbessern [26]. Die o.g. Vorstudie hatte das Ziel, die genetische Variabilität der Zervixkarzinome zu modellieren und die Beständigkeit von onkologischen Markern anhand der Probenanzahl zu ermitteln (n=32). Die Autoren weisen allerdings darauf hin, dass dieser Ansatz anhand einer größeren Studie validiert werden sollte. Ein wichtiger Aspekt, gegeben durch die Variabilität zwischen den verschiedenen Stadien gemäß der „Federation International of Gynaecology and Obstetrics“ (FIGO), fand in dieser Analyse jedoch keine Berücksichtigung. Eine Untersuchung der verschiedenen Tumorstadien könnte aber ein weitaus sensitiveres Niveau der prognostischen Eigenschaften von potentiell allozierten onkologischen Markern bieten. Die Überlagerung der verschiedenen Genexpressionsprofile mit den FIGO Stadien könnte deshalb sogar das Methodenspektrum in dem Sinne erweitern, das jedes Stadium möglicherweise auch genetisch festgelegte und daher sehr spezifische Therapiemaßnahmen erfordert.

### 1.3.5 Tumor-Subtypisierungsstudie

Tumorentitäten lassen sich mittels spezifischer Genexpressionsprofile klassifizieren [1]. Die Suche nach Mustern in Genexpressionsprofilen von Krebspatienten [49] hat bereits viele Methoden für die Untersuchung der verschiedenen Tumorentitäten etabliert [50–54]. Gensignaturen finden extensive Verwendung in klinischen Studien [55], vor allem seit ein Zusammenhang zwischen Genexpression und Therapieausgang als erwiesen gilt [56, 57]. Darüber hinaus sind Genexpressionssignaturen außerordentlich wertvoll für das

Verständnis der komplexen, dem Krebs zugrundeliegenden Prozesse sowie deren molekularen Eigenschaften [58, 59]. Besonders Genexpressionssignaturen im Eisenstoffwechsel [60] und der molekularen Prozesse der Hypoxie [61] zeigen bemerkenswert prädiktive Eigenschaften. Daher geht der Nutzen von Genexpressionsprofilen über das Entdecken neuer Tumorsubklassen, den Vergleich von Tumorentitäten und der Vorhersage des Therapieausgangs hinaus [20]. Obschon eine Vielzahl erfolgreicher Studien diese Aussage stützt [62], ist jedoch die Mehrzahl der molekularen Grundlagen der krebspezifischen Prozesse weitestgehend unerforscht, vor allem solche die für das Überleben des Patienten während der Therapie entscheidend sind, z.B. Metastasen und Chemoresistenz. Daher ist es wünschenswert, so genannte Biomarker für die Früherkennung der Krankheit, ihren potentiellen Ausgang und für das Ableiten von Therapiemöglichkeiten zu etablieren. Routinemäßig wird bei Brustkrebspatienten der Status der Rezeptoren für Östrogen, Progesteron und dem humanen epidermalen Wachstumsfaktor (HER2) vor der Therapie immunhistochemisch festgehalten. Der Rezeptorstatus beeinflusst die möglichen Therapieoptionen, z.B. ist bei Überexpression des HER2 Rezeptors eine Behandlung mit Trastuzumab vorgesehen. Andererseits wird über den Rezeptorstatus auch ein möglicher Therapieausgang festgelegt, z.B. korreliert die Überexpression von HER2 mit einer schlechten Prognose [63]. Ferner manifestiert sich durch Abwesenheit der Rezeptoren auch ein bestimmter Therapieausgang. Insbesondere der Phänotyp, dem alle drei Rezeptoren fehlen, der sogenannte „triple-negative“ vereint beides, die wenigsten Behandlungsmöglichkeiten und die schlechteste Prognose [64]. Triple-negativer Brustkrebs fällt in das sogenannte „multi-class“ Problem und kann in fünf [65, 66] oder sogar sechs [67] Subtypen untergliedert werden. Obwohl bereits mehrere Methoden für die Entdeckung der multi-class Subtypen postuliert wurden [68, 69], stellte sich nunmehr heraus, dass dieses Problem nicht trivial und daher nicht leicht zu lösen ist [70, 71]. Ein Vergleich der Resultate von Studien, die molekulare Klassifikation von Brustkrebspatienten mit hierarchischer Clusteranalyse betrieben, lieferten nur geringe Übereinstimmung [72]. Um dennoch eine Klassifikation von Brustkrebspatienten zu ermöglichen, wurden neue Herangehensweisen entwickelt, die bevorzugt proliferationsbasierte Gene für die Unterscheidung von Tumorklassen verwenden [73, 74]. Da nun aber ausschließlich proliferationsbasierten Gene verwendet werden [75], könnten der Therapie essentielle krebspezifische, biologische Informationen vorenthalten werden [76].

## 2 Zielsetzung dieser Arbeit

Microarraydaten haben sich in der letzten Dekade in den Lebenswissenschaften etabliert. Die Analyse der Microarraydaten rückt zunehmend auch in den Fokus klinischer Forschung. Verschiedene Aspekte müssen jedoch im Vorfeld berücksichtigt werden, um zu gewährleisten, dass der technologische Nutzen adaptiert und ein Erkenntnistransfer gewährleistet werden kann. Die Anforderungen an Software und Analysemethoden, an die Datenqualität selbst sowie an den experimentellen Entwurf sind enorm, um längerfristig auch die personalisierte Medizin zu etablieren. Die Analyse von Genexpressionsdaten ist nicht länger exklusive Aufgabe der Bioinformatikexperten. Das Erzielen statistisch signifikanter Ergebnisse gestaltet sich beim Analysieren von Microarraydaten besonders anspruchsvoll und erfordert neben rechen technischer Fachkenntnis auch biologisches Hintergrundwissen. Daher soll in dieser Arbeit zunächst die Softwareentwicklung der Microarrayanalysemethoden und die Qualitätssicherung von Microarraydaten behandelt werden, und erst anschließend werden exemplarisch drei Microarraystudien vorgestellt, in denen auf die verschiedenen Aspekte im experimentellen Entwurf eingegangen werden soll. So wird anhand einer Wirkstoffstudie gezeigt, wie mittels Microarraydaten ein Wirkmechanismus analysiert werden kann. Des Weiteren soll in einer Studie mittels Microarraydaten die Tumorprogression analysiert werden. In einer anderen Studie erfolgt eine Tumorsubtypisierung mittels Microarraydaten.

**Softwareentwicklung für die Analyse von Microarraydaten** In dieser Arbeit wurde eine freie Software, namens maRt (microarray Reporting tool) für die intuitive Analyse von Microarraydaten erstellt [77]. Während der Entwicklung von maRt lag der besondere Fokus auf einfachster Bedienung und hohem Datendurchsatz. Die Software integriert die neuesten Entwicklungen in der parallelen Datenverarbeitung und nutzt die objektorientierte und Betriebssystemunabhängige Programmiersprache Java. Anstatt „das Rad neu zu erfinden,“ verwendet maRt die GPL basierte, quelloffene Software des Bioconductor Projekts [34]. Die Programmiersprache Java bietet die Möglichkeit Daten parallel zu verarbeiten, die Statistiksprache R ist auf statistisches Rechnen ausgerichtet und bietet bereits bewährte Bibliotheken verschiedenster Funktionalität. In

maRt werden die Kerneigenschaften von Java kombiniert mit der statistischen Leistungsfähigkeit der Statistiksprache R. Zudem vereint die Software transparenten Zugriff auf aktuelle Internetdienste und Datenbanken und stellt darüber hinaus integrierte Reporting-Funktionalitäten in Standard HTML zur Verfügung. Daher ist maRt ein besonders attraktives Werkzeug für Nutzer ohne besonders fundierten bioinformatischen Hintergrund. Außerdem schafft maRt durch den modularen Aufbau eine vielversprechende Basis für die Entwicklung neuer Microarray Analysenwerkzeuge.

**Qualitätssicherung von Microarraydaten** Qualitätskontrolle und Normalisierung haben einen immensen Stellenwert im Verlauf der Analyse von Microarraydaten. Gegenwärtig gibt es verschiedene Methoden, die eine Qualitätssicherung von Microarraydaten ermöglichen. Die Visualisierung der erhobenen Qualitätsparameter stellt sich in der Praxis jedoch als problematisch dar. Daher wurde in dieser Arbeit eine Methode zur Visualisierung von Standard Qualitätskontrollen mit Hilfe von Circos [78] entwickelt [79]. Diese Methode legt die verschiedenen Qualitätsparameter in zirkulär angelegten Segmenten an. Auf diese Weise kann sowohl die Qualität der Proben, als auch potentielle Ausreißer in einem gesamten Microarraydatenset visualisiert werden. Durch diese Art der Visualisierung der verschiedenen Parameter individueller Arrays kann schnell ein erster Gesamtüberblick entstehen. Eine Besonderheit der Darstellungsweise ist die Kondensierung der Informationen auf schlichtweg problematische Werte, d.h. das Arrays, deren Kontrollproben unterdurchschnittliche Qualität aufweisen, auf diese Art direkt im Fokus stehen. Die entwickelte Methode ist derzeit für Microarrays der Plattformen GPL96 und GPL570 anwendbar. Dies sind die mit Abstand am häufigsten in öffentlichen Datenbanken anzutreffenden Formate.

**Wirkstoffstudie** Die Wirkung von freiem und liposomalem Cisplatin (CDDP) an sensitiven und resistenten ovariellen Karzinomzellen wurde in dieser Arbeit auf transkriptioneller Ebene verglichen. Mit Hilfe von Microarraydaten konnte erstmals gezeigt werden, dass liposomales CDDP eine von freiem CDDP völlig verschiedene Genexpression zur Folge hat. Diese Genexpressionsdaten illustrieren, dass liposomales CDDP selektiv Gene induziert, die auf DNA-Schäden und Apoptose in resistenten A2780cis Zellen hinweisen. Die Ergebnisse verdeutlichen ersten Erkenntnisse bezüglich des molekularen Wirkmechanismus, der liposomales CDDP befähigt, auch in CDDP resistenten Zellen zu wirken. Daher könnte liposomales CDDP künftig zu einer wichtigen Ressource für die klinische Anwendung an Patienten mit wiederkehrendem ovariellen Karzinom werden.

**Tumorprogressionsstudie** Gegenwärtig sind keine allgemein akzeptierten Biomarker für die Charakterisierung von Zervixkarzinomen verfügbar. In dieser Arbeit wurde daher die Varianz-Komponenten Analyse [26] für die Detektion von Biomarkern angewendet. Die Methode wurde in dieser Arbeit auf Microarray Genexpressionsdaten aus vier öffentlich verfügbaren Zervixkarzinom Studien angewandt (n=126) [80]. Die wegberreitende Studie von Bachtary *et al.* inspirierte zu der Annahme, dass sowohl intra Tumor- als inter Stadien-Heterogenität die Zuverlässigkeit onkologischer Marker maßgeblich beeinflusst. Zielführend in der vorliegenden Arbeit war die Entwicklung von sowohl statistisch als auch klinisch relevanten Genexpressionsprofilen von zervikalen Tumoren. Zusätzlich zur Varianz-Komponenten Analyse soll die so genannte „Gene Set Enrichment Analysis“ (GSEA) [81] für die weiterführende Untersuchung der erhaltenen Genexpressionsprofile verwendet werden.

**Tumor-Subtypisierungsstudie** Ein weiterer Schwerpunkt dieser Arbeit ist eine großangelegte Subtypisierungsanalyse von Microarray-Genexpressionsdaten aus vier öffentlich verfügbaren Brustkrebsstudien (n=514) [82]. Zu diesem Zweck wurden sogenannte „Voting“-Methoden aus der Literatur [83–85] berücksichtigt und ein neuer Algorithmus entwickelt, welcher die Extraktion von Genen, deren Bewertung und letztlich die Wahl eines Klassifizierers zur Aufgabe hat. Der Algorithmus basiert auf maschinellen Lernverfahren und statistischen Klassifikationsalgorithmen, die dem Bioconductor Paket CMA entnommen wurden [86]. Das sogenannte „multi-class“ Problem wurde über Konsensus-Clustering [87, 88] gelöst. Diese Methode zeigte beachtlichen Erfolg in Wilkerson *et al.* [89]. Die erhaltenen Gensignaturen wurden mit Hilfe der „Gene Set Enrichment Analysis“ (GSEA) auf ihren biologischen Wert untersucht [81], die klinischen Eigenschaften der Signaturen konnten mit GOBO (*i.e.* „gene expression-based outcome for breast cancer online“) analysiert werden.



# 3 Material und Methoden

## 3.1 Softwareentwicklung für die Analyse von Microarraydaten

### 3.1.1 Die Statistiksprache R

R ist eine quelloffene Variante der Statistiksprache S und eignet sich für die Lehre, die Datenanalyse und zur Methodenentwicklung [90]. Derzeit sind im „Comprehensive R Archive Network“ (CRAN) über 3500 Pakete gelistet, die frei zugängliche und quelloffene Software unterschiedlichster Funktionalität anbieten. Neben verschiedenen [R Handbüchern](#) gibt es ein eigenes [R Journal](#), eine aktive Entwicklergemeinschaft und alljährlich international stattfindende R Konferenzen ([useR!](#)). R ist eine imperative, prozedurale, scriptbasierte Sprache und bietet eine intuitive Syntax, sowie eine klar strukturierte Semantik. Die primitiven wie auch die komplexen Datentypen sind syntaktisch in ihrer Typisierung auf die Berechnung von Matrizen optimiert. Selbst einfache arithmetische Rechenoperationen werden intern daher als Matrizenoperationen realisiert. So können bei der Berechnung großer Datenmengen enorme Performanzen erreicht werden, die zudem auch auf viele Nachkommastellen genau sind.

### 3.1.2 Das Bioconductorprojekt

Das auf der Statistiksprache R basierende Bioconductorprojekt bietet eine inhaltlich auf die Analyse von genomischen Hochdurchsatzdaten abgestimmte, quelloffene Softwareumgebung [34]. Aufgrund der enormen Vielzahl an R Paketen, die statistisch anspruchsvolle Rechenoperationen übernehmen, konnte im Bioconductorprojekt auf grundlegende Funktionalität zurückgegriffen werden und daher insgesamt auf die Entwicklung von Funktionen und Paketen fokussiert werden, die für die Lebenswissenschaften bestimmt sind. Derzeit umfasst das Projekt über 550 verschiedene Bibliotheken,

die sich in dedizierte Aufgabenbereiche untergliedern lassen. Ein Großteil der Pakete jedoch findet in der Analyse und Annotation von Microarraydaten Anwendung.

### 3.1.3 Die Software Entwicklungsumgebungen NetBeans und Eclipse

In dieser Arbeit wurden für die Entwicklung von Software die in der objektorientierten Hochsprache Java geschriebenen Software Entwicklungsplattformen [Eclipse](#) [91] und [NetBeans](#) [92] verwendet. Beide Entwicklungsumgebungen wurden ehemals von bekannten Softwareunternehmen entwickelt und sind nun quelloffen sowie freigeben.

**Die Rich Client Plattform (RCP)** Eclipse und auch NetBeans können als sogenannte Rich Client Plattform für die Entwicklung und Realisation eigener Projekte genutzt werden. Dieser Ansatz bietet den Vorteil, dass durch die RCP ein lauffähiges und stabiles Programmrahmenwerk gegeben ist, welches zunächst nur die minimalen Anforderungen erfüllt. Alle Funktionen, die das Programm später anbieten soll, werden über eigenständige Module eingebunden. Aus diesem Ansatz heraus ist es daher möglich, eine eigenständige, völlig unabhängige Anwendung zu erstellen, die aufgrund des OSGi Rahmens modularisiert und transparent im Bezug auf die zugrundeliegende Betriebssystemschicht ist. Die [OSGi](#) ist eine Allianz verschiedener großer Soft- und Hardware Firmen, die mit der Java Virtual Machine, eine hardwareunabhängige Softwareplattform entwickeln. Neben den Vorzügen der Integration eigener und bereits bestehender Module, bieten RCP basierte Anwendungen daher einen hohen Grad an Qualität und Fehlertoleranz, da die Module eigenständige Instanzen einer RCP darstellen [93]. Die Eclipse Plattform bietet neben der vollständigen Modularisierung der Software, auch einen auf die Erstellung von Modulen basierten Prozess zur Softwareentwicklung, durch das sogenannte Eclipse Modeling Framework ([EMF](#)) [94].

**Methoden der agilen Software Entwicklung** Die testunggetriebene Entwicklung von Software („test driven development“ (TDD) [95]) ist neben dem sogenannten „extreme Programming“ eine Form der agilen Softwareentwicklung. Im TDD wird das Ziel verfolgt, zunächst einen Test für die zu implementierende Funktion zu schreiben und erst anschließend die eigentliche Funktion. Diese Vorgehensweise kann durch die Software Testumgebung [JUnit](#) [96] realisiert werden. In der praktischen Durchführung wird zunächst der Konstruktor der Funktion geschrieben und die Eigenschaften bzw.

Merkmale festgelegt, diese werden jedoch erst später implementiert. An dieser Stelle werden von JUnit alle zu implementierenden Tests vordefiniert, die in einen Soll- und Istwertabgleich resultieren. Wenn nun die zuvor gewünschten Eigenschaften implementiert werden, kann JUnit in einem Testlauf der neuen Funktion Soll und Istwert abgleichen, d.h. die erstellte Routine wird so lange geändert, bis sie das gewünschte Ergebnis zurückgeben kann. Da die agile Softwareentwicklung ebenfalls die rasche Implementierung neuer Module bedeutet, muss die Integrität des Gesamtprojekts stetig überprüft werden. Diese Überprüfung kann über die kontinuierliche Integration mit [Hudson](#) [97] gewährleistet werden. Das Gesamtprojekt wird mit Hudson auf einem dedizierten Server für unterschiedliche, emulierte Betriebssysteme erstellt und getestet. Die Resultate können kontinuierlich überwacht werden. Das Prinzip der kontinuierlichen Integration weist große Ähnlichkeit mit der TDD auf und kann als Fortführung dieses Prinzips auf die Gesamtprojektebene betrachtet werden. Das Software Projekt Management Werkzeug [Maven](#) [98] übernimmt in diesem Prozess die Aufgabe, alle im Projekt vorhandenen internen Bibliotheken zur richtigen Abfolge bereitzustellen. Des Weiteren löst Maven zugleich die Versions Kompatibilität der externen Bibliotheken auf.

**Versionsverwaltung** Da Softwareentwicklung die ständige Änderung des Projekts bedeutet, wird eine Versionsverwaltung benötigt, mit der Änderungen nachvollzogen, verfolgt und auch wieder rückgängig gemacht werden können. Zusätzlich bietet eine Versionsverwaltung auch eine Form der Datensicherung, wenn diese an eine weitere Maschine delegiert wird. Subversion ([SVN](#)) ist eine frei verfügbare Versionsverwaltung [99]. Das Softwareprojekt `maRt` wird über SVN auf den Google Code Server gespiegelt und der Quellcode ist unter [code.google.com/p/martool/source/browse](http://code.google.com/p/martool/source/browse) einzusehen. Eine andere Versionsverwaltung, die zudem neue Funktionalitäten bietet, die dem Software Projektmanagement Rechnung tragen, bietet [Git](#) [100]. Der Quellcode für das `Circos Microarray Qualitätskontrolle` Projekt ist über Git auf Github gespiegelt und unter [circos-arrayQC](#) zu finden. Auch der Quellcode der `FECSnV` Methode verwendet Git als Versionsverwaltung und ist in Github abgelegt unter [FECSnV](#).

## 3.2 Statistische Methoden und Maschinelle Lernverfahren

### 3.2.1 Statistische Methoden und Tests

In dieser Arbeit werden statistische Standardtests als Klassifizierer und zur Extraktion von Genen verwendet. Diese sind der t-Test, der Welch-Test, der F-Test, und der Kruskal-Test. Diese statistischen Tests unterscheiden sich durch die Annahme der zugrunde liegenden Verteilung und auch formal in der mathematischen Ausführung. Das logische Prinzip jedoch ist für alle hier genannten Methoden durchaus vergleichbar. Allgemein wird der Mittelwert der Stichprobe genutzt, um somit den Mittelwert der Grundgesamtheit mit einem vorgegebenen Wert zu vergleichen. Das Resultat dieses Vergleichs ist an die Wahrscheinlichkeit einer eingangs erwogenen Hypothese geknüpft. Diese wird als sogenannte Nullhypothese formuliert und tritt für den Fall ein, dass der Mittelwert der Grundgesamtheit einem vorgegebenem Wert entspricht. Eine Alternativhypothese wird formuliert, wenn dem nicht so ist und der Mittelwert der Grundgesamtheit nicht dem Erwartungswert entspricht. Der Welch-Test vergleicht nach diesem Schema die Mittelwerte zweier Stichproben auf Identität gegen die Alternative, dass einer der beiden Mittelwerte kleiner ist. Der F-Test unterscheidet sich vom Welch-Test lediglich durch die Annahme, dass die Daten nicht normalverteilt, sondern der F-Verteilung entsprechen. Im Kruskal-Test wird die Nullhypothese angenommen, wenn die zugrunde liegende Verteilung einer Chi-Quadrat Verteilung gleichkommt.

**Die False Discovery Rate (FDR)** Die Identifikation differenziert-exprimierter Gene ist das Ziel der meisten Microarraystudien. Gemeinhin wird ein Schwellwert definiert unterhalb jenem differenziert-exprimierte Gene vermutet werden, dieser wird häufig aus einem statistischen Test abgeleitet. In Microarrayexperimenten werden große Mengen von Genexpressionswerten auf Signifikanz getestet und üblicherweise gilt ein p-Wert unter 0,05 als signifikant. Dies beschränkt die Rate der falsch positiv selektierten Gene in den Resultaten auf unter 5%. Da jedoch die Anzahl zu testender Gene immens ist, liefert eine solche Beschränkung inakzeptable Ergebnisse. Eine gangbare Alternative ist daher die Korrektur der resultierenden Genlisten durch die sogenannte False Discovery Rate (FDR) [101]. Das Konzept der FDR ist abgeleitet aus der von Benjamini und Hochberg vorgestellte Family Wise Error Rate (FWER) [102]. Über die FDR kann beim multiplen Hypothesentesten, die Zahl der fälschlicherweise verworfenen Nullhypothesen korrigiert werden. Die FDR gibt demnach die Rate der falsch-positiven Ergebnisse

an. Obwohl die FDR gegenwärtig als die Methode der Wahl gilt, sollte beachtet werden, dass ein zu strikt gewählter Wert den Ausschluss biologisch interessanter Proben zur Folge haben könnte. Des Weiteren sollte auch beachtet werden, dass statistische Signifikanz nicht notwendigerweise an biologische Signifikanz gebunden ist [103].

**Varianzanalyse mit einem gemischten Modell** Bachtary *et al.* untersuchten mit einem gemischten Modell die intra Tumorheterogenität in Genexpressionsprofilen von Zervixkarzinompatienten [26]. Diese wegberreitende Studie wurde in dieser Arbeit auf die verschiedenen Stadien von Zervixkarzinomen ausgeweitet. Der R Code wurde freundlicherweise nach Korrespondenz mit B. Bachtary von M. Pintilie bereitgestellt. Die Methode führt auf ein gemischtes Modell zurück, in welchem sowohl ein fester Anteil als auch ein zufälliger Anteil modelliert werden. Gemeinhin werden solche Modelle in Messreihen verwendet, die Änderungen in einem stetigen Wert registrieren sollen, z.B. bei der maschinellen Fertigung. In diesem Ansatz ist der stetige Wert durch die Genexpression repräsentiert und der zufällige Wert durch die verschiedenen Stadien gegeben. Mit diesem Modell wurde die Varianz innerhalb eines Stadiums ( $W$ ) quantifiziert und in Relation zur Gesamtvarianz zwischen den Stadien gesetzt ( $B$ ). Die Gesamtvarianz kann nun mit der Formel 3.1 wiedergegeben werden:

$$T = W + B \tag{3.1}$$

Durch Formel 3.1 ist mit dem Verhältnis  $W/T$  ein Maß für die Tumorheterogenität zwischen den Stadien gegeben.

**Hauptkomponentenanalyse (PCA)** Die Hauptkomponentenanalyse wird mathematisch als Singulärwertzerlegung von  $X$  formuliert, wie durch die Formel 3.2 gegeben ist:

$$X = UDV^T \tag{3.2}$$

Wobei  $U$  eine orthogonale Matrix ist, deren Spalten die linksseitigen Einzelvektoren und  $V$  ebenfalls eine orthogonale Matrix ist, deren Spalten jedoch die rechtsseitigen Einzelvektoren darstellen. Die Variable  $D$  ist eine diagonale Matrix, welche durch die Eigenwerte definiert ist. Die Spalten in  $UD$  geben die Hauptkomponenten von  $X$  wieder. Ein besonderes Merkmal der Hauptkomponentenanalyse ist durch die Linearkombination  $X_{v1}$  gegeben, diese beinhaltet die größte Varianz unter allen möglichen Line-

arkombinationen von  $\mathbf{X}$  [104]. Für die Analyse der Genexpression in Microarraydaten wird folgende Summenformel angewandt:

$$x^{(e)} = \sum_{k=1}^e U_k D_k V_k^T \quad (3.3)$$

Durch die Formel 3.3 lassen sich die Werte für die Hauptkomponenten in den Genexpressionsdaten berechnen. Die Matrix  $\mathbf{UD}$  gibt die Werte für die Gene und die Matrix  $DV^T$  die Werte für die berücksichtigten Microarrays wieder. Ziel der Hauptkomponentenanalyse ist eine Projektion der Daten durch wenige Linearkombinationen und unter Ausschluss von Informationseinbußen. Zugleich werden dabei Redundanzen in den Daten durch Korrelationen in den Datenpunkten zusammengefasst. Es ist davon auszugehen, dass statistisch normalverteilte Daten durch Normierung, Mittelwert und Kovarianz vollständig charakterisierbar sind. Daher können normalverteilte Daten durch die Hauptkomponenten als statistisch unabhängige Größen repräsentiert werden, die wesentliche den Daten inne wohnende Aspekte zusammenfassen.

**Der Brier Score** G. Brier entwickelte 1950 eine Messgröße, um die Genauigkeit der Vorhersage von Wahrscheinlichkeiten in einer Reihe von potentiellen Ausgängen zu bewerten, indem er Prognose und tatsächlichen Ausgang direkt miteinander verglich [105]. Die Bewertung der Genauigkeit der Vorhersagekraft einer Methode findet gewöhnlich bei der binären Klassifikation Verwendung. Das Brier-Maß kann wie folgt formuliert werden:

$$\frac{1}{n} \sum_{k=1}^n (P_k - O_k)^2 \quad (3.4)$$

In Formel (3.4) ist der Zusammenhang zwischen prognostizierter Klassenzugehörigkeit  $\mathbf{P}$  und tatsächlicher Klasse  $\mathbf{O}$  aufgezeigt. Daraus folgt, dass ein niedriges Brier-Maß eine höhere Genauigkeit und in diesem Zusammenhang die optimale Effizienz durch den Wert 0 darstellt.

**Der Matthews's Korrelationskoeffizient** Der von Matthews entwickelte Korrelationskoeffizient (MCC) lässt sich direkt über die folgende Konfusionsmatrix berechnen:

$$MCC = \sqrt{\frac{TPxTN - FPxFN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.5)$$

Mit TP: Wahr Positiv, TN : Wahr Negativ, FP: Falsch Positiv und FN: Falsch Negativ. Vor allem bei unterschiedlichen Klassengrößen liefert dieses Maß eine realistische

Einschätzung über die Genauigkeit der Klassifikation, da alle Parameter berücksichtigt werden. Im Gegensatz dazu bietet die Präzision nur eine Aussage über die wahr positiv gemachten Aussagen, dass führt bei ungleichen Klassengrößen zu einer verzerrten Bewertung der Klassifikation und wird hier daher nicht verwendet.

### 3.2.2 Statistische Klassifikatoren

Statistische Methoden ermöglichen eine Klassifizierung der Eingangsdaten. Wie noch später gezeigt wird, unterscheiden sich die statistischen Methoden von den maschinellen Lernverfahren weitestgehend dadurch, dass diese nicht eine vorherige Lernperiode absolvieren.

**Diskriminanzanalyse** Des Weiteren wurden in dieser Arbeit verschiedene Varianten der Diskriminanzanalyse angewendet. Diese sind die diagonale Diskriminanzanalyse (DDA), die Diskriminanzanalyse nach Fisher (FDA), die lineare Diskriminanzanalyse (LDA) und die „Partial Least Squares“ Variante PLS-LDA [70]. Ebenso wie die Hauptkomponentenanalyse bewirkt die Diskriminanzanalyse eine Dimensionsreduzierung, wobei jedoch auch die Klassenzugehörigkeiten der Eingangsdaten berücksichtigt werden können. Daher zählen die heute bekannten Variationen der DA zu den ursprünglichsten Klassifikationsverfahren. In diesen Verfahren bringt eine Diskriminanzfunktion den wesentlichen Merkmalsunterschied zweier Populationen zum Ausdruck, oftmals ist dieser bereits durch die Varianz der Populationen gegeben. Schließlich kamen auch für Microarraydaten spezifische Methoden zum Einsatz. Diese sind die Klassifikation über Lineare Modelle mit dem Bioconductorpaket *Limma* [106] und das Klassifizierungsverfahren nach Golub [1].

### 3.2.3 Klassifikation über maschinelle Lernverfahren

Das Methodenspektrum in der Analyse von Microarraydaten konnte durch maschinelle Lernverfahren substantiell erweitert werden. Diese Verfahren lernen die Gesetzmäßigkeiten innerhalb der Datenstruktur ohne dabei den Datensatz auswendig zu lernen, denn dies führt zum sogenannten „Overfitting“. Auf diese Weise verallgemeinert der Algorithmus die interne Datenstruktur, sodass eine Generalisierung der gelernten Erkenntnisse auf unbekannte Daten erfolgen kann. Methoden, die eine vorherige Kategorisierung des Lerndatensatzes erfordern, zählt man zu Verfahren die überwachtes Lernen

anwenden. Die bekanntesten Vertreter dieser Gattung sollen im folgenden vorgestellt werden.

**Support Vector Machine (SVM)** Die sogenannte SVM [107] kann neue Klasseninstanzen differenzieren nachdem eine Lernphase mit bereits kategorisierten Beispielpklassen absolviert wurde. Geometrisch betrachtet konstruiert die Methode eine Hyperebene zwischen Vektor Repräsentationen der Genexpressionsdaten, um die Klassen voneinander zu trennen. Im einfachsten Fall sind die Daten linear trennbar, d.h. die Trennfläche kann maximiert werden, ohne dabei die Daten in eine höhere Dimension zu transformieren. Im nicht-linear trennbaren Fall aber muss ein sogenannter Kerneltrick angewendet werden, der eine Transformation der Daten in eine höhere Dimension bewirkt, in der die Daten einfacher zu trennen sind. Ein Beispiel für ein Kernelfunktion findet sich in der radialen Basisfunktion (RBF), welche nachfolgend beschrieben ist durch:

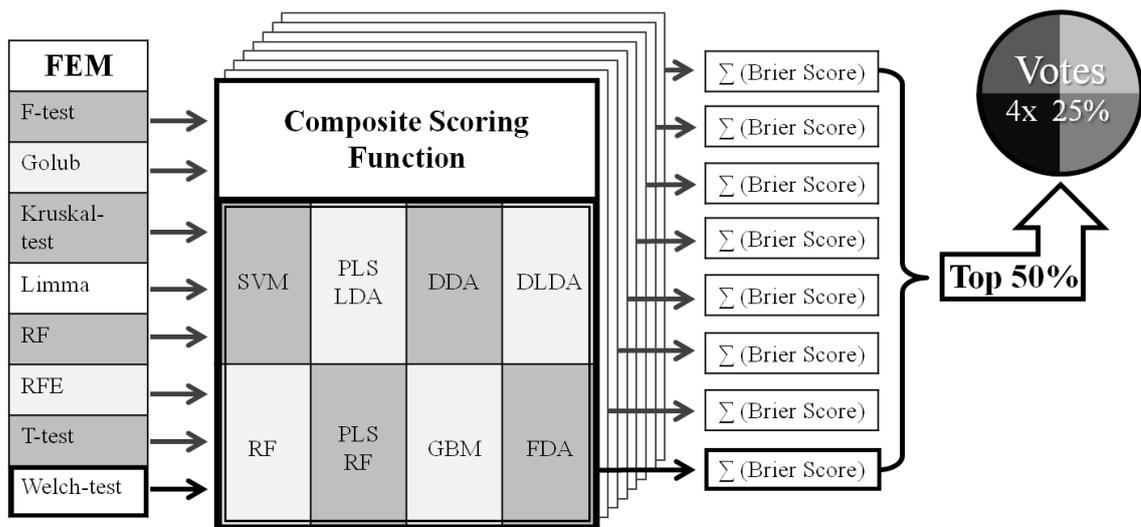
$$k(x_i, x_j) = \exp -\frac{1}{2}\sigma^2(\|x_i - x_j\|^2) \quad (3.6)$$

Diese Funktion variiert durch den Abstand zu einem festgelegtem Koordinatenursprung, demnach sind die durch eine solche Funktion beschriebenen Werte radialsymmetrisch. Eine vollständige Berechnung der Transformation über die Kernelfunktion ist jedoch aufwendig und daher werden nur die Vektoren berücksichtigt, die als Stützvektoren der Hyperebene Verwendung finden. In nicht-linearen Fällen, wenn eine Trennung mit Hilfe des Kerneltricks dennoch problematisch ist, wird eine sogenannte Softmargin parametrisiert. Diese kann Fehler in der Klassifikation zulassen, ohne dabei das Gesamtergebnis zu beeinträchtigen. Guyon *et al.* entwickelten eine für die Klassifikation von Genexpressionsprofilen ausgelegte SVM Variante, die Recursive Feature Elimination (RFE) genannt wird [108]. In diesem Algorithmus wird die SVM verwendet, um Features zu eliminieren, die für eine Klassifizierung nur marginale Bedeutung aufweisen.

**Random Forest (RF)** Der Random Forest Algorithmus ist ein sogenanntes Ensemble Lernverfahren mit dem eine Klassifikation oder eine Regression durchgeführt werden können [109]. Die grundlegenden Lernerinstanzen sind über Entscheidungsbäume realisiert, welche durch sogenanntes Bootstrapping die initialen Trainingsobjekte abbilden. Die Vorteile gegenüber der SVM sind zunächst durch eine eingebaute Feature Selektion und die Möglichkeit zur Parallelisierung gegeben, und nicht zuletzt ergeben sich auch durch die Handhabung Vorzüge, vor allem bei der Parametersuche. RF kann auch Verwendung als eine Erweiterung der „Partial Least Squares“ Methode [110] finden, um so

einen PLS-RF Algorithmus zu etablieren [111].

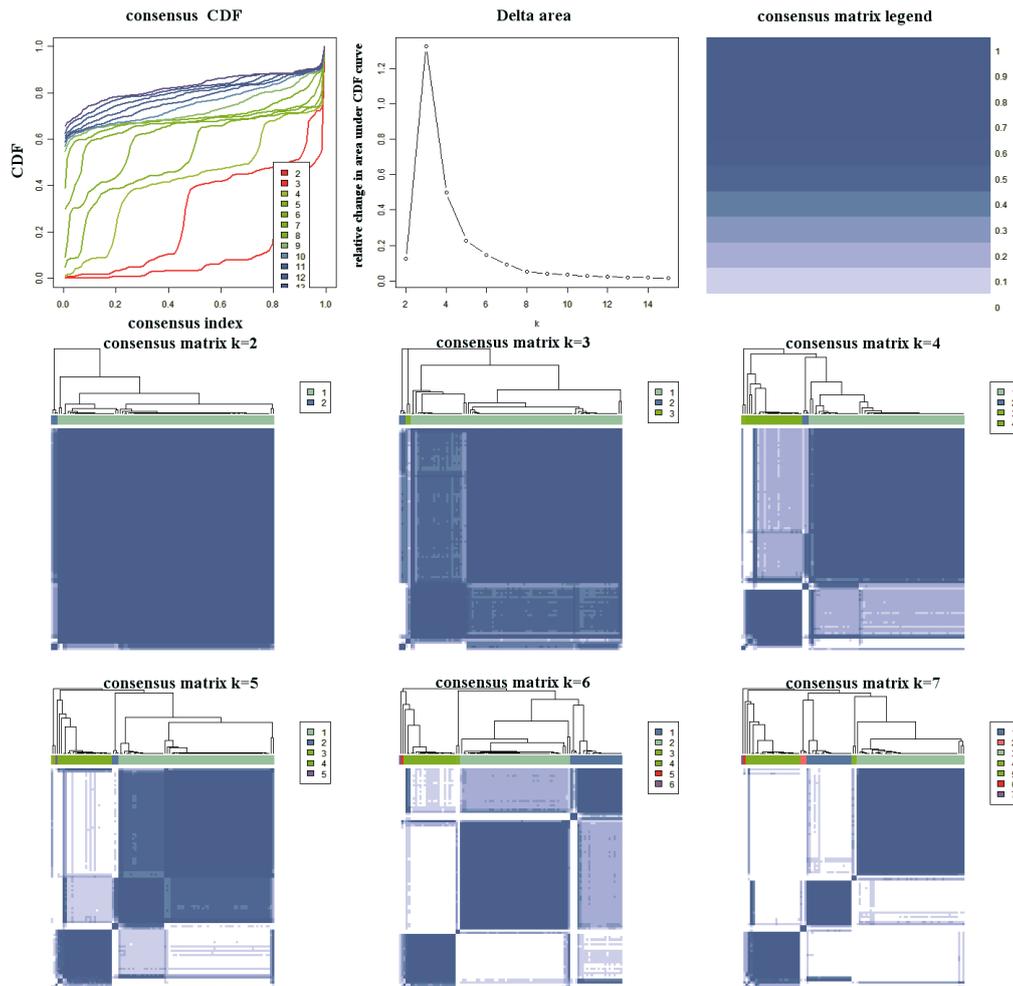
**Gradient Boosting Machine GBM** Ein weiterer Algorithmus, der ebenso Entscheidungsbäume berücksichtigen kann, ist durch die „Gradient Boosting Machine“ von Friedman [112] gegeben. Durch die GBM lässt sich eine Gruppe sogenannter schwacher Lerner zu einem stark in der Generalisierung verbesserten Ensemble von Klassifizierern steigern. Die Methode bedient sich einer sogenannten Verlustfunktion, mit welcher schrittweise eine aus den Trainingsdaten bekannte Zielfunktion approximiert wird. Auf diese Weise wird gleichzeitig die Verlustfunktion minimiert. Für den Fall, dass die Methode Entscheidungsbäume berücksichtigt, muss eine von Friedman vorgeschlagene Änderung im Algorithmus implementiert werden, dabei gilt es den Parameter  $\gamma$ , welcher die Koeffizienten der Verlustfunktion multipliziert nicht auf den gesamten Baum hin zu optimieren, sondern jeweils verschiedene  $\gamma$  für die einzelnen Bereiche der Entscheidungsbäume zu optimieren.



**Abbildung 3.1:** Die Abbildung 3.1 stellt ein Schema des in dieser Arbeit entworfenen Feature Extraktion, Composite Scoring und Voting (FECSnV) Algorithmus dar.

**Feature Extraktion über einen Composite Scoring und Voting Algorithmus** Die Abbildung 3.1 zeigt ein Schema zum prinzipiellen Ablauf des Feature Extraktion, Composite Scoring und Voting (FECSnV) Algorithmus [82]. Die vorgestellte Methode bezieht die verwendeten Algorithmen über das Bioconductorpaket *CMA* [86], welches

eine standardisierte Schnittstelle zu insgesamt 21 Maschinellen Lernverfahren und statistischen Klassifikatoren anbietet. Gegenwärtig verwendet die FECSnV Methode acht Feature Extraktion Algorithmen (*i.e.* F-test, Golub, Kruskal-test, Limma, RF, RFE, T-test und Welch-test) sowie insgesamt acht Maschinelle Lernverfahren und statistische Klassifikatoren zur internen Bewertung der extrahierten Features (*i.e.* SVM, PLS-LDA, DDA, DLDA, RF, PLS-RF, GBM und FDA). Der modulare Aufbau der Methode gestattet die Integration weiterer Algorithmen, um diese vorläufige Version nachzurüsten. Im Allgemeinen basiert die Methode auf drei grundlegenden Schritten, der Feature Extraktion, deren Bewertung und anschließender Auswahl der diversesten Features. Der Ablauf der Methode soll folgend im Detail erläutert werden. Im ersten Schritt wird eine zuvor definierte Anzahl an Genen über einen der acht vorgegebenen statistischen oder maschinellen Lernalgorithmen extrahiert. Diese Liste aus Genen wird im zweiten Schritt über ein Konsortium aus acht verschiedenen Klassifizierern jeweils mit einem Brier-Maß bewertet, das zurückgegebene Resultat besteht aus der Summe aller erhaltenen Brier-Werte. In jedem Klassifizierungsschritt wird intern eine fünffache Kreuzvalidierung durchgeführt. Die interne Kreuzvalidierung unterteilt die Arrays in zwei verschieden große Trainings- und Testsets. Das Trainingsset dient dem Schätzen der Parameter des Klassifikators, das Testset wird für die Berechnung des Klassifikationsfehlers verwendet. Nach fünf Wiederholungen sind die Parameter und die Performanz des resultierenden Modells optimiert. Allgemein wird durch Kreuzvalidierung die Generalisierbarkeit des Modells nachhaltig verbessert. Wie bereits erwähnt vergeben die Klassifizierer in der Bewertung jeweils einen Brier-Wert zwischen null und eins, d.h. die aus der Bewertung resultierende Summe der Brier-Werte kann theoretisch zwischen null und acht liegen. Der Ablauf wird für alle acht im ersten Schritt festgelegten Feature Extraktion Algorithmen wiederholt. Im abschließenden, dritten Schritt werden die acht insgesamt erhaltenen Genlisten auf die Hälfte reduziert, also vier Genlisten, die jeweils die Gene mit den insgesamt niedrigsten Brier-Werten enthalten. Anschließend wird aus diesen Genlisten jeweils das viertel Gene selektiert, das intern die niedrigst bewerteten Gene enthält. Aus dieser letzten Auslese wird eine neue, aus der Methode resultierende, Genliste erstellt. Inhalt dieser Genliste sind diejenigen Gene, die im Bezug auf die eingangs festgelegten Klassen grundverschieden exprimiert sind. In Zukunft soll die hier entwickelte Methode in ein eigenständiges Bioconductorpaket integriert werden. Darüber hinaus bietet es sich an, einige der internen Abläufe nebenläufig zu implementieren, da z.B. der erste Schritt acht unabhängig agierende Algorithmen umfasst. Auch die nachfolgende Bewertung ist unabhängig und kann daher parallel implementiert werden.



**Abbildung 3.2:** Die Abbildung 3.2 präsentiert die Ergebnisse des Consensus-Clustering.

**Consensus Clustering** Monti *et al.* entwickelten das sogenannte Consensus Clustering. Dies ist eine neue Methode für das Aufdecken von Clustern und ist speziell auf die Analyse von Genexpressionsdaten zugeschnitten [87]. Der Algorithmus wiederholt mehrere Läufe, in denen die Daten zunächst aufgrund von Ähnlichkeit angeordnet werden, so dass sich Cluster bilden. Anschließend werden die Daten perturbiert und zufällig angeordnet, danach startet der Algorithmus den nächsten Lauf. Nach jedem Lauf werden die Ergebnisse der Probenanordnungen aufgrund ihrer Ähnlichkeit in der sogenannten Consensus Matrix festgehalten. Die resultierende Consensus Matrix kann für unterschiedliche Werte von  $k$ , welche die Anzahl potentieller Cluster implizieren, als sogenannte Heatmap visualisiert werden. Die Visualisierung ist zudem auch we-

sentlich für die Erfassung der Stabilität der entstehenden Cluster. Dieser Prozess ist nicht unproblematisch, denn für das Etablieren neuer Cluster muss ein vorangehender geteilt werden. Consensus Clustering wurde in dieser Arbeit in 5000 Wiederholungen von hierarchischen Clustern, basierend auf der Pearson Korrelations Matrix ausgeführt, der Algorithmus wird über das Bioconductorpaket *ConsensusClusterPlus* [88] angeboten.

### 3.2.4 Methoden der Microarrayanalyse

**Gene Set Enrichment Analyse (GSEA)** 2005 stellten Subramanian *et al.* die Gene Set Enrichment Analyse, einen wissenschaftlichen Algorithmus vor [81]. Der Algorithmus testet, ob eine gegebene Gensignatur signifikant innerhalb eines zuvor definierten Gensets angereichert ist. Der Ursprung solcher vordefinierten Gensets liegt in Stoffwechselweg Datenbanken wie KEGG [113], Biocarta [114] und Reactome [115], es können aber auch die drei Gen Ontologien und in wissenschaftlichen Publikationen hinterlegte Signaturen auf Anreicherung geprüft werden. Die in GSEA verfügbaren Signaturen sind in der Molecular Signatures Datenbank ([MSigDB](#)) hinterlegt. Die MSigDB ist in fünf Bereiche gegliedert, welche die verfügbaren Signaturen thematisch geordnet für die Einbindung in die Analyse anbieten. Die GSEA berechnet den Wert der Anreicherung (*i.e.* enrichment score ES) einer gegebenen Signatur in den aus der MSigDB bezogenen Signaturen. Der ES ist begleitet von einer Abschätzung der Signifikanz mittels p-Wert, welcher wiederum über die sogenannte „False Discovery Rate“ (FDR) für multiples Hypothesentesten angepasst wird. In dieser Arbeit wurde ein Genset basierter Permutationstest mit 1000 Permutationen angewendet, welcher die Proben über Student's t-Statistik bewertet. Gensets die einen nominalen p-Wert unterhalb 0,01 und eine FDR unterhalb 0,05 aufwiesen, wurde als signifikant angenommen. Zusätzlich wurden die GSEA Resultate auf Gensets beschränkt, die einen Anreicherungs Wert erzielten, der größer als  $|0,6|$  ist.

**Gene expression-based Outcome for Breast cancer Online (GOBO)** Auf Genexpression basierende, klinische Endpunkte von Brustkrebspatienten konnten über den Webservice [GOBO](#) analysiert werden. GOBO wird von Ringnér *et al.* als ein Projekt der Lund Universität bereitgestellt [116]. In dieser Ressource sind die Genexpressionsprofile von 1881 Brustkrebspatienten vereint. Zusätzlich ist auch eine detaillierte klinische Annotation verfügbar *i.e.* diverse Endpunkte sowie Rezeptor- und Lymphknotenstatus. Weiterhin ist eine Stratifikation auch im Bezug auf Alter der Patienten,

Tumor Grad und Tumor Größe möglich. Die Genexpressionsprofile in GOBO enthalten auch die Profile von 326 Tamoxifen behandelten Patienten aus fünf eigenständigen und somit unabhängigen Studien. Für präzise Details über die verfügbaren Endpunkte sei auf die umfangreiche Dokumentation in GOBO verwiesen. In dieser Arbeit werden die untersuchten Tumor spezifischen Gensets nach Östrogen Rezeptor- und Lymphknoten Status stratifiziert, für die übrigen GOBO Parameter werden die vorgegebenen Werte übernommen, d.h. Tumorselektion: Alle, Zahl der Gruppen: 3 Quantillen und Zensierung nach 10 Jahren.

**Gen Ontologien** Eine formelle Herangehensweise, um die Vielzahl der biologischen Phänomene zu strukturieren, wurde von Ashburner und dem GO Konsortium vorgestellt [117]. Ursprünglich der Metaphysik entstammend wurde das Konzept der Ontologien in den frühen siebziger Jahren in den Forschungsbereich der künstlichen Intelligenz integriert. Heute ist die Analyse der Gen Ontologie ein fester Bestandteil in den Microarrayanalysemethoden, um einen globalen Eindruck über die einer Analyse zugrunde liegenden biologischen Prozesse zu erhalten. Die formale, willkürliche Kategorisierung in „biologischer Prozess“, „molekulare Funktion“ und „zelluläre Komponente“ bietet verschiedene Schichten der Repräsentation von biologischen Daten. Diese Formalisierung beruht auf dem Prinzip, dass die biologischen Kernfunktionen von allen Eukaryoten geteilt werden [117]. Die Gen Ontologie ermöglicht den Abgleich von großen, hierarchisch geordneten Schichten biologischer Information mit den resultierenden Genlisten aus einer typischen Microarrayanalyse. Jedoch ist bei dieser Art Analyse Vorsicht geboten, damit wertvolle Erkenntnisse aus Gen Ontologien bezogen werden können und nicht etwa falsch positive Resultate. Es ist bereits erwähnt worden, dass die statistischen Standardmethoden nur begrenzt in der Microarrayanalyse Einsatz finden. Ein weiteres Problem resultiert aus der hierarchischen Struktur der Ontologien. Häufig werden Analysemethoden dieser Struktur nicht gerecht, da in den meisten Herangehensweisen lediglich eine Term für Term Einzelanalyse durchgeführt wird [118]. Eine Lösung, in der der hierarchischen Struktur Rechnung getragen wird findet sich in der Ontologizer Software von Grossmann *et al.* [118]. Die Autoren bewiesen den Nutzen ihres sogenannten „Parent-Child“ Algorithmus an realen Microarraydaten und zeigten darüber hinaus, dass der Term für Term Einzelansatz zu falsch positiven Resultaten führt. Ein detaillierter Vergleich der verfügbaren GO Termanalyse Werkzeuge ist in Khatri und Draghici gegeben [119].

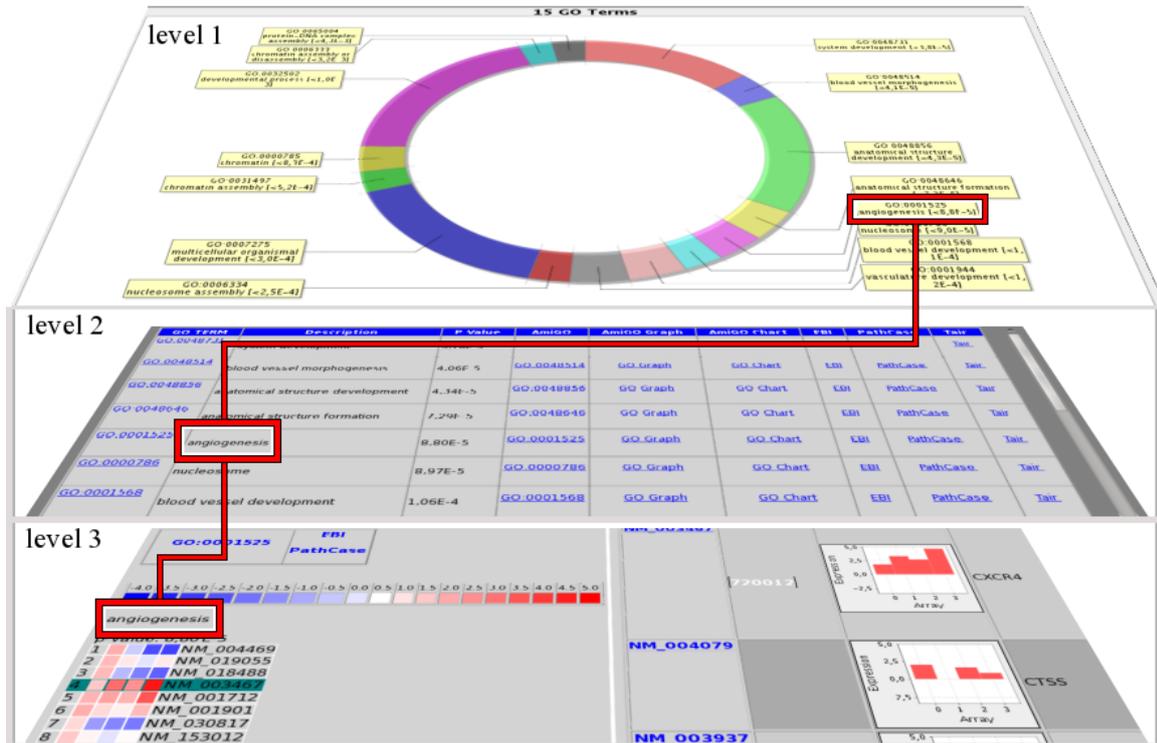
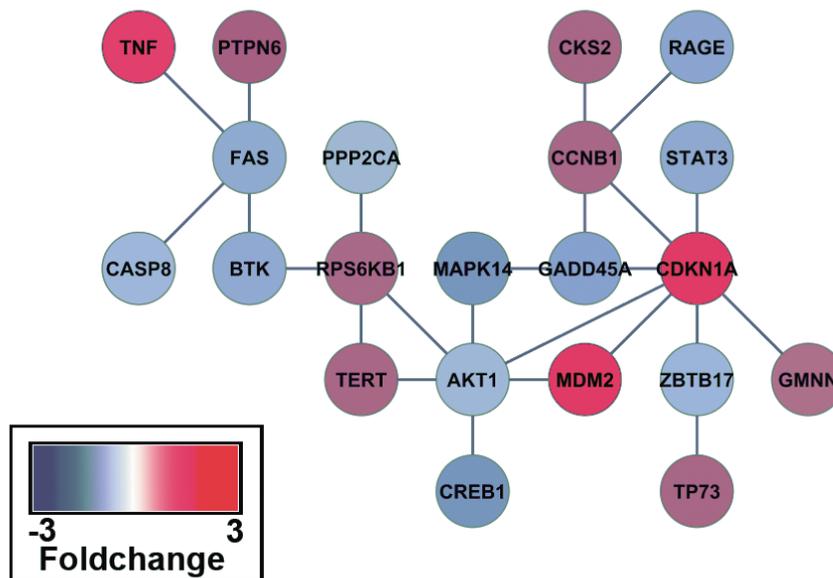


Abbildung 3.3: Die Abbildung 3.3 stellt drei Schichten eines mit maRt generierten GO Term Reports graphisch dar.

**Pathway Analyseressourcen und Werkzeuge** Derzeit bieten sich für die Charakterisierung von Genexpressionsprofilen eine Vielzahl online verfügbarer, biologischer Ressourcen an. Neben den bereits etablierten biologischen Pathwaydatenbanken wie KEGG [113] und BioCarta [114] entstehen neue Pathwayressourcen, die hoch qualitative biologische Information öffentlich zugänglich machen, wie z.B. die Reactome [115] Datenbank. Hier bieten sich online eine Vielzahl an Optionen, *i.e.* das Durchsehen aller verfügbaren Pathways, die Möglichkeit verschiedene Identifier in den Stoffwechselwegen zu finden oder sogar der Vergleich aller bekannten Pathways zweier verschiedener Spezies. Der Webdienst [Pathguide](#) gibt eine aktuelle Übersicht der gegenwärtigen Bemühungen in der Erstellung von biologischen Pathwayressourcen und neu aufkommenden Datenbanken. Heutzutage ist es relativ einfach, sich diverser Datenbank Ressourcen zu bedienen und die Microarrayanalyse mit hoch qualitativen Pathwayinformationen anzureichern. Diese zusätzliche Information bietet hilfreiche Details für das Verständnis der gegenwärtig aktiven Geschehnisse in einer erhaltenen Liste von Genexpressionswerten. Diese Entwicklung wurde in die Wege geleitet durch die heute als System Biologie bezeichnete Disziplin. Die System Biologie unterstützte auch die Entwicklung einer

gemeinsam genutzten Computer Sprache, in dem die zu untersuchenden biologischen Phänomene beschrieben und untereinander getauscht werden können (<http://sbml.org> [SMBL](#)). Die Möglichkeit verschiedene biologische Schichten übereinander zu lagern und die zellulären Aktivitäten zu simulieren ist eines der Hauptanliegen der System Biologie. In Pathway-Express stellen Draghici *et al.* [120] eine Systembiologische Methode vor, um Muster in Pathways zu analysieren. Die Methode beruht auf einer Wirkungsfaktor Analyse (*i.e.* impact factor), welche die Relevanz biologischer Parameter einbezieht. Anhand drei realer Microarraydatensets konnte gezeigt werden, dass der Algorithmus die gängigen Pathwayanalyse Methoden übertrumpft. Ein System biologischer Ansatz erweist den Nutzen und die Gangbarkeit der Pathwayanalyse durch die Kombination verschiedener Schichten biologischer Information. Künftig sollte in der Analyse von Microarraydaten die Verwendung verschiedener biologischer Ressourcen und Informationsschichten zur gewöhnlichen Strategie werden. Dies könnte auch Vorteile für die Reproduzierbarkeit und die Vergleichbarkeit der Resultate bedeuten. Darüber hinaus würde auch die Breite und die Tiefe des Verstehens der zugrunde liegenden biologischen Prozesse zunehmen. Dieses Verständnis kann auch durch Anwendung verschiedener Werkzeuge und Ressourcen vertieft werden. Werkzeuge die in dieser Arbeit für die Pathwayanalyse Verwendung finden sind neben Pathway-Express auch Cytoscape [121] und MetaCore<sup>TM</sup> [122, 123].

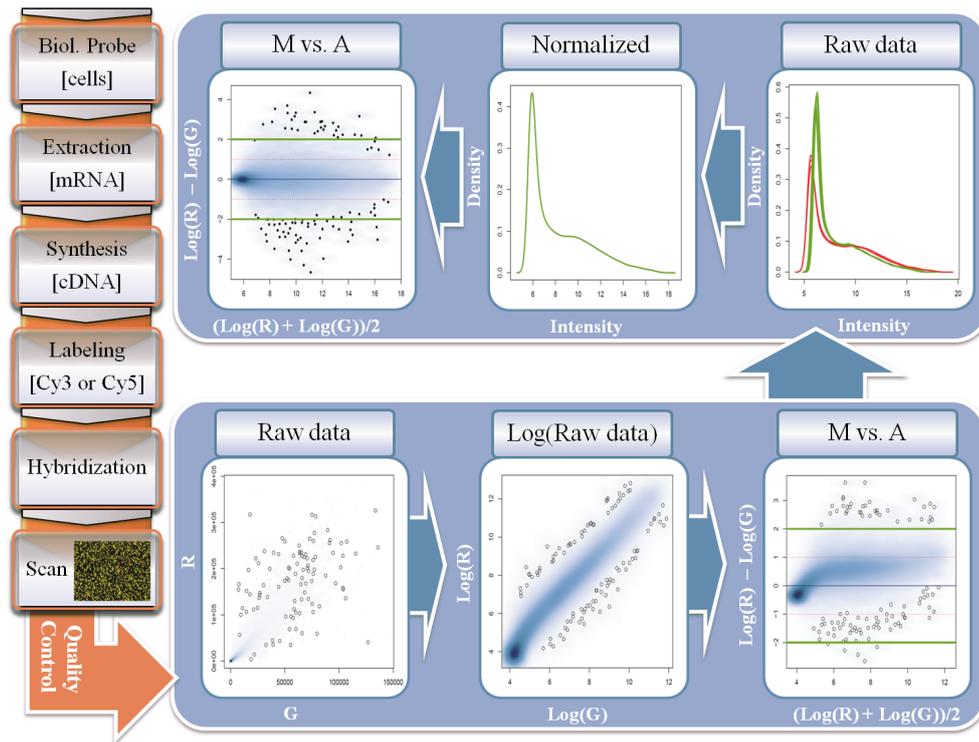


**Abbildung 3.4:** In Abbildung 3.4 ist das Resultat einer Pathway Analyse gezeigt, typischerweise werden die Genexpressionswerte auf den Knoten des Graphen farblich dargestellt.

## 3.3 Microarray-Auswertung und -Visualisierung

### 3.3.1 Microarrayarbeitsfluss

In Abbildung 3.5 ist der Microarrayarbeitsfluss schematisch dargestellt. Heutzutage können Microarraydaten aus verschiedenen Quellen bezogen werden, die zudem verschiedene biologische Informationsschichten charakterisieren. Da alle Microarrayformate den selben technologischen Ursprung aufweisen, sind die Arbeitsabläufe, die zur Gewinnung der Daten dienen, größtenteils vergleichbar. An dieser Stelle wird ein allgemeiner Überblick gegeben, in dem die einzelnen Schritte eines typischen Microarrayexperiments sukzessive vorgestellt werden. Der erste Schritt dient der Gewinnung von Probenmaterial. Meist wird das Material als Zellkultur etabliert, wie in der Abbildung 3.5 gezeigt. Im nächsten Schritt wird die zu untersuchende biologische Schicht extrahiert, welche häufig angereichert und in eine stabilere Form gebracht werden muss. Für den Fall, dass z.B. die aktive Expression des Transkriptom untersucht werden soll, wird das mRNA Probenmaterial zunächst in die stabilere copy DNA (cDNA) umsynthetisiert. Im Anschluss wird das genetische Material mit einem Farbstoff markiert. Zwei unter-



**Abbildung 3.5:** Die Microarraytechnologie erfasst die Genexpression biologischer Proben. Die Abbildung 3.5 stellt den experimentellen Ablauf links im Bild und die Verarbeitung der gewonnenen Rohdaten rechts im Bild dar. [15]

schiedliche Farbstoffe werden benötigt, wenn Zweikanal-Microarrays verwendet werden. Das fluoreszenzmarkierte Material wird anschließend auf die Microarrayoberfläche hybridisiert. Die unspezifische Hybridisierung kann anschließend abgewaschen werden und nach einer kurzen Inkubationszeit kann ein Laser den Farbstoff zur charakteristischen Emission anregen und schließlich ein Bild der Intensität der Fluoreszenz aufgezeichnet werden. Die aufgezeichneten Intensitätswerte dienen im Nachhinein als vertretend für die Genexpressionswerte. Nach den obligatorischen herstellereinspezifischen Qualitätskontrollen der so gewonnenen Rohdaten bezeichnet man die Intensitäten als „R“ (*i.e.* rot) und „G“ (*i.e.* grün). Werden Einkanal-Microarrays verwendet, so ist das erste Array „R“ und das folgende, als Referenz dienende Array „G“. In größeren Einkanal-Datensets ist es üblich, statt eines einzigen Arrays, den Median aller Arrays als Referenz festzulegen. Ein Scatterplot der Rohintensitäten resultiert in einem schlecht skalierten Bild der Expressionswerte und ist im unteren Teil der Abbildung 3.5 gezeigt. Das nachfolgende Bild zeigt, dass durch logarithmische Transformation die Werte eine vergleichbare Skala erhalten. Jedoch zeigt dieses Bild auch, dass die Verteilung verfälscht ist, da ei-

ne winkelhalbierende Gerade zu erwarten wäre. Die Verzerrung wird deutlicher, wenn die Expressionswerte auf beide Skalen doppelt logarithmisch in einem sogenannten M (*i.e.* minus) versus A (*i.e.* add) Plot aufgetragen werden. Die Expressionswerte werden durch die doppelt logarithmische Skalierung in die Horizontale transformiert. Dabei repräsentieren die horizontalen Koordinaten die durchschnittliche Genexpression und die vertikalen Koordinaten entsprechen dem Grad der differentiellen Genexpression als der Veränderungsrate (*i.e.* fold-change). Die Intensitätswerte weisen aufgrund von technologischen und experimentellen Faktoren Abweichungen auf. Daher ist eine weitere Datenverarbeitung notwendig, um die Daten vergleichbar zu machen. Mit der Quantilnormalisierung werden die gemessenen Intensitätswerte transformiert. Jedes Quantil der R-Intensität und G-Intensität wird dazu absteigend sortiert und schließlich durch den Median der Expressionswerte ersetzt, um so identische Verteilungen zu erhalten. Das bedeutet also, dass die eigentlich gemessenen Intensitäten verworfen werden und die Genexpression durch den Median der gemessenen Genexpression ausgetauscht wird, die Gennamen jedoch bleiben von diesem Austausch unberührt. Im ersten Bild der Abbildung 3.5 ist im oberen Teil der M- versus A-Plot der normalisierten Genexpressionswerte gezeigt. Wie erwartet bildet sich eine horizontale Linie, da ein Großteil der Gene als nicht differentiell exprimiert angenommen wird. In der Literatur sind verschiedene Strategien zur Normalisierung von Microarraydaten zu finden, die meisten jedoch sind spezifisch für Einkanal- [124, 125] oder Zweikanal- [126] Microarrays. Zusätzlich finden sich weitere herstellerepezifische Herangehensweisen, um Artefakte in Microarraydaten zu entfernen und die Arrays letztlich miteinander vergleichbar zu machen [127–130].

### 3.3.2 Qualitätssicherung von Microarraydaten

**Öffentlich verfügbare Microarray Studien in GEO** Microarrayrohdaten öffentlich verfügbarer Studien wurden aus der GEO Datenbank [22] bezogen. In Tabelle 3.1 sind alle Studien gelistet, welche einer qualitativen Untersuchung unterzogen wurden.

**Tabelle 3.1:** Die Tabelle fasst die untersuchten Studien aus GEO zusammen.

GEO ID	Anzahl	Referenz	Öffentlich seit
GSE9801	6	[131]	Nov 07, 2011
GSE32700	46	[132]	Okt 07, 2011
GSE9936	105	[133]	Feb 07, 2008

Alle untersuchten Studien beruhen auf dem Affymetrix eigenen Human Genome U133A Format und wurden im Jahr 2011 veröffentlicht, mit Ausnahme von Studie GSE9936

diese ist bereits seit 2008 öffentlich verfügbar.

**Datenvorverarbeitung, Normalisierung und Hauptkomponentenanalyse** Die Aufbereitung der Microarraydaten wurden mit R [134], Version 2.15 und Bioconductor [34] Version 2.10 durchgeführt. Microarray Rohdaten wurden über das Bioconductorpaket *GEOquery* [135] geladen und mit dem Paket *affy* [136] als Batch verarbeitet. Zusätzlich wurde *affy* auch dazu benutzt, Kandidaten aufzudecken, die potentiell RNA-Degradation aufweisen. Die nachfolgenden Qualitätskontrollen wurden mit dem *yaqcaffy*<sup>1</sup> Paket vorgenommen. Dieses ist speziell auf Affymetrix Arrays ausgerichtet. Es wurde berechnet, ob Ausreißerarrays in Bezug auf die durchschnittliche Hintergrundintensität oder den durchschnittlichen Rauschwert vorliegen. Zusätzlich wurden Ausreißer, die in den beiden Kontrollproben für Haushaltsgene (*i.e.*  $\beta$ -Actin und GADPH) vorlagen, und auch Ausreißer in den internen, sogenannten spike-in Proben, sowie den Poly-A Kontrollen ermittelt. Schließlich wurden alle Microarrays einer Quantilnormalisierung ohne Hintergrundabgleich mit dem *RMA* Paket [137] durchgeführt. Im Anschluss wurde mit den normalisierten Werten eine Hauptkomponentenanalyse über das *pcaMethods* Paket [138] durchgeführt. Die Werte der ersten, über den Mittelwert zentrierten Hauptkomponente wurden für die darauffolgende Visualisierung gespeichert.

**Visualisierung der Qualitätsparameter mit Circos** Die Ergebnisse der hier vorgestellten Qualitätskontrollen wurden mit Hilfe dedizierter R Skripte zusammengefasst. Darüber hinaus wurden für Circos die spezifischen Eingabedateien generiert, welche die Koordinaten der Darstellungen beinhalten. Circos [78], Version 0.6 und Perl<sup>2</sup> Version 5.1.16 wurden verwendet, um die zirkulären Qualitätsabbildungen zu erstellen. Zusätzlich wurde auch eine für Circos spezifische Konfigurationsdatei erstellt, in der über Befehle festgelegt ist, wie und in welcher Reihenfolge die Daten präsentiert werden. Der R Code und eine Demonstration mit öffentlichen, in GEO abgelegten Daten ist auf GitHub verfügbar unter der Adresse: <http://github.com/buzzmak/circos-arrayQC>.

### 3.3.3 Das „Microarray Analysis and Reporting Tool: maRt“

An dieser Stelle soll ein typischer Datenanalyseablauf für Oligonukleotid-Microarrays nachgezeichnet werden. Dabei werden die verfügbaren Funktionen und die verwendeten Bibliotheken vorgestellt.

---

<sup>1</sup><http://www.bioconductor.org/packages/release/bioc/html/yaqcaffy.html>

<sup>2</sup><http://strawberryperl.com>

**Import und Normalisierung von Microarraydaten** Die Software unterstützt derzeit neben Textdateien, die durch Tabulator begrenzt sind, die Formate folgender Hersteller: Affymetrix, Agilent und Illumina. Zusätzlich lassen sich Microarraydaten auch über die durch die Array-Express Datenbank vergebene Referenznummer direkt importieren. Künftig soll auf diese Weise auch die Gene Expression Omnibus Datenbank GEO integriert werden. Darüber hinaus bestehen Pläne, die auch den Import und die Analyse von aCGH Datensets ermöglichen sollen. Derzeit werden die Microarraydaten über nebenläufige Verarbeitung importiert und liegen anschließend als Kopie im Arbeitsverzeichnis der Software. Diese Maßnahme gewährleistet, dass eine explorative Datenanalyse die Originaldaten unversehrt belässt. Die Normalisierung der Daten ist über dedizierte Bioconductor [34] Pakete gegeben, die solche Funktionalität für Einkanal- (affy [136]) und Zweikanal- (limma [126]) Microarrays bieten. Diese Pakete bieten vielseitige Strategien für die Datenvorverarbeitung und werden über Menüpunkte im Dialog Normalisierung angeboten. Über diesen Dialog lassen sich die Standard Routinen zur Normalisierung von Einkanal- (RMA und *expresso*) und Zweikanal- (*quantile* und *Aquantile*) Microarrays aufrufen. Das Speichern der normalisierten Arrays und auch der gegenwärtigen Sitzung ist über das Dateimenü möglich. Der aktuelle Zustand der Sitzung wird archiviert und kann in einer neuen Sitzung wieder aufgerufen werden.

**Filterung von Microarraydaten** Die Software filtert Microarraydaten, spaltenbasiert nach einzelnen Werten, einer oder zwei Bedingungen oder aufgrund einer vordefinierten Liste von Genen. Einfache Textdateien, die durch Tabulator begrenzt sind, oder Dateien im „gene association“ Format der Gene Ontologie (GO) Datenbank [117] können in *maRt* als Suchkriterien verwendet werden. Auf diese Weise können alle Genexpressionswerte spezifischer GO Termini in einem Datenset abgefragt werden. Eine Molekülliste, die der „Pathway Interaction“ Datenbank (PID) [139] entstammt, kann ebenso mit *maRt* verarbeitet werden. Auf diese Weise können leicht alle Genexpressionswerte zu den Genen eines spezifischen Stoffwechselweges erhalten werden. Das Screening großer Datensets basierend auf spezifischen Stoffwechselwegen kann zur Aufdeckung von Genen führen, die eine Schlüsselrolle in eben diesem Stoffwechselweg übernehmen. Die Filterfunktion in *maRt* ist über ein „Producer-Konsumer“ Muster [140] gelöst, welches intern eine binäre Suche implementiert. Zusätzlich konnte an dieser Stelle auch durch nebenläufige Datenverarbeitung die Geschwindigkeit gesteigert und auf den Einsatz der ehemals verwendeten, nativen Java Datenbank (*Derby*) verzichtet werden.

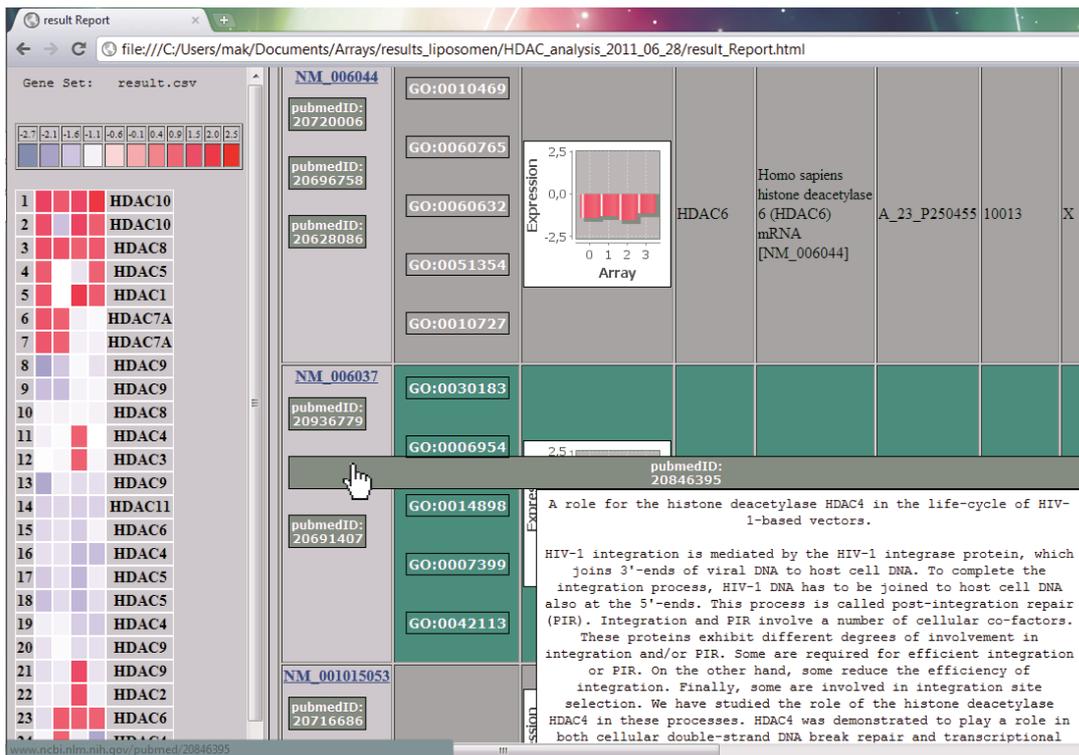
**Annotation von Transkripten** Die Annotation der microarrayresidenten Transkripte, also die Zuordnung von Genproben und Gennamen wird völlig transparent als ein im Hintergrund ausgeführter Webdienst über die BioMart Datenbank [141] realisiert. Dieser Dienst bietet Zugriff auf lebenswissenschaftliche Datenressourcen wie dem Swiss-Prot Teil der UniProt Wissensdatenbank und erhält somit die aktuellsten Beschreibungen. Durch die transparente Einbindung des BioMart Webservice werden die Anforderungen an den Nutzer im Bezug auf Kenntnisse und Interaktionen auf ein Minimum reduziert.

**Zugriff auf Gen Ontologie und Stoffwechselwege** Einerseits kann maRt den DAVID Webservice für funktionelle Annotation [25] intern aufrufen, um schnellen und automatischen Zugriff zu gewähren. Andererseits kann über maRt auch direkt die DAVID Webseite aufgerufen werden, in der das aktuelle Datenset bereits vorgeladen ist. Dort kann der Nutzer von den Optionen und Möglichkeiten des sehr gut dokumentierten DAVID Webdienstes profitieren. Sobald die online generierten Ergebnisse der funktionellen Annotationsanalyse wieder in maRt geladen sind, bietet die Software GO Terme und Stoffwechselwege aus BioCarta [114] und KEGG[113].

**Die Generierung eines HTML-Reports** Derzeit können drei verschiedene HTML-Reports erstellt werden, die jeweils unterschiedliche Inhalte zusammenfassen:

- Ein Report bietet eine sogenannte Heatmap, in der alle zuvor selektierten Gene dargestellt werden. Zusätzlich sind detaillierte Annotationen bezüglich aller Proben verfügbar, welche von den letzten drei aktuellen Publikationen aus der PubMed Datenbank begleitet werden, soweit diese vorhanden sind. Alle Proben sind ebenfalls mit der Nucleotid Datenbank von Entrez Gene verbunden.
- maRt ist in der Lage aus den eingespeisten DAVID Daten einen dreischichtigen GO Term Report zu erstellen, welcher eine Übersicht über alle GO Terme in der Analyse ermöglicht. Dieser Report reicht bis zu den Expressionsniveaus einzelner Gene, welche als Heatmap für jeden spezifischen GO Term erstellt werden, wie in Abbildung 3.3 dargestellt.
- Anhand der DAVID Ergebnisse wird darüber hinaus ein weiterer Report erstellt, der alle gefundenen Stoffwechselwege mit darauf bezogenen p Werten und Genen extrahiert.

Auf diese Weise kann die Software über ein einfaches und intuitiv bedienbares graphisches Interface GO Termini mit Genexpressionsprofilen verknüpfen. Dies war kürzlich

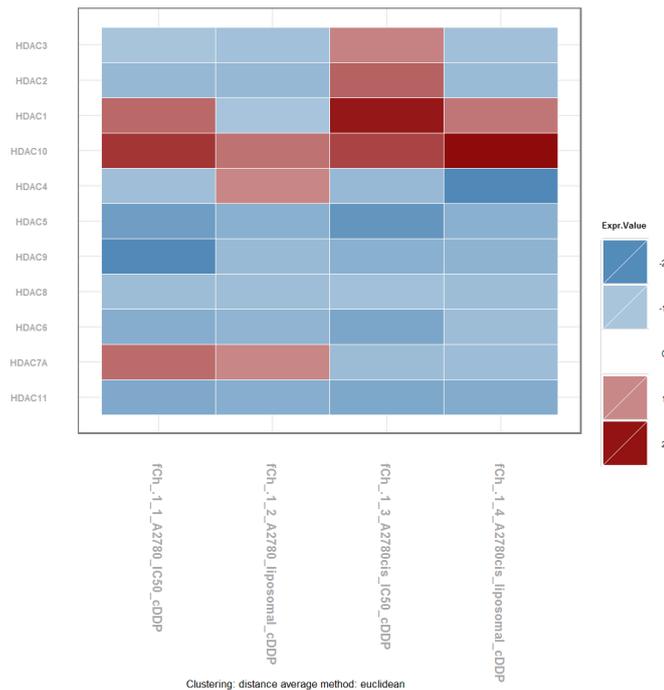


**Abbildung 3.6:** Die Abbildung zeigt, wie die aus einer Microarrayanalyse resultierenden Genlisten als HTML-Report gespeichert werden, um einen Zugang zu weiterführender Information zu erhalten.

nur über die Kommandozeile möglich und jenen Experten überlassen, die über die nötigen Kenntnisse in R Programmierung verfügten. Die Resultate einer maRt Analyse sind im HTML Format gespeichert. Dieses Format wurde ausgewählt, da Webbrowser weitestgehendst vorhanden sind, die Ergebnisse mit anderen geteilt werden können und außerdem weiterführende Informationen nur einen Klick entfernt sind.

**Clusteranalyse** Einfaches dendrogrambasiertes Clustern ist in maRt über native R Funktionen gegeben, wie zum Beispiel hierarchisches Clustern oder Clustern über K-Means. Zusätzlich ist ein statistisches Clustern über das Paket Pvcclus [142] gegeben.

**Lineare Modelle für die Microarraydatenanalyse mit limma** maRt bietet ebenso eine Analyse der differentiellen Genexpression über lineare Modelle mit *limma* [106]. Die Resultate solcher Analysen können mit dem *ggplot2* Paket [143] als hoch qualitative Vorschau abgebildet werden.



**Abbildung 3.7:** Die Abbildung visualisiert über das Paket *ggplot2* differentiell exprimierte Gene in den vertikal angeordneten Proben. Die logarithmierten Expressionswerte sind rot dargestellt, wenn das Gen induziert und blau, wenn das betreffende Gen reprimiert ist.

**Automatisierung** In *maRt* ist eine Routine zur Automatisierung namens *automaRt* implementiert. Somit ist eine einfache Methode zur Wiederholung bereits etablierter Analyseschemata gegeben. Ein weiterer Aspekt ist, dass die Lebenswissenschaften ein hoch dynamisches Forschungsgebiet sind, in dem der aktuelle Wissensstand täglich vorangebracht wird. Dieser Dynamik unterliegen damit auch alle Annotationen von Microarrayproben sowie die Zugehörigkeit der Proben zu bestimmten Stoffwechselwegen. Daher könnte es notwendig sein, besonders ältere Datensets durch eine Wiederholung der Analyse auf den aktuellsten Kenntnisstand der Forschung zu bringen.

**Explorative Analyse mit R** Einige Nutzer werden den R Code inspizieren wollen, der für die statistischen Berechnungen in *maRt* ausgeführt wird. Zusätzlich soll auch die Möglichkeit gegeben sein, neue Eigenschaften hinzuzufügen. Dies ist über die R Konsole in *maRt* möglich, alle vorgenommenen Änderungen können permanent gemacht werden. Die Möglichkeit, den neu erstellten Code mit anderen Nutzern zu teilen, wäre über ein Internetforum gegeben, welches in Zukunft etabliert werden kann.

**Exemplarische Datensets** Da maRt eine breite Auswahl an Bioconductorpaketen [34] integriert, können auch Daten wie z.B. das bekannte akute lymphatische Leukämie Datenset aus Ross *et al.* [144] per Mausklick geladen werden. So können die Eigenschaften, die maRt bietet an einem bekannten Datenset ausgiebig getestet werden.

**Ausgabedateien für Software von Drittanbietern** Die Software ist in der Lage, Eingabedateien für verschiedene Arten bioinformatischer Software zu generieren. So ist es z.B. möglich, die Analyse in MeV [23], Cytoscape [121] und BioLayout Express 3D [145] weiterzuführen. Diese Werkzeuge haben eigene interessante Analysestrategien und erweitern somit die Breite der Möglichkeiten. Darüber hinaus gibt es in maRt einen dedizierten Exportdialog, welcher es dem Nutzer gestattet, die Daten nach seinen Vorlieben zu arrangieren. Abschließend können also mit Hilfe der Vorschau und Reporting-Fähigkeiten von maRt, wertvolle Ergebnisse in hoher Qualität erstellt und gesichert werden.

## 3.4 Wirkstoffstudie mit liposomalem Cisplatin

### 3.4.1 Herstellung der Liposomen

Die Liposomen wurde von Frau Dr. Michaela Krieger aus dem Arbeitskreis Prof. Bendas hergestellt. Die Details sind bereits in Krieger *et al.* [45] beschrieben. An dieser Stelle sollen nur die wesentlichen Arbeitsschritte genannt werden.

Ein getrockneter Lipidfilm, welcher die angegebene Lipidzusammensetzung enthält (SPC/Chol/mPEGPE/cyanurPEG-PE; 65/30/3/2, molares Verhältnis) wurde bei 65 °C mit einer 0.9 prozentigen NaCl Lösung hydriert, die 8 mg/mL Cisplatin enthält. Anschließend wurden eine Ultrabeschallung zur Größenhomogenisierung angewandt. Holotransferrin wurde wie beschrieben mit den CyanurPEG-PE Ankern der Liposomen gekoppelt und anschließend über Gel-Permeations-Chromatographie (Sephadex G50) aufgereinigt, um das ungebundene Holotransferrin und das nicht in Liposomen eingeschlossene Cisplatin zu ersetzen. Die Partikelgröße der Liposomen wurde mittels dynamischer Lichtstreuung (DLS) erfasst. Die Quantifizierung der intrazellulären Platinakkumulierung wurde durchgeführt wie in [45] beschrieben. Es folgt eine Übersicht über die verwendeten Chemikalien. Soja Phosphatidylcholine (SPC) wurde durch die Lipoid AG (Ludwigshafen, Deutschland) bereitgestellt. Polyethyleneglycolphosphatidylethanolamine (mPEG-PE) wurde über Avanti Polar Lipids (Alabaster,

AL, USA) bezogen. Cyanur-PEG-PE als Lipid Anker für die Protein Anbindung wurde synthetisiert wie in [146] beschrieben. Cyanur Chlorid, CDDP, Holotransferrin (holo), Cholesterol (chol), Trichostatin (TSA) und Sephadex G-50 wurden bei Sigma-Aldrich Chemie (Steinheim, Deutschland) gekauft. DMSO und MTT wurden über AppliChem (Darmstadt, Deutschland) und Salpetersäure 65 % (Suprapur<sup>®</sup>) wurde über Merck (Darmstadt, Deutschland) bezogen. Das QiAamp<sup>®</sup>-Blood-Mini-Kit ist von Qiagen (Hilden, Deutschland), alle übrigen Chemikalien wurden bei AppliChem erworben (Darmstadt, Deutschland) soweit nicht anders vermerkt.

### 3.4.2 Zellkultur

Die Ovarialkarzinomzelllinie A2780 und ihre CDDP resistente Variante A2780cis wurden über die ECACC erworben. Die ECACC listet diese Zelllinien unter den Nummern 93112519 (A2780) und 93112517 (A2780cis). Die Zellen wurden von Dr. Michaela Krieger in RPMI 1640 Medium (PANTM-biotech, Passau, Deutschland) kultiviert und den folgenden Bedingungen entsprechend inkubiert: 10% fetales Kälberserum (FKS) (Sigma-Aldrich), 5% (50  $\mu\text{g}/\text{mL}$  Streptomycin and 50 U/mL Penicillin G, Sigma-Aldrich) und 1,5% (365  $\mu\text{g}/\text{mL}$  L-Glutamine, Sigma-Aldrich) bei 37°C und 5 % CO<sub>2</sub>.

### 3.4.3 MTT-Zellzytotoxizitäts-Test

Frau Dr. Michaela Krieger hat die im Folgenden beschriebenen MTT-Zytotoxizitäts-Tests durchgeführt, um die zytotoxische Aktivität des freien CDDP sowie des liposomalen CDDP zu messen. Die Zellen wurden in 96 Well-Platten mit transparentem Boden über Nacht präinkubiert, zu je  $2 \cdot 10^4$  Zellen pro Well in 90  $\mu\text{L}$  Medium. Am nächsten Tag wurden die Zellen mit jeweils ansteigenden Konzentrationen von CDDP oder liposomalem CDDP für 72 Stunden bei 37°C und 5 % CO<sub>2</sub> inkubiert. Anschließend wurden die Zytotoxizitäts-Tests durchgeführt, wobei zu jedem Well 20  $\mu\text{L}$  MTT Reagens hinzugefügt wurden (5 mg MTT / 1 mL PBS), für etwa eine Stunde, bis Formazankristalle entstanden. Der Überstand wurde entfernt und die Zellen in 100  $\mu\text{L}$  DMSO aufgelöst. Die Absorption des Farbstoffes wurde bei 570 nm mit Hintergrundnivellierung bei 690 nm über einen sogenannten Plate Reader gemessen (Thermomultiscan EX, Thermo, Schwerte, Deutschland). Die TSA enthaltenden MTT Zytotoxizitäts-Tests fanden ebenso in 96 Well-Platten mit transparentem Boden zu je  $2 \cdot 10^4$  Zellen pro Well, jedoch in 80  $\mu\text{L}$  Medium statt. Die Zellen wurden über Nacht kultiviert und zwei Stunden

zuvor mit 10  $\mu\text{L}$  TSA (100 nM) inkubiert [147] Anschließend wurden die Zellen mit jeweils ansteigenden Konzentrationen von CDDP oder liposomalem CDDP behandelt, wie oben beschrieben.

### 3.4.4 Messung der DNA Platinierung

Die Messung der DNA Platinierung wurde von Frau Dr. Michaela Krieger und Dennis Alex unter Anleitung von Dr. Alex Hilger im Universitätsklinikum Essen durchgeführt. Für die Messung wurden  $1 \cdot 10^6$  A2780 und A2780cis Zellen, über Nacht in sechs Well Platten kultiviert. Am nächsten Tag wurde das Medium gewechselt und CDDP oder liposomales CDDP hinzugefügt, in finaler Konzentration von jeweils 20  $\mu\text{M}$  pro Well und wieder für 24 Stunden bei 37°C und 5 % CO<sub>2</sub> inkubiert. Nach der Inkubation wurde das Medium schnell abgeschöpft und die Zellen einmal in 1 mL eiskaltem PBS (4°C) gewaschen. Anschließend wurde das PBS entfernt, die Zellen für 2 Minuten mit Trypsin verdaut und in kaltem Medium (4°C) resuspendiert und in 2 mL Eppendorf-Röhrchen bei 1500 g für eine Minute zentrifugiert. Der Überstand wurde verworfen und das Zellpellet in 1 mL eiskaltes PBS resuspendiert und bei 24.100 g für eine Minute zentrifugiert. Dieser Überstand wurde ebenso verworfen und das Pellet in 1 mL eiskaltem PBS gewaschen und erneut bei 24.100 g zentrifugiert. Letztlich wurde der Überstand entfernt und das Pellet bei -20°C eingefroren. Nach dem Auftauen wurde die DNA über Fest-Phasen Extraktion isoliert (QIAamp<sup>®</sup>, Qiagen, Hilden, Deutschland). Über UV Photometrie wurde der so erhaltene DNA-Gehalt gemessen (UV/VIS-Spectrophotometer, Model Genesys 6; Thermo Electron, Dreieich, Deutschland). Zur Charakterisierung des Platingehalts wurden die Proben mit dem Speed-Vac<sup>®</sup> Univa-po 100 H getrocknet (UniEquip Laborgerätebau and Vertriebs GmbH, Martinsried, Deutschland). Die getrocknete DNA wurde später für 24 Stunden bei 70°C in 1 % HNO<sub>3</sub> (suprapure<sup>®</sup>) lysiert. Die Proben wurden schließlich über Massenspektrometrie mit induktiv gekoppeltem Plasma (ICP-MS) analysiert [148]. Platin Konzentrationen [ $\mu\text{g}$ ] wurden über die Relation der Gesamt DNA Konzentration berechnet [ $\mu\text{g}$ ].

### 3.4.5 RNA Isolierung und Microarraydaten Gewinnung

Die Versuche wurden von Dr. Michaela Krieger und Norbert Brenner am Center of Advanced European Studies and Research (caesar) übernommen. Dazu wurden die Zellen bei Populationsdichten von  $2 \cdot 10^6$  (A2780) und  $3 \cdot 10^6$  (A2780cis) in 15 mL RPMI Medium über Nacht in T-75 Flaschen inkubiert. Das Zellkulturmedium wurde

anschließend durch 15 mL auf 37°C angewärmtes Medium ersetzt, welches entweder freies CDDP oder liposomales CDDP oder leere Liposomen enthielt. Dieser Ansatz wurde 72 Stunden bei 37°C und 5 % CO<sub>2</sub> inkubiert. Für die RNA Isolierung wurden unbehandelte A2780 und A2780cis Zellen, sowie Zellen, die einer IC<sub>50</sub> Konzentration an freiem CDDP ausgesetzt waren verwendet, wie oben beschrieben. Ebenso wurden die Zellen, welche mit liposomalem Wirkstoff inkubiert wurden, mit IC<sub>50</sub> Konzentrationen behandelt. Für die Microarraydatengewinnung wurden auch leere Liposomen eingesetzt, um den Liposomeneffekt auszubalancieren. Dazu wurden A2780 und A2780cis Zellen mit leeren holotransferrin-gekoppelten Liposomen inkubiert. Die folgenden acht Experimente wurden von Norbert Brenner aus dem Arbeitskreis Dr. Royer angesetzt, die letztlich auf vier Zweikanal-Microarrays gegeben wurden.

- A2780 unbehandelt versus A2780 behandelt mit CDDP
- A2780cis unbehandelt versus A2780cis behandelt mit CDDP
- A2780 leere Liposomen versus A2780 beh. mit liposomalem CDDP
- A2780cis leere Liposomen versus A2780cis beh. mit liposomalem CDDP

A2780 Zellen wurden mit 1,72  $\mu M$  und A2780cis Zellen wurden mit 8,94  $\mu M$  CDDP behandelt. Diese Konzentration wurde auch in Liposomen erreicht. Die Lipidkonzentrationen der Liposomen sind 0,80  $\mu mol$  (A2780) und 4,15  $\mu mol$  (A2780cis). Für die Isolation der RNA wurde das Medium entfernt und die Zellen mit 10 mL PBS gewaschen und schließlich mit DTT-haltigem RLT-Puffer lysiert. Die Zelllysate wurden bei -80°C aufbewahrt bis die RNA-Extraktion durchgeführt wurde. Agilent 44k Human Genome Zweikanal-Microarrays wurden mit den oben beschriebenen Proben hybridisiert. Zur Kontrolle wurden die RNA Proben (je 500 ng) amplifiziert und die unbehandelten Proben mit Cy3<sup>®</sup>-CTP, die behandelten Proben jedoch mit Cy5<sup>®</sup>-CTP gefärbt, um cRNA zu erhalten (gemäß Protokoll des Herstellers). Das Verhältnis der Farbstoffeinkonzentration wurde mit einem Nanodrop-Photometer gemessen und betrug mehr als 8 pmol Farbstoff pro  $\mu g$  cRNA (Kisker, Steinfurt, Deutschland). Für die Array Hybridisierung wurden jeweils 825 ng Cy3<sup>®</sup> gefärbte Kontrolle gegen 825 ng Cy5<sup>®</sup> gefärbte behandelte Probe gemischt und gemäß der Anweisungen im Protokoll des Herstellers inkubiert. Die erhaltenen Arrays wurden mit einem Agilent Microarray Scanner ausgelesen. Die Genexpressionswerte konnten mit der Agilent Feature Extraktion Software ermittelt werden.

## 3.5 Tumorprogressionsstudie anhand von Zervixkarzinomen

### 3.5.1 Microarraydaten

Microarrayrohdaten aus vier öffentlich verfügbaren Zervixkarzinom Studien wurden aus der [GEO](#) Datenbank [22] bezogen. Die folgenden Studien wurden auf qualitativ hochwertige Arrays evaluiert.

Studie **A**: Referenznummer GSE5787, Bachtiry *et al.*, Microarrays im Format Human Genome U133A Plus 2.0 [26]. Ziel dieser Studie war die Untersuchung der intra Tumorerheterogenität von Zervixkarzinompatienten. Als Ergebnis wurde ein Modell erstellt, welches das Verhältnis aus intra und inter Tumorvariabilität bildet. So ergibt sich ein Maß für die intra Tumorerheterogenität, dieses Maß kann ebenso für die Zuverlässigkeit von Biomarkern herangezogen werden.

Studie **B**: Referenznummer GSE6791, Pyeon *et al.*, Microarrays im Format Human Genome U133A Plus 2.0 [149]. Diese Studie hatte das Ziel Genexpressionsprofile von HPV- und HPV+ Kopf-Hals-Karzinomen mit Genexpressionsprofilen von Zervixkarzinomen zu vergleichen.

Studie **C**: Referenznummer GSE9750, Scotto *et al.*, Microarrays im Format Human Genome U133A [150]. Das Ziel dieser Studie war eine funktionelle Untersuchung von Mutationen des langen Arms von Chromosom 20 in den frühen Stadien von Zervixkarzinomen. Neben der funktionellen Charakterisierung wurde auch nach potentiellen Biomarkern geforscht. Diese Studie enthielt ebenso Referenzproben aus gesundem Zervixgewebe.

Studie **D**: Referenznummer GSE26511, Nordhuis *et al.*, Microarrays im Format Human Genome U133A Plus 2.0 [151]. Diese Studie umfasst eine Untersuchung der frühen Stadien von Zervixkarzinomen und identifiziert tumorcharakteristische Stoffwechselwege die im Zusammenhang mit dem Lymphknotenstatus stehen.

### 3.5.2 Zervixkarzinompatienten Charakteristika

Die Vereinigung der oben genannten Studien ergab insgesamt Microarraydaten von 102 Zervixkarzinombiopsien und 24 Microarrays, die Genexpression von gesundem Zervix aufzeichnen. Alle Patienten sind zwischen 27 und 77 Jahren alt (median 42,5) und wurden somit retrospektiv untersucht.

**Tabelle 3.2:** Die Tabelle fasst die Anzahl untersuchter Patienten in FIGO Stadien zusammen.

NC	IB	IB1	IB2	IIA	IIB	IIIB
24	12	21	15	10	25	19

In Tabelle 3.2 ist die Anzahl der Patienten in Verbindung mit ihrer Einordnung gemäß der „Federation International of Gynaecology and Obstetrics“ (FIGO) zusammengefasst.

### 3.5.3 Datenvorverarbeitung

Vor der eigentlichen Datenverarbeitung und Analyse der Microarraydaten muss zunächst eine Normalisierungsprozedur durchgeführt werden, welche den Großteil der technischen Variation entfernt und darüber hinaus den Vergleich der Expressionswerte ermöglicht. Normalisierung ausführende Algorithmen sind weitestgehend im quelloffenen, R-basierten Bioconductorprojekt implementiert. Für die folgenden Analysen wurde R in der Version 2.15 verwendet. Da Microarraydaten aus vier verschiedenen Laboren und sogar von unterschiedlichen Herstellerserien verwendet wurden, oblagen die Abläufe in Vorverarbeitung und anschließender Normalisierung oberster Sorgfalt. Zuerst wurde eine Qualitätskontrolle mit dem *Simpleaffy* Paket [152] durchgeführt. Diese führte dazu, dass Arrays die eine höhere Hintergrundintensität als 75 aufwiesen verworfen wurden. Zudem wurden Arrays von der weiteren Analyse ausgeschlossen, deren Expressionswerte der Proben für die Kontroll-Haushaltsgene den empfohlenen Schwellwert mehr als dreifach überschreiten [152]. Anschließend wurde eine RNA Degradationskontrolle ausgeführt. Die auf diese Weise vorverarbeiteten Daten konnten nun mit der Quantilnormalisierung des Pakets *RMA* [136] weiterverarbeitet werden, ohne dabei einen Hintergrundabgleich durchzuführen wie in [153]. Abschließend werden die bereits normalisierten Datensets mit Hilfe des Pakets *CONOR* [154] zu einer Patientenmatrix vereint.

## 3.6 Triple-negativ Brustkrebs Subtypisierungsstudie

### 3.6.1 Microarraydaten aus öffentlichen Brustkrebs-Studien

Microarraydaten verschiedener Brustkrebs-Studien, die triple-negative Patienten beinhalten, wurden aus den öffentlich verfügbaren Datenbanken [Array-Express](#) und [21] [GEO](#) [22] bezogen. Die folgenden Studien wurden auf qualitativ hochwertige Arrays hin evaluiert.

Studie **A**: Referenznummer GSE19615, Li *et al.*, 115 Microarrays im Format Human Genome U133A Plus 2.0 [155]. Das Ziel dieser Studie war die Identifikation tumorspezifischer Gene, die auf bestimmte Chemotherapeutika ansprechen. Die Autoren finden die Amplifikation von Genen auf Chromosomenarm 8q22, welche mit dem Rezidiv verknüpft sind, obschon eine Chemotherapie eingeleitet wurde.

Studie **B**: Referenznummer GSE20194, Popovici *et al.* und Shi *et al.*, 278 Microarrays im Format Human Genome U133A [33, 84]. Diese Studie ist Teil des MAQC-II Projekts. Die Daten wurden dazu verwendet der Fragestellung nachzugehen, welchen Einfluss die Wahl der Auswahlmethode für die Klassifizierung hat und in welchem Zusammenhang dies zur Performanz des erhaltenen Prediktors steht.

Studie **C**: Referenznummer GSE2603, Minn *et al.*, 121 Microarrays im Format Human Genome U133A [56]. Die Autoren dieser Studie identifizierten brustkrebspezifische Gene, welche Metastasen in den Lungen verursachen und validierten dies in einem Maus-Modell.

Studie **D**: Referenznummer E-TABM-157, Neve *et al.*, 115 Microarrays im Format Human Genome U133A [156]. Das Ziel dieser Studie war die Etablierung eines zellkultur-basierten Modellsystems für Brustkrebs.

Zusätzlich sollen an dieser Stelle bereits Studien aufgelistet werden, die erst im Ergebnisteil Erwähnung finden. Diese sind Studien mit den Referenznummern GSE1456 [62], GSE3493 [157], GSE6532 [73] und GSE7390 [158].

### 3.6.2 Brustkrebspatienten

Retrospektiv wurden die Microarraydaten aus den oben genannten Studien vereint und insgesamt konnten so insgesamt 514 Genexpressionsprofile von Primärtumoren aus Patienten und zusätzlich auch 51 Genexpressionsprofile aus verschiedenen Brustkrebs Zellkulturen gewonnen werden. Die Patienten sind zwischen 27 und 77 Jahre alt

und die Tabelle 4.8 listet die Patienten nach dem jeweiligen Rezeptorstatus auf. Für die Zellkulturen jedoch konnte nur zwischen triple-negativ und Referenz Brustkrebs unterschieden werden, da ein immunohistochemischer Nachweis des Rezeptorstatus nicht verfügbar ist.

### 3.6.3 Daten Vorverarbeitung

Die Analyse der Microarraydaten erfolgt mit Hilfe von Algorithmen, die weitestgehend im quelloffenen, R-basierten Bioconductorprojekt [34] implementiert sind. R in der Version 2.15 und Bioconductor in Version 2.10 wurden für alle statistischen Analysen und die Datenverarbeitung verwendet. Obschon die verwendeten Studien auch normalisierte Microarraydaten bereitstellen, wurde die Analyse jedoch von den Rohdaten ausgehend begonnen, um verschiedene Studien und Plattformen zu vergleichen. Zuerst wurden die Arrays einer Qualitätskontrolle mit dem Paket *simpleaffy* [152] unterzogen, welche zum Ausschluss von Arrays führte, die eine durchschnittliche Hintergrundintensität von mehr als 75 aufweisen und deren Kontroll-Haushaltsgene über dem Schnitt lagen ( $> 3$ -fach induziert). Anschließend wurde ein RNA-Verdautest durchgeführt. Als Nächstes wurden die einzelnen Datensets jeweils mit der Quantilnormalisierung des *RMA* Pakets [136] ohne eine Hintergrundanpassung verarbeitet. Schließlich konnten die aus den Studien erworbenen Daten zu einer einzigen Matrix mit dem CONOR Paket zusammengeführt werden [154].



# 4 Ergebnisse und Diskussion

## 4.1 Softwareentwicklung

### 4.1.1 Das Microarray-Analysewerkzeug maRt

Die Software besteht aus quelloffenen Java und R Bibliotheken, sowie einer Vielzahl an Bioconductorpaketen. Alle statistischen Routinen werden über R abgewickelt, da diese Sprache für statistisches Rechnen entworfen wurde. Ein weiterer Vorteil von R ist der Zugriff auf das R basierte Bioconductorprojekt und die darin enthaltenen Pakete und Annotationsdatenbanken. In der aktuellen Version 0.8.9 ist die ehemals verwendete JRI Bibliothek obsolet geworden. Diese realisierte die Schnittstelle zwischen Java und R. In der aktuellen Version wird diese Aufgabe von der nativen Java Klasse ProcessBuilder übernommen, welche eine transparente Einbindung externer Programme ermöglicht. Die Software wird frei für akademische Zwecke als ausführbare Vollversion für MAC, PC und Unix basierte Betriebssysteme angeboten, in der alle benötigten, externen Bibliotheken enthalten sind. Es ist davon auszugehen, dass Bioinformatiker bereits die notwendigen externen Bibliotheken, sowie ein Java Entwicklungskit (JDK) in Version 1.6 und R in Version 2.15 installiert haben und nur an der maRt Bibliothek interessiert sind. Dieser Leserkreis sei auf den frei verfügbaren Quellcode verwiesen, welcher unter folgender URL [code.google.com/p/martool](http://code.google.com/p/martool) zu finden ist. Die offizielle Dokumentation und verschiedene Tutorien in Bezug auf maRt sind auf den Webseiten der Bioinformatik Abteilung des Instituts für Pharmazeutische Chemie zu finden ([pharma.uni-bonn.de/www/mart](http://pharma.uni-bonn.de/www/mart)).

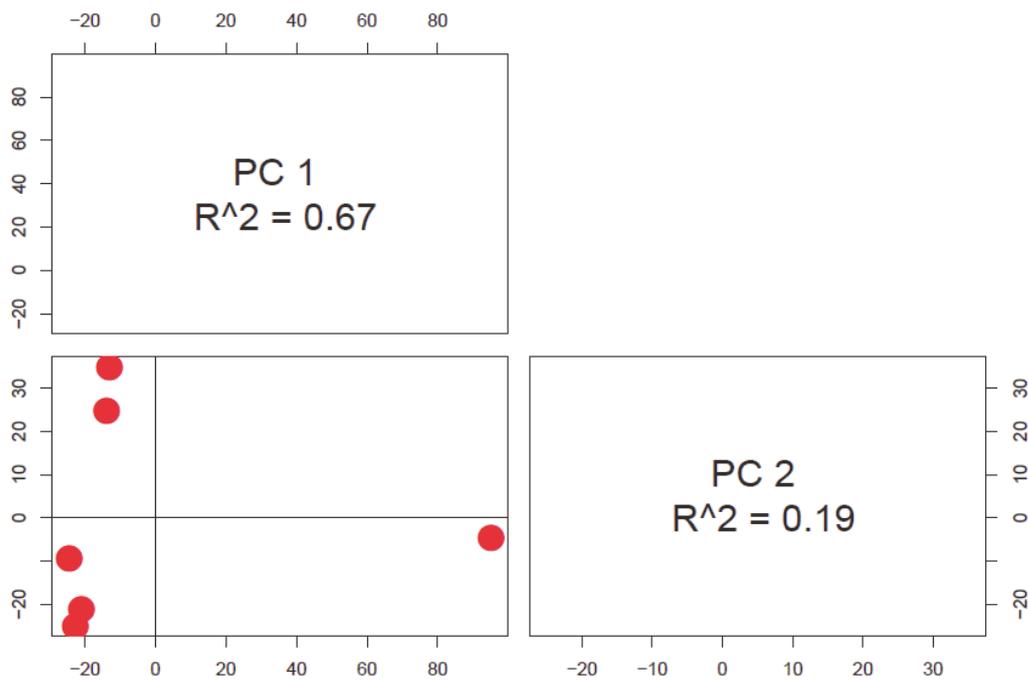
**Fazit** Obwohl bereits eine Vielzahl an Microarray-Analysewerkzeugen existiert, bieten die wenigsten den Komfort einer grafischen Bedienerschnittstelle, eine flexible und modulare sowie individuelle Erweiterbarkeit, eine Unterstützung der aktuellen Multi-Core Prozessoren und eine gänzlich betriebssystemunabhängige Anbindung. In maRt

finden sich dadurch gleich mehrerer Alleinstellungsmerkmale, die die Software zu einer attraktiven Microarray-Analyseplattform machen.

### 4.1.2 Qualitätssicherung von Microarraydaten

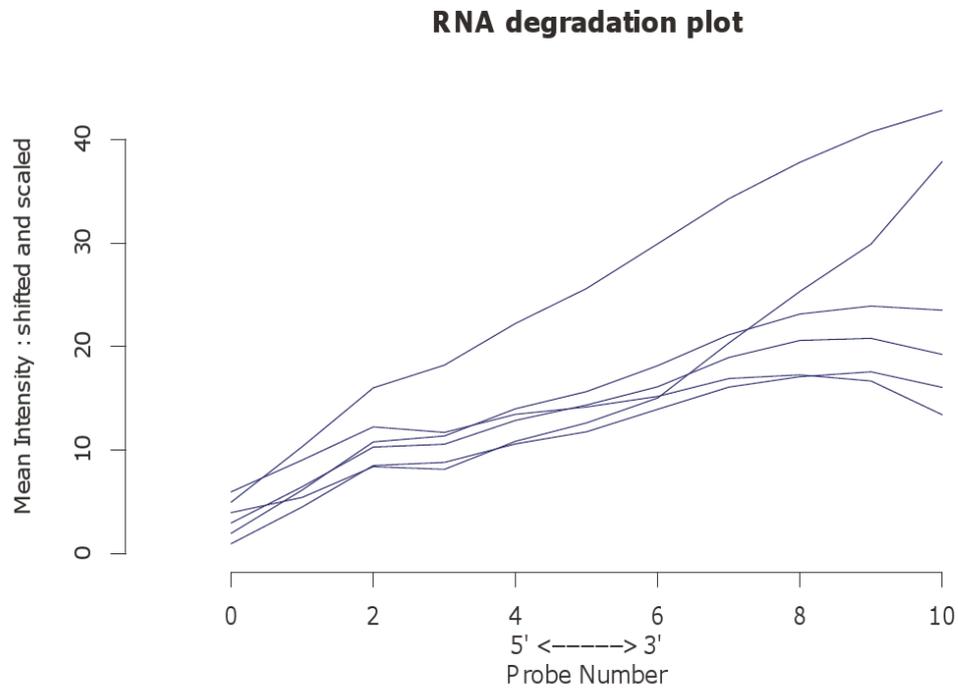
#### 4.1.2.1 Erhebung der Qualitätsparameter diverser in GEO verfügbarer Studien

Gegenwärtig ist die Mehrzahl der in GEO abgelegten Microarraydatensets im Affymetrix eigenen Format. Daher liegt der Fokus der Untersuchungen auf Human Genome U133A Arrays (d.h. laut GEO die Plattform GPL96). Drei repräsentative Beispieldatensets mit unterschiedlicher Anzahl an Arrays wurden selektiert, um die Visualisierungskapazitäten dieses Vorgehens zu demonstrieren. Die prinzipielle Anwendung wird nun anhand sechs GPL96 basierter Arrays der Studie GSE9801 gezeigt.



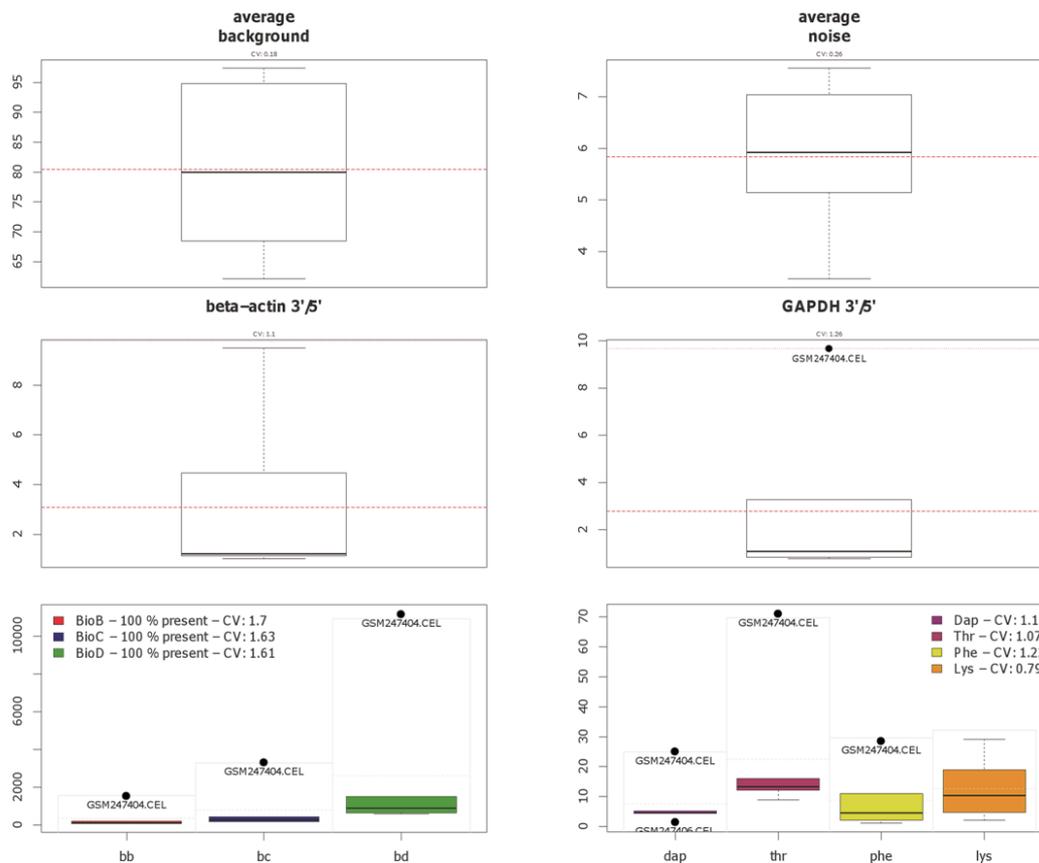
**Abbildung 4.1:** Die Abbildung stellt die ersten beiden Hauptkomponenten der sechs Arrays aus dem Datenset GSE9801 graphisch dar.

Die Ergebnisse der Hauptkomponentenanalyse, welche mit dem Paket *pcaMethods* erstellt wurden, sind in der Abbildung 4.1 dargestellt. Die Abbildung bildet die erste gegen die zweite Hauptkomponente ab. Hier grenzen sich zwei Cluster von einem Array ab, welches als potentieller Ausreißer auffällig wird.



**Abbildung 4.2:** Die Abbildung stellt die Intensitäten von elf Affymetrix spezifischen Kontrollproben dar, die potentielle RNA-Degradation messen.

In Abbildung 4.2 sind Mittelwerte der Intensitäten von elf Kontrollproben dargestellt, die vorangegangenen RNA-Verdau und damit einhergehende potentielle RNA-Degradation detektieren. Das Array mit den höchsten Intensitätswerten stellt einen Ausreißer dar.

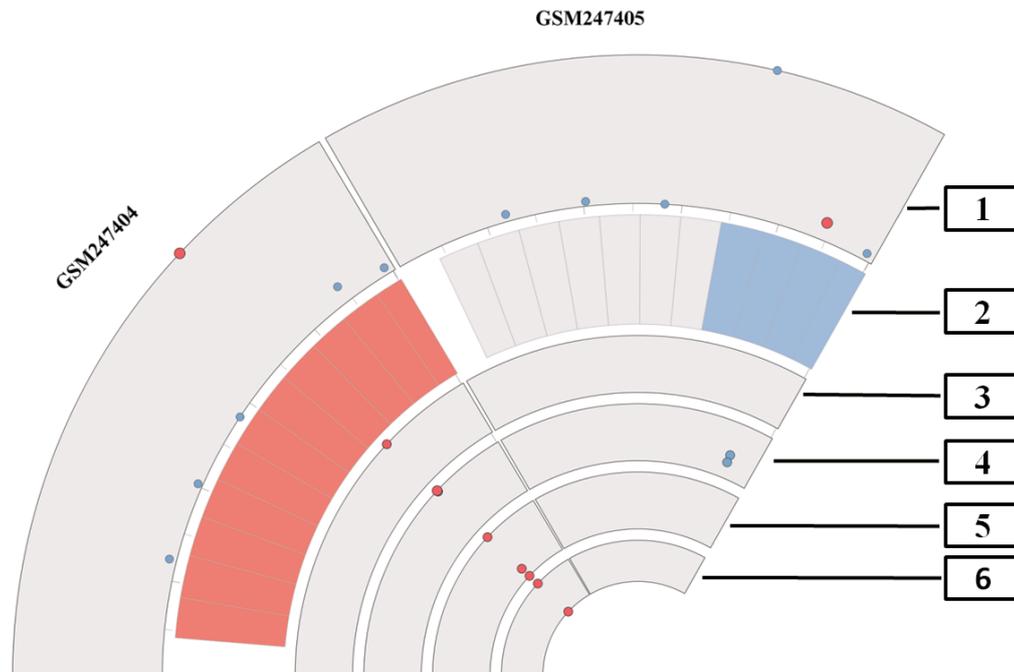


**Abbildung 4.3:** Die Abbildung präsentiert die unterschiedlichen Qualitätsparameter die mit dem *yaqcaffy* Paket abgefragt werden können. Der obere und mittlere Teil der Graphik zeigt Boxplots der beiden Haushaltsgene und der durchschnittlichen Intensität im Bezug auf Hintergrund und Rauschen. Der untere Teil stellt die Resultate der Affymetrix internen Kontrollproben dar, diese sind die spike-in Proben im linken Teil und die Poly-A Farbstoff spezifischen Proben im rechten Teil der Graphik.

Das Bioconductorpaket *yaqcaffy* bietet unterschiedliche Qualitätskontrollen für Affymetrix Arrays an. Die Abbildung 4.3 fasst die Ergebnisse der Qualitätsuntersuchungen mit dem *yaqcaffy* Paket des Datensets GSE9801 graphisch in sogenannten Boxplots zusammen. Die Mehrzahl im Paket angebotener Methoden zur Untersuchung der Qualität werden demonstriert, sowie die vorgegebenen Schwellwerte. Die oberen vier Boxplots zeigen neben der durchschnittlichen Hintergrundintensität und dem durchschnittlichen Rausch-Niveau auch beide Proben, die für die Haushaltsgene  $\beta$ -Actin und GAPDH spezifisch sind. Im Boxplot, der die Intensität des GAPDH Gens über das gesamte Datenset misst, kann das Array GSM247404 als Ausreißer identifiziert werden. Die unteren Boxplots zeigen auf der linken Seite die internen, sogenannten spike-in Proben und auf

der rechten Seite die Poly-A Kontrollen. Auch in diesen, speziell für Affymetrix Arrays typischen Kontrollen, ist Array GSM247404 eindeutig als Ausreißer auszumachen.

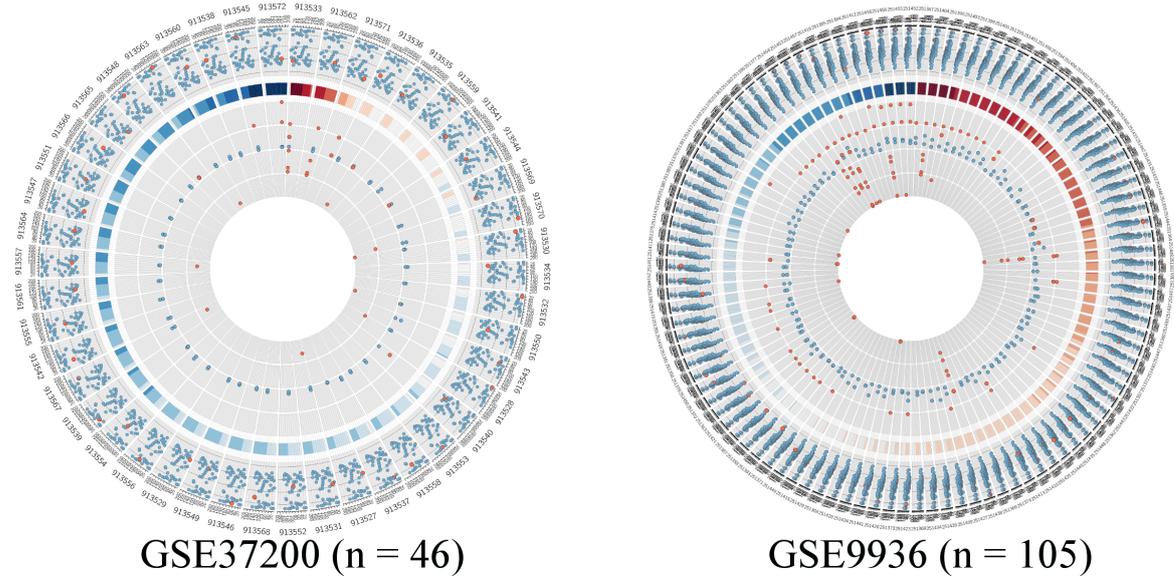
#### 4.1.2.2 Visualisierung der Qualitätsparameter mit Circos



**Abbildung 4.4:** Die Abbildung fasst die verschiedenen Qualitätsmerkmale mit Hilfe von Circos in einer zirkulären Abbildung zusammen. Dargestellt sind zwei Arrays aus dem Datenset GSE9801. Der Index denotiert folgendes: 1 = erste Hauptkomponente, 2 = RNA-Degradation, 3 = Hintergrund und Rauschen, 4 =  $\beta$ -Actin und GADPH, 5 = interne spike-in Proben, 6 = Poly-A Kontrollen.

In Abbildung 4.4 bietet sich eine Übersicht der in dieser Arbeit entwickelten Darstellungsweise für Qualitätsmerkmale in Microarraydatensets. Die Abbildung ist eine Kombination aus den bisher vorgestellten Qualitätsmerkmalen und integriert zusätzlich alle individuellen Arrays eines Datensets in zirkulärer Art und Weise. Die Informationen bezüglich der Qualität wurden fast ausschließlich auf das Vorhandensein von Ausreißern reduziert, somit fällt der Fokus schnell auf fehlerhafte Arrays. Im Folgenden werden die einzelnen Segmente aus Abbildung 4.4 vorgestellt. Das erste Segment bildet die Werte der ersten, über den Mittelwert zentrierten Hauptkomponente ab. Dabei wird das aktuelle Array als roter Punkt und die übrigen als blaue Punkte dargestellt. Die übliche

Vorgehensweise, die erste gegen die zweite Hauptkomponente darzustellen, erwies sich als weniger praktikabel. Zudem erfasst die erste Komponente bereits den Großteil der erklärten Varianz, sodass hier nur die erste Hauptkomponente Berücksichtigung findet. In der Abbildung finden daher Arrays zusammen, die ähnliche Varianz in der Gesamtexpression aufweisen, wohingegen Arrays unterschiedlicher Varianz separat angeordnet werden. Das nächste Segment bildet die Intensitätswerte der elf Degradationskontrollen ab. Hier ist eine geringe Qualität als rotes Rechteck und Fälle, die offensichtlich keine Degradation aufweisen, in blau dargestellt. Mittlerer RNA-Abbau wird in weiß dargestellt, diese Proben haben Intensitätswerte im Bereich der Toleranz. Der besseren Übersicht halber sind die Arrays aufgrund der RNA-Degradationswerte sortiert, Arrays von ähnlicher RNA-Qualität werden so nebeneinander angeordnet. Der dritte Kreis stellt potentielle Ausreißer hinsichtlich der Schwellenwerte für die durchschnittliche Hintergrundintensität und das Rausch-Niveau dar. Diese werden, so vorhanden, als rote Punkte visualisiert. Der vierte Ring bietet eine Übersicht über die Expression der beiden Haushaltsgene ( $\beta$ -Actin und GADPH). Die Expression dieser Proben wird blau dargestellt und sollte im Expressionsniveau nicht weit auseinander liegen. Im Fall, dass die Proben den Grenzwert übersteigen, werden diese als roter Punkt abgebildet. Die anschließenden zwei Kreissegmente sind für die Darstellung der Affymetrix eigenen technischen Kontrollen reserviert. Affymetrix setzt hier die sogenannte spike-in RNA ein. Diese besteht aus einer genormten Menge an RNA, die dem Experiment beim Hybridisieren beigemischt wird und kann später durch die technischen Kontrollen wieder auffindig gemacht werden. Der fünfte Ring stellt eben dieses Vorhandensein fest und nur Proben, die nicht nachweisbar sind, werden als rote Punkte dargestellt. Auf dem innersten Ring sind Ausreißer in den sogenannten Poly-A Kontrollen dargestellt, welche Auskunft über die Güte des Fluoreszenzmarkierungs Schritt geben. Die Abbildung 4.4 stellt zwei Arrays des Datensets GSE9801 dar, hier sind ein Array von geringwertiger Qualität (GSM247404) und ein hochqualitatives Array (GSM247405) dargestellt. Das Array GSM247404 weist Ausreißerwerte in jedem Qualitätsmerkmal auf, dies spiegelt auch die erste Hauptkomponente wider. Im Gegensatz dazu erweist sich das Array GSM247405 unauffällig im Bezug auf Qualitätseinbußen, sogar die beiden Proben der Haushaltsgene zeigen Expression im erwarteten Bereich.



**Abbildung 4.5:** Qualitätsmerkmale zweier in Tabelle 3.1 gelisteter Datensets mit größerer Probenanzahl. Die linke Graphik zeigt ein Microarraydatenset, welches 46 Arrays beinhaltet, die Visualisierung der vorgestellten Qualitätsparameter stellt sich hier unproblematisch dar. Die rechte Graphik fasst die Qualitätsparameter eines 105 Arrays beinhaltenden Datensets zusammen, in der Darstellung wird offenbar die Grenze der Visualisierungskapazität der vorgestellten Methode erreicht.

Im nächsten Schritt werden die Grenzen der Visualisierungskapazität der vorgestellten Methode ausgelotet. Daher wurden zwei der in Tabelle 3.1 genannten Datensets in Abbildung 4.5 dargestellt. Der linke Teil der Abbildung erfasst die Studie GSE37200 mit 46 Arrays, der rechte Teil stellt 105 Arrays aus Datenset GSE9936 dar. Daraus folgt für Datenset GSE37200, dass es drei Arrays von fraglicher Qualität enthält. Zwei dieser Arrays sind demnach tatsächlich qualitativ bedenklich, da RNA-Degradation festgestellt ist und zusätzlich die durchschnittliche Intensität im Hintergrund zu hoch und die Richtwerte für das Hintergrundrauschen überschritten sind. Zudem weist der zweit innerste Ring alle drei spike-in Proben als fehlend auf. Die Mehrheit der Arrays in diesem Set ist jedoch von akzeptabler Qualität, bis auf wenige Ausreißer, die in den Poly-A Kontrollen zu finden sind. Das Datenset GSE9936 weist einige Kandidaten für RNA-Degradation auf, jedoch ist nur einer davon zudem auch Ausreißer in allen anderen Bereichen. Die Mehrzahl der RNA-Abbaukandidaten hat lediglich eine höhere durchschnittliche Hintergrundintensität. Die übrigen Arrays, in denen RNA-Abbau unproblematisch erscheint, weisen jedoch unzulässige Werte für durchschnittliche Hintergrundintensität und ein stärkeres Hintergrundrauschen auf. Zudem haben diese Arrays auch grenzwertige Messwerte für die Expression der Haushaltsgene und

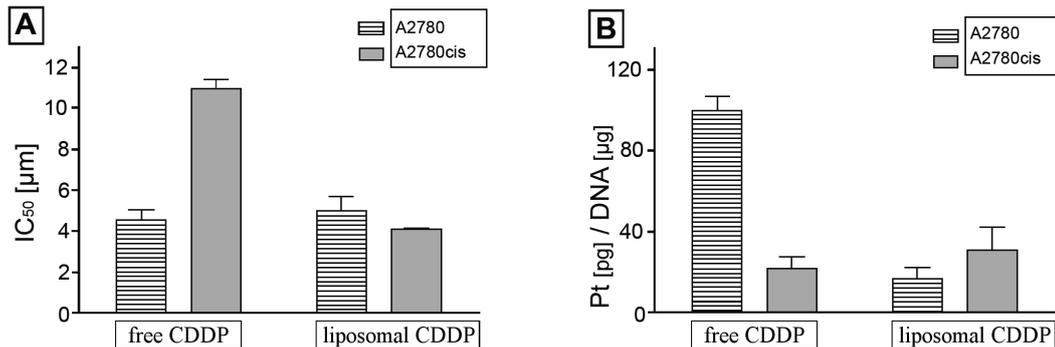
weitere Ausreißer in den technischen spike-in und Poly-A Kontrollen. Insgesamt wären in diesem Datenset sieben Arrays von der weiteren Analyse auszuschließen. Dies unterstreicht die Tatsache, dass Qualitätssicherung durch eine Mehrzahl an Methoden umzusetzen und an verschiedenen Merkmale festzumachen ist, um ein sicheres Ausschlussverfahren zu gewähren. Zusammengefasst bietet die hier vorgestellte Methode eine Qualitätssicherung für den Großteil an Datensets in GEO, die auf dem GPL96 und dem dazu verwandten GPL570 und GPL571 Format basieren. Jedoch sollte die Probenanzahl nicht zu weit über 100 Arrays liegen, da Datensets ab einer Größe von 130 Microarrays nicht genügend detailliert dargestellt werden können. Dennoch findet die hier vorgestellte Methode eine breite Anwendung, da die meisten Datensets eben in dieser Größenordnung liegen. Auch sei darauf hingewiesen, dass die resultierenden Circos Abbildungen als skalierbare Vektorgrafiken vorliegen und daher für die ressourcenschonende Verwendung im Internet geeignet sind.

**Fazit** Qualitätskontrolle ist einer der wichtigsten vorbereitenden Schritte für die Analyse von Microarraydaten. Jedoch sind bislang keine generellen Methoden verfügbar, um die Ergebnisse einer qualitativen Evaluierung zu visualisieren, wobei auch einzelne Arrays in großen Datensets berücksichtigt werden. Daher wurde in dieser Arbeit mit Hilfe von Circos eine bequeme Methode für die Darstellung der Ergebnisse der Qualitätskontrolle von Microarraydaten entwickelt.

## 4.2 Microarray Analysen

### 4.2.1 Liposomales Cisplatin Wirkstoffstudie

#### 4.2.1.1 DNA Platinierungseffizienz von freiem und liposomalem CDDP in Ovarialkarzinom Zellen



**Abbildung 4.6:** Die Abbildung 4.2.1.3 stellt die Effizienz der Zytotoxizität und DNA Platinierung des freien und liposomalen CDDP nach 72 Stunden in A2780 und A2780cis Ovarialkarzinomzellen anhand zweier Graphen dar. Die hier gezeigten Graphen repräsentieren den Mittelwert  $\pm$  SEM. Die zugrunde liegenden Daten fassen die Ergebnisse aus drei bis sechs unabhängigen Experimenten zusammen, die jeweils in Triplicaten durchgeführt wurden. Panel A stellt die über MTT Experimente gemessene Viabilität der Zellen (jeweils  $2e^4$  [Zellen pro Well]) nach 72 Stunden dar. Panel B zeigt die über Massenspektrometrie mit induktiv gekoppeltem Plasma (ICP-MS) gemessene DNA Platinierung nach Behandlung der A2780 und A2780cis Zellen mit freiem bzw. liposomalem CDDP.

CDDP induziert durch seine DNA interkalierende Eigenschaften Platin-DNA Addukte, die Antitumorwirkung haben. Kürzlich konnte gezeigt werden, dass mit liposomalem CDDP die Cisplatin Resistenz von Ovarialkarzinomzellen umgangen werden kann [45]. In dieser Arbeit sollen die DNA Platinierungseffizienz von liposomalem und freiem CDDP in A2780 und A2780cis Zellen untersucht werden. Die experimentellen Daten, die in dieser Arbeit analysiert werden, wurden von Dr. Michaela Krieger und Dr. Alex Hilger zur Verfügung gestellt. Die Versuchsdurchführung wurde von Dr. Michaela Krieger geleitet, dazu wurden A2780 und A2780cis Zellen mit identischen Konzentrationen CDDP über 24 Stunden behandelt und über Massenspektrometrie mit induktiv gekoppeltem Plasma (ICP-MS) analysiert. Dies ermöglicht eine Quantifizierung der DNA

Platinierung in Picogramm Platinmoleküle pro  $\mu\text{g}$  DNA wie in Abbildung 4.2.1.3 in Teil B gezeigt. Die Behandlung der A270 Zellen mit freiem Wirkstoff erreicht hohe Platinierungsraten, mit 100 Picogram pro  $\mu\text{g}$  DNA (gestreifte Spalte). Im Gegensatz dazu werden mit freiem Wirkstoff in A2780cis Zellen nur niedrige Platinniveaus erzielt, wie in der grauen Spalte gezeigt. Es ist erwiesen, dass die CDDP Resistenz wenigstens zum Teil durch intrazelluläre Mechanismen erklärt werden kann, die eine Platinierung der DNA verhindern. Das Ausmaß der CDDP Resistenz von A2780cis Zellen wurde durch MTT Zytotoxizitäts-Tests festgestellt, wie in Abbildung 4.2.1.3 in Teil A gezeigt. Demnach erreicht freies CDDP  $IC_{50}$  Werte von  $11\mu\text{M}$  in A2780cis Zellen und  $4\mu\text{M}$  in A2780 Zellen. Nun soll die DNA Platinierungseffizienz in A2780 und A2780cis Zellen mit identischen Konzentrationen an liposomalem CDDP verglichen werden. Die Ergebnisse zeigen, dass liposomales CDDP nicht signifikante DNA Platinierung in A2780 oder A2780cis Zellen induziert, wie in Abbildung 4.2.1.3 in den rechten Spalten von Teil B gezeigt. Daraus folgt, dass der freie und liposomale Wirkstoff einen wesentlichen Unterschied in der Fähigkeit aufweist, DNA Addukte zu induzieren. Dennoch kann liposomales CDDP die Cisplatinresistenz der A2780cis Zellen überwinden, wie in Abbildung 4.2.1.3 in Teil A in der rechten Spalte gezeigt. Also verursacht liposomales Cisplatin zytotoxische Effekte in A2780cis Zellen, sogar in Abwesenheit detektierbarer DNA Platinierung.

#### 4.2.1.2 Die Aktivität des freien und liposomalem CDDP unter Einfluss des Histon-Deacetylase Inhibitors TSA

**Tabelle 4.1:** Die Tabelle listet die  $IC_{50}[\mu\text{m}]$  der untersuchten Ovarialkarzinomzellen auf, die über die MTT Experimente erhalten wurden. Die Daten repräsentieren die Mittelwerte aus drei bis sechs Experimenten mit jeweils  $2e^4$  Zellen für 72 Stunden.

Behandlung	A2780	A2780cis
Freies CDDP	4,53	10,95
Freies CDDP + TSA	2,07	8,05
Liposomales CDDP	4,97	4,09
Liposomales CDDP + TSA	3,92	4,24

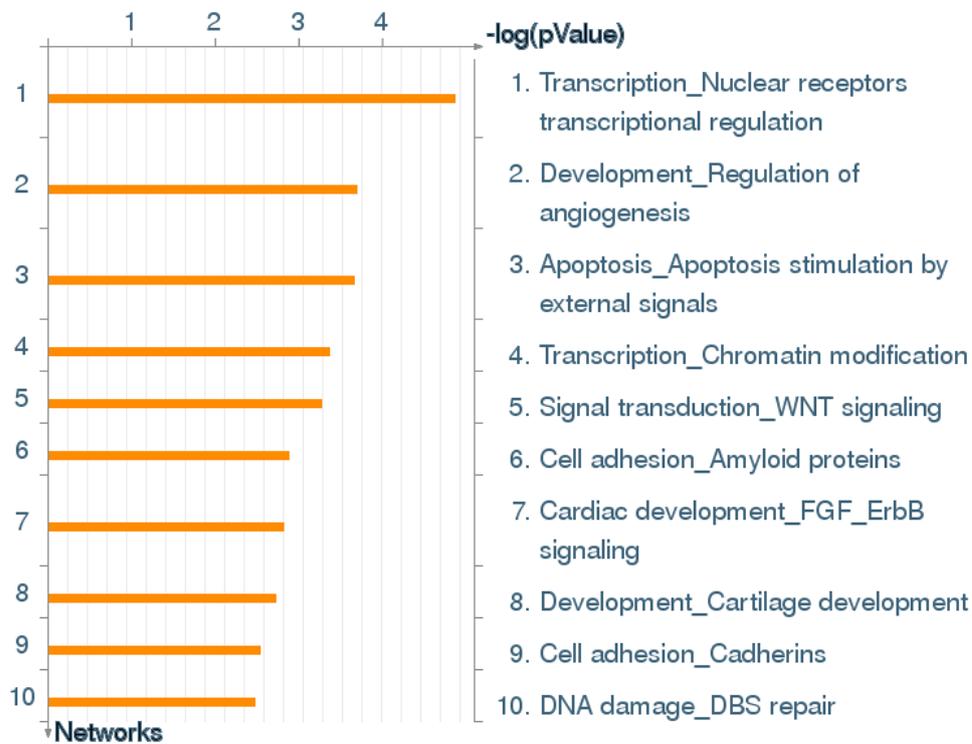
In der Literatur wurde beschrieben, dass Trichostatin A (TSA) zumindest teilweise den CDDP resistenten Status von A2780cis Zellen überwinden kann, indem der intrinsische Apoptose-Pathway eingeschaltet wird [147]. Trichostatin A ist ein Histon-Deacetylase (HDAC) Inhibitor und spezifisch für HDACs der Klassen I, IIa, IIb und IV [159]. Aufgrund dieses Befundes wurde zunächst das Apoptose auslösende Potential von TSA

auf A2780 und A2780cis Zellen untersucht, in Kombination mit einer Behandlung mit freiem oder liposomalem Wirkstoff, siehe Tabelle 4.1. Die Versuchsdurchführung wurde von Dr. Michaela Krieger realisiert. Einerseits konnte die Aktivität des freien CDDP an A2780 und A2780cis Zellen unter Zugabe von TSA verstärkt werden, andererseits wird die Aktivität des liposomalen Wirkstoffs in A2780 und A2780cis Zellen durch TSA nicht beeinflusst. Diese Resultate weisen darauf hin, dass es einen wesentlichen Unterschied zwischen dem freien und dem liposomalen Wirkstoff gibt, inwiefern diese intrazellulär wahrgenommen und verstoffwechselt werden. Es bleibt zunächst jedoch unbeantwortet, ob dieser Befund an die zelluläre Aufnahme des liposomalen CDDP im Vergleich zum freien Wirkstoff gekoppelt ist, oder ob die Art und Weise der intrazellulären Verarbeitung entscheidend ist. Da offenbar ein komplexes Wechselspiel diverser zellulärer Netzwerke und Pathways involviert ist, muss eine systematische Evaluierung erfolgen, um ein Verstehen für die verschiedenen Aktivitäten zu entwickeln. Die Analyse der Genexpressionsprofile von A2780 und A2780cis Zellen, nach Behandlung mit freiem oder liposomalem CDDP könnte Aufschluss über die auf genetischer Ebene ablaufenden Prozesse geben, um aufzuklären wie liposomales CDDP den resistenten Status der A2780cis Zellen überwinden kann.

### **4.2.1.3 Analyse der Microarraydaten von A2780cis Zellen erlaubt Einblick in die Resistenz Mechanismen**

Eine Analyse der mit Cisplatinresistenz assoziierten Gene innerhalb des Genexpressionsprofils von A2780cis Zellen, könnte aufschlussreich für die Identifizierung des Mechanismus sein, der liposomales CDDP befähigt, eben diese Resistenz zu umgehen. Genexpressionssignaturen der resistenzassoziierten Gene wurden von Norbert Brenner mit Zweikanal Agilent Microarrays aufgezeichnet, wie im Materialteil beschrieben. Nach Hybridisierung wurde die Agilent Feature Extraktion Software benutzt, um die Genexpressionsniveaus in A2780 und A2780cis Zellen zu erhalten und die Veränderungsrate (*i.e.* fold-change) in den resistenten A2780cis Zellen zu bestimmen. Die resultierenden Gensignaturen wurden im Detail mit MetaCore<sup>TM</sup> analysiert, einem Pathway Analyse und Data-Mining Werkzeug für die Charakterisierung und Analyse von Genexpressionswerten (GeneGo, Carlsbad, CA, USA). Diese Analyse wurde in Kooperation mit Dr. Hans-Dieter Royer des Instituts für Humangenetik Düsseldorf durchgeführt. MetaCore<sup>TM</sup> kann die relevanten Pathways, Netzwerke und zellulären Prozesse identifizieren, welche den CDDP resistenten Status der A2780cis Zellen manifestieren. Eine Charakterisierung der Transkriptom-assoziierten CDDP Resistenz erfolgte über den

Enrichment Analyse Modus der Software und war auf die Anreicherung der GeneGo Prozessnetzwerke ausgerichtet, wie in Abbildung 4.7 gezeigt.



**Abbildung 4.7:** Abbildung 4.7 vermittelt einen Eindruck über die Prozessnetzwerke, denen Bedeutung bei der Untersuchung der Chemoresistenz zukommt. Die Prozessnetzwerke werden von Hand kuriert und entstammen der Gen-Ontologie sowie den GeneGo Pathway Karten.

Die erhaltenen GeneGo Prozesse beschreiben Netzwerkmodelle der signifikant induzierten zellulären Prozesse. GeneGo Prozessnetzwerke werden von Hand kuriert und beruhen auf der Basis von GO Termini und GeneGo Pathway Karten. Die am höchsten angereicherten Netzwerke beschreiben DNA-Schaden und DNA-Doppelstrangreparatur. Dies weist darauf hin, dass in A2780cis Zellen auf transkriptioneller Ebene ein DNA-Schadenreparaturprogramm initiiert wird. In A2780cis Zellen ist dies offenbar ein wichtiger CDDP Resistenzmechanismus. Zusätzlich findet sich eine Anreicherung von Netzwerken, die den Zellzyklus, die Transkription, die Zelladhäsion und zelluläre Signalisierung, sowie Prozesse der Entwicklungsbiologie auf der Ebene der Transkription regulieren. Diese Resultate demonstrieren die transkriptionelle Aktivierung diverser zellulärer Netzwerke, die während der Entwicklung der CDDP Resistenz in Ovarialkar-

zinomzellen in Erscheinung treten. Angesichts dessen ist es besonders interessant, das Netzwerk „Transkription und Chromatin Modifikation“ auf Rang 4 in Abbildung 4.7 zu finden. Die Chromatinmodifikation und -Neugestaltung sind bedeutende Prozesse in der Regulation der Transkription. Chromatinmodifizierung ist ebenso verknüpft mit den Experimenten zur Inhibition von HDAC, wie oben gezeigt. Diese Regulierung ist ein wesentlicher Unterschied zwischen der Wirkweise von freiem und liposomalem CDDP, und offensichtlich auch ein Schlüsselhinweis für die Begründung, auf welche Weise liposomales CDDP die Resistenz überwindet, daher wird im Folgenden dieser Prozess im Detail analysiert.

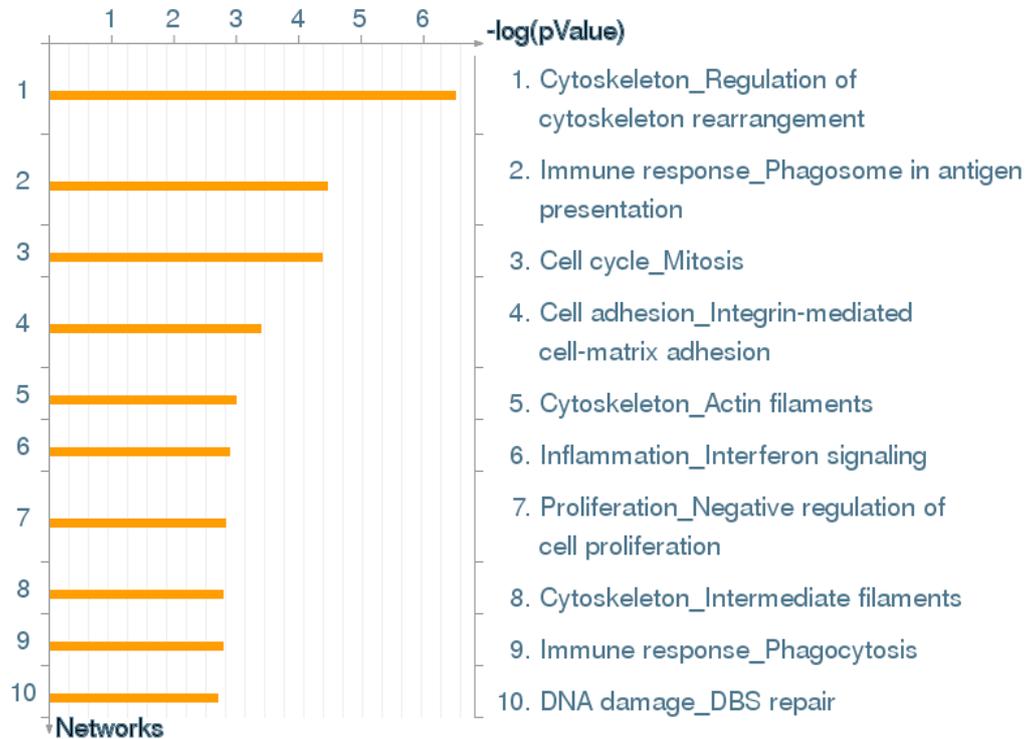
Die Tabelle 7.4 fasst die Gene zusammen, die in A2780cis Zellen im GeneGo Prozessnetzwerk „Transkription und Chromatin Modifikation“ induziert sind. Diese Gene beinhalten Histon-modifizierende Gene, Histone selbst und Sirtuine (SIRT1, SIRT3 und SIRT7), diese gelten als für die Stress-Resistenz verantwortlich. Aus der Literatur geht hervor, dass Chromatinproteine eine große Rolle in der Wahrnehmung und Reparatur von DNA-Schäden ausführen. Diese Resultate erklären zum Teil, wie die Akquisition eines neuen transkriptionellen Programms in CDDP resistenten Zellen etabliert werden kann.

**Tabelle 4.2:** Die Tabelle listet die Gene auf, die im GeneGo Prozessnetzwerk „Transkription und Chromatin Modifikation“ induziert sind.

Gen Symbol	RefSeq ID	Gen Beschreibung
SMYD3	<a href="#">NM_022743</a>	SET and MYND domain containing 3
BAF180	<a href="#">NM_018165</a>	polybromo 1
HIST3H3	<a href="#">NM_003493</a>	histone cluster 3, H3
SET8	<a href="#">NM_020382</a>	SET domain containing 8
EPC1	<a href="#">NM_025209</a>	enhancer of polycomb homolog 1 (Drosophila)
SATB1	<a href="#">NM_001131010</a>	SATB homeobox 1
SIRT1	<a href="#">NM_012238</a>	sirtuin 1
TRRAP	<a href="#">NM_003496</a>	transformation/transcription associated protein
SIRT3	<a href="#">NM_012239</a>	sirtuin 3
BAZ1A	<a href="#">NM_013448</a>	bromodomain adjacent to zinc finger domain, 1A
H3F3A	<a href="#">NM_002107</a>	H3 histone, family 3A
MSL3L1	<a href="#">NM_006800</a>	male-specific lethal 3 homolog (Drosophila)
DOT1	<a href="#">NM_032482</a>	DOT1-like, histone H3 methyltransferase
EMSY	<a href="#">NM_020193</a>	chromosome 11 open reading frame 30
RERE	<a href="#">NM_001042681</a>	arginine-glutamic acid dipeptide repeats
PCAF	<a href="#">NM_003884</a>	"p300/CBP-associated factor", PCAF
MTA2	<a href="#">NM_004739</a>	metastasis associated 1 family, member 2
SIRT7	<a href="#">NM_016538</a>	sirtuin 7
BAF60c	<a href="#">NM_001003801</a>	actin dependent regulator of chromatin
HIST2H2AC	<a href="#">NM_003517</a>	histone cluster 2, H2ac
Histone H2A	<a href="#">NM_003516</a>	histone cluster 2, H2aa3
BCoR	<a href="#">NM_017745</a>	BCL6 corepressor
G9A	<a href="#">NM_006709</a>	euchromatic histone-lysine N-methyltransferase 2
HIST1H4B	<a href="#">NM_003544</a>	histone cluster 1, H4b

**Genexpressionsanalyse der chemoresistenten A2780cis Zellen in Respons auf freies CDDP** Als Nächstes wird das Transkriptionsprofil von A2780cis Zellen auf den freien Wirkstoff untersucht. Die Analyse der A2780cis Zellen, wie oben beschrieben, ergründet zunächst den initialen Status der CDDP Resistenz. Diese Analyse liefert wichtige Informationen für den Vergleich der Effekte von liposomalem und freiem CDDP in resistenten A2780cis Zellen. Für diesen Vergleich wurden A2780cis Zellen mit einer  $IC_{50}$  Dosis CDDP von ( $20\mu Mol$ ) für 48 Stunden behandelt. Für die Zweikanal-Microarrayanalyse wurden daher gleiche Mengen Cy3<sup>®</sup> und Cy5<sup>®</sup> markierter cRNA aus A2780cis Zellen und mit CDDP behandelte A2780cis Zellen auf sogenannte „whole genome“ Microarrays hybridisiert, wie im Material- und Methodenteil beschrieben. Die resultierende CDDP spezifische (akute Respons) Gensignatur wurde anschließend mit der MetaCore<sup>™</sup> Software nach Anreicherung auf GeneGo Prozessnetzwerke hin

untersucht. Die am höchsten angereichert gefundenen Prozessnetzwerke sind in der Abbildung 4.8 dargestellt.



**Abbildung 4.8:** In Abbildung 4.8 sind die zehn am meisten angereicherten Prozesse aufgezeigt, die in A2780cis Zellen nach Behandlung mit freiem CDDP induziert sind.

Diese beinhalten das Zytoskelett, die Zelladhäsion, die Immunantwort und die Mitose. Das Netzwerk „Zytoskelett und Regulation“ der Zytoskelett Neugestaltung beinhaltet eine große Anzahl Gene, die in A2780cis Zellen durch freies CDDP induziert werden. Diese Gene gehören zur Familie der Tubuline, Actine, Gelsoline, Paxilline, Vinculine und Zyxine, um einige zu nennen. Eine Literatursuche in der PubMed Datenbank brachte jedoch nicht besonders signifikante Assoziationen mit Cisplatin Resistenz zu Tage, mit Ausnahme diverser Tublin Gene. Es übersteigt jedoch bei Weitem den Rahmen dieser Arbeit, eine detaillierte Analyse aller über GeneGo bezogenen Prozess Netzwerke durchzuführen, die in resistenten Zellen mittels freiem CDDP induziert werden.

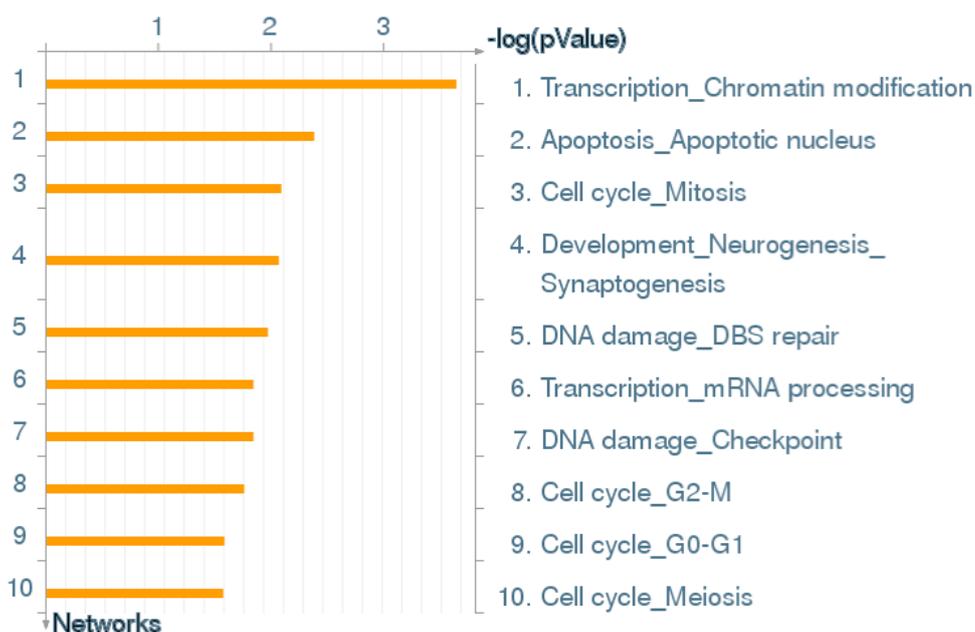
Daher wird der Fokus auf das GeneGo Prozessnetzwerk „Transkription und Chromatin Modifikation“ beschränkt. Wie bereits oben beschrieben, sind viele Gene in diesem

Netzwerk in chemoresistenten A2780cis Zellen induziert. Vor allem drei verschiedene Sirtuine werden durch freies Cisplatin induziert, aber auch weitere Gene, deren Proteine Histone modifizieren können. Alle hier induzierten Gene finden sich in der Tabelle 4.3.

**Tabelle 4.3:** Die Tabelle listet die Histon modifizierenden Gene auf, die in A2780cis Zellen nach Gabe von freiem CDDP induziert sind.

Gen Symbol	RefSeq ID	Gen Beschreibung
CHAF1B	<a href="#">NM_005441</a>	chromatin assembly factor 1, subunit B (p60)
H3F3A	<a href="#">NM_002107</a>	H3 histone, family 3A
HDAC10	<a href="#">NM_032019</a>	histone deacetylase 10
HDAC1	<a href="#">NM_004964</a>	histone deacetylase 1
HIST1H1C	<a href="#">NM_005319</a>	histone cluster 1, H1c
HIST1H2BC	<a href="#">NM_003526</a>	histone cluster 1, H2bc
HIST1H2BE	<a href="#">NM_003523</a>	histone cluster 1, H2be
HIST1H2BF	<a href="#">NM_003522</a>	histone cluster 1, H2bf
HIST1H2BH	<a href="#">NM_003524</a>	histone cluster 1, H2bh
HIST1H2BL	<a href="#">NM_003519</a>	histone cluster 1, H2bl
HIST3H3	<a href="#">NM_003493</a>	histone cluster 3, H3
PML	<a href="#">NM_033247</a>	promyelocytic leukemia

Die Expression der Gene HDAC1 und HDAC10 erscheint an dieser Stelle besonders erwähnenswert. Das HDAC1 Gen ist im Hinblick auf die Literatur [147] bemerkenswert, da bereits ein Zusammenhang zwischen CDDP Resistenz und HDAC1 Expressionsstatus besteht. Jedoch ist der GeneGo Prozess „Transkription und Chromatin Modifikation“ nicht unter den ersten zehn angereicherten Gen Signaturen, die durch freies CDDP in A2780cis Zellen induziert sind, wie in Abbildung 4.8 gezeigt. Gleichwohl erscheinen Gene dieses Netzwerkes aber wichtig, um die Unterschiede bezüglich der Aktivität zwischen liposomalem und freiem CDDP zu erklären. Daher extrahieren wir die Gene aus dem GeneGo Prozess „Transkription und Chromatin Modifikation“, welche in A2780cis Zellen induziert sind, nach Behandlung mit  $20\mu M$  freiem Cisplatin (siehe Tabelle 4.3). Es ist evident, dass der freie Wirkstoff die Expression der Histon Deacetylasen HDAC1 und HDAC10 induziert und ebenso den Histon- und Chromatin-Aufbau Faktor CHAF1B.



**Abbildung 4.9:** In Abbildung 4.9 sind die zehn am meisten angereicherten Prozesse aufgezeigt, die in A2780cis Zellen nach Behandlung mit liposomalem CDDP induziert sind.

**Expressionsanalyse chemoresistenter A2780cis Zellen nach Behandlung mit liposomalem CDDP** Die  $IC_{50}$  Werte in Abbildung Teil A bestätigen, dass liposomales CDDP in der Lage ist, die Chemoresistenz der A2780cis Zellen zu überwinden [45]. In dieser Arbeit wurde gezeigt, dass die intrazelluläre Aufnahme des liposomalen CDDP in A2780cis Zellen, im Vergleich zu freiem Cisplatin einen anderen Effekt hinsichtlich der Formation von Platinaddukten aufweist, wie in Abbildung 4.2.1.3 in der rechten Spalten von Teil B gezeigt. Es ist interessant zu ergründen, auf welche Weise liposomales CDDP die Resistenz überwinden kann, obwohl insgesamt nur wenig DNA-Platinaddukte detektiert werden. Daher wurden Genexpressionsprofile mit Hilfe der Microarraytechnologie erstellt, um diesen Sachverhalt auf transkriptioneller Ebene zu beleuchten. A2780cis Zellen wurden mit einer  $IC_{50}$  Dosis liposomalem CDDP für 48 Stunden behandelt. Als Kontrolle dienten A2780cis Zellen, die zwar über den gleichen Zeitraum, jedoch nur mit leeren Liposomen behandelt wurden. Gleiche Mengen RNA wurde aus diesen Zellen extrahiert, in cDNA umgeschrieben, mit äquivalenten Mengen an Cy3<sup>®</sup> und Cy5<sup>®</sup> Farbstoff markiert und anschließend auf einen Microarray hybridisiert, wie im Material- und Methodenteil beschrieben. Die resultierende liposomale CDDP spezifische Gensignatur (353 Gene) wurde mit der MetaCore<sup>™</sup> Software

nach Anreicherung in GeneGo spezifischen Prozessnetzwerken untersucht (GO Lokalisation, GO Molekulare Funktion und GO Prozesse). Der signifikanteste Eintrag in der Kategorie GO Lokalisation war das Nukleosom ( $p = 8.203e^{-17}$ ). Weitere signifikante GO Lokalisationen sind durch intrazelluläres Organell ( $p = 9.4858e^{-10}$ ), Chromatin ( $p = 4.262e^{-9}$ ), Chromosom ( $p = 1.257e^{-6}$ ), und weitere gegeben (siehe Appendix, Abbildung 7.1). Die Analyse der Kategorie GO Molekularen Funktion identifizierte die Terme Bindung, Protein Bindung und DNA Bindung als signifikant (siehe Appendix, Abbildung 7.2). Die Analyse der Kategorie GO Prozess liefert Nukleosom Aufbau ( $p = 3.175e^{-19}$ ), Protein DNA Komplex Aufbau und Chromatin Aufbau als die signifikantesten Terme (siehe Appendix, Abbildung 7.3).

**Tabelle 4.4:** Die Tabelle fasst die koexprimierten Gene aus dem Prozess „Transkription und Chromatin Modifikation“ zusammen, die nach der Behandlung mit liposomalem CDDP induziert sind.

Gen Symbol	RefSeq ID	Gen Beschreibung
SIRT7	<a href="#">NM_016538</a>	sirtuin 7
HIST2H2BF	<a href="#">NM_001024599.3</a>	Histone 2, H2BF
HIST1H2AA	<a href="#">NM_170745</a>	Histone cluster 1, H2AA
HDAC2	<a href="#">NM_001527</a>	Histone deacetylase 2
HIST2H2BE	<a href="#">NM_003528</a>	H2B, histone family, member Q
PML	<a href="#">NM_033247</a>	promyelocytic leukemia
HIST1H1B	<a href="#">NM_005322</a>	H1, histone family, member 5
HIST2H3C	<a href="#">NM_021059</a>	Histone cluster 2, H3c
H3F3A	<a href="#">NM_002107</a>	H3, histone family 3A

Des Weiteren identifizierte die MetaCore<sup>TM</sup> Software Anreicherung in verschiedenen GeneGo Prozessnetzwerken wie in Abbildung 4.9 gezeigt. Die dabei erhaltenen und am höchsten angereicherten Prozesse sind die Netzwerke „Transkription und Chromatin Modifikation“ (Sirtuin7, Histone H2, Histone H2A, Histone deacetylase class II, Histone H2B, PML, Histone H3, Histone H1, Histone H3.3) und ein Apoptose-spezifischer Prozess „Apototic Nucleus“ (p21, Histone H2B, GADD45 alpha, Caspase-9, Histone H3, Histone H1, Lamin A/C) sind die signifikantesten angereicherten Prozesse. Die Tabelle 4.4 fasst die koexprimierten Gene aus dem Prozess „Transkription und Chromatin Modifikation“ zusammen. Sehr interessant sind die signifikant angereicherten GeneGo Prozesse „DNA Damage DBS Repair“ (Sirtuin7, PP2A catalytic, RAD51C, Sirtuin, PP2A regulatory, Histone H3), und „DNA Damage Checkpoint“ (p21, 14-3-3 gamma, Heme oxygenase 1, GADD45 alpha, BTG2, 14-3-3). Diese Ergebnisse bieten signifikante Beweise dafür, dass die  $IC_{50}$  Dosis liposomales Cisplatin in resistenten A2780cis Zellen, assoziiert mit der Expression spezifischer Gene, die an der Nukleosom Formierung

partizipieren, sowie an der Reparatur und Überprüfung der DNA. In diesem Kontext ist es wichtig zu erwähnen, dass der Nukleosomkern der aus den Histonproteinen H2A, H2B, H3 und H4 besteht, ebenfalls ein molekulares Ziel von Cisplatin ist.

**Analyse der p53 induzierten Gene in A2780cis Zellen durch freies und liposomales Cisplatin** Es ist bekannt, dass ein intakter p53 Tumorsuppressor für eine Cisplatin induzierte Apoptose in ovariellen Krebszellen notwendig ist [160]. A2780 und auch A2780cis Zellen sind hinsichtlich des p53 Status wildtyp [161]. Der Tumorsuppressor p53 induziert nach Behandlung mit Cisplatin reaktive Sauerstoffspezies (ROS) und aktiviert somit den p38 MAPK Pathway [162]. Um die Auswirkung von p53 in A2780cis Zellen die mit freiem oder liposomalem Cisplatin behandelt wurden zu analysieren, wurden Untersuchungen des Transkriptom mit einem speziellen Suchmodus der MetaCore<sup>TM</sup> Software vorgenommen, der auf die Analyse von Transkriptionsfaktoren und Zielgenen spezialisiert ist. Diese Analyse weist die Expression von 89 Genen nach, die mit p53 verbunden sind und durch einen  $IC_{50}$  an freiem Cisplatin induziert werden, wie in Abbildung 4.10 dargestellt. Hinsichtlich der Behandlung mit einem  $IC_{50}$  an liposomalem Wirkstoff konnten bei dieser Analyse 60 induzierte Gene nachgewiesen werden, dies ist in Abbildung 4.11 gezeigt. Beide Abbildungen zeigen ein Netzwerk aus unterschiedlichen Genen, in dem p53 im Zentrum abgebildet ist. Eine Beschreibung der im Netzwerk verwendeten Symbole befindet sich im Appendix in Abbildung 7.4. Die MetaCore<sup>TM</sup> Datenbank, die im Hintergrund einer solchen Analyse abgefragt wird beinhaltet umfassende Datensets aus der Literatur und auch Datenbestände aus einer proprietären Datenbank. Positive Effekte werden als grüne Verknüpfungen angezeigt, wohingegen negative Effekte durch rote Verknüpfungen untermalt sind. Die Position der Pfeilspitze deutet daraufhin, ob die unterschiedlichen Netzwerkobjekte in Abbildung 4.10 und 4.11 von p53 reguliert oder ob diese selbst p53 regulieren.



**Abbildung 4.10:** Netzwerk der Gene, die mit p53 verbunden sind und in Response auf  $IC_{50}$  freies Cisplatin induziert sind. Details zu dieser Abbildung finden sich im Text. Eine Beschreibung der im Netzwerk verwendeten Symbole befindet sich im Appendix in Abbildung 7.4.

Anschließend werden die p53 Netzwerke gemäß der statistischen Parametern der Netzwerke analysiert. Dies erlaubt eine Auswahl der signifikantesten GO Prozesse innerhalb der Netzwerke zu treffen. Bei dieser Analyse geht hervor, dass die Behandlung von A2780cis Zellen mit einem  $IC_{50}$  an freiem Cisplatin ( $11 \mu M$ ) hoch signifikant Gene induziert ( $p = 3.953e^{-20}$ ), die dem GO Prozess „Regulation of Apoptosis“ zu zuordnen sind (IER3, BCAR1, FGFR3, BLOC1S2, JAG2, LGALS1, BAX, BID, DUSP2, LYN, CASP, TRAF4, CARD10). Zusätzlich sind Gene aus dem GO Prozess „Response to Stress“ signifikant über-exprimiert ( $p = 4.786e^{-18}$ ) in Cisplatin behandelten A2780cis Zellen (NMU, HSP90AA1, ZFP36, HSPA8, MAPK11, HSPB1, HSP90AB1).

Weitere Untersuchungen durch textmining ergaben einen interessanten Zusammenhang einer signifikant hohen Anzahl an über-exprimierten Genen und dem p38 MAPK Pathway (BAX, BID, CASP9, EGR1, HSP27, JAK1, JUNB, JUND, LYN, MEK3, PARP1, PPM1D, STAT6). Der p38 MAPK Pathway agiert als ein Sensor für reaktive Sauerstoffspezies (ROS) [163], zudem ist bekannt, dass freies Cisplatin Apoptose induziert über eine p53 vermittelte Aktivierung des p38 MAPK Pathways durch ROS Generierung [162]. Als Nächstes wurde das p53 Netzwerk von A2780cis Zellen nach



eine  $IC_{50}$  Dosis liposomales Cisplatin effektiv resistente A2780cis Zellen und ist dabei niedriger konzentriert. Die Microarraydaten zeigen, dass durch liposomales Cisplatin Gene des extrinsischen Apoptosewegs induziert sind, diese sind TNFRSF10B – DR5, c-FLIP-L und CD70-TNFSF7. Daher liegt die Schlussfolgerung Nahe, dass freies Cisplatin über oxidativen Stress (ROS) und in Konsequenz davon den p53 abhängigen intrinsischen Apoptoseweg induziert. Der liposomale Wirkstoff hingegen induziert über DNA Schaden effektiv den extrinsischen (Rezeptor-vermittelten) Apoptoseweg. In dieser Arbeit wurde daher ein molekularer Mechanismus beschrieben, der aufzeigt wie liposomales Cisplatin die Cisplatinresistenz umgehen kann.

### 4.2.2 Abschließende Diskussion der Wirkstoffstudie

Seit nunmehr über drei Dekaden werden Liposomen als ein vielversprechendes Wirkstoff Transportsystem in verschiedenen zielgerichteten Therapieansätzen berücksichtigt. In der Krebstherapie wird dieses Konzept zusätzlich durch den Effekt der erhöhten Permeabilität und der Retention (EPR) stark begünstigt. Dieser Effekt begünstigt eine passive Akkumulierung des Wirkstoffs in Tumorgewebe [164] und verringert somit unerwünschte Nebeneffekte an gesundem Gewebe und Organen. Die Liposomenaktivität auf dem zellulärem Niveau wurde bislang so verstanden, dass Liposomen durch ihre Beladung nach dem Zelleintritt eine höhere intrazelluläre Wirkstoffkonzentration verursachen. Diese Hypothese wurde in einer vorangegangenen Arbeit [45] daraufhin untersucht, ob eine Überwindung der Chemoresistenz von Tumorzellen erzielt werden kann. Tatsächlich konnte gezeigt werden, dass liposomales Cisplatin die Cisplatinresistenz in A2780cis Zellen durch höhere intrazelluläre Cisplatinkonzentration überwindet [45]. Dies hebt die oben genannte Hypothese hervor und bestätigt, dass eine verminderte Cisplatinaufnahme ein dominierender Faktor der Cisplatinresistenz ist.

Die Konzentration der Platinaddukte jedoch, die in dieser Studie erhoben wurden, welche von liposomalem Cisplatin verursacht werden, weisen allerdings auf einen anderen Wirkmechanismus hin. Die Anfangs erwogene Hypothese, dass liposomales Cisplatin durch erhöhte intrazelluläre Wirkstoffkonzentration die Cisplatinresistenz überwinden kann, erscheint nun zu einfach. Leere Liposomen sind per se in der verwendeten Konzentration nicht zytotoxisch [45], so dass die Möglichkeit einer additiven Wirkung auszuschließen ist. Die Liposomenaktivität in A2780cis Zellen muss daher über eine Genexpressionsanalyse nachvollzogen werden. A2780cis Zellen wurden mit den entsprechenden  $IC_{50}$  Dosen von freiem oder liposomalem Cisplatin behandelt, um jegliche Effekte, die auf unterschiedlichen Wirkstoffaktivitäten basieren auszuschließen. Auf diese Wei-

se werden die qualitativen Unterschiede im apoptotischen Potential der Wirkstoffe auf der Transkriptionsebene nachvollziehbar. Eine globale Analyse der deregulierten Gene, welche funktionell durch die am meist affektierten GeneGo Netzwerke wiedergegeben werden, demonstriert eindrucksvoll, dass liposomales Cisplatin ein einzigartiges Set zellulärer Stoffwechselwege induziert, die gänzlich verschieden von jenen sind, die durch den freien Wirkstoff induziert werden. Im Hinblick auf die Frage, wie liposomales Cisplatin die Resistenz in A2780cis Zellen überwindet, sind jene Befunde am wichtigsten, welche Stoffwechselwege ausgehend von p53 differentiell induzieren. Es ist bekannt, dass Cisplatin über eine p53 Aktivierung und über die Bildung reaktiver Sauerstoffspezies (ROS) eine Aktivierung des p38 MAPK Pathways und somit die über die Mitochondrien vermittelte, intrinsische Apoptose induziert [162]. Die erhobenen Microarraydaten zeigen für freies Cisplatin, dass Schlüsselgene aus der Apoptose-vermittelnden Bcl-2 Familie wie zum Beispiel BAX und BID herauf-reguliert sind. Diese Situation ist grundverschieden im Fall von liposomalem Cisplatin. Verschiedene Komponenten des extrinsischen Apoptose Pathways werden durch liposomales Cisplatin induziert. Die Aktivierung des alternativen Apoptose Weges kann die höhere Wirksamkeit von liposomalem Cisplatin in A2780cis Zellen bei geringerer Dosis im Vergleich zu freiem Cisplatin erklären. Es ist davon auszugehen, dass die mit der Zeit erworbenen und durch evolutionärem Druck hervorgebrachten Verteidigungsmechanismen von A2780cis Zellen im Bezug auf Cisplatin durch den liposomalen Wirkstoff umgangen werden. Ob das Umgehen der Bildung von ROS und damit die Verhinderung einer Induktion des p38 MAPK Pathways durch liposomales Cisplatin auch auf andere Tumorzellen übertragbar und somit auch von den insgesamt weniger toxischen Nebeneffekten profitiert werden kann, ist Grundlage für weitere Forschung. Unter Berücksichtigung der Ergebnisse aus Tabelle 4.1 in der gezeigt wurde, dass TSA resistente A2780cis Zellen auf freies Cisplatin sensitiviert, nicht aber die Wirkung von liposomalem Cisplatin beeinflusst, kann nun durch obige Beobachtungen erklärt werden, zumal TSA den intrinsischen Apoptose Pathway sensitiviert, wie beschrieben in Muscolini *et al.* [147]. Liposomales Cisplatin induziert dagegen nicht den intrinsischen Pathway, und daher wird auch kaum eine durch Kombination vermittelte, additive Wirkung von TSA und liposomalem Cisplatin beobachtet. Obwohl diese Erkenntnis zu erklären vermag, wie liposomales Cisplatin den resistenten Status von A2780cis Zellen überwinden kann, fehlt eine Antwort hinsichtlich der Relation auf das verminderte DNA-Platinaddukt Niveau und der gesteigerten apoptotischen Aktivität. Es ist wohl bekannt, dass Cisplatin innerhalb des Zytosols verschiedene Zielstrukturen neben der DNA aufweist und somit verschiedene subzelluläre Kompartimente angreift, mit weitreichenden Konsequenzen für unterschiedliche funktionelle und signalgebende Pathways [165]. All diese subzellular-

lären Komponenten werden von Cisplatin angegriffen und steuern somit einen Beitrag zum beginnenden globalen Apoptose Szenario bei. Die Wirkung von liposomalem Cisplatin unterscheidet sich grundlegend von der Wirkung freien Wirkstoffs, da nach der endozytotisch vermittelten Aufnahme andere subzelluläre Komponenten adressiert werden, mit ebenfalls verschiedenen Konsequenzen für die globale zelluläre Antwort. Dies könnte auch erklären, wie eine marginale Änderung in der räumlichen intrazellulären Verteilung von Cisplatin, gravierende Auswirkungen für das Gleichgewicht der zellulären apoptotischen Aktivität aufweist, ohne eine Korrelation zur DNA-Platinierung aufzuweisen.

**Fazit** Ungeachtet der Tatsache, dass viele offene Fragen hinsichtlich der intrazellulären Aufnahme der Liposomen verbleiben, ist dies jedoch die erste Arbeit, die mechanistische Informationen bietet und in Aussicht stellt, dass Liposomen ein erfolgversprechendes Werkzeug zur Überwindung der Chemoresistenz darstellen. Eine der wichtigsten Aussagen dieser Arbeit ist, dass liposomales CDDP im Unterschied zum freien CDDP, über den Tumorsuppressor p53 eine Aktivierung der extrinsisch vermittelte Apoptose einleiten kann, unabhängig von Resistenzstatus und Effektivität der DNA Platinierung. Dieser Wirkmechanismus dezimiert effektiv chemoresistente A2780cis Zellen und eröffnet somit neue Diskussionen für den kumulativen Effekt liposomal induzierter Wirkstoffe, die über eine bloße Carrier Funktion hinausreichen.

### 4.2.3 Resultate der Zervixkarzinom Tumor Progressionsstudie

#### 4.2.3.1 Qualitätskontrolle und Normalisierung der Studien aus GEO führte auf eine Patientendatenmatrix

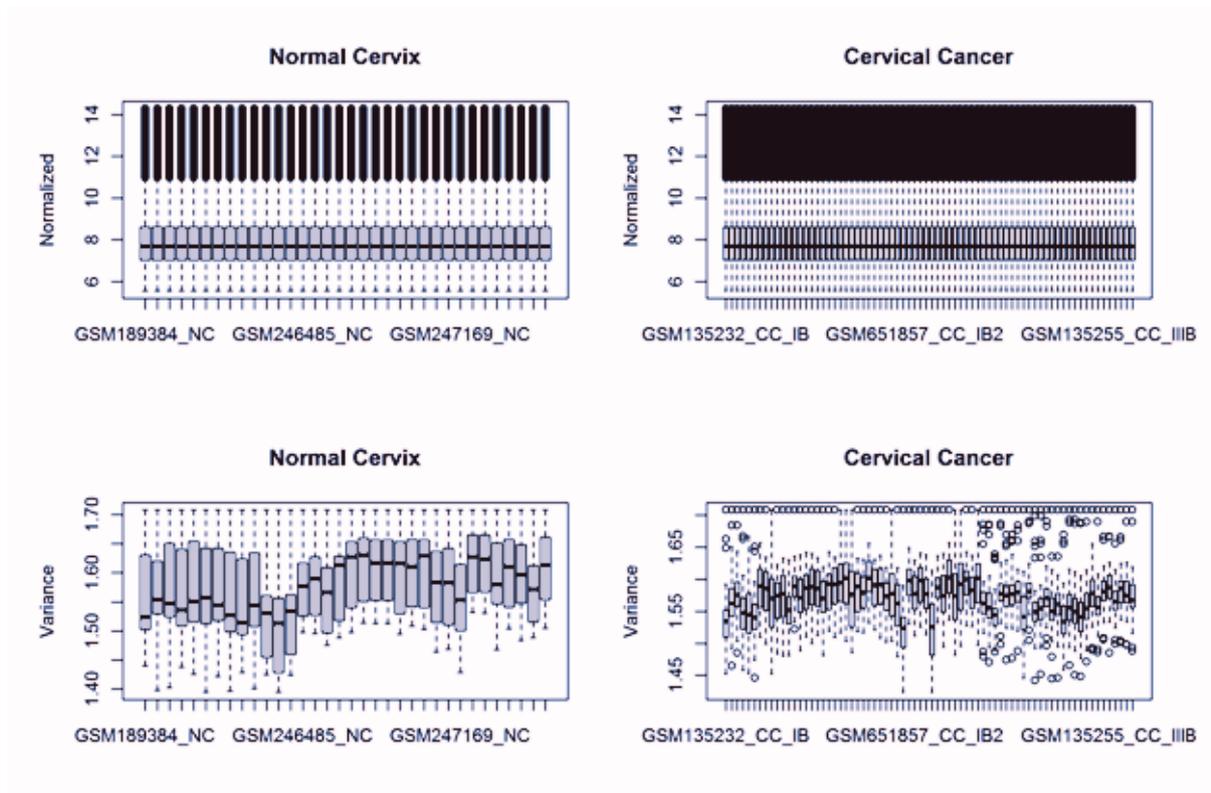
Die Erhebung und Bewertung der Qualitätsparameter von Microarraydaten ist ebenso wichtig wie die Normalisierung der Daten selbst, daher inspizieren wir die Studien im Folgenden separat.

Die Studie **GSE5787** enthielt 33 Arrays, von denen nur das Array GSM135260 eine Ausreißerkontrollprobe aufwies, die übrigen Kontrollen waren unauffällig.

In Studie **GSE6791** fanden nur 18 von 27 Arrays Verwendung in der nachfolgenden Analyse, die übrigen mussten aufgrund von Qualitätseinbußen verworfen werden. Sogar die 18 verwendeten Arrays wiesen geringfügige, qualitative Schwankungen auf. Weil jedoch die durchschnittliche Hintergrundintensität und die gemessenen Signale im Bereich der Toleranz waren, wurden sie dennoch berücksichtigt.

Acht der 57 Arrays aus Studie **GSE9750** mussten aufgrund des kaum vorhandenen Signals von der Analyse ausgeschlossen werden (*i.e.* GSM246422, GSM246423, GSM246484, GSM246485, GSM246486, GSM246487, GSM246488 and GSM246489). Im Gegensatz dazu konnten alle 39 Arrays der Studie **GSE26511** berücksichtigt werden, da diese von überragender Qualität sind.

Als Nächstes wurden alle Studien einem zusätzlichen RNA-Degradationstest unterzogen wobei jedoch kein weiteres Array auffällig erschien. So konnten 131 Arrays extrahiert werden, die qualitativ hohen Standards entsprechen. Sinngemäß mussten jedoch Studien, die in den Studien nur durch einzelne Proben vertreten waren, ebenso von der Analyse ausgeschlossen werden, da so geringe Probenzahlen als nicht repräsentativ gelten. Dazu zählten Proben aus den Stadien IA, IIA und ebenso die drei Proben aus dem späten und daher seltenen IVB Stadium. Insgesamt verblieben somit 126 Arrays für die weitere Bearbeitung. Die Tabelle 3.2 listet die Arrays nach Stadium bezüglich FIGO quantitativ auf. Alle Arrays wurden zunächst innerhalb der Studien über die Quantilnormalisierung des Pakets *RMA* [136], ohne Hintergrundabgleich vornormalisiert. Anschließend wurden diese in drei aufeinanderfolgenden Schritten mit dem CONOR Paket [154] zu einer einzigen Patientenmatrix gebündelt.

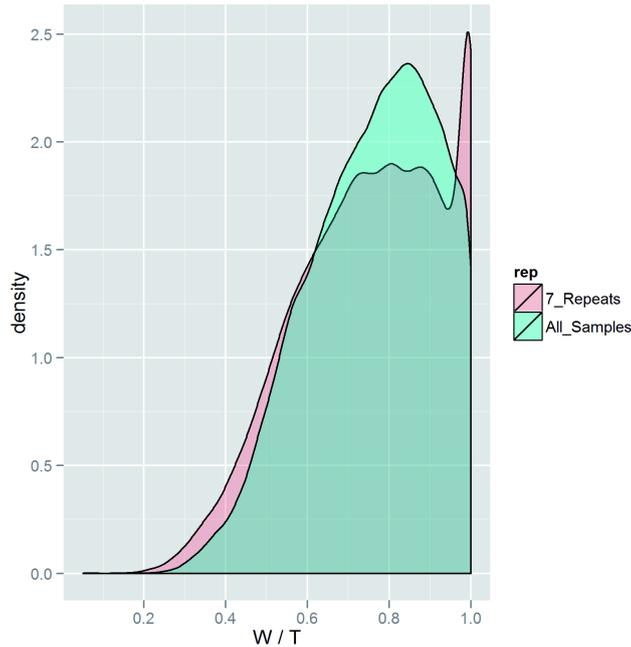


**Abbildung 4.12:** Die Abbildung 4.12 stellt die Verteilungen der normalisierten Intensitäten als Boxplots dar. Der untere Teil der Abbildung zeigt die Varianz innerhalb der einzelnen Arrays.

Die normalisierten Microarrays wurden als Boxplots in der Abbildung 4.12, neben der intra Array Varianz dargestellt. Die erhaltenen Genexpressionswerte sind gleichverteilt, nachdem die Quantilnormalisierung angewendet wurde. Die Varianz jedoch bleibt erhalten, wie in Abbildung 4.12 gezeigt.

#### 4.2.3.2 Varianz Komponentenanalyse der Zervixkarzinom Stadien

Die Patientenmatrix enthält 22 277 Proben für jedes der 126 Arrays. Diese repräsentieren eine gesunde und sechs histologische Gewebefunde zervikaler Karzinome durch die FIGO. Diese Proben wurden einer Varianz-Komponentenanalyse unterzogen, um die weniger heterogen exprimierten Proben zu erhalten.



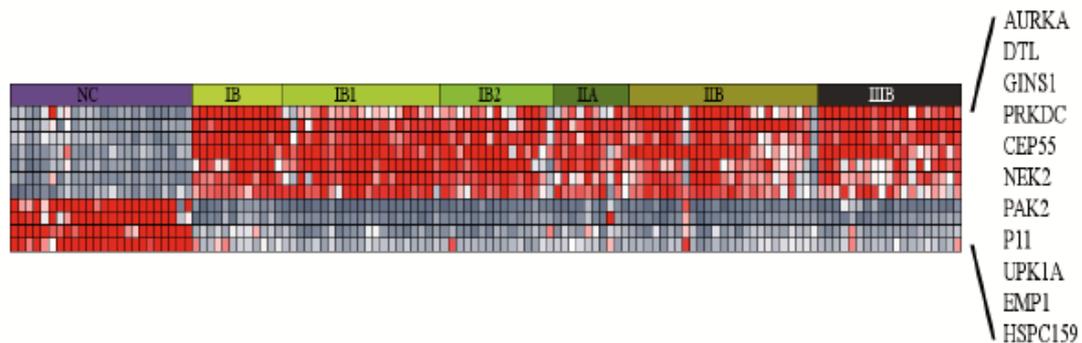
**Abbildung 4.13:** Abbildung 4.13 stellt die resultierenden Dichteverteilungen aus der Varianz-Komponentenanalyse verschiedener Probenanzahlen graphisch dar. Während die Analyse mit sieben, die verschiedenen Stadien repräsentierenden Proben einen steilen Anstieg in der Variabilität bedeutet, kann unter Verwendung aller Proben eine geringere Steigung erzielt werden. Dies weist darauf hin, dass für die Erfassung von wenig heterogen exprimierten Genen eine viel größerer Probenanzahl benötigt wird.

In der Abbildung 4.13 ist die resultierende Verteilung des W/T Verhältnisses der untersuchten Stadien gezeigt. Die schwarze Linie beschreibt die Dichteverteilung für sieben zufällig gewählte Replikate der Studien, die rote Linie hingegen bildet die Verteilung aller mit der Varianz-Komponentenanalyse untersuchten Arrays ab. Die Varianz-Komponentenanalyse der sieben zufällig ausgewählten Replikate führt zu einem steilen Anstieg der Proben, die ein W/T Verhältnis von 1 aufweisen. Ein kleinere Verhältniszahl zeigt jedoch Proben auf, die innerhalb der Studien weniger heterogen exprimiert sind. Dies zeigt, dass die Verwendung aller verfügbaren Proben zu einem besseren Ergebnis, also einer geringeren Gesamtheterogenität führt. Der Peak der unter Verwendung aller Proben erhaltenen Dichteverteilung liegt bei einem W/T Verhältnis von 0,8.

**Schwellwert der weniger heterogenen Proben** Im nächsten Schritt definierten wir einen Schwellwert, um weniger heterogene und daher konstitutiv über die Stadien ex-

primierte Proben zu erhalten. Im folgenden werden Proben, die einen W/T Anteil von 0,75 und weniger besitzen ausgewählt. So können 9873 Proben für die weitere Verarbeitung gewonnen werden.

#### 4.2.3.3 Konstitutiv über alle Studien exprimierte, wenig heterogenen Proben



**Abbildung 4.14:** Abbildung 4.14 stellt die Gene dar, die nach der Varianz-Komponentenanalyse, die geringste Variabilität innerhalb der Stadien aufweisen. Diese Gene sind konstitutiv in den Stadien exprimiert und für normales bzw. karzinogenes Gewebe spezifisch. Das UPK1A Gen kodiert für das Gewebe spezifische Uroplakin Protein. Dieses hat Gewebe begrenzende Funktion und könnte somit eine Markerfunktion für gesundes Gewebe einnehmen und im Umkehrschluss für die Lokalisierung von Tumorgewebe eingesetzt werden.

Eine Cluster-Analyse der 9783 Proben bringt wenig heterogene Gene, die konstitutiv über alle Studien exprimiert werden zum Vorschein. Diese Proben sind induziert oder reprimiert in Zervixkarzinomen und haben einen W/T Anteil zwischen 0,18 und 0,38. Sieben dieser Proben sind in allen Zervixkarzinom Stadien induziert. Dies sind die folgenden Gene GINS1, PAK2, DTL, AURKA, PRKDC, NEK2 und CEP55. Die übrigen Proben sind ausschließlich in gesundem Gewebe induziert, jedoch in Zervixkarzinomen über alle Stadien reprimiert. Die Tabelle 4.5 führt alle über die Varianz-Komponentenanalyse erhaltenen potentiellen Biomarker auf.

**Tabelle 4.5:** Die Tabelle listet alle potentiell verwendbaren Biomarker und deren Annotation.

Nr.	W/T	Symbol	Beschreibung	RefSeq ID
1	0,34	GIN51	GIN5 complex subunit 1 (Psf1 homolog)	<a href="#">NM_021067</a>
2	0,35	PAK2	p21 protein (Cdc42/Rac)-activated kinase 2	<a href="#">NM_002577</a>
3	0,35	DTL	denticleless homolog (Drosophila)	<a href="#">NM_016448</a>
4	0,37	AURKA	aurora kinase A	<a href="#">NM_198437</a>
5	0,38	PRKDC	protein kinase, DNA-activated	<a href="#">NM_006904</a>
6	0,38	NEK2	NIMA -related kinase 2	<a href="#">NM_002497</a>
7	0,38	CEP55	centrosomal protein 55kDa	<a href="#">NM_018131</a>
8	0,18	P11	26 serine protease	<a href="#">NM_006025</a>
9	0,24	EMP1	epithelial membrane protein 1	<a href="#">NM_001423</a>
10	0,26	UPK1A	uroplakin 1A	<a href="#">NM_007000</a>
11	0,26	HSPC159	galectin-related protein	<a href="#">NM_014181</a>

#### 4.2.3.4 GSEA des aus der Varianz-Komponentenanalyse erhaltenen Genexpressions Sets

9783 Proben haben einen W/T Anteil geringer als 0,75, diese wurden nach Anreicherung von bekannten Gensignaturen mit GSEA untersucht. Zunächst jedoch wurde nur die Genexpression gegen die Gesamtheit der Proben aus Stadium I mit GSEA verglichen. Die Tabelle 4.6 listet alle in dieser Analyse angereicherten Gensets auf. Die Mehrzahl der angereicherten Prozesse in Zervikalkarzinomen stehen im Zusammenhang mit der Instandhaltung der DNA. Zusätzlich finden sich Signaturen, die Proliferation bedeuten oder auch Metastasen einleiten. Die Gensignatur „Cervical Cancer Proliferation Cluster“ enthält 105 in den Daten angereicherte Gene, deren P-Wert und FDR unter  $1e^{-4}$  liegt. Diese Signatur wurde von Rosty *et al.* [166] geprägt, allerdings umfasst die vollständige Signatur 163 Gene, da die Autoren auch Gene, welche hoch variabel sind, berücksichtigt haben. Dieser Anteil jedoch wurde in der hier vorliegenden Analyse ausgelassen. Die Gensignatur „Bidus, Metastasis up“ [167] verweist auf die Publikation von Bidus *et al.*, die Autoren untersuchten die Vorhersage von Lymphknoten Metastasen im Endometriumkarzinom mittels Genexpressionssignaturen. Diese Signatur ist in der Patientenmatrix auf hohem Signifikanz Niveau angereichert, da der erhaltene P-Wert und die FDR unterhalb von  $1e^{-4}$  liegen. Die Ergebnisse der Voruntersuchung zeigen auf, dass die vorgeschlagene Methode Phänotyp abhängige und zugleich wenig heterogene Gensignaturen extrahiert. Daher werden nun auch die funktionellen Eigenschaften jedes einzelnen Stadiums untersucht, dies geschieht über den Vergleich eines jeden Stadiums mit dem nachfolgendem Stadium.

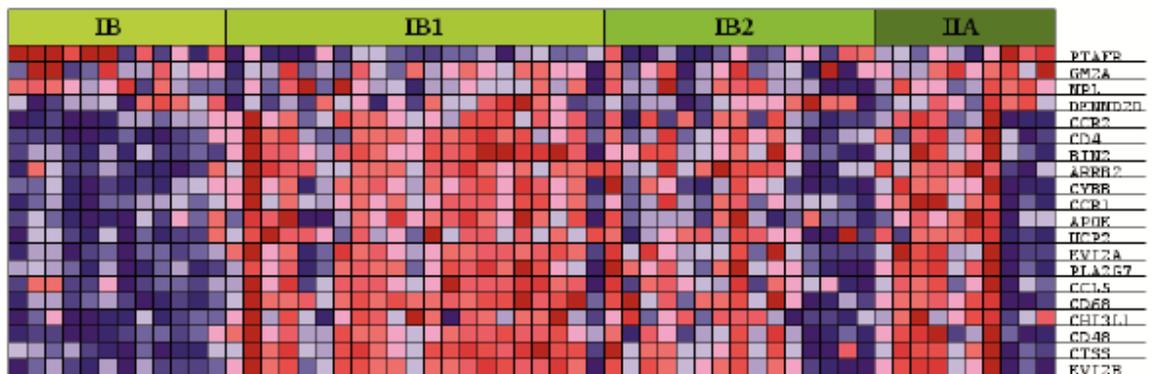
**Tabelle 4.6:** Die Tabelle listet alle signifikant angereicherten Gene in Stadium I auf.

IB	Beschreibung	Gene	ES	p-Wert	FDR
1	GO, RNA export from nucleus	15	0,72	$1e^{-4}$	$1e^{-2}$
IB1	Beschreibung	Gene	ES	p-Wert	FDR
1	Wilensky, Response to darapladib	20	0,90	$1e^{-4}$	$1e^{-4}$
2	Gaurnier, PSMD4 targets	23	0,89	$1e^{-4}$	$1e^{-4}$
3	Flechner, biopsy kidney transplant rejected	44	0,83	$1e^{-4}$	$1e^{-4}$
4	Wieland, Up by HBV infection	56	0,80	$1e^{-4}$	$1e^{-4}$
5	Browne, Interferon responsive genes	35	0,80	$1e^{-4}$	$1e^{-4}$
6	Reactome, Chemokine receptors	23	0,78	$1e^{-4}$	$1e^{-4}$
7	KEGG, Graft versus host disease	18	0,77	$1e^{-4}$	$1e^{-4}$
8	Reactome, TCR signaling	26	0,76	$1e^{-4}$	$1e^{-4}$
9	GO, Chemokine activity	16	0,76	$1e^{-4}$	$4e^{-3}$
10	Reactome, Downstream TCR signaling	19	0,75	$1e^{-4}$	$1e^{-4}$
11	GO, G protein coupled receptor binding	17	0,74	$1e^{-4}$	$7e^{-3}$
12	GO, Chemokine receptor binding	18	0,74	$1e^{-4}$	$8e^{-3}$
IB2	Beschreibung	Gene	ES	p-Wert	FDR
1	Rozanov, MMP14 targets subset	15	0,85	$1e^{-4}$	$1e^{-4}$
2	KEGG, ECM receptor interaction	43	0,80	$1e^{-4}$	$1e^{-4}$
3	Reactome, NCAM1 interactions	28	0,78	$1e^{-4}$	$1e^{-4}$
4	GO, Regulation of cell migration	15	0,75	$6e^{-3}$	$2e^{-2}$

**Über GSEA erhaltene Prozesse in Zervixkarzinomen** Die Tabellen 4.6 und 4.7 geben einen Überblick über die durch GSEA erhaltenen Signaturen. Unter den höchst signifikanten und meist angereicherten Prozessen sind “Wilensky, Response to Darapladib” (ES 0,90) und “Gaurnier, PSMD4 Targets” (ES 0,89) in Stadium IB1 zu finden. Diese Signaturen werden ebenfalls in Stadium IIA angereichert und erhalten eine Anreicherung (ES) von 0,84 und 0,76 respektive. Die Gaurier Signatur ist nochmals in Stadium IIB exprimiert und erhält erneut eine hohe Anreicherung (ES 0,80).

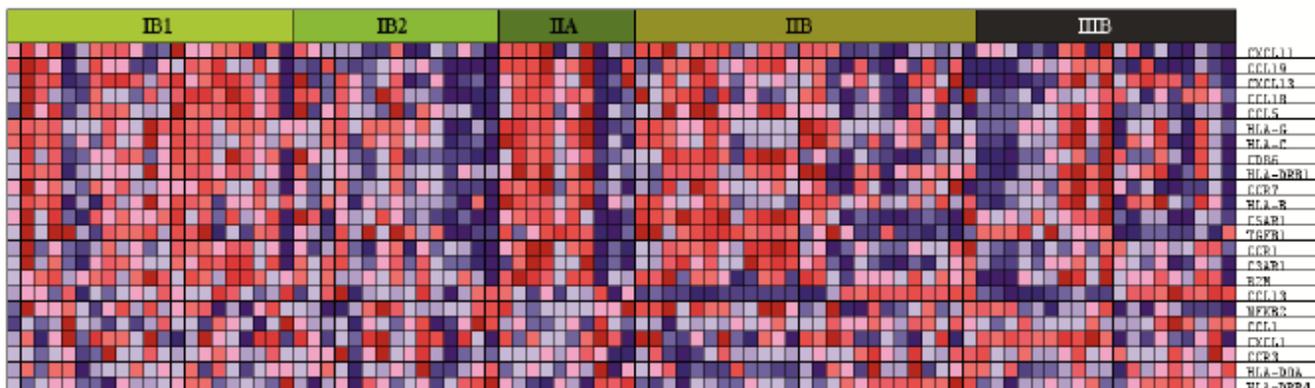
In der Literatur beschreiben Wilensky *et al.* [168] eine Gensignatur, welche nach der Behandlung mit dem Lipoprotein-assoziierte Phospholipase A2 (Lp-PLA2) Inhibitor darapladib reprimiert wird. Die Autoren postulieren, dass der Lp-PLA2 Inhibitor Arteriosklerose bedingte Läsionen in den Gefäßwänden reduziert und eine antiinflammatorische Wirkung ausübt, indem eine 24 Gene umfassende Signatur reprimiert wird, welche mit der Rekrutierung von Makrophagen und T-Zellen assoziiert ist.

In Stadium IB1 und IIA sind 20 Gene, die der Wilensky Signatur angehören induziert. Diese sind darüber hinaus signifikant angereichert (P-Wert und FDR  $< 1e^{-4}$ ), wie in Abbildung 4.15. gezeigt.



**Abbildung 4.15:** Die Abbildung 4.15 stellt die Expressionswerte der Wilensky Signatur als Heatmap dar. Diese Signatur ist in Zervixkarzinomen in Stadium IB1 und Stadium IIA induziert, und geht phänotypisch offenbar mit der Bildung von Läsionen einher. Darapladib könnte die Expression dieser Gene reprimieren und somit auch den Gewebeveränderungen entgegen wirken.

Die Gaurnier Signatur besteht insgesamt aus 59 Genen, von denen 23 als signifikant angereichert in den Stadien IB1, IIA und IIB vorliegen, wie in der Abbildung 4.16 gezeigt



**Abbildung 4.16:** Die Abbildung 4.16 zeigt 23 Gene der Gaurnier Signatur in einer Heatmap. Diese Signatur ist in den Stadien IB1, IIA und IIB induziert und oftmals in Begleitung gravierender immunologischer Signaturen.

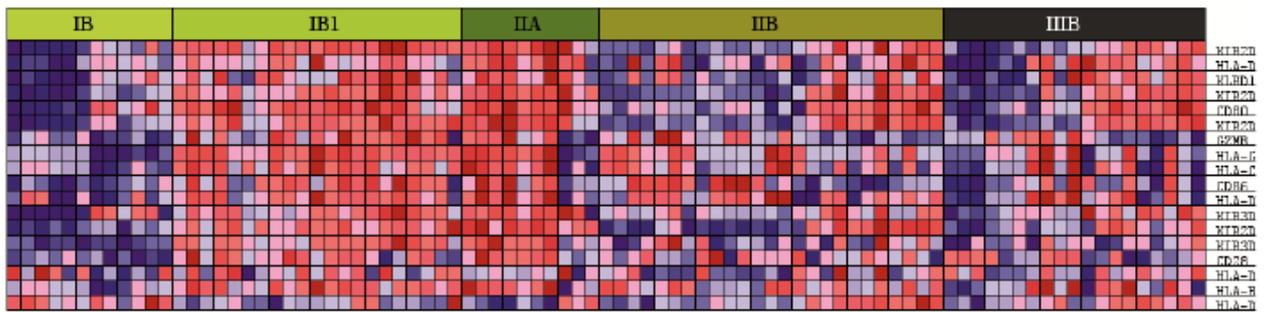
Die Autoren konnten belegen, dass Angiocidin, das Protein Produkt des PSMD4 Gens,

die Aktivierung von antitumoralen, MMP9 sezernierenden Monozyten zur Folge hat [169]. Jedoch konnte hier nur ein Bruchteil der Gaurnier Signatur als signifikant exprimiert nachgewiesen werden, mit P-Wert und FDR  $< 1e^{-4}$ . Die Gaurnier Signatur ist stets co-exprimiert zu einer in der KEGG Datenbank vorhandenen Signatur, die auf einem dramatischen immunologischen Prozess basiert, “KEGG, Graft versus Host Disease”.

**Tabelle 4.7:** Die Tabelle listet alle signifikant angereicherten Gene in Stadium II und Stadium III auf.

<b>IIA</b>	<b>Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>	<b>FDR</b>
1	Wilensky, Response to darapladib	20	0,84	$1e^{-4}$	$1e^{-4}$
2	Browne, Interferon responsive genes	35	0,79	$1e^{-4}$	$1e^{-4}$
3	GO, Humoral immune response	15	0,77	$1e^{-4}$	$3e^{-2}$
4	Gaurnier, PSMD4 targets	23	0,76	$1e^{-4}$	$1e^{-3}$
5	Biocarta, BCR pathway	17	0,73	$1e^{-4}$	$3e^{-2}$
6	KEGG, Graft versus host disease	18	0,72	$1e^{-4}$	$1e^{-4}$
7	Reactome, downstream TCR signaling	19	0,72	$4e^{-3}$	$4e^{-2}$
8	Reactome, TCR signaling	26	0,71	$1e^{-4}$	$1e^{-2}$
9	GO, Chemokine activity	16	0,70	$1e^{-4}$	$1e^{-2}$
<b>IIB</b>	<b>Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>	<b>FDR</b>
1	Gaurnier, PSMD4 targets	18	0,80	$1e^{-4}$	$1e^{-4}$
2	GO, Metalloendopeptidase activity	19	0,78	$1e^{-4}$	$1e^{-4}$
3	KEGG, Graft versus host disease	23	0,71	$2e^{-3}$	$2e^{-2}$
4	Flechner, Biopsy kidney transplant rejected	44	0,70	$1e^{-4}$	$1e^{-4}$
<b>IIIB</b>	<b>Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>	<b>FDR</b>
1	Slebos, Head and neck cancer with HPV up	30	0,78	$1e^{-4}$	$1e^{-4}$
2	KEGG, Drug metabolism cytochrome P450	32	0,70	$1e^{-4}$	$1e^{-4}$

Eine Heatmap dieser Signatur ist in der Abbildung 4.17 dargestellt. Die sogenannte Graft-versus-Host-Reaktion kann durch drei bestimmte Schritte charakterisiert werden. In der ersten Phase werden inflammatorische Cytokine ausgeschüttet, gefolgt von Interferon gamma (IFN- $\gamma$ ) und schließlich initiieren zuvor aktivierte CTL und NK Zellen die Apoptose über Fas-Fas Ligand Interaktionen. Brown *et al.* postulieren eine 68 Gene beinhaltende Signatur, welche spezifisch für auf Interferon-alpha ansprechende Gene ist [170], 35 dieser Gene konnten als signifikant exprimiert in den Stadien IB1 (ES 0.80) und IIA (ES 0.79) angereichert nachgewiesen werden.



**Abbildung 4.17:** Die Abbildung 4.17 zeigt die „KEGG, Graft vs. Host Disease“ Signatur in einer Heatmap, diese ist simultan zur Gaurnier Signatur induziert.

Eine weitere immunologisch relevante Signatur wurde von Flechner *et al.* [171] beschrieben, diese ist in Stadium IB1 (ES 0.83) und Stadium IIB (0.70) signifikant angereichert. Die Signatur beinhaltet Gene, deren Überexpression nach einer Nieren Transplantation der Abstoßung des Organs vorausgehen („Biopsy Kidney Transplant Rejected vs. Ok Up“).

**Heatmap „Graft vs. Host Disease“** Eine Heatmap der durch GSEA erhaltenen, aus KEGG stammenden „Graft vs. Host Disease“ Signatur ist in der Abbildung 4.17 dargestellt.

**Stadien spezifische Prozesse in Zervixkarzinomen** In Stadium IB liegt als einziger Prozess der „Export von RNA aus dem Zellkern“ innerhalb des Signifikanzniveaus, welches durch FDR und ES gegeben ist. Die weiteren gefundenen Prozesse in Stadium IB konnten nicht als signifikant angesehen werden. Eine Signatur jedoch enthielt Anreicherung (ES 0,56) von Genen, welche bereits von Peyon *et al.* beschrieben wurden und deren Daten durch Arrays aus GSE6791 Eingang in diese Arbeit fanden. In Stadium IB1 ist eine 107 Gene betragende Signatur angereichert, welche während der Klärung einer HBV Infektion induziert ist, diese wurde von Wieland *et al.* beschrieben [172]. 56 dieser Gene sind signifikant induziert und haben P-Werte und FDR unterhalb  $1e^{-4}$ . Die Untersuchung des Stadiums IB2 führte auf Gensignaturen, die nicht in den übrigen Stadien zu finden waren. Rozanov *et al.* stellten die „MMP14 Targets Subset“ Signatur in [173] vor, diese ist in Stadium IB2 angereichert. Die Autoren postulierten, dass die

Genexpressionssignatur auf eine MT1-MMP abhängige Migration und Invasion schließen lässt. Zudem ist ein Extrazellulärmatrix Rezeptorinteraktionsprozess, wie in der KEGG Datenbank hinterlegt und der NCAM1 Interaktionsprozess aus der Reactome Datenbank signifikant angereichert (P-Wert,  $FDR < 1e^{-4}$ ), wie auch die Regulation der Zellmigration aus der Gen Ontologie (P-Wert  $6e^{-3}$ ). In Stadium IIA sind immunologisch relevante Genexpressionssignaturen, d.h. die humorale Immunantwort, T-Zell vermittelte Prozesse wie der TCR Signalweg und auch tiefer gelegene, dem Signalweg assoziierte Prozesse angereichert. Abgesehen von den bereits in Stadium IIB erwähnten Prozessen, weist der GO Terminus „Metalloendopeptidase Aktivität“ eine signifikante Anreicherung (ES 0,78) auf. In Stadium IIIB gab es insgesamt zwei Prozesse, die den Signifikanzniveaus gerecht wurden. Der erste Prozess ist eine von Slebos *et al.* vorgestellte Signatur, welche auf 89 Gene verweist und in HPV positiven Kopf-Hals-Karzinomen exprimiert ist [174]. Die zweite signifikante Gensignatur ist durch die in KEGG abgelegte „Drug Metabolism Cytochrome P450“ Signatur gegeben, deren Anreicherung 0,70 beträgt. Die mit GSEA ermittelten Gene aus der Varianz-Komponentenanalyse erbrachten eine Charakterisierung des transkriptionellen Programms jedes Zervixkarzinom Stadiums. Die Gensignaturen, die aus dieser Vorgehensweise resultieren überlappen zwar oftmals mit den Signaturen aus anderen Studien, jedoch nur in Bereichen robuster Genexpression, da hoch variable und damit für die Charakterisierung unzuverlässige Gene ausgelassen werden.

### 4.2.4 Abschließende Diskussion der Zervixkarzinom Progressionsstudie

In dieser Arbeit wurde eine funktionelle Untersuchung der Genexpressionsprofile von Zervixkarzinom Stadien und eine Identifikation potentiell klinisch relevanter Biomarker durchgeführt. Zu diesem Zweck wurden Genexpressionsprofile FIGO basierter Kohorten retrospektiv aus vier öffentlich verfügbaren Microarraydaten vereint. Nach Qualitätskontrolle und Normalisierung verblieben Genexpressionsprofile von 126 Patienten für die weitere Verarbeitung durch Varianz-Komponentenanalyse und GSEA. Diese Vorgehensweise vermied die Analyse hoch variabler Gene, welche auf spekulative Genexpressionssignaturen hin führen. Stattdessen wurde eine konservative Vorgehensweise angewandt, welche zuverlässige Biomarker prospektiert, die konstitutiv exprimiert sind.

**Biomarker** Die Varianz-Komponentenanalyse ermittelte elf Gene, welche potentiell als Biomarker für die Detektion von Zervixkarzinomen fungieren können. Sieben dieser Proben sind konstitutiv induziert in allen Zervikalkarzinom-Stadien. Diese sind die vier Kinasen: PAK2, NEK2, AURKA und PRKDC. Des Weiteren zählt dazu die Ligase DTL und ein Gen, welches für das zentrosomale Protein 55 (CEP55) codiert. Zudem das GINS1 Gen, welches essentiell für die DNA Replikation ist. GINS1 ist in Hefe und in Xenopus Eiern [175] evolutiv konserviert. Diese Gene stehen grundsätzlich im Zusammenhang mit Zellzyklus und DNA betreffenden Prozessen. Andererseits finden sich vier Gene, welche durchgängig reprimiert in Zervixkarzinom Patienten sind, diese sind ausschließlich in gesundem Gewebe induziert. Zu diesen gehören der Tumor-Marker P11 [176] und der Tumor-Suppressor EMP1 [177]. Ein weiteres, ebenso reprimiertes Gen ist UPK1A. Dieses kodiert ein zelloberflächenlokalisierendes Protein und HSPC159, ein dem Galectin verwandtes Protein. Schlussfolgernd betrachtet muss der Nutzen der hier vorgestellten Biomarker durch künftige Studien erst noch belegt werden, da diese jedoch aus Patientendaten stammen und von geringer Variabilität sind, bieten sie ausgezeichnete Kandidaten.

**Stadienspezifische Prozesse** Die GSEA umfasste eine über Varianz-Komponentenanalyse erhaltene Patientenmatrix. Diese enthielt 9783 Proben, welche einen W/T Anteil von weniger als 0,75 aufwiesen. Es ist besonders auffällig, dass die hier erhaltenen Resultate Überlappungen mit Ergebnissen von drei aus der Literatur stammenden Studien aufweisen [149, 166, 174]. Zudem gibt es überlappende Genexpressionssignaturen mit zwei weiteren Virus verwandten Genexpressionsstudien (HBV und HCMV) [170, 172]. Diese Resultate bezeugen, dass die hier vorgestellte Methode in Bezug auf die konservative Exploration variabler Genexpressionsprofile anwendbar ist. Dies ist im Hinblick auf die jüngere Literatur besonders aussichtsreich, da vergleichbare Studien häufig nur in marginal überlappenden Gensignaturen resultieren.

**Darapladib** Die hier vorgestellten Ergebnisse rechtfertigen die Vermutung, dass in darapladib eine neue Option für die adjuvante Therapie von Zervixkarzinomen gegeben sein könnte, da wiederholt Genexpressionssignaturen auftreten, die eine Behandlung mit darapladib nahe legen [168]. Obwohl dieses Resultat einen signifikanten Befund darstellt gilt es jedoch auch zu bedenken, dass die Untersuchungen auf den transkriptionellen Status von Zervixkarzinomen beschränkt ist. Ob dieser Befund tatsächlich klinische Auswirkungen hat, wäre erst in unabhängigen klinischen Studien zu zeigen. Der Lipoprotein-assoziierte Phospholipase A2 (Lp-PLA2) Inhibitor darapladib reprimiert

miert 24 Gene, von denen 20 in Zervixkarzinompatienten exprimiert sind. Wilensky *et al.* zeigten, dass darapladib Läsionen in atherosklerotischen Gefäßen reduziert und zudem auch eine antiinflammatorische Wirkung ausübt [168]. HPV Neuinfektionen manifestieren sich als zervikale intraepitheliale Neoplasie (CIN) oder werden innerhalb von 36 Monaten geklärt [178]. Daher wäre eine Studie, ob die Ausbildung von Läsionen in Zervixkarzinomen mit darapladib reduziert werden kann aussichtsreich, auch im Hinblick darauf, ob infolgedessen eine Klärung der Infektion vorangetrieben werden kann.

**Angiocidin** Eine weitere Behandlungsmodalität für die Therapie von Zervixkarzinomen wäre durch den neuen Angiogenese Inhibitor angiocidin gegeben. Gaurnier *et al.* beschrieben als erste, dass das Protein Produkt des PSMD4 Gens zur Expression von insgesamt 59 Genen in THP-1 Zellen führt, welche die Zellen zu antitumoralen MMP9 sezernierenden Monozyten transformiert [169]. In dieser Arbeit konnten jedoch nur 23 dieser Gene signifikant induziert in den Stadien IB1, IIA und IIB nachgewiesen werden. Eine sich wiederholende Transkription dieser Gene könnte eine Reaktion des Immunsystems der Patienten bedeuten, welches auf Virenzyklus inhärente Signale reagiert. Zudem könnte dies auch die Aktivierung immunologischer Verteidigungsmechanismen nach sich ziehen, die eine Klärung der Infektion veranlassen. Da allerdings nicht alle 59 Gene als transkribiert gefunden werden, ist davon auszugehen, dass auch die immunologische Reaktion nur teilweise aktiviert scheint. Als Grundlage für Spekulation könnte der Einfluss von viralen Proteinen auf den Wirts-Zellzyklus dienen. Wie in der Literatur bereits beschrieben, ist das virale E6 Protein in der Lage, die Wirtsphysiologie zu einem Großteil zu beeinflussen, z.B. indem der Tumorsuppressor P53 abgebaut wird [166]. Die Gaurnier Signatur wird jedoch stets von Genexpressionssignaturen begleitet, welche dramatische immunologische Prozesse nach sich ziehen. Diese sind die Graft-versus-Host-Reaktion und Abstoßung des Organs nach Nierentransplantation. Die oben genannten immunologischen Prozesse könnten eine konkretere Basis für die Fehlfunktion in der Aktivierung der Monozyten bieten. Grundlage für diese Untersuchung waren Genexpressionsprofile von Zervixkarzinom Patienten. Daher gilt es zu bedenken, bevor eine klinische Auswirkung prognostiziert werden kann, dass die getroffenen Aussagen erst durch eine umfangreiche Überprüfung und weitere klinische Studien verifiziert werden müssen. Diese sollte zunächst feststellen, ob die gefundenen Effekte lediglich auf der Änderung von biologischen Markern beruhen, oder ob tatsächlich eine Änderung der Anzahl und Zusammensetzung der Diversität in den akkumulierenden immunologischen Zellen beobachtet werden kann. Letzteres würde auf

eine tatsächlich stattfindende Auseinandersetzung des Wirts-Immunsystems mit den Folgen der HPV Infektion hinweisen.

**Funktionelle Untersuchung des Stadiums IB2** Das Stadium IB2 weist im Gegensatz zu den sich wiederholenden Genexpressionsmustern der übrigen Stadien ein gänzlich eigenes Expressionsprofil auf. In Stadium IB2 findet sich eine von Rozanov beschriebene Gensignatur, welche auf Matrix-Metalloprotease Aktivität hinweist („MMP14 Targets Subset“) [173]. Die Autoren wiesen nach, dass die Expression des MT1-MMP Gens direkt in Assoziation zu Invasion und Metastase im Allgemeinen steht. Zudem wurden hoch signifikante Gensignaturen, die auf zelluläre Lokomotion zurückzuführen sind, in Stadium IB2 gefunden. Diese sind die in GO hinterlegte Regulation der Zell-Migration und NCAM1 Interaktionen aus der Reactome Datenbank. Ein Hinweis auf die transkriptionelle Aktivierung von invasiven Prozessen kann in Gensignaturen die auf ECM Rezeptor Interaktionen hindeuten gefunden werden. Wie Rozanov *et al.* bereits zeigten induziert die MT1-MMP Aktivierung auch eine Aktivierung der ECM Instandhaltung. Die oben genannten Prozesse weisen einen Zusammenhang auf, da Tumorzellen den Metalloproteasen verursachten Verdau der zellulären Matrix durch die Aktivierung der ECM Instandhaltung ausbalancieren [173]. Andererseits wäre dies auch ein Beweis für ein transkriptionelles Programm, welches die Tumorzellen aus dem Stadium IB2 nach Stadium II transponieren, in der eine Invasion in die unterhalb des Uterus liegenden Gewebsschichten stattfindet. Eine Analyse der Genexpressionsprofile von Zervikalkarzinom Patienten bietet also in dieser Hinsicht eine neue Perspektive im Hinblick auf HPV vermittelte Transkriptionsprozesse.

**Ausblick** Es gibt eine stetig wachsende Vielzahl publizierter Genexpressionssignaturen und biologischer Stoffwechselwege die künftig eine breitere Basis für das Verständnis der Prozesse bieten, die HPV befähigt die Immunabwehr permanent zu umgehen. Dieses Wissen kann auch Implikationen für andere karzinogene Viren bieten. Daher ermöglichen bereits etablierte Genexpressionssignaturen offenbar die nächste größere Behandlungsmodalität als Chance in der Krebstherapie.

**Fazit** In dieser Arbeit konnten elf hoch zuverlässige Proben mit potentiellen Biomarker Eigenschaften gefunden werden. Des Weiteren bezeugen breite Überlappungen in den gefundenen Gensignaturen mit Signaturen aus Studien in der Literatur den Nutzen dieser Methode. Das augenfälligste Resultat jedoch waren Genexpressionssignaturen, die mit Signaturen assoziiert sind, die nach einer Behandlung mit angiocidin

und darapladib exprimiert sind. Vorausgehende Studien belegen, dass angiocidin die Differenzierung von Monozyten zu Makrophagen induziert, die antitumorale Wirkung ausüben. Darapladib hingegen übt eine antiinflammatorische Wirkung aus. Daher ist anzunehmen, dass angiocidin und darapladib sich als neue Optionen in der Therapie von Zervixkarzinomen erweisen werden.

## 4.2.5 Triple-negativ Brustkrebs Subtypisierungsstudie

### 4.2.5.1 Qualitätskontrolle und Normalisierung führen auf eine molekulare subtypspezifische Genexpressionsmatrix

**Tabelle 4.8:** Die Tabelle 4.8 listet alle klinisch relevanten Merkmale der Brustkrebspatienten sowie der Zellkulturen auf. Insgesamt gibt es Microarray Expressionsdaten von 367 Patienten, davon haben 99 den Status TNBC. Zusätzlich gibt es 20 TNBC Zellkultur Proben, die verschiedene Brustkrebs Zelllinien repräsentieren.

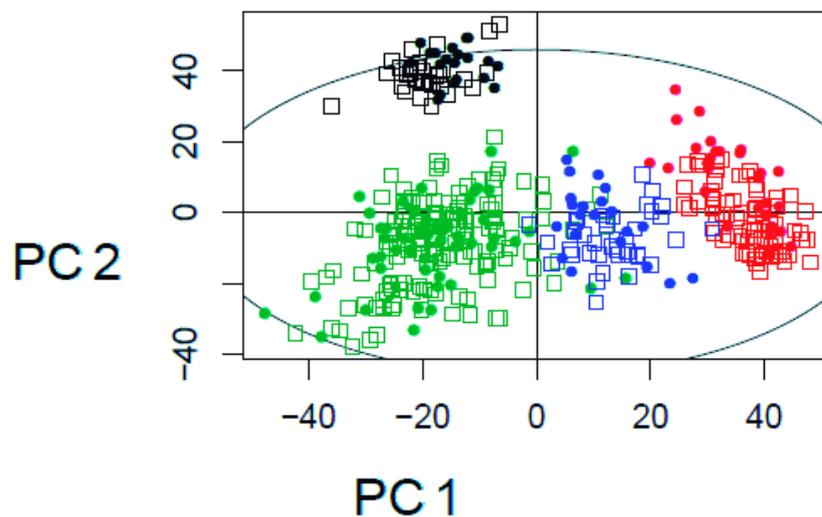
	GSE19615	GSE20194	GSE2603	E-TABM-157	Gesamt
<b>ER Status</b>					
Negative	27	51	21	20	119
Positive	57	114	30	-	201
Missing	-	-	-	31	31
<b>PGR Status</b>					
Negative	3	39	7	20	69
Positive	54	75	23	-	152
Missing	-	-	-	31	31
<b>HER2 status</b>					
Negative	42	101	23	20	186
Positive	15	13	7	-	35
Missing	-	-	-	31	31

Die Qualität der Rohdaten wurde vor der Normalisierung evaluiert, so dass nur hoch qualitativ erscheinende Arrays in die Analyse aufgenommen wurden. Insgesamt bleiben 367 Arrays für die weitere Bearbeitung. Diese setzen sich anhand ihres Rezeptorstatus wie folgt zusammen. Es wurden 99 Arrays von triple-negativen Patienten (*i.e.* „NNN“) und 35 Arrays von triple positiven Patienten (*i.e.* „PPP“) erhalten. Ferner sind 49 Arrays von doppelt negativen mit Östrogen positivem Rezeptor (*i.e.* „PNN“) und 45 Arrays von doppelt negativen Patienten mit positivem HER2 Status (*i.e.* „NNP“) enthalten. Obwohl es drei weitere rezeptorvariante Subtypen gibt, erreichten diese jedoch

nicht eine statistisch repräsentative Anzahl. Für die nachfolgend durchgeführten Klassifizierungen wurden die molekularen Subtypen auf 35 Array große Sets stratifiziert, die alle 22 277 Proben je Array enthielten. Auf diese Weise wurden die molekularen Subtypen gegen den triple-negativen Fall verglichen.

#### 4.2.5.2 Hauptkomponentenanalyse aller Brustkrebs Studien

Eine graphische Darstellung der Hauptkomponentenanalyse ist in Abbildung 4.18 gegeben. Verschieden Formen und Farben kennzeichnen die Herkunft der Daten und auch den charakterisierenden Rezeptorstatus. Leere Quadrate symbolisieren die Referenz Brustkrebs Proben, wohingegen gefüllte Kreise die triple-negativen Proben andeuten.



**Abbildung 4.18:** Die Abbildung 4.18 stellt die ersten beiden Hauptkomponenten graphisch dar. Die Hauptkomponenten weisen deutlich eine Aggregation nach Studienherkunft auf, obwohl die Genexpressionswerte der einzelnen Studien vereint und normalisiert worden sind. Die erklärte Varianz durch die erste Hauptkomponente beträgt 18%, die zweite Hauptkomponente erklärt 11% der Varianz.

#### 4.2.5.3 Klassifizierung der Expressionsprofile von Referenz Brustkrebs gegen triple-negative Brustkrebs

Zunächst muss überprüft werden, ob Klassifizierungen mit den erhobenen Daten möglich sind, da eine Hauptkomponentenanalyse zwar eine Trennung der Daten vermochte,

diese jedoch für den Rezeptor unspezifisch und sogar nach Ursprung der Studien (siehe Abbildung 4.18) erfolgte. Daher wurde zunächst von Markus Hanl aus dem Arbeitskreis Prof. Wiese mit dem Random Forest (RF) Algorithmus die Kompatibilität der Studien erforscht. Jedes Datenset wurde zunächst einzeln verwendet, um nach Erstellung des RF-Modells die Klassenzugehörigkeit der übrigen Daten vorherzusagen. Die Vorhersage der Patientendaten mittels auf Zellkulturen basierenden Daten ergab ein unzureichendes Ergebnis, d.h. die Matthews's Korrelationskoeffizienten waren kleiner als 0,59. Dies war ein Ausschlusskriterium für die weitere Verwendung. Die Vorhersage der Patientendaten beruhend auf den weiteren Patientendaten erzielte annehmbare Ergebnisse, mit Ausnahme der Vorhersage von Studie GSE19615. Der umgekehrte Fall jedoch, wenn Daten aus Studie GSE19615 zur Vorhersage verwendet wurden, erbrachte vertretbare Resultate.

**Tabelle 4.9:** In Tabelle 4.9 sind die von Markus Hanl erhaltenen Ergebnisse der Klassifizierung mit dem Random Forest Algorithmus aufgelistet. Die Gewichtung der intrinsischen Eigenschaften der Datensets wurde evaluiert durch den Zusammenschluss der Ergebnisse von jeweils 50 RF Modellen a 2000 Entscheidungsbäumen.

	<b>E-TABM-157</b>	<b>GSE19615</b>	<b>GSE2603</b>	<b>GSE20194</b>
<b>E-TABM-157</b>	1	0.58	0.56	0.59
<b>GSE19615</b>	0.73	0.95	0.82	0.79
<b>GSE2603</b>	0.65	0.55	1	0.78
<b>GSE20194</b>	0.36	0.59	0.60	0.92

Die Klassifizierungen wurden mit einer fünffachen Leave-Group-Out Kreuzvalidierung durchgeführt und die Trainings- und Testsets wurden auf den allgemeinen Rezeptorstatus generalisiert, d.h. RBC versus TNBC. Anschließend wurden die erhaltenen RF-Modelle dazu verwendet, den Rezeptorstatus der übrigen Datensets vorherzusagen. Die Klassifizierungen anhand der zellkulturbasierten Microarraydaten waren, wie bereits erwähnt problematisch. Daher mussten diese von der übrigen Analyse ausgeschlossen werden. Im Gegensatz dazu gestaltete sich die Klassifizierung der Expressionsprofile auf Basis der Patientendaten mit den entsprechenden RF-Modellen. Diese Klassifizierung erhielt ein hohes Maß an Genauigkeit ( $> 0,8$ ), geringe Fehlerraten ( $< 0,1$ ) und ein statistisch relevantes Maß der Matthews's Korrelationskoeffizienten für die Klassifizierung von triple-negativen und Referenz Brustkrebs Expressionsprofilen (siehe Tabelle 4.9). Die Anzahl der Entscheidungsbäume in den jeweiligen Modellen war 2000 und es wurden bei jedem Split 149 Gene berücksichtigt. Um das Klassenungleichgewicht auszugleichen wurde ein sogenanntes Bootstrapping während der Modellgenerierung durchgeführt, um die Probenanzahl in den Klassen anzugleichen.

#### 4.2.5.4 GSEA der Genexpressionsprofile von triple-negativen Patienten, die über die Klassifikation mit dem Random Forest erhalten wurden

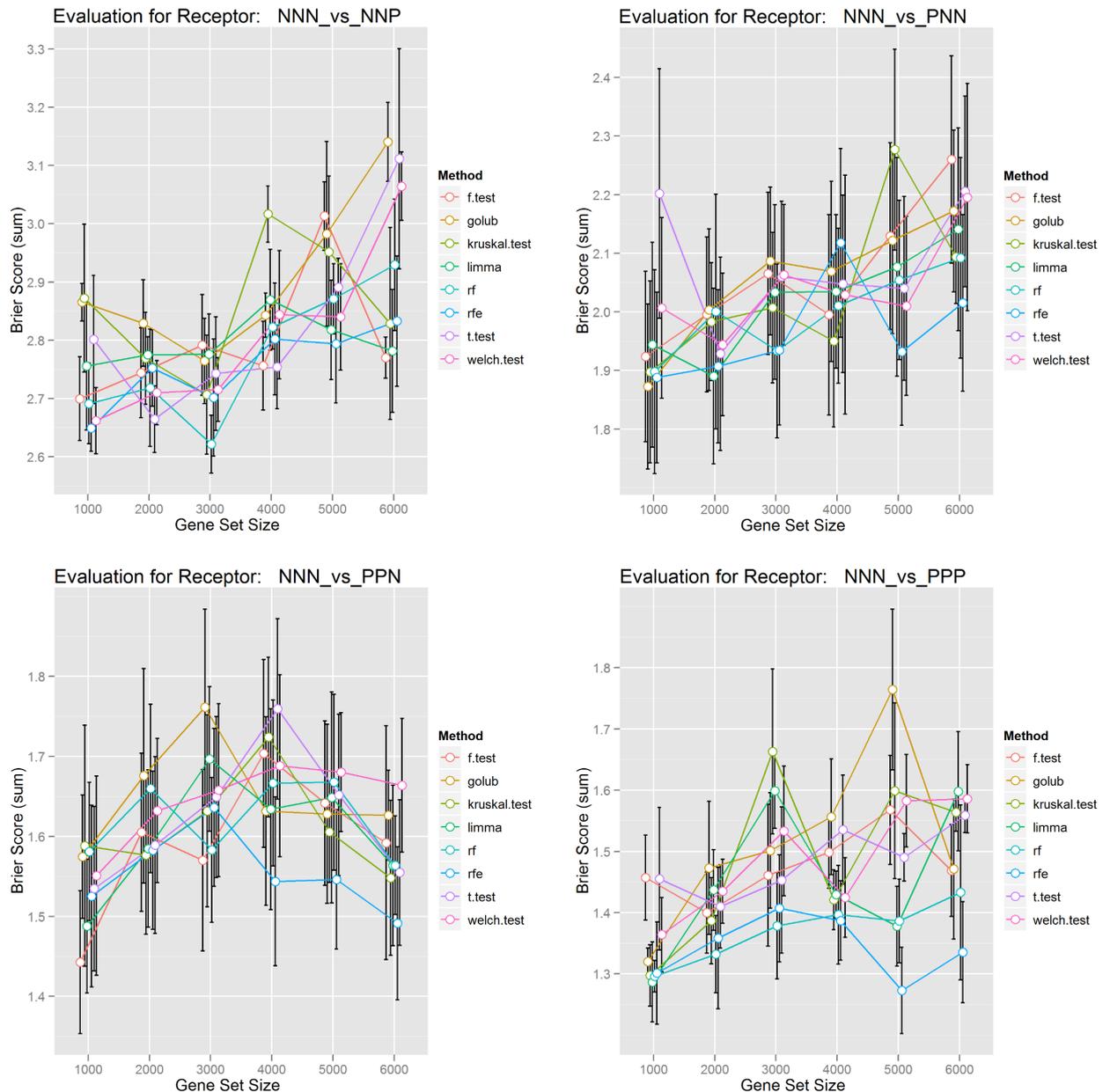
**Tabelle 4.10:** Die Tabelle 4.10 listet 25 signifikant angereicherte Gensignaturen in triple-negativem Brustkrebs auf, die über die Klassifikation mit dem Random Forest Algorithmus erhalten werden. Die ersten zehn Prozesse beschreiben ESR1 beinhaltende Signaturen, die reprimiert sind. Oftmals sind jedoch nur Bruchteile der in der Datenbank für molekulare Signaturen (MSigDB) hinterlegten Signaturen angereichert.

Beschreibung	Gene	ES	p-Wert	FDR
Yang, Breast cancer ESR1 DN	15 (19)	-0.84	$1e^{-4}$	$1e^{-4}$
Doane, Breast cancer ESR1 DN	45 (48)	-0.83	$1e^{-4}$	$1e^{-4}$
Smid, Breast cancer relapse in brain UP	31 (41)	-0.80	$1e^{-4}$	$1e^{-4}$
Vantveer, Breast cancer ESR1 DN	130 (223)	-0.79	$1e^{-4}$	$1e^{-4}$
Turashvili, Breast ductal carcinoma vs. lobular	26 (72)	-0.79	$1e^{-4}$	$1e^{-4}$
Graham, Normal quiescent vs normal dividing	29 (89)	-0.78	$1e^{-4}$	$1e^{-4}$
Smid, Breast cancer basal UP	290 (676)	-0.78	$1e^{-4}$	$1e^{-4}$
Smid Breast cancer relapse in bone DN	171 (337)	-0.77	$1e^{-4}$	$1e^{-4}$
Smid, Breast cancer luminal B DN	252 (599)	-0.77	$1e^{-4}$	$1e^{-4}$
Doane, Breast cancer classes DN	18 (34)	-0.76	$1e^{-4}$	$1e^{-4}$
Vantveer, Breast cancer metastasis DN	45 (107)	-0.76	$1e^{-4}$	$1e^{-4}$
Flechner, Biopsy kidney transplant rejected	25 (88)	-0.76	$1e^{-4}$	$1e^{-4}$
Yang, Breast cancer ESR1 laser DN	17 (37)	-0.75	$1e^{-4}$	$1e^{-4}$
Hinata, NF $\kappa$ B targets keratinocyte UP	24 (71)	-0.75	$1e^{-4}$	$1e^{-4}$
Whiteford, Pediatric cancer markers	31 (92)	-0.74	$1e^{-4}$	$1e^{-4}$
Benporath, ES core nine correlated	47 (100)	-0.74	$1e^{-4}$	$1e^{-4}$
Sotiriou, Breast cancer grade 1 vs. 3 UP	50 (153)	-0.73	$1e^{-4}$	$1e^{-4}$
Rosty, Cervical cancer proliferation cluster	44 (140)	-0.72	$1e^{-4}$	$1e^{-4}$
Bidus, Metastasis UP	43 (217)	-0.68	$1e^{-4}$	$1e^{-4}$
Benporath, Proliferation	52 (147)	-0.68	$1e^{-4}$	$1e^{-4}$
Turashvili, Breast ductal carcinoma vs. ductal	77 (202)	-0.67	$1e^{-4}$	$1e^{-4}$
Senese, HDAC2 targets DN	32 (131)	-0.66	$1e^{-4}$	$1e^{-4}$
Poola, Invasive breast cancer UP	77 (304)	-0.65	$1e^{-4}$	$1e^{-4}$
Charafe, Breast cancer luminal vs. basal DN	144 (456)	-0.64	$1e^{-4}$	$1e^{-4}$
Schuetz, Breast cancer ductal invasive UP	102 (355)	-0.63	$1e^{-4}$	$1e^{-4}$

#### 4.2.5.5 Entwicklung der FECSnV Methode

Die Charakterisierung der biologischen Eigenschaften von triple-negativen Brustkrebs Expressionssignaturen erfolgte mit einer GSEA unter Verwendung der 3000 wichtigsten, durch den RF Klassifizierer zugewiesenen Genen. Die Tabelle 4.10 listet 25 der meist signifikanten Gensignaturen als Charakteristika des triple negativen molekularen Subtyps auf. Wie erwartet enthalten die ersten Gensets Signaturen, die auf Repression

des Östrogenrezeptors (ESR1) hinweisen. Dieser ist gegenwärtig der stärkste klinische Prediktor. Zusätzlich finden sich verschiedene Brustkrebs relevante Signaturen in unterschiedlich angereichertem, jedoch signifikantem Zustand. Diese weisen auf Prozesse hin, die Metastasenbildung, einen spontanen Rückfall der Krankheit (Relapse) und sogar Stammzellcharakteristika beinhalten.

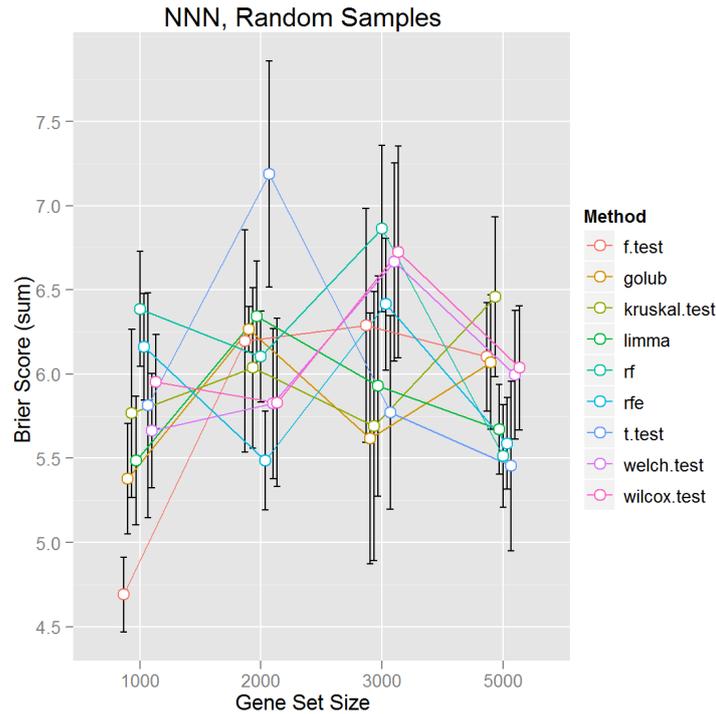


**Abbildung 4.19:** Die Abbildung 4.26 stellt die Ergebnisse der Klassifizierung graphisch anhand des Mittelwerts der Brier-Werte dar. Alle vier berücksichtigten Rezeptor Subtypen werden jeweils in fünf unabhängigen Klassifizierungsrunden mit dem triple-negativen Subtyp verglichen. Dabei werden jeweils Klassengrößen von 35 Subtypen eingesetzt, die zu Gensets von 1000 bis 6000 Gene selektiert werden.

Zunächst wurden die molekularen Subtypen mit dem Random Forest (RF) Klassifizierer getrennt. Die Performanz des RF Klassifizierers war jedoch unzureichend, die molekularen Brustkrebs Subtypen anhand des Rezeptorstatus zu unterscheiden (hier

nicht gezeigt). Daher wurde in dieser Arbeit eine Methode entwickelt, welche drei für die Klassifizierung essentiellen Schritte durchführt. Diese sind die Extraktion von Trennungsmerkmalen, den sogenannten Features, deren Bewertung mit einer sogenannten Scoring Funktion und die Reduktion der erhaltenen Trennungsmerkmale über eine Voting Funktion partizipiert. Ein Schema dieser drei sukzessive ausgeführten Schritte ist in Abbildung 3.1 gezeigt [82]. Die vorgestellte Methode integriert verschiedene Algorithmen aus maschinellen Lernverfahren und der statistischen Klassifizierung. Die Robustheit der Methode wurde anhand zufällig gewählter Proben analysiert, die eine Klassifizierung zweier Subtypen, bestehend aus völlig unterschiedlichen Klassenzugehörigkeiten simuliert. Die Ergebnisse aus fünf unabhängigen Simulationen dieser Art zufälligen Klassifikation sind in der Abbildung 4.20 gezeigt. Wie erwartet rangiert die Summe der Brier-Punktwerte zwischen fünf und sieben. In den meisten Fällen sind die Summen unabhängig von der eingangs verwendeten Gensetgröße. In der Abbildung 4.26 ist die Evaluierung anhand tatsächlicher molekularer Subtypen für verschiedene Gensetgrößen gezeigt.

Es ist auffällig, dass viel niedrigere Brier-Punktwerte erreicht werden, wenn der Methode zusammenhängende Klassen zugeführt werden. Des Weiteren ist auch ein Ansteigen der Punktwerte für größere Gensets zu beobachten. Alle Klassifizierungen wurden gegen eine gemeinsame Referenz, den triple-negativen Rezeptorsubtyp durchgeführt („NNN“). Insgesamt betrachtet nehmen die Brier-Punktwerte auch ab, je mehr die molekularen Subtypen voneinander divergieren, d.h. von doppelt negativ zu niedrigeren Werten im Vergleich triple positiv vs. triple-negative. Innerhalb der Klassifizierungen ist jedoch auch eine Zunahmen der Punktwerte ab einer Gensetgröße von über 3000 Genen zu verzeichnen, daher sollten für künftige Klassifizierungen Gensets mit nicht mehr als 3000 Genen verwendet werden. Außerdem ist anzumerken, dass zudem innerhalb der extrahierten Gene auch Redundanzen vorhanden sind.



**Abbildung 4.20:** Die Abbildung 4.20 zeigt die Mittelwerte der Brier-Werte aus fünf unabhängigen Runden für die Klassifikation mit zufällig ausgewählten Proben.

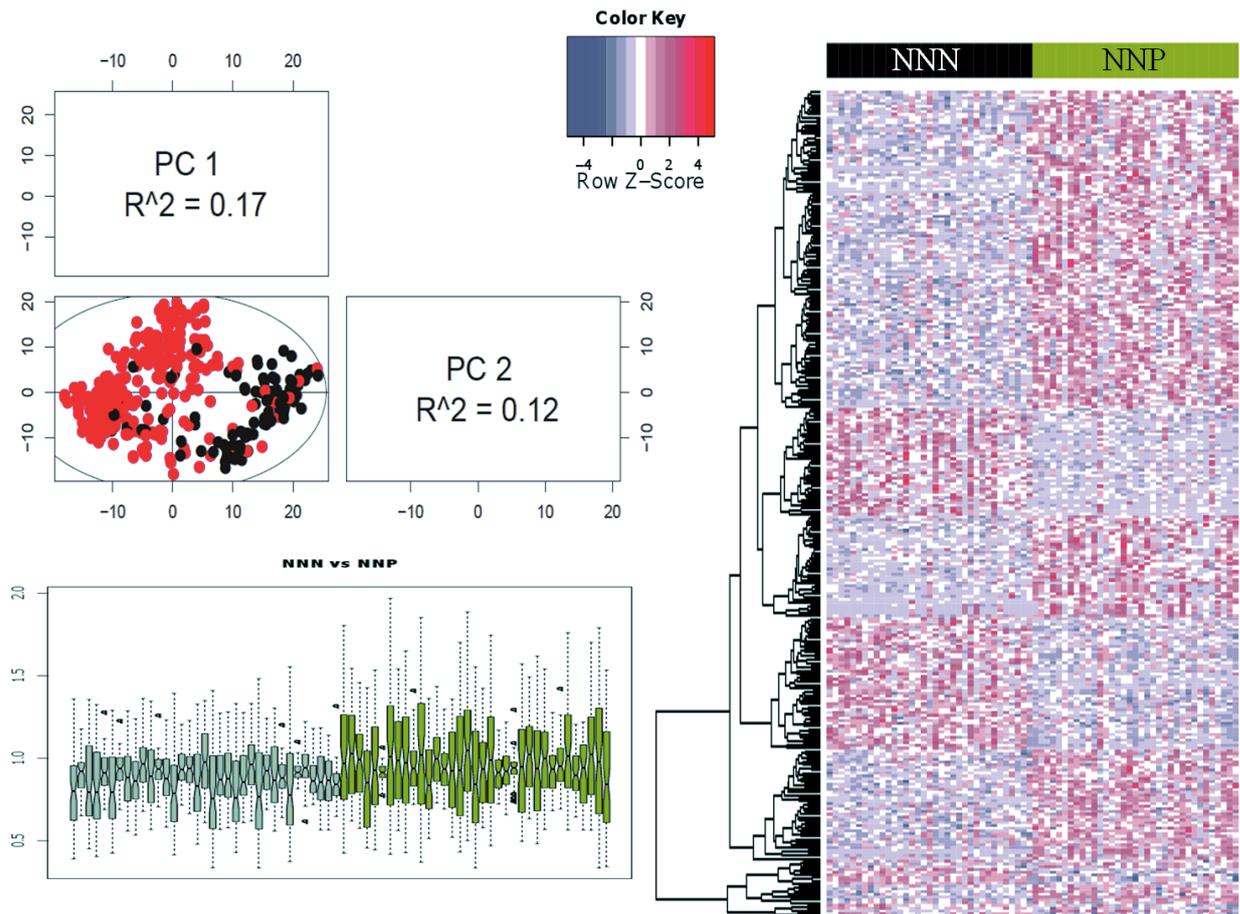
**Test auf Robustheit der FECSnV Methode** Das Resultat fünf unabhängiger Läufe zur Klassifikation von zufällig gewählten Proben, die daher keine geschlossenen Klassen repräsentieren ist in der Abbildung 4.20 dargestellt. Wie erwartet wird insgesamt eine Zunahme in den Brier-Werten beobachtet. Diese rangieren nun zwischen fünf und sieben. Zusätzlich scheint es keine Änderung in Relation auf die verwendete Gensetgröße zu geben, die erhaltenen Werte sind nahezu linear.

#### 4.2.5.6 GSEA der molekularen subtypspezifischen Signaturen, die mit der FECSnV Methode erhalten wurden

**Tabelle 4.11:** Die Tabelle 4.11 listet alle signifikant angereicherten Genexpressionsprofile in Brustkrebspatienten auf, die mit der FECSnV Methode erhalten wurden. Die hier dargestellten Prozesse sind signifikant in den molekularen Subtypen angereichert, d.h. der P-Wert ist kleiner 0,01 und die FDR unter 0,05. Die ersten zehn Prozesse sind induzierte, Brustkrebs relevante Signaturen, gefolgt von neun reprimierten Signaturen, die ebenso eine Relevanz in Brustkrebs aufweisen. Fälle in denen eine Signatur in dem betreffenden Subtypen nicht induziert ist, werden mit dem Buchstaben „X“ gekennzeichnet.

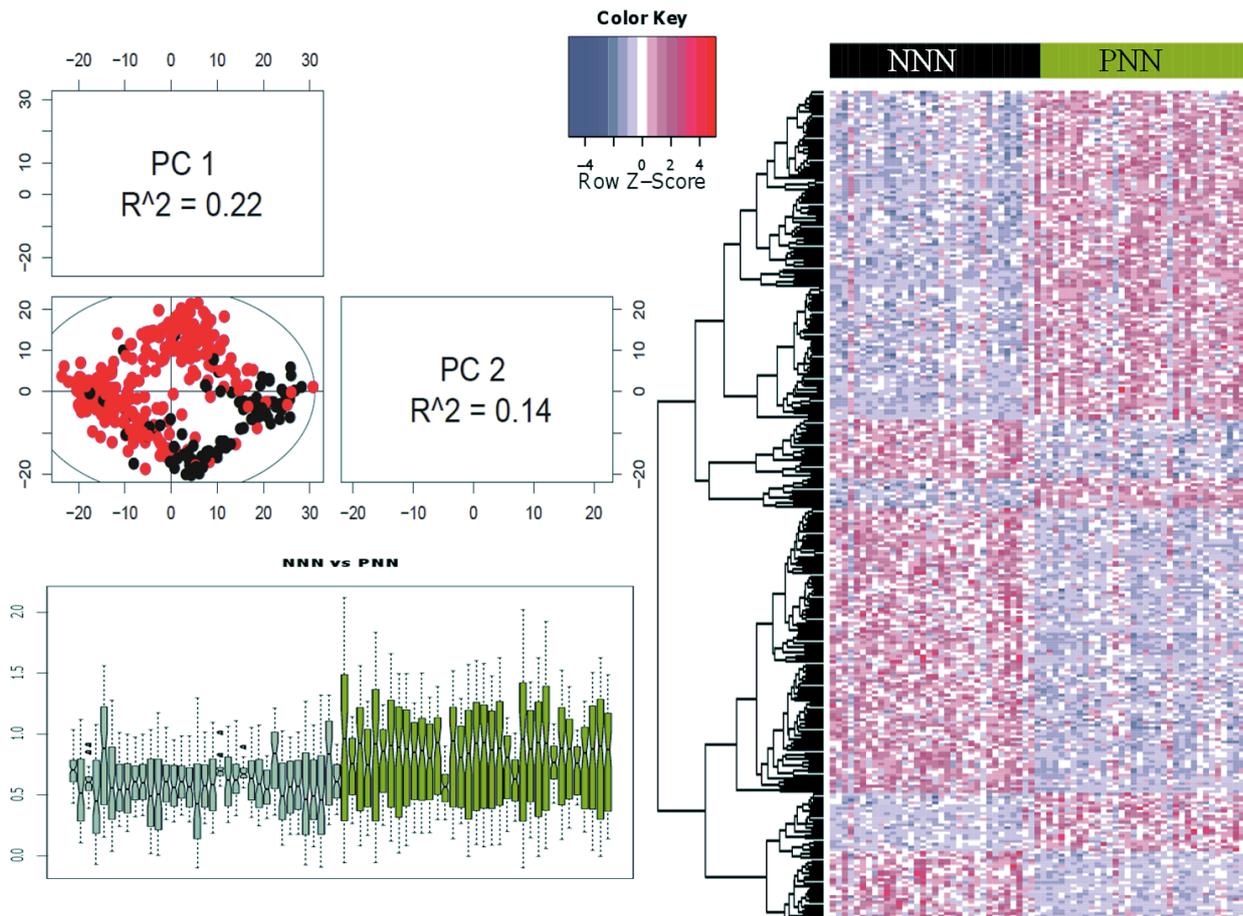
Beschreibung	PPP	PPN	PNN	NNP
Doane, Breast cancer classes UP	0.78	0.77	0.90	0.75
Yang, Breast cancer ESR1 UP	0.84	0.88	0.90	0.72
Doane, Breast cancer ESR1 UP	0.81	0.82	0.85	0.68
Vantveer, Breast cancer ESR1 UP	0.73	0.77	0.80	<b>X</b>
Smid, Breast cancer relapse in bone UP	0.79	0.81	0.87	0.71
Charafe, Breast cancer luminal vs. mesenchymal UP	0.62	0.64	0.70	0.64
Vantveer, Breast cancer metastasis UP	<b>X</b>	0.78	0.76	<b>X</b>
Smid, Breast cancer luminal B UP	0.74	0.77	0.82	<b>X</b>
Charafe, Breast cancer luminal vs. basal UP	0.65	0.71	0.73	<b>X</b>
Smid, Breast cancer luminal A UP	0.72	0.77	0.63	<b>X</b>
Massarweh, Response to estradiol	<b>X</b>	0.69	0.69	<b>X</b>
Smid, Breast cancer relapse in brain DN	0.78	0.81	0.84	0.68
Lien, Breast carcinoma metaplastic vs. ductal DN	0.83	0.87	0.86	0.73
Smid, Breast cancer basal DN	0.73	0.80	0.78	0.61
Gozgit, ESR1 targets DN	0.60	0.61	<b>X</b>	<b>X</b>
Schuetz, Breast cancer ductal invasive DN	0.62	0.70	<b>X</b>	<b>X</b>
Riggins, Tamoxifen resistance DN	0.62	0.63	<b>X</b>	<b>X</b>
Huang, Dasatinib resistance DN	0.77	0.78	0.72	<b>X</b>
Massarweh, Tamoxifen resistance DN	<b>X</b>	0.64	0.60	<b>X</b>
Smid, Breast cancer relapse in lung DN	0.81	<b>X</b>	0.80	<b>X</b>

Die Tabelle 4.11 listet für die molekularen Subtypen alle mit GSEA als angereichert gefundenen Signaturen auf. Die ersten zehn Signaturen weisen auf die Induktion Brustkrebs relevanter Prozesse hin, welche den ESR1 Typ sowie Luminal A und B enthalten. Die nachfolgenden neun Prozesse beschreiben ebenso Brustkrebs verwandte Prozesse, die jedoch reprimiert sind. Darunter finden sich Signaturen, die den Basalen Typ charakterisieren, sowie Tamoxifen und Dasatinib Resistenz. Tamoxifen wird generell in der Therapie von Östrogenrezeptor negativem Brustkrebs verwendet, währende Dasatinib an HER2 positiven Brustkrebspatienten derzeit in klinischer Phase II getestet wird [179].



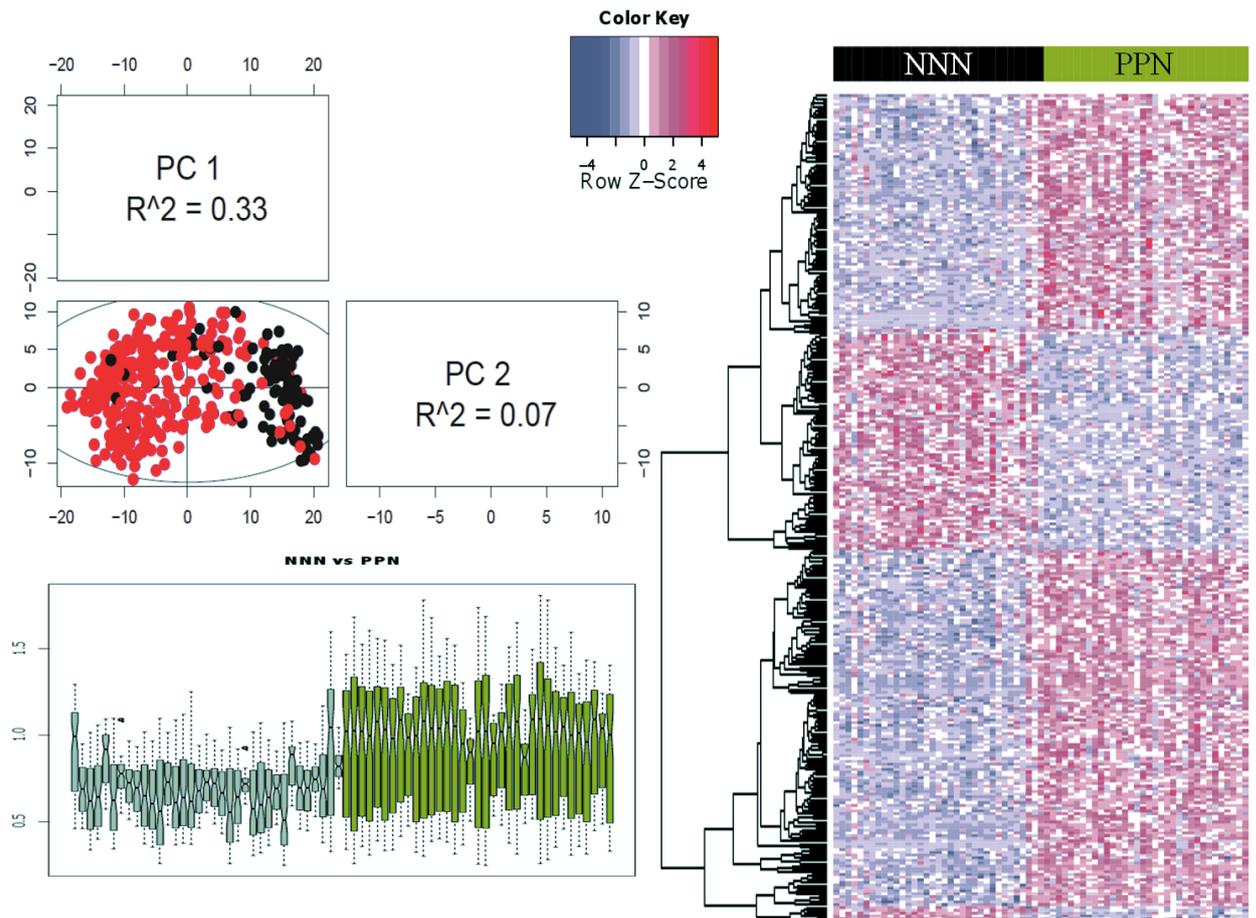
**Abbildung 4.21:** Die Abbildung 4.21 stellt die Resultate der FECSnV Klassifizierung des molekularen Subtyps NNP graphisch dar.

**Expressionsprofil des NNP Subtyps** Die Klassifizierung des NNP Subtyps gegen den triple-negativen Fall mit der FECSnV Methode ergab 1926 Gene, deren Expressionsprofil in der Abbildung 4.21 gezeigt ist. Das linke obere Diagramm stellt die ersten beiden Hauptkomponenten graphisch einander gegenüber, wobei die triple negativen Fälle als rote Kreise dargestellt werden. Das untere Diagramm zeigt Box-Plots der Varianz der berücksichtigten Proben und die Heatmap im rechten Teil zeichnet die Genexpressionswerte der Gene, die als Klassifizierer gelten.



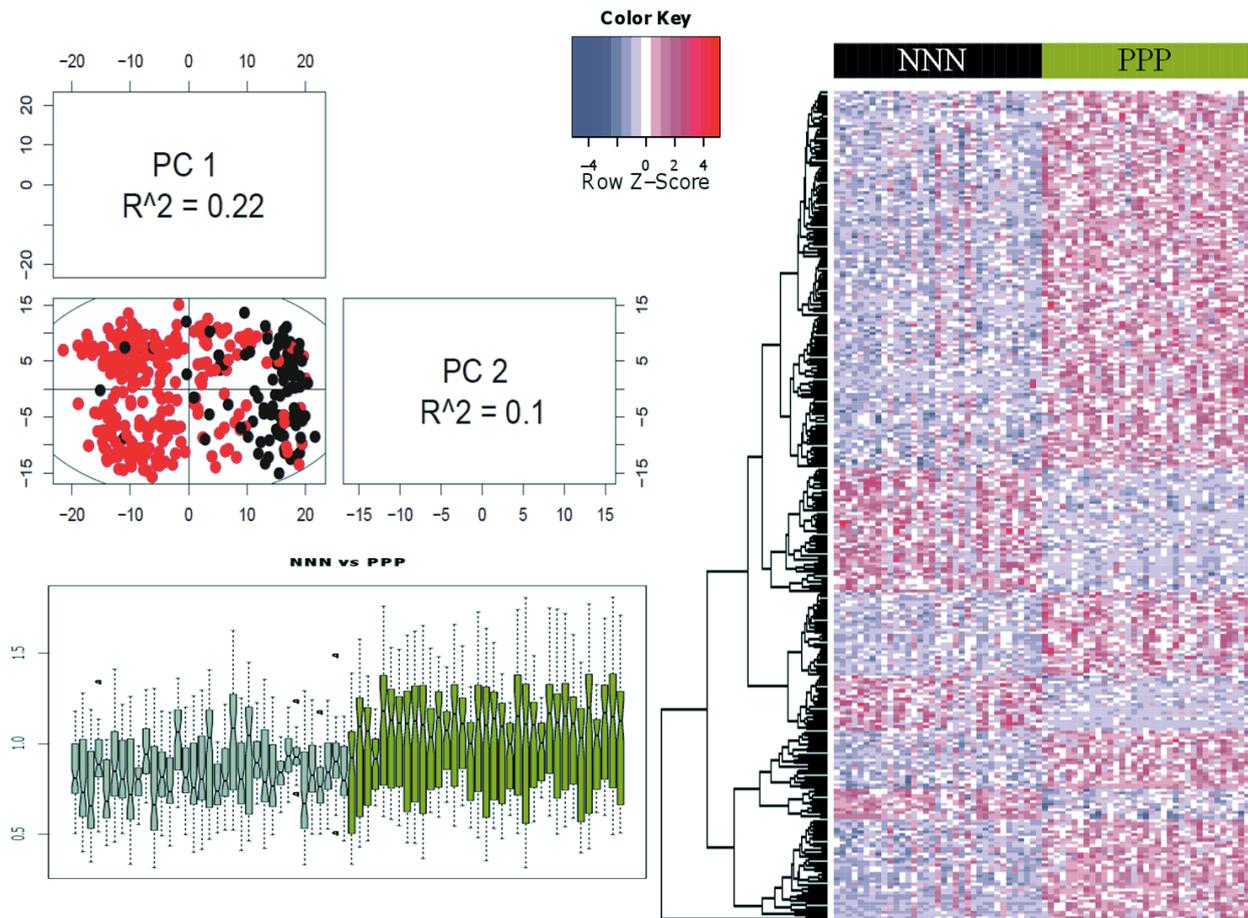
**Abbildung 4.22:** Die Abbildung 4.22 stellt die Resultate der FECSnV Klassifizierung des molekularen Subtyps PNN graphisch dar.

**Expressionsprofil des PNN Subtyps** In Abbildung 4.22 ist das Resultat der Klassifizierung des PNN Subtyps gegen den triple-negativen Fall mit der FECSnV Methode gezeigt. Hier konnten 1945 Gene erhalten werden, die als Klassifizierer gelten. Das Expressionsprofil dieser Gene ist in der Abbildung 4.22 gezeigt. Das linke obere Diagramm stellt die ersten beiden Hauptkomponenten graphisch einander gegenüber, die triple-negativen Fälle sind als rote Kreise dargestellt und das untere Diagramm zeigt Box-Plots der Varianz der berücksichtigten Proben.



**Abbildung 4.23:** Die Abbildung 4.23 stellt die Resultate der FECSnV Klassifizierung des molekularen Subtyps PPN graphisch dar.

**Expressionsprofil des PPN Subtyps** Die Klassifizierung des PPN Subtyps gegen den triple-negativen Fall mit der FECSnV Methode ergab 1119 Gene, die als Klassifizierer gelten. Das Expressionsprofil dieser Gene ist in der Abbildung 4.23 gezeigt. Das linke obere Diagramm stellt die ersten beiden Hauptkomponenten graphisch einander gegenüber, wobei die triple negativen Fälle als rote Kreise dargestellt werden. Das untere Diagramm zeigt Box-Plots der Varianz der berücksichtigten Proben.



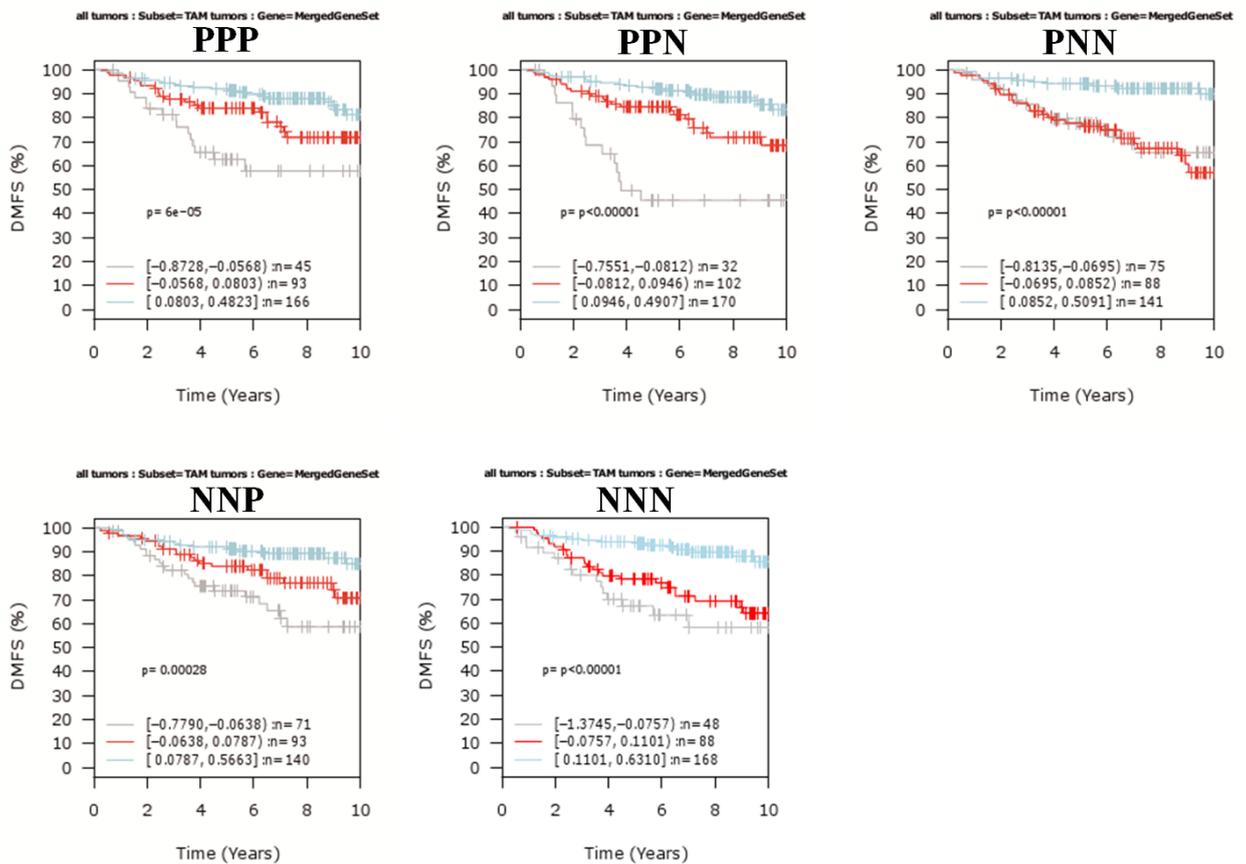
**Abbildung 4.24:** Die Abbildung 4.24 stellt die Resultate der FECSnV Klassifizierung des molekularen Subtyps PPP graphisch dar.

**Expressionsprofil des PPP Subtyps** Die Klassifizierung des PPN Subtyps gegen den triple-negativen Fall mit der FECSnV Methode ergab 1650 Gene, deren Expressionsprofil in der Abbildung 4.24 gezeigt ist. Das linke obere Diagramm stellt die ersten beiden Hauptkomponenten graphisch einander gegenüber, wobei die triple negativen Fälle als rote Kreise dargestellt werden. Das untere Diagramm zeigt Box-Plots der Varianz der berücksichtigten Proben und die Heatmap im rechten Teil zeichnet die Genexpressionswerte der Gene, die als Klassifizierer gelten.

#### 4.2.5.7 Klinische Endpunkt-Analyse mit dem Webservice GOBO

GOBO ( *i.e.* „Gene expression-based Outcome for Breast cancer Online“) ist ein Webservice, der seit 2011 von Ringner *et al.* zur öffentlichen Verfügung bereit gestellt wird

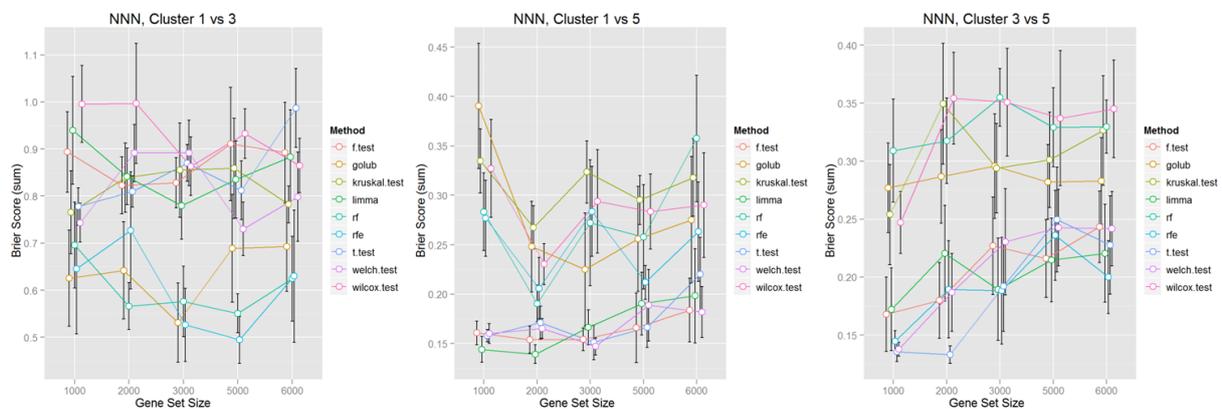
[116]. Dieser Webservice ist spezialisiert auf die Prognose des Krankheitsverlaufes anhand von Genexpressionssignaturen. Die GOBO Datenbank ermöglicht eine präklinische Validierung von Gensignaturen und somit auch eine Charakterisierung der molekularen Subtypen. Die Vorhersagekraft der erhaltenen Signaturen wurde an zwei unabhängigen Studien untersucht, die Genexpression und klinischen Endpunkt (DMFS) Tamoxifen behandelter Patienten bereit stellen (Studie GSE12093 [179] und Studie GSE6532 [180]). Die untersuchten Genexpressionsprofile der molekularen Subtypen, basierend auf den Resultaten der FECSnV Methode führten auf hoch prädiktive, Subtyp spezifische Gensignaturen ( $p$ -Wert  $< 2e^{-4}$ ) für den klinischen Endpunkt DMFS Tamoxifen behandelter Patienten, wie in der Abbildung 4.25 gezeigt.



**Abbildung 4.25:** Die Abbildung 4.25 stellt Kaplan-Meier Diagramme der untersuchten Signaturen für die molekularen Subtypen dar. Hier sind die klinischen Verläufe aller fünf untersuchten molekularen Subtypen im Bezug auf den Endpunkt DMFS gezeigt. Für alle Subtypen bedeutet eine herunter regulierte Expression der untersuchten Gensignaturen eine schlechte Prognose.

### 4.2.5.8 Consensus-Clusteranalyse der triple-negativen Brustkrebs Genexpressionsprofile

Die Klassifizierung der triple-negativen Subtypen wurde mit den aus FECSnV erhaltenen Signaturen über Consensus-Clustering [87] mit dem Bioconductorpaket *ConsensusClusterPlus* [88] betrieben. In Abbildung 3.2 sind die Resultate der Consensus-Clusteranalyse als Folge von Diagrammen zu sehen. Im Panel (a) ist die kumulative Verteilungsfunktion dargestellt (CDF); Die CDF weist keine weitere Flächenzunahme auf für Werte von k über sieben. Das Panel (b) zeigt das delta der Fläche der CDF Kurve. Auch hier gibt es keine nennenswerte Änderung der Fläche unter der CDF für k gleich sieben. Das Panel (c) visualisiert die Consensus-Matrix, diese zeigt die Robustheit der potentiellen Cluster für ansteigende Werte von k. In der graphischen Darstellung der Consensus-Matrix ist, ab einem k größer fünf, das Formieren von drei größeren Clustern zu beobachten. Die relative Änderung in der Deltafläche wird generell genutzt, um ein Maß für den bestmöglichen Wert für k zu erhalten. Hier bedeuten kleinere Änderungen robustere Cluster. Wie in Abbildung 3.2 zu sehen tritt der Fall, dass nur noch marginale Änderungen in der CDF Kurve erreicht werden, ab einem k von sieben auf. Dies bedeutet jedoch, dass die Datenmatrix auf drei große Cluster separiert wird, mit Cluster 1 (n=20), Cluster 3 (n=48) und Cluster 5 (n=24).

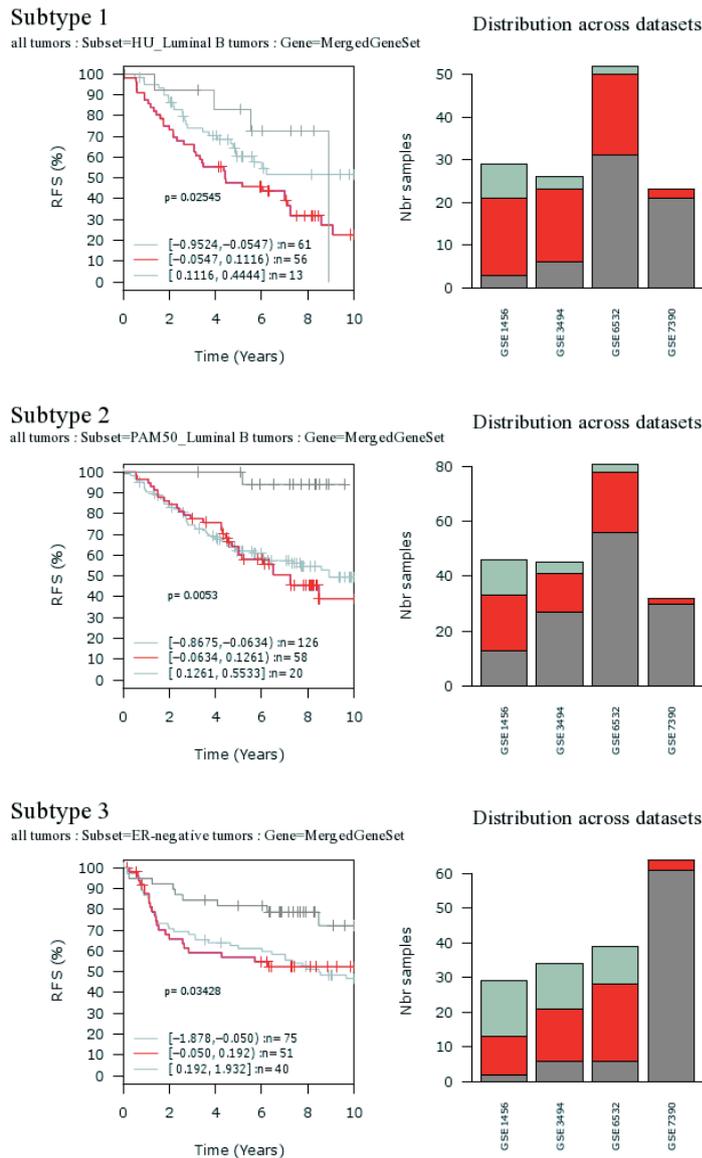


**Abbildung 4.26:** Die Abbildung 4.26 stellt die Resultate der Klassifizierung der Subtypen graphisch dar. Der Mittelwert der Brier-Werte, erhalten aus fünf unabhängigen Runden mit jeweils 20 Proben ist in den drei Abbildungen als Graph aufgetragen. Niedrige Werte zeigen an, dass besonders homogene Klassen untersucht wurden.

**Klassifizierung potentieller triple-negativer Subtypen mittels der FECSnV Methode** Die Consensus-Cluster Methode konnte basierend auf den FECSnV Signaturen die Klassenzugehörigkeiten potentieller TNBC Subtypen erstellen. Diese Zugehörigkeiten unterteilen die 99 initialen TNBC Arrays in drei Gruppen, welche insgesamt 92 Arrays umspannen. Die verbleibenden sieben Arrays wurden über Consensus-Clustering in vier zusätzliche Subtypen arrangiert, jedoch waren diese statistisch unterrepräsentiert und daher von der weiteren Analyse ausgeschlossen. Im nächsten Schritt wurden zunächst Subtyp 1 als Referenz gegen Subtyp 3 und auch Subtyp 5 verglichen und letztlich Subtyp 3 mit Subtyp 5 verglichen. Die Klassifizierungen bilden jeweils fünf Läufe, welche 20 zufällig gewählte Arrays eines Subtyps gegen 20 zufällig gewählte Arrays des Referenz Subtyps beinhalten, sowie alle 22 277 Proben eines jeden Arrays. In Abbildung 4.26 sind die Ergebnisse der Klassifizierung der triple-negativen Brustkrebs Subtypen mit der FECSnV Methode gezeigt. Insgesamt kann ein sehr geringer Brier-Punktwert erzielt werden, besonders die Vergleiche Subtyp 1 gegen Subtyp 5 und Subtyp 3 gegen Subtyp 5, erzielen niedrige Brier-Werte.

#### 4.2.5.9 Klinische Endpunkt Analyse der potentiellen triple-negativen Subtyp Signaturen mit GOBO

Letztlich soll eine Analyse der klinischen Eigenschaften der Signaturen erfolgen, die für die potentiellen TNBC Subtypen charakterisierend sind. Die Kaplan-Meier Diagramme in Panel (a) der Abbildung 4.27 resultieren aus einer Analyse der oben beschriebenen Signaturen mit Hilfe des GOBO Webservice. Die Signaturen welche die TNBC Subtypen charakterisieren, sind prädiktiv für das rezidivfreie Überleben (RFS) von Brustkrebspatienten in vier unabhängigen Studien. Die prädiktiven Eigenschaften der erhaltenen Signaturen sind signifikant für Tumoren des Typs Luminal B und ER negativ ( $p$ -Wert  $< 0,03$ ). In Subtyp 1 und Subtyp 3 finden sich Expressionswerte, die in Luminal B Tumoren eine schlechte Prognose im Fall der Repression dieser Gene bedeutet. Die Subtyp 5 charakterisierenden Signaturen sind prädiktiv für den Ausgang von Brustkrebs Tumoren vom Typ ER negativ, auch hier bedeutet eine reprimierte Signatur einen schlechten Krankheitsverlauf.



**Abbildung 4.27:** Die Abbildung 4.27 stellt die Resultate der klinischen Verläufe in Bezug auf RFS der untersuchten Subtypen in Kaplan-Meier Diagrammen dar. In allen Fällen ist die Expression der charakterisierenden Gensignaturen mit einer schlechten Prognose verknüpft.

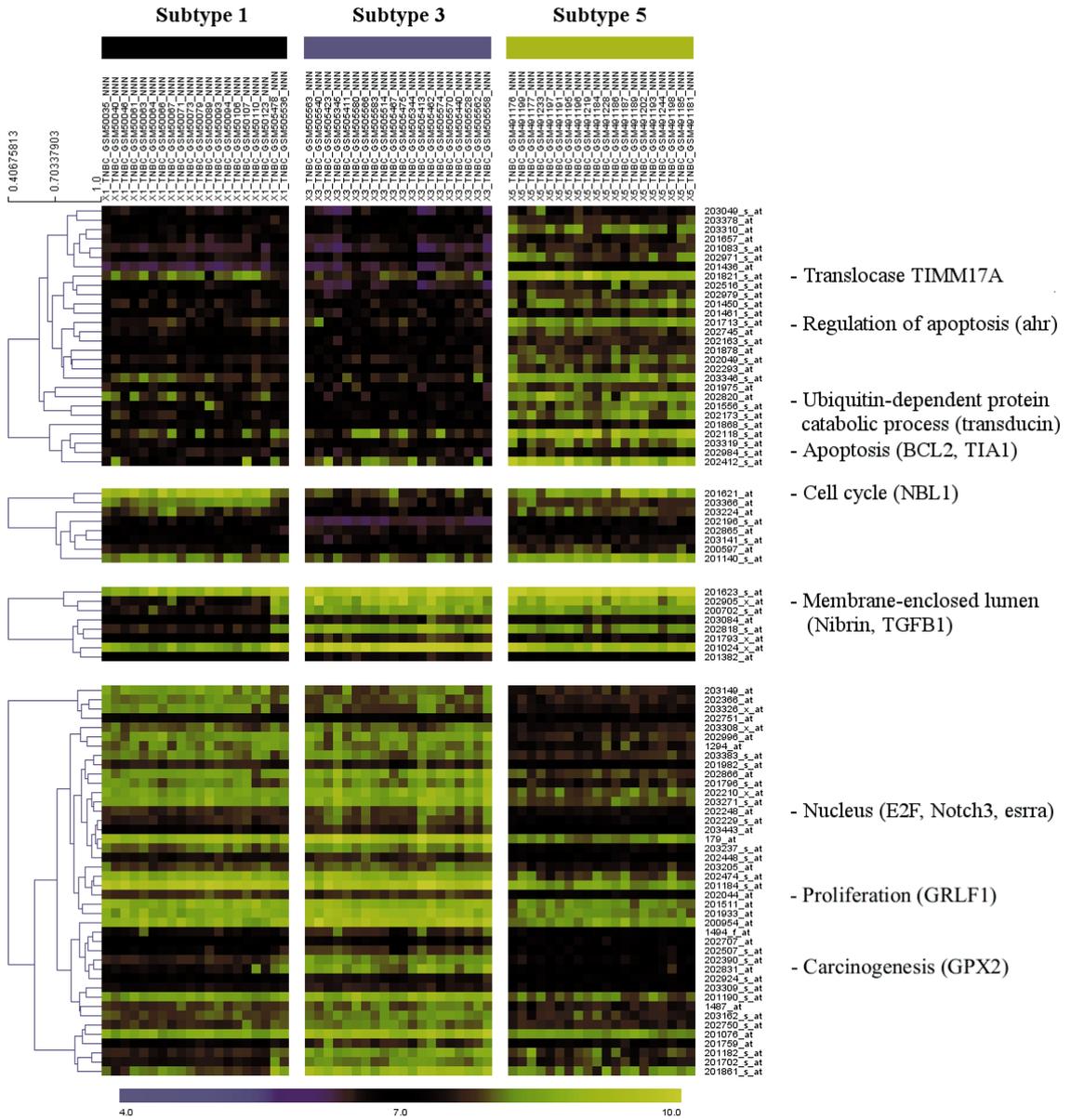
## 4.2.5.10 GSEA der potentiellen triple-negativen Subtypen

**Tabelle 4.12:** Die Tabelle listet alle über GSEA erhaltenen Prozesse für die untersuchten Subtypen auf, die **FDR** ist jeweils  $< 1e^{-4}$ .

<b>Subtyp 1: Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>
Turashvili, Breast lobular carcinoma vs. lobular normal dn	27 (71)	0.88	$1e^{-4}$
Turashvili, Breast lobular carcinoma vs. ductal normal up	24 (67)	0.88	$1e^{-4}$
Kegg, ECM receptor interaction	22 (84)	0.86	$1e^{-4}$
Vecchi, Gastric cancer advanced vs. early up	39 (167)	0.86	$1e^{-4}$
Chiang, Liver cancer subclass CTNNB1 dn	23 (145)	0.86	$1e^{-4}$
Piccaluga, Angioimmunoblastic lymphoma up	65 (207)	0.83	$1e^{-4}$
Yao, Temporal response to progesterone cluster 16	25 (83)	0.83	$1e^{-4}$
Wilcox, Presponse to progesterone dn	23 (66)	0.83	$1e^{-4}$
Proteinaceous, Extracellular matrix	37 (98)	0.82	$1e^{-4}$
Extracellular, Matrix part	25 (57)	0.82	$1e^{-4}$
Landis, ERBB2 breast tumors 324 dn	27 (158)	0.81	$1e^{-4}$
Wu, Cell migration	31 (186)	0.79	$1e^{-4}$
Wang, SMARCE1 targets up	42 (166)	0.75	$1e^{-4}$
Kegg, Focal adhesion	45 (201)	0.72	$1e^{-4}$
<b>Subtyp 3: Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>
Hahtola, Sezary syndrom up	20 (101)	-0.80	$1e^{-4}$
Graham, CML quiescent vs. normal quiescent up	18 (96)	-0.72	$1e^{-4}$
Graham, CML dividing vs. normal quiescent up	20 (187)	-0.69	$1e^{-4}$
Bild, CTNNB1 oncogenic signature	19 (80)	-0.65	$1e^{-4}$
Jaatinen, Hematopoietic stem cell dn	30 (232)	-0.64	$1e^{-4}$
Chandran, Metastasis up	22 (97)	-0.62	$1e^{-4}$
Vecchi, Gastric cancer advanced vs. early dn	15 (139)	-0.62	$1e^{-4}$
Rhein, ALL glucocorticoid therapy up	18 (80)	-0.60	$2e^{-4}$
Charafe, Breast cancer luminal vs. basal up	35 (383)	-0.55	$1e^{-4}$
<b>Subtyp 5: Beschreibung</b>	<b>Gene</b>	<b>ES</b>	<b>p-Wert</b>
Turashvili, Breast lobular carcinoma vs. ductal normal up	28 (67)	-0.80	$1e^{-4}$
Izadpanah, Stem cell adipose vs. bone dn	21 (108)	-0.75	$1e^{-4}$
Chen, HOXA5 targets 9hr up	23 (228)	-0.72	$1e^{-4}$
Sengupta, Nasopharyngeal carcinoma with LMP1 up	35 (399)	-0.71	$1e^{-4}$
Turashvili, Breast lobular carcinoma vs. lobular normal dn	26 (71)	-0.71	$1e^{-4}$
Wang, LMO4 targets dn	23 (349)	-0.70	$1e^{-4}$
Mili, Pseudopodia haptotaxis up	23 (552)	-0.69	$1e^{-4}$
Schuetz, Breast cancer ductal invasive up	98 (355)	-0.68	$1e^{-4}$
Gary, CD5 targets dn	24 (442)	-0.65	$1e^{-4}$
Onder, CDH1 targets 2 up	54 (257)	-0.65	$1e^{-4}$
Kim, WT1 targets dn	36 (471)	-0.63	$1e^{-4}$

### **GSEA der Expressionsprofile der drei potentiellen triple-negativen Subtypen**

Die Tabelle 4.12 listet die GSEA Resultate der potentiellen triple-negativen Subtypen auf, deren Signaturen mit der FECS<sub>n</sub>V Methode erhalten wurden. Die angereicherten Gensignaturen vertreten unterschiedliche Tumorentitäten, jedoch stellte sich heraus, dass diese nur Bruchteile der in MSigDB hinterlegten Gesamtsignaturen sind. Einige Brustkrebs spezifische Prozesse zeichnen sich jedoch ab, die an dieser Stelle zusammengestellt werden: Im ersten potentiellen Subtyp finden sich Signaturen, die im Zusammenhang mit Reaktion auf ERBB2 und Progesteron stehen. Der zweite Subtyp enthält Prozesse, die für Stammzellen sowie für metastasierende Zellen charakteristisch sind. Im dritten Subtyp sind Signaturen zu finden, die auf die Aktivität verschiedener Zielgene hinweisen, diese sind: HOXA5, LMO4, CD5, SMARCA2 und WT1.



**Abbildung 4.28:** Die Abbildung 4.28 stellt die Resultate der FECSnV Analyse mit Dendrogrammen dar. Es sind vier spezifische Cluster-Segmente dargestellt, die unterschiedlichen Prozessen entsprechen. Unterhalb der Heatmap sind die Ergebnisse der Hauptkomponentenanalysen für die einzelnen Subtypen graphisch dargestellt.

Die Genexpressionsprofile der potentiellen TNBC Subtypen wurden über GSEA nach potentiellen Unterschieden hinsichtlich biologischer Merkmale untersucht. In Tabelle 4.12 sind alle erhaltenen Gensignaturen und Prozesse zusammengefasst, die Resultat dieser Analyse waren. Da diese Resultate jedoch nur Bruchteile der vollständigen in MSigDB hinterlegten Signaturen sind, lässt auch deren Aussage, besonders im Hinblick auf Generalisierbarkeit zu wünschen übrig. Daher wurden die Subtyp spezifischen Signaturen mit dem funktionellen Annotations Clustering der DAVID Datenbank analysiert [25].

Wie in Abbildung 4.28 gezeigt, gibt es einige regulierte, Brustkrebs relevante Prozesse. Die abgebildete Heatmap stellt die Expressionswerte der Subtypen dar und ist in vier Segmente untergliedert. Das erste Segment beinhaltet Gene, die ausschließlich in Subtyp 5 induziert sind. Dies sind das mitochondriale Translokase Gen TIMM17A, dessen Protein die mtDNA Stabilität gewährleistet und der DNA Schaden signalisierende Arylhydrogen Rezeptor (AhR), dessen Protein auf die Regulation der Apoptose einwirkt. Des Weiteren ist das Oncogenese und Ubiquitinierung assoziierte Transducin beta-like 1 X-linked Gen (TBL1X) aktiviert und TIA1, ein Apoptose fördernden Faktor, ebenso exprimiert. Das zweite Segment in der Abbildung 3 4.28 in Panel (b) beinhaltet das potentielle Tumor-Supressor und Zellzyklus regulierende Gen NBL1, welches in Subtyp 1 und Subtyp 5 induziert ist. Das dritte Segment enthält das Brustkrebs prädisponierende Gen Nibrin (NBN), dieses ist in Subtyp 3 und 5 induziert. Das letzte Segment enthält drei Gene, deren Proteinprodukte im Nukleus lokalisieren, unter diesen ist der Transkriptionsfaktor E2F4, der Proliferation assoziierte Glucocorticoid Rezeptor DNA Binfaktor (GRLF1) und das Karzinogenese verwandte GPX2 Gen. Diese Gene sind induziert in den Subtypen 1 und Subtypen 3. Das Ergebnis einer Hauptkomponentenanalyse ist in Abbildung 4.28 in Panel (c) gezeigt. Die ersten drei Hauptkomponenten weisen spezifische Cluster auf, welche die potentiellen Subtypen repräsentieren, so dass impliziert wird, dass diese ebenso bestimmte Entitäten darstellen.

### **4.2.6 Abschließende Diskussion der Tumor Subtypisierungsstudie**

Die Genexpressionsprofile von Brustkrebs Patienten eignen sich für eine Prognose des klinischen Ausgangs der Krankheit. Eine molekulare Stratifikation der Tumor und Tumor Subtypen bleibt jedoch ausstehend. Im Allgemeinen werden maschinelle Lernverfahren und statistische Klassendetektionalgorithmen angewandt, um phänotypspezifische Gensignaturen zu erhalten. In vielen Fällen und unabhängig vom eingesetzten

Algorithmus, beinhalten solche Analysemethoden eine Routine für die Parameter Anpassung, in dem der Algorithmus weitestgehend auf die Daten abgestimmt wird. Objektiv jedoch ist eine Vorgehensweise, in der viele verschiedene Algorithmen, ohne eine Anpassung der Parameter genutzt werden. Um diesen Ansatz zu verfolgen, wurden in dieser Arbeit die sogenannten Voting Methoden aus der Literatur [84] neu berücksichtigt und eine Methode entwickelt, welche die essentiellen Schritte einer Klassifikation, über Extraktion, Bewertung und direkte Wahl vereint. Die neue Methode extrahiert Gensignaturen, die spezifisch für molekulare Subtypen sind, wie anhand von Signaturen aus der MSigDB festgestellt werden kann. Zusätzlich gelingt eine Stratifikation über den Webservice GOBO, da die über FECSnV erhaltenen Signaturen spezifisch sind für die 10 Jahreswahrscheinlichkeit des fernmetastasenfreien Überlebens (DMFS) von Tamoxifen behandelten Brustkrebspatienten. Diese Resultate rechtfertigen die Methode und führten dazu, eine Subtypisierung von triple-negativen Brustkrebs Patienten anhand der Expressionsdaten vorzunehmen. Gegenwärtig kann mit der Methode jedoch nicht direkt eine Bestimmung der Klassenzugehörigkeiten vorgenommen werden. Daher wurde mit Hilfe von Consensus-Clustering eine Klassenzuordnung prognostiziert, unter Verwendung der TNBC spezifischen Signaturen, die zuvor mit FECSnV erhaltenen wurden. Eine Evaluation der 99 TNBC Genexpressionsprofile ergab jedoch nur drei statistisch repräsentative Subtypen. In der Literatur haben Lehmann *et al.* [67] jedoch nach Analyse von 587 TNBC Expressionsprofilen sechs Subtypen vorgeschlagen, welche wie folgt untergliedert sind: zwei Basal-artige (BL1, BL2), ein Immunomodulatorischer (IM), ein Mesenchymaler (M), ein Mesenchymal Stammzell-artiger (MSL) und einer vom Typ Luminal Androgen Rezeptor (LAR). Um diesem Sachverhalt Rechnung zu tragen, sollte die FECSnV Methode anhand eines umfangreicheren Genexpressionsdatensets verwendet werden. Dann sind mit hoher Wahrscheinlichkeit auch die übrigen Subtypen repräsentativ vertreten.

**Biologische Eigenschaften der Subtyp charakterisierenden Signaturen** Eine Analyse der biologischen Eigenschaften brachte die Erkenntnis, dass in Subtyp 1 und Subtyp 3 Gene reprimiert sind, welche Apoptoseregulation zur Aufgabe haben. Zusätzlich induzieren diese Subtypen Gene, die Proliferation und Karzinogenese bedeuten. Beide Subtypen exprimieren das E2F4 Gen, ein Mitglied der E2F Genfamilie, welche mit Tamoxifen Resistenz assoziiert ist. Subtyp 1 und Subtyp 3 unterscheiden sich jedoch im Aktivitätszustand der Expression von Signaturen, die Zellzyklus Kontrolle und DNA Instandhaltung bewirken. Diese Subtypen weisen Gemeinsamkeiten mit den Basal-artigen BL1 und BL2 auf, beschrieben in [67]. Im Gegensatz dazu weist Sub-

typ 5 andere molekulare Charakteristika auf, da alle bisher diskutierten Gene induziert werden, jedoch mit Ausnahme des Proliferations Clusters. Stattdessen wird eine Signatur induziert deren Folgen Oncogenese und Ubiquitinierung sind, die auch Expression des TBL1X Gens beinhalten [181]. Li *et al.* zeigten das TBL1X essentiell für die Rekrutierung von  $\beta$ -Catenin und Aktivierung des Wnt Signalweges sind. Daher könnte dieser Subtyp Ähnlichkeit zum Mesenchymalen Subtyp in [67] haben. Zusätzlich reguliert Subtyp 5 die Aktivität des E2F4 Gens herunter, welches mit dem BRCA1 Gen interagiert [182]. Die Nukleotid Sequenz der E2F Genfamilie ist hoch konserviert im Promoter Bereich des BRAC1 Gens [183]. Daher sollte evaluiert werden, ob es eine Gemeinsamkeit bei Brustkrebspatienten gibt, die eine geringe Expression des BRCA1 Gens aufweisen, da diese Patienten besonders gutes Ansprechverhalten auf platinbasierte Chemotherapie aufzeigen [184].

**Ausblick** Insgesamt konnte in dieser Arbeit gezeigt werden, dass die Stratifikation von Genexpressionsprofilen neben klinischen Aspekten auch neue Tumor Subtypen aufdecken kann. Die resultierenden Gensignaturen konnten zum Teil sogar eine Erklärung für das Scheitern konventioneller Therapie, aber auch neue Möglichkeiten und Strategien in der Krebstherapie aufzeigen.

**Fazit** Die Identifikation und Charakterisierung von Tumor Subtypen auf Basis von Genexpressionsprofilen bietet neue Aspekte, aus denen sich sogar therapeutische Möglichkeiten für triple-negativen Brustkrebs ableiten können. Die Analyse und Klassifizierung von Microarraydaten über molekulare Subtypen erfolgte mit dem in der Arbeit vorgestellten FECSnV Algorithmus [82]. Die biologische Charakterisierung und die klinische Stratifikation wurden über GSEA, funktionelles Annotations Clustering und die Endpunktanalyse mit GOBO durchgeführt. Daraus ergaben sich zunächst Subtyp spezifische Signaturen, welche prädiktiv für das Fernmetastasenfreie Überleben von Tamoxifen behandelten Brustkrebspatienten sind. Über Consensus-Clustering und unter Verwendung der FECSnV Methode konnten drei triple-negative Subtypen identifiziert werden. Während zwei dieser Subtypen Gemeinsamkeiten mit den Basal-artigen BL1 und BL2 Subtypen aufwiesen, zeigte der dritte Ähnlichkeit zum Mesenchymalen Subtyp in [67]. Im Gegensatz zu den Basal-artigen, ist die Expression des E2F4 Gens im Mesenchymalen Subtyp reprimiert und der Wnt Signalweg über  $\beta$ -Catenin aktiviert. Die Klassifizierung der Genexpressionsprofile von Brustkrebspatienten führt auf neue Tumor Subtypen, welche eine klinische Stratifizierung ermöglichen. Zudem beinhaltet eine Analyse der Subtypen charakterisierenden Signaturen nicht nur biologische, tumor-

spezifische Aspekte, sondern auch Möglichkeiten für die Entwicklung neuer Strategien in der patientenbasierten Therapie.



# 5 Zusammenfassung

**Softwareentwicklung** Die Analyse von Genexpressionsprofilen fällt nicht länger exklusiv in den Aufgabenbereich von Bioinformatik Experten. Jedoch gestaltet sich der Erhalt von statistisch signifikanten Resultaten schwierig und erfordert neben biologischem Hintergrundwissen auch weitreichende rechentechnische Fertigkeiten. In dieser Arbeit wird ein neues, anwenderfreundliches Microarray Reportingwerkzeug namens maRt vorgestellt [77]. Die Software bietet Zugang zu bioinformatischen Ressourcen, wie der Gen Ontologie und aktuellen biologischen Stoffwechselwegen durch den Zugriff auf die Datenbanken DAVID und BioMart. Die Ergebnisse werden in strukturierten HTML Reports zusammengefasst, wobei verschiedene Schichten an Informationen preisgegeben werden. In diesen Reports sind Inhalte aus diversen internetbasierenden Quellen integriert und verlinkt. Die Software zieht Nutzen aus der ubiquitären „multi-core“ Technologie moderner Arbeitsplatzrechner über parallele Datenverarbeitung. Aufgrund der RCP basierten internen Infrastruktur von maRt bietet die Software eine günstige Basis für die Entwicklung neuer R basierter Anwendungen. Verfügbarkeit: Routinen für die Installation unter diversen Betriebssystemen, Dokumentation und verschiedene Tutorien sind LGPL Lizenz basiert auf den Webseiten des Pharmazeutischen Instituts zu finden unter: <http://www.pharma.uni-bonn.de/www/mart>. Die Software ist für akademische Zwecke frei zugänglich.

**Qualitätssicherung von Microarray Daten** Die Qualitätssicherung und die Verfahren zur Normalisierung von Microarray Daten nehmen einen hohen Stellenwert im Verlauf der Analyse von Genexpressionsprofilen ein. Gegenwärtig sind verschiedene Methoden verfügbar, die Qualitätserfassung von Microarray Datensets anbieten. Jedoch scheint keine Standardvisualisierung vorhanden zu sein, mit deren Hilfe auch individuelle Qualitätsparameter von Microarrays abgebildet werden können. In dieser Arbeit wird eine bequeme Methode präsentiert, die das Visualisieren von Standard Qualitätsparametern mit Circos ermöglichen [79]. Die Darstellung der verschiedenen Parameter und potentiellen Ausreißer ist zirkulär angeordnet, um so jedes einzelne Array eines Datensets visuell zu erfassen. Die vorgestellte Methode erbringt vielversprechende Resultate für die Mehrheit der öffentlich verfügbaren Datensets, welche häufig auf dem von Affymetrix lizenzierten Human Genome (GPL 96, GPL570 und GPL571) Formaten basieren. Zukünftig könnten die vorgeschlagenen Circos basierten Qualitätsabbilder als allgemeiner Standard in Datenbanken, die Genexpressionsdaten bereitstellen übernommen werden, um die rasche Erstbegutachtung der Microarraydatensets zu gewährleisten.

**Wirkstoffstudie** Kürzlich erst wurde in einer Studie berichtet, dass liposomales Cisplatin die Resistenz von ovariellen A2780cis Zellen überwindet [45]. In dieser Arbeit wurde gefunden, dass die zytotoxische Aktivität von liposomalem Cisplatin nicht an die DNA Platinierung dieser Zellen geknüpft ist. Diese Tatsache suggeriert, dass der Wirkmechanismus von liposomalem Cisplatin grundsätzlich verschieden von freiem Cisplatin ist. Um einen Einblick in die zugrundeliegenden Mechanismen zu erhalten, müssen Resistenz assoziierte Gensignaturen auf der Ebene des Transkriptom von A2780 und A2780cis Zellen nach Behandlung mit  $IC_{50}$  Dosen an freiem und liposomalem Cisplatin analysiert werden. Eine Analyse der Prozessnetzwerke der deregulierten Gene bestätigte, dass liposomales Cisplatin ein signifikant unterschiedliches Transkriptionsmuster aufweist im Vergleich zu freiem Cisplatin. Der Tumorsuppressor p53 ist ein Schlüsselgen in der differentiellen Transkription als Antwort auf freies bzw. liposomales Cisplatin. Freies Cisplatin induziert über die Aktivierung von p38 MAPK eine Vielzahl an Genen, die als Schlüsselgene im intrinsischen (Mitochondrialen) Apoptose Pathway agieren, z.B. BAX, BID, CASP9. Liposomales Cisplatin induziert jedoch Gene, die im extrinsischen Apoptoseweg partizipieren (TNFRSF10B–DR5, ,CD70-TNFSF7). Dieser Mechanismus ist entscheidend für die Fähigkeit liposomalen Cisplatins, um eine Überwindung der Cisplatinresistenz von A2780cis Zellen zu erzielen. Nach dem Zelleintritt beeinflusst der liposomale Wirkstoff die Aktivität von Cisplatin an den subzellulären Komponenten im Gegensatz zu freiem Cisplatin auf gravierende Weise und verursacht daher eine komplexe Signalwirkung die in ein apoptotisches Szenario mündet. Dies wirft neues Licht auf die Ansätze, welche Liposomen als Wirkstoff-Transporter in der Krebstherapie einsetzen und suggeriert, dass liposomales Cisplatin eine vielversprechende Strategie in für die Behandlung von Cisplatinresistenten Ovarialkarzinomen.

**Tumor Progressionsstudie** Gegenwärtig sind keine allgemein akzeptierten Biomarker für die Charakterisierung von Zervixkarzinomen verfügbar. In dieser Arbeit wurde daher die Varianz-Komponenten Analyse [26] für die Detektion von Biomarkern eingesetzt. Die Methode wurde in dieser Arbeit auf Microarray Genexpressionsdaten aus vier öffentlich verfügbaren Zervixkarzinom Studien angewandt (n=126) [80]. Die wegbereitende Studie von Bachtary *et al.* inspirierte zu der Annahme, dass sowohl intra Tumoral als inter Stadien-Heterogenität die Zuverlässigkeit onkologischer Marker maßgeblich beeinflusst. Zielführend in der vorliegenden Arbeit war die Entwicklung sowohl statistisch als auch klinisch relevanter Genexpressionsprofile von zervikalen Tumoren. Zusätzlich zur Varianz-Komponenten Analyse soll die so genannte „Gene Set Enrichment Analysis“ (GSEA) [81] für die weiterführende Untersuchung der erhaltenen Genexpressionspro-

file verwendet werden. Insgesamt wurden 22.277 Gene mit der Varianz-Komponenten Analyse untersucht, elf Gene zeichnet eine besonders niedrige Variabilität aus, d.h. ihre intra Stadien Varianz lag im Verhältnis zur Gesamtvarianz (W/T) zwischen 0,18 und 0,38. Sieben dieser Proben waren in den verschiedenen Zervixkarzinom Stadien induziert, diese sind GINS1, PAK2, DTL, AURKA, PRKDC, NEK2 und CEP55. Die übrigen vier Gene sind nur in normalem Zervix Gewebe exprimiert, diese sind P11, EMP1, UPK1A und HSPC159. Eine GSEA der 9.873 Gene, die ein W/T Verhältnis von unter 0,75 aufwiesen, ergab signifikant angereicherte Genexpressionssignaturen, welche auf eine Behandlung durch angiocidin und darapladib hinwiesen. Zusätzlich konnten immunologisch relevante Gensignaturen gefunden werden, die gravierende Prozesse beschreiben, wie z.B. die Graft versus Host oder die akute Nierentransplantat Abstossungsreaktion. Des Weiteren konnten Gensignaturen in Stadium IB2 gefunden werden, welche auf MT1-MMP abhängige Migration und Invasivität hinweisen. Diese Gensignaturen sind in Begleitung einer Gen Expressionssignatur, die ECM Rezeptor vermittelte Interaktionen bedeuten. Schlussfolgernd bedeutet die Analyse von Zervixkarzinom Gen Expressionsprofilen eine neue Perspektive auf HPV vermittelte Transkriptionsprozesse. Diese neue Aussicht birgt ein tiefgehendes Verständnis der karzinogenen Konsequenzen und kann sogar eine Verbesserung der therapeutischen Möglichkeiten bieten.

**Tumor Subtypisierungsstudie** Ein weiterer Schwerpunkt dieser Arbeit war eine großangelegte Subtypisierungsanalyse von Microarray Genexpressionsdaten aus vier öffentlich verfügbaren Brustkrebs Studien (n=514) [82]. Zu diesem Zweck wurden sogenannte „Voting“-Methoden aus der Literatur [83–85] berücksichtigt und ein neuer Algorithmus entwickelt, welcher die Extraktion von Genen, deren Bewertung und letztlich die Wahl eines Klassifiziers zur Aufgabe hat. Der Algorithmus basiert auf maschinellen Lernverfahren und statistischen Klassifikationsalgorithmen, die dem Bioconductorpaket CMA entnommen wurden [86]. Das sogenannte „multi-class“ Problem wurde über Consensus-Clustering [87, 88] gelöst. Diese Methode zeigte beachtlichen Erfolg in Wilkerson *et al.* [89]. Die erhaltenen Gensignaturen wurden mit Hilfe der „Gene Set Enrichment Analysis“ (GSEA) auf ihren biologischen Wert untersucht [81], die klinischen Eigenschaften der Signaturen konnten mit „gene expression-based outcome for breast cancer online“, (GOBO) analysiert werden. Durch die Klassifizierung der molekularen Subtypen konnten spezifische Gensignaturen erhalten werden, die biologische und klinische Relevanz aufweisen. Die Subtyp spezifischen Signaturen sind hoch prädiktiv für DMFS von Tamoxifen behandelten Brustkrebspatienten. Zudem konnte über Consensus-Clustering und der vorgestellten Klassifizierungsmethode eine Charakterisierung von triple nega-

tiven Subtypen erzielt werden, in der drei potentielle Fälle nachgewiesen wurden. Ein potentieller Subtyp wies eine geringe E2F4 Expression auf und die charakterisierende Signatur war prädiktiv für RFS von Östrogen negativen Brustkrebspatienten. Die Gen Expressionssignaturen der übrigen Subtypen haben Ähnlichkeiten mit luminal B Tumoren. Die Klassifikation der Expressionsprofile von Brustkrebspatienten enthüllt potentielle, neue Tumorsubtypen, die klinische Auswirkungen beinhalten. Des Weiteren bietet ein Verstehen der komplexen und abberanten Biologie von Brustkrebs zusätzliches Potential für neue Strategien in der klinischen Therapie.



## 6 Summary

**Software Development** Abstract: Analysis of gene expression profiles is no longer exclusively a task for bioinformatic experts. However, gaining statistically significant results is challenging and requires both biological knowledge and computational know-how. Here we present a novel, user-friendly microarray reporting tool called maRt [77]. The software provides access to bioinformatic resources, like gene ontology terms and biological pathways by use of the DAVID and the Bio- Mart web-service. Results are summarized in structured HTML reports, each presenting a different layer of information. In these report, contents of diverse sources are integrated and interlinked. To speed up processing, maRt takes advantage of the multi-core technology of modern desktop computers by using parallel processing. Since the software is built upon a RCP infrastructure it might be an outset for developers aiming to integrate novel R based applications. Availability: Installer, documentation and various kinds of tutorials are available under LGPL license at the website of our institute <http://www.pharma.uni-bonn.de/www/mart>. This software is free for academic use.

**Assessing microarray data quality** Quality control and normalization is considered the most important step in the analysis of microarray data. At present there are various methods available for quality assessments of microarray datasets. However there seems to be no standard visualization routine, which is depicting also individual microarray quality. Here we present a convenient method for visualizing the results of standard quality control tests using Circos plots [79]. In these plots various quality measurements are drawn in a circular fashion, thus allowing for visualization of the quality and all outliers of each distinct array within a microarray dataset. The proposed method shows promising results for the majority of publicly available datasets, which are frequently based on the Affymetrix Human Genome U133A (GPL 96) platform. In future the proposed Circos quality measurement plots might be adopted as common standard in databases providing microarray raw data, for quick initial and visual assessment of the whole dataset.

**Drug Delivery Study** Previously we reported that liposomal cisplatin (CDDP) overcomes CDDP resistance of ovarian A2780cis cancer cells [45]. Here we find that the cytotoxic activity of liposomal CDDP is not associated with DNA platination in these cells. This suggests that the mode of action of liposomal CDDP is different from the free drug. To gain insight into the resistance gene signature and underlying mechanisms of liposomal activity, we performed a transcriptome analysis of untreated A2780cis cells, and A2780cis cells in response to the exposure with  $IC_{50}$  values of free vs. liposomal CDDP. A process network analysis of the deregulated genes confirmed that liposomal CDDP induces a significantly different transcriptional activation pattern compared to free CDDP. p53 was identified as a key point directing the different transcriptional response to free vs. liposomal CDDP. Free CDDP induces numerous genes of potential key players in the intrinsic (mitochondrial) apoptosis pathway, e.g. BAX, BID, CASP9 via p38 MAPK activation. This is not evident in response to liposomal CDDP, which induces genes of the extrinsic pathway of apoptosis (TNFRSF10B – DR5, CD70-TNFSF7). This appears as crucial mechanism of CDDP liposomes to overcome the resistance in A2780cis cells. The liposomal agent seriously influences after cell entry the targeting of CDDP to subcellular components vs. free CDDP and thus affects a complex signaling, which ends in an apoptotic scenario. This sheds a new light on liposomal drug carrier approaches in cancer and suggests liposomal CDDP as promising strategy for the treatment of CDDP resistant ovarian carcinomas.

**Tumor Progression Study** Purpose: To assign functional properties to gene expression profiles of cervical cancer stages and identify clinically relevant biomarker genes. Experimental Design: Microarray samples of 24 normal and 102 cervical cancer patients from four publicly available studies were pooled and evaluated (n=126) [80]. High quality microarrays were normalized using the CONOR package from the Bioconductor project. Gene expression profiling was performed using variance-component analysis for accessing most reliable probes, which were subsequently processed by Gene Set Enrichment Analysis. Results: Of 22.277 probes that were subject to variance-component analysis eleven probes had low heterogeneity, i.e. a W/T ratio between 0.18 and 0.38. Seven of these probes are induced in all cervical cancer stages these are GINS1, PAK2, DTL, AURKA, PRKDC, NEK2 and CEP55. The other four probes are induced in normal cervix P11, EMP1, UPK1A and HSPC159. We performed GSEA of 9.873 probes which had a W/T ratio of  $< 0.75$ . Repeatedly, significant gene expression signatures were found that are related to treatment using angiocidin and darapladib. Additionally expression signatures from immunological disease signatures were found e.g. graft

versus host disease and acute kidney rejection. Another finding comprises a gene expression signature in stage IB2 that refers to MT1-MMP dependent migration and invasion. This gene signature is accompanied by gene expression signatures which refer to ECM receptor mediated interactions. Conclusion: Analysis of cervical cancer patient gene expression data reveals a novel perspective on HPV mediated transcription processes. This novel point of view contains a better understanding and even might provide improvements to cancer therapy.

**Tumor Subtyping Study** Purpose: Identification, characterization and validation of tumor subtypes using gene expression patterns from triple negative breast cancer (TNBC) patients [82]. Experimental Design: Microarray data from breast cancer (BC) patients of four publicly available studies (n = 541) were pooled and evaluated. High quality microarray data were normalized using RMA and CONOR packages from the Bioconductor project. Molecular subtype classification was performed using random forest and a novel classification algorithm, which allows for feature extraction via composite scoring and voting (FECSnV). Moreover potential triple negative subtypes were calculated by consensus clustering and subtype specific gene signatures were obtained by the proposed FECSnV method. Subsequently we evaluated the biological and clinical properties of the derived signatures, i.e. via gene set enrichment analysis (GSEA), functional annotation clustering using DAVID and gene expression-based outcome for breast cancer online (GOBO). Results: Classification of receptor variant molecular subtypes yields specific gene signatures, comprising biological and clinical relevance. The subtype specific signatures are highly predictive for DMFS of Tamoxifen treated BC patients. Moreover, using consensus clustering and the proposed classification algorithm, a characterization of triple negative subtypes yielded three distinct TNBC subtypes. One subtype exhibits low E2F expression and its characterizing gene expression signature is predictive for RFS of ER negative BC patients. The gene expression signatures of the additional subtypes are sharing commonalities with luminal B tumors. Conclusion: Classification of BC patient gene expression profiles may reveal potential novel tumor subtypes, which comprise clinical impact. Furthermore a comprehension of the complex and aberrant biology of cancer might additionally hold potential for novel strategies in cancer therapy.



# 7 Appendix

## 7.1 Genexpressions Studien aus “Gene Expression Omnibus” (GEO)

Die folgenden Studien fanden in der vorliegenden Arbeit Verwendung:

**Tabelle 7.1:** Die Tabelle fasst die untersuchten Studien aus GEO zusammen.

GEO ID	Anzahl	Referenz	Datum
GSE9801	6	[131]	Nov 07, 2011
GSE32700	46	[132]	Okt 07, 2011
GSE9936	105	[133]	Feb 07, 2008

## 7.2 Wirkstoff Studie

**Analyse der p53 induzierten Gene in A2780cis Zellen** Eine Analyse der durch den Transkriptionsfaktor p53 induzierten Gene führte auf zwei im Appendix gelisteten Tabellen, welche die Gene auflisten, die durch die verschiedene CDDP Behandlung und durch den Transkriptionsfaktor p53 induziert werden.

Gen Symbol	RefSeq ID	Gen Beschreibung
BAI1	<a href="#">NM_001702</a>	brain-specific angiogenesis inhibitor 1 (BAI1)
BAX	<a href="#">NM_138764</a>	BCL2-associated X protein (BAX), transcript variant epsilon
BID	<a href="#">NM_197966</a>	BH3 interacting domain death agonist (BID), transcript variant 1
BTG3	<a href="#">NM_006806</a>	BTG family, member 3 (BTG3)
CARD10	<a href="#">NM_014550</a>	caspase recruitment domain family, member 10 (CARD10)

CDK7	<a href="#">NM_001799</a>	cyclin-dependent kinase 7 (MO15 homolog, <i>Xenopus laevis</i> , cdk-activating kinase) (CDK7)
COL18A1	<a href="#">NM_030582</a>	collagen, type XVIII, alpha 1 (COL18A1), transcript variant 1
DUSP2	<a href="#">NM_004418</a>	dual specificity phosphatase 2 (DUSP2)
EGR1	<a href="#">NM_001964</a>	early growth response 1 (EGR1)
EGR2	<a href="#">NM_000399</a>	early growth response 2 (Krox-20 homolog, <i>Drosophila</i> ) (EGR2)
EI24	<a href="#">NM_004879</a>	etoposide induced 2.4 mRNA (EI24), transcript variant 1
ELK3	<a href="#">NM_005230</a>	ELK3, ETS-domain protein (SRF accessory protein 2) (ELK3)
FBXO11	<a href="#">NM_025133</a>	F-box protein 11 (FBXO11), transcript variant 1
FBXO11	<a href="#">NM_012167</a>	F-box protein 11 (FBXO11), transcript variant 3
FGFR3	<a href="#">NM_000142</a>	fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism) (FGFR3), transcript variant 1
FHL2	<a href="#">NM_201555</a>	four and a half LIM domains 2 (FHL2), transcript variant 2
G3BP2	<a href="#">NM_203505</a>	Ras-GTPase activating protein SH3 domain-binding protein 2 (G3BP2), transcript variant 1
GP1BB	<a href="#">NM_000407</a>	glycoprotein Ib (platelet), beta polypeptide (GP1BB)
GPX1	<a href="#">NM_201397</a>	glutathione peroxidase 1 (GPX1), transcript variant 2
GSN	<a href="#">NM_198252</a>	gelsolin (amyloidosis, Finnish type) (GSN), transcript variant 2
HDAC1	<a href="#">NM_004964</a>	histone deacetylase 1 (HDAC1)
HSD17B1	<a href="#">NM_000413</a>	hydroxysteroid (17-beta) dehydrogenase 1 (HSD17B1)
HSP90AA1	<a href="#">NM_005348</a>	heat shock protein 90kDa alpha (cytosolic), class A member 1 (HSP90AA1), transcript variant 2
HSP90AB1	<a href="#">NM_007355</a>	heat shock protein 90kDa alpha (cytosolic), class B member 1 (HSP90AB1)

HSPA8	<a href="#">NM_153201</a>	heat shock 70kDa protein 8 (HSPA8), transcript variant 2
HSPB1	<a href="#">NM_001540</a>	heat shock 27kDa protein 1 (HSPB1)
ID3	<a href="#">NM_022743</a>	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein (ID3)
ID4	<a href="#">NM_001546</a>	inhibitor of DNA binding 4, dominant negative helix-loop-helix protein (ID4)
IER3	<a href="#">NM_003897</a>	immediate early response 3 (IER3), transcript variant short
IGFBP1	<a href="#">NM_000596</a>	insulin-like growth factor binding protein 1 (IGFBP1), transcript variant 1
JAG2	<a href="#">NM_002226</a>	jagged 2 (JAG2), transcript variant 1
JUND	<a href="#">X56681</a>	Human junD mRNA. [X56681]
LYN	<a href="#">NM_002350</a>	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog (LYN)
MAD2L1	<a href="#">NM_002358</a>	MAD2 mitotic arrest deficient-like 1 (yeast) (MAD2L1)
MAPK11	<a href="#">NM_002751</a>	mitogen-activated protein kinase 11 (MAPK11), transcript variant 1
MICA	<a href="#">NM_000247</a>	MHC class I polypeptide-related sequence A (MICA)
MT2A	<a href="#">NM_005953</a>	metallothionein 2A (MT2A)
NMU	<a href="#">NM_006681</a>	neuromedin U (NMU)
NR4A1	<a href="#">NM_002135</a>	nuclear receptor subfamily 4, group A, member 1 (NR4A1)
PARP1	<a href="#">NM_001618</a>	poly (ADP-ribose) polymerase family, member 1 (PARP1)
PBK	<a href="#">NM_018492</a>	PDZ binding kinase (PBK)
PCNA	<a href="#">NM_002592</a>	proliferating cell nuclear antigen (PCNA), transcript variant 1
PHLDA2	<a href="#">NM_003311</a>	pleckstrin homology-like domain, family A, member 2 (PHLDA2)
PPM1D	<a href="#">NM_003620</a>	protein phosphatase 1D magnesium-dependent, delta isoform (PPM1D)
RECQL4	<a href="#">NM_004260</a>	RecQ protein-like 4 (RECQL4)

RHOC	<a href="#">NM_175744</a>	ras homolog gene family, member C (RHOC)
RRM2	<a href="#">NM_001034</a>	ribonucleotide reductase M2 polypeptide (RRM2)
S100A2	<a href="#">NM_005978</a>	S100 calcium binding protein A2 (S100A2)
S100A4	<a href="#">NM_002961</a>	S100 calcium binding protein A4 (S100A4), transcript variant 1
SALL4	<a href="#">NM_020436</a>	sal-like 4 (Drosophila) (SALL4)
SCD	<a href="#">NM_005063</a>	stearoyl-CoA desaturase (delta-9-desaturase) (SCD)
SDF2L1	<a href="#">NM_022044</a>	stromal cell-derived factor 2-like 1 (SDF2L1)
SIDT2	<a href="#">NM_001040455</a>	SID1 transmembrane family, member 2 (SIDT2)
STEAP3	<a href="#">NM_182915</a>	STEAP family member 3 (STEAP3), transcript variant 1
SUMO2	<a href="#">NM_006937</a>	SMT3 suppressor of mif two 3 homolog 2 (SUMO2), transcript variant 1
TOP2A	<a href="#">NM_001067</a>	topoisomerase (DNA) II alpha 170kDa (TOP2A)
TRAF4	<a href="#">NM_004295</a>	TNF receptor-associated factor 4 (TRAF4), transcript variant 1
TUBA1C	<a href="#">NM_032704</a>	tubulin, alpha 1c
TUBA3D	<a href="#">NM_080386</a>	tubulin, alpha 3d
TUBA4A	<a href="#">NM_006000</a>	tubulin, alpha 4a
TUBB2A	<a href="#">NM_001069</a>	tubulin, beta 2A (TUBB2A)
TUBB2C	<a href="#">NM_006088</a>	tubulin, beta 2C (TUBB2C)
TUBB3	<a href="#">NM_006086</a>	tubulin, beta 3 (TUBB3)
TUBB4	<a href="#">NM_006087</a>	tubulin, beta 4 (TUBB4)
VCL	<a href="#">NM_014000</a>	vinculin (VCL), transcript variant 1

**Tabelle 7.2:** Die Tabelle listet die Gene auf, die durch p53 induziert und mit freiem CDDP aktiviert werden.

Gen Symbol	RefSeq ID	Gen Beschreibung
ANXA4	<a href="#">NM_001153</a>	annexin A4 (ANXA4)
ASCC3	<a href="#">NM_022091</a>	activating signal cointegrator 1 complex subunit 3 (ASCC3), transcript variant 2
BLOC1S2	<a href="#">NM_001001342</a>	biogenesis of lysosome-related organelles complex-1, subunit 2 (BLOC1S2), transcript variant 2

BTG2	<a href="#">NM_006763</a>	BTG family, member 2 (BTG2)
CASP9	<a href="#">NM_001229</a>	caspase 9, apoptosis-related cysteine peptidase (CASP9), transcript variant alpha
CDKN1A	<a href="#">NM_078467</a>	cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A), transcript variant 2
CDKN1A	<a href="#">NM_000389</a>	cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A), transcript variant 1
DDB2	<a href="#">NM_000107</a>	damage-specific DNA binding protein 2, 48kDa (DDB2)
DDR1	<a href="#">NM_013994</a>	discoidin domain receptor family, member 1 (DDR1), transcript variant 3
FDXR	<a href="#">NM_024417</a>	ferredoxin reductase (FDXR), nuclear gene encoding mitochondrial protein, transcript variant 1
GDF15	<a href="#">NM_004864</a>	growth differentiation factor 15 (GDF15)
HDAC10	<a href="#">NM_032019</a>	histone deacetylase 10 (HDAC10)
HIST1H1C	<a href="#">NM_005319</a>	histone 1, H1c (HIST1H1C)
HIST1H3D	<a href="#">NM_003530</a>	histone 1, H3d (HIST1H3D)
HMOX1	<a href="#">NM_002133</a>	heme oxygenase (decycling) 1 (HMOX1)
ID2	<a href="#">NM_002166</a>	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (ID2)
ITGB1	<a href="#">NM_133376</a>	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) (ITGB1)
MMP1	<a href="#">NM_002421</a>	matrix metalloproteinase 1 (interstitial collagenase) (MMP1)
MVP	<a href="#">NM_017458</a>	major vault protein (MVP), transcript variant 1
PLK3	<a href="#">NM_004073</a>	polo-like kinase 3 (Drosophila) (PLK3)
PML	<a href="#">NM_033247</a>	promyelocytic leukemia (PML), transcript variant 8
RAD51C	<a href="#">NM_002876</a>	RAD51 homolog C (S. cerevisiae) (RAD51C), transcript variant 2
RPS19	<a href="#">NM_001022</a>	ribosomal protein S19 (RPS19)
RPS27	<a href="#">NM_001030</a>	ribosomal protein S27 (metallopanstimulin 1) (RPS27)
RPS27L	<a href="#">NM_015920</a>	ribosomal protein S27-like (RPS27L)

STARD4	<a href="#">NM_139164</a>	START domain containing 4, sterol regulated (STARD4)
SULF2	<a href="#">NM_018837</a>	sulfatase 2 (SULF2), transcript variant 1
TP53I3	<a href="#">NM_004881</a>	tumor protein p53 inducible protein 3 (TP53I3), transcript variant 1
TRIAP1	<a href="#">NM_016399</a>	TP53 regulated inhibitor of apoptosis 1 (TRIAP1)
TUBB6	<a href="#">NM_032525</a>	tubulin, beta 6 (TUBB6)
TYRP1	<a href="#">NM_000550</a>	tyrosinase-related protein 1 (TYRP1)

**Tabelle 7.3:** Die Tabelle listet die Gene auf, die durch p53 induziert und sowohl mit freiem als auch mit liposomalem CDDP aktiviert werden.

Gen Symbol	RefSeq ID	Gen Beschreibung
APOE	<a href="#">NM_000041</a>	apolipoprotein E (APOE)
ATP6V1D	<a href="#">NM_015994</a>	ATPase, H <sup>+</sup> transporting, lysosomal 34kDa, V1 subunit D (ATP6V1D)
CALD1	<a href="#">NM_033138</a>	caldesmon 1 (CALD1), transcript variant 1
CD70	<a href="#">NM_001252</a>	CD70 molecule
CDK9	<a href="#">NM_001261</a>	cyclin-dependent kinase 9 (CDC2-related kinase) (CDK9)
CREM	<a href="#">NM_183013</a>	cAMP responsive element modulator (CREM), transcript variant 19
CREM	<a href="#">NM_001881</a>	cAMP responsive element modulator (CREM), transcript variant 2
CREM	<a href="#">NM_182772</a>	cAMP responsive element modulator (CREM), transcript variant 16
CTSL1	<a href="#">NM_001912</a>	cathepsin L1 [ <i>NM001912</i> ]
DCP1B	<a href="#">NM_152640</a>	DCP1 decapping enzyme homolog B ( <i>S. cerevisiae</i> ) (DCP1B)
DDX5	<a href="#">NM_004396</a>	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 (DDX5)
EIF5	<a href="#">NM_001969</a>	eukaryotic translation initiation factor 5 (EIF5), transcript variant 1
ELL	<a href="#">NM_006532</a>	elongation factor RNA polymerase II (ELL)
GADD45A	<a href="#">NM_001924</a>	growth arrest and DNA-damage-inducible, alpha (GADD45A)

HEXIM1	<a href="#">NM_006460</a>	hexamethylene bis-acetamide inducible 1 (HEXIM1)
LIF	<a href="#">NM_002309</a>	leukemia inhibitory factor (cholinergic differentiation factor) (LIF)
LMNA	<a href="#">NM_005572</a>	lamin A/C (LMNA), transcript variant 2
MFN2	<a href="#">NM_014874</a>	mitofusin 2 (MFN2), nuclear gene encoding mitochondrial protein
MST1	<a href="#">NM_020998</a>	macrophage stimulating 1 (hepatocyte growth factor-like) (MST1)
OGT	<a href="#">NM_181672</a>	O-linked N-acetylglucosamine (GlcNAc) transferase (OGT), transcript variant 1
PARP3	<a href="#">NM_001003935</a>	poly (ADP-ribose) polymerase family, member 3 (PARP3), transcript variant 3
PLK2	<a href="#">NM_006622</a>	polo-like kinase 2 (Drosophila) (PLK2)
PRKAB2	<a href="#">NM_005399</a>	protein kinase, AMP-activated, beta 2 non-catalytic subunit (PRKAB2)
RCHY1	<a href="#">NM_015436</a>	ring finger and CHY zinc finger domain containing 1 (RCHY1), transcript variant 1
SAT2	<a href="#">NM_133491</a>	spermidine/spermine N1-acetyltransferase 2 (SAT2)
SERTAD1	<a href="#">NM_013376</a>	SERTA domain containing 1 (SERTAD1)
SIRT7	<a href="#">NM_016538</a>	sirtuin (silent mating type information regulation 2 homolog) 7 (SIRT7)
STX6	<a href="#">NM_005819</a>	syntaxin 6 (STX6)
TNFRSF10B	<a href="#">NM_003842</a>	tumor necrosis factor receptor superfamily, member 10b (TNFRSF10B), transcript variant 1
VHL	<a href="#">NM_000551</a>	von Hippel-Lindau tumor suppressor (VHL)
YWHAG	<a href="#">NM_012479</a>	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide (YWHAG)
ZFAT1	<a href="#">NM_020863</a>	ZFAT zinc finger 1

**Tabelle 7.4:** Die Tabelle listet die Gene auf, die durch p53 induziert und mit liposomalem CDDP aktiviert werden.

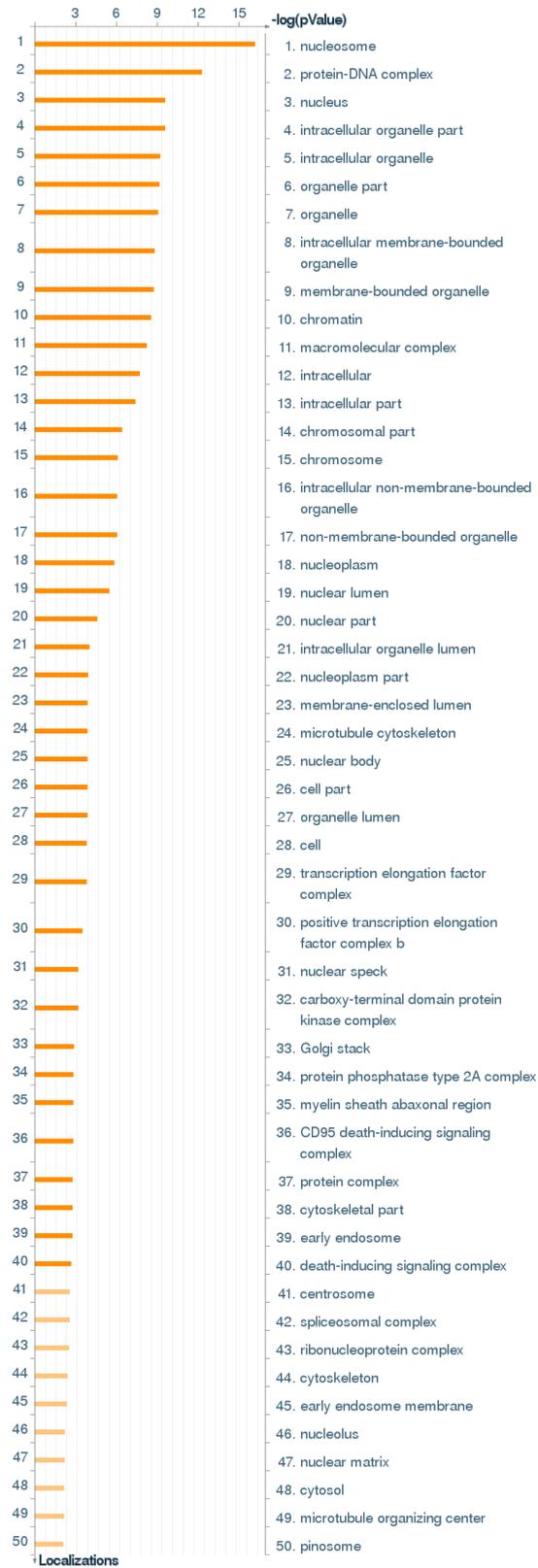


Abbildung 7.1: Die Abbildung stellt signifikante Netzwerke der Kategorie GO Lokalisation dar.



Abbildung 7.2: Die Abbildung stellt signifikante Netzwerke der Kategorie GO Molekularen Funktion dar.

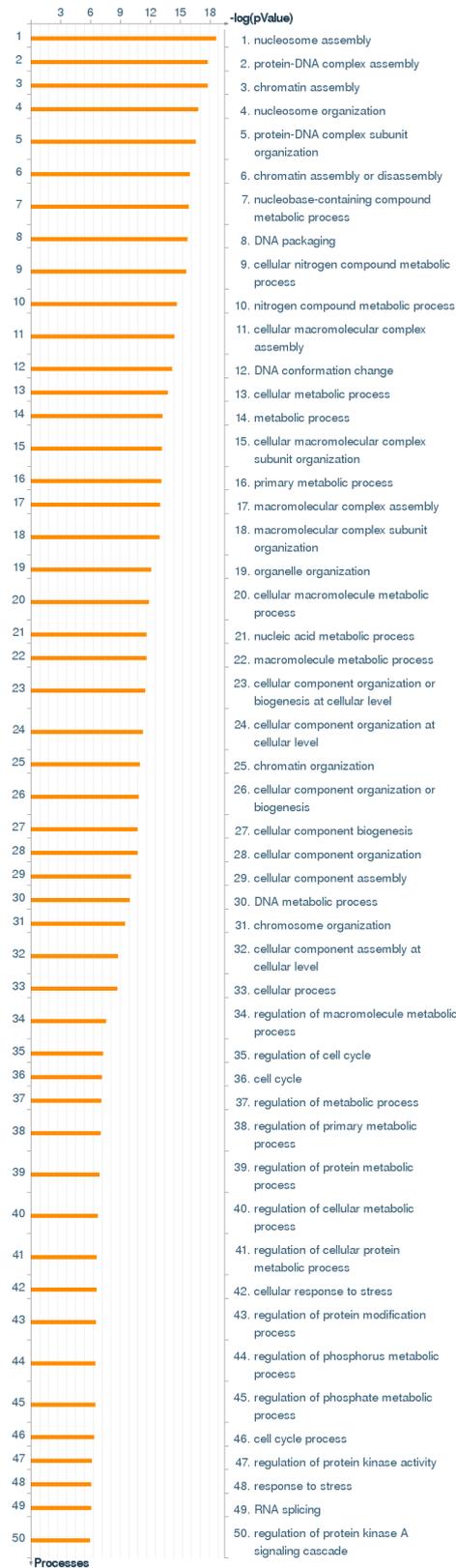


Abbildung 7.3: Die Abbildung stellt signifikante Netzwerke der Kategorie GO Prozess dar.

Description of network objects	
	Generic kinase
	Generic enzyme
	Protein kinase
	Generic protease
	Metalloprotease
	Transporter
	Generic receptor
	Receptors with enzyme activity
	Receptor ligand
	Transcription factor
	Generic binding protein
	Protein
	Overexpressed gene(s)
	Positive effect
	Negative effect
	Unspecified effect

**Abbildung 7.4:** Die Abbildung stellt eine Beschreibung der im Netzwerk verwendeten Symbole dar.



## 8 Literaturverzeichnis

- [1] Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537.
- [2] Van De Vijver, M.; Yudong, D. H.; Vant Veer, L. J.; Dai, H.; Hart, A.; Voskuil, D. W.; et al., A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med.* **2002**, *347*, 1999–2009.
- [3] Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467–470.
- [4] Lipshutz, R. J.; Morris, D.; Chee, M.; Hubbell, E.; Kozal, M. J.; Shah, N. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **1995**, *19*, 442–447.
- [5] Hughes, T. R.; Mao, M.; Jones, a. R.; Burchard, J.; Marton, M. J.; Shannon, K. W. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol.* **2001**, *19*, 342–347.
- [6] Schraml, P.; Schwerdtfeger, G.; Burkhalter, F.; Raggi, A.; Schmidt, D.; Ruffalo, T.; et al., Combined array comparative genomic hybridization and tissue microarray analysis suggest PAK1 at 11q13.5-q14 as a critical oncogene target in ovarian carcinoma. *Am J Pathol.* **2003**, *163*, 985–992.
- [7] Zilberman, D. The human promoter methylome. *Nat Genet.* **2007**, *39*, 442–443.
- [8] Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J. M. High density DNA methylation array with single CpG site resolution. *Genomics* **2011**, *98*, 288–295.
- [9] Thomson, J.; Parker, J.; Perou, C. A custom microarray platform for analysis of microRNA gene expression. *Nat Meth.* **2004**, *1*, 1–7.

- [10] Wellhausen, R.; Seitz, H. Facing current quantification challenges in protein microarrays. *J Biomed Biotechnol.* **2012**, *2012*, 831347.
- [11] Kononen, J.; Bubendorf, L.; Kallionimeni, A.; Bärklund, M.; Schraml, P.; Leighton, S. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **1998**, *4*, 844–847.
- [12] Brown, P. O.; Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat Genet.* **1999**, *21*, 33–37.
- [13] Hu, G. K.; Madore, S. J.; Moldover, B.; Jatkoe, T.; Balaban, D.; Thomas, J.; Wang, Y. Predicting splice variant from DNA chip expression data. *Genome Res.* **2001**, *11*, 1237–1245.
- [14] Dalma-Weiszhausz, D. D.; Warrington, J.; Tanimoto, E. Y.; Miyada, C. G. The affymetrix GeneChip platform: an overview. *Methods Enzymol.* **2006**, *410*, 3–28.
- [15] Koch, M.; Wiese, M. Accessing cancer metabolic pathways by the use of microarray technology. *Curr Pharm Des.* **2012**, 1–16.
- [16] Veer, L. V.; Dai, H.; Vijver, M. V. D.; Yudong, D. H.; Hart, A. A. M.; Mao, M.; et al., Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536.
- [17] Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* **2004**, *351*, 2817–2826.
- [18] Carlson, B. What the Devil Is Personalized Medicine? *Biotechnol Healthc.* **2008**, *5*, 17–19.
- [19] Naylor, S. Unraveling human complexity and disease with systems biology and personalized medicine. *Per Med.* **2010**, *7*, 275–289.
- [20] Weigelt, B.; Baehner, F. L.; Reis-filho, J. S. The contribution of gene expression profiling to breast cancer classification , prognostication and prediction : a retrospective of the last decade. *J Pathol.* **2010**, *220*, 263–280.
- [21] Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Vilo, J.; Abeygunawardena, N.; et al., ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **2003**, *31*, 68–71.

- [22] Barrett, T.; Suzek, T. O.; Troup, D. B.; Wilhite, S. E.; Ngau, W.-C.; Ledoux, P. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res.* **2005**, *33*, D562–566.
- [23] Saeed, A.; Sharov, V.; White, J.; Li, J.; Liang, W.; Bhagabati, N. TM4: A Free, Open-Source System for Microarray Data Management and Analysis. *Biotechniques* **2003**, *34*, 374–378.
- [24] Al-Shahrour, F.; Díaz-Uriarte, R.; Dopazo, J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* **2005**, *21*, 2988–2993.
- [25] Huang, B. D. W.; Lempicki, R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* **2009**, *4*, 44–57.
- [26] Bachtary, B.; Boutros, P. C.; Pintilie, M.; Shi, W.; Bastianutto, C.; Li, J.-H.; et al., Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clin Cancer Res.* **2006**, *12*, 5632–5640.
- [27] Marshall, E. Getting the noise out of gene arrays. *Science* **2004**, *306*, 630–631.
- [28] Allison, D. B.; Cui, X.; Page, G. P.; Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet.* **2006**, *7*, 55–65.
- [29] Yauk, C. Review of the literature examining the correlation among DNA microarray technologies. *Environ Mol Mutagen.* **2006**, *394*, 380–394.
- [30] Jeffery, I. B.; Higgins, D. G.; Culhane, A. C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* **2006**, *7*, 359.
- [31] Shi, L.; Reid, L. H.; Jones, W. D.; Shippy, R.; Warrington, J. a.; Baker, S. C.; et al., The MicroArray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. *Nat Biotechnol.* **2006**, *24*, 1151–1161.
- [32] Shi, L.; Perkins, R. G.; Fang, H.; Tong, W.; et al., Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol.* **2008**, *19*, 10–18.

- [33] Shi, L.; Campbell, G.; Jones, W. D. W.; Campagne, F.; Wen, Z.; Walker, S. J.; et al., The MicroArray Quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature* **2010**, *28*, 827–838.
- [34] Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.
- [35] McClure, J.; Wit, E. Post-normalization quality assessment visualization of microarray data. *Comp Funct Genomics.* **2003**, *4*, 460–467.
- [36] Powell, S. N.; Bindra, R. S. Targeting the DNA damage response for cancer therapy. *DNA Repair* **2009**, *8*, 1153–1165.
- [37] Behrens, B. C.; Hamilton, T. C.; Masuda, H.; Bohrens, B. C.; Grotzinger, K. R.; Whang-peng, J.; et al., Characterization of a cis -Diamminedichloroplatinum ( II ) -resistant Human Ovarian Cancer Cell Line and Its Use in Evaluation of Platinum Analogues. *Cancer Res.* **1987**, *47*, 414–418.
- [38] Linton, K. J.; Higgins, C. F. Structure and function of ABC transporters: the ATP switch provides flexible control. *Pflugers Arch.* **2007**, *453*, 555–567.
- [39] Francia, G.; Green, S. K.; Bocci, G.; Man, S.; Emmenegger, U.; Ebos, J. M. L. Down-regulation of DNA mismatch repair proteins in human and murine tumor spheroids: implications for multicellular resistance to alkylating agents. *Mol Cancer Ther.* **2005**, *4*, 1484–1494.
- [40] Pasello, M.; Michelacci, F.; Scionti, I.; Hattinger, C. M.; Zuntini, M.; Caccuri, A. M.; et al., Overcoming glutathione S-transferase P1-related cisplatin resistance in osteosarcoma. *Cancer Res.* **2008**, *68*, 6661–6668.
- [41] Yin, M.; Yan, J.; Martinez-Balibrea, E.; Graziano, F.; Lenz, H.-J.; Kim, H.-j. ERCC1 and ERCC2 Polymorphisms Predict Clinical Outcomes of Oxaliplatin-Based Chemotherapies in Gastric and Colorectal Cancer : A Systemic Review and Meta-analysis. *Clin Cancer Res.* **2011**, *17*, 1632–1640.
- [42] Cole, S. P. C.; Deeley, R. G. Transport of glutathione and glutathione conjugates by MRP1. *Trends Pharmacol Sci.* **2006**, *27*, 438–446.
- [43] Cheng, J. Q.; Jiang, X.; Fraser, M.; Li, M.; Dan, H. C.; Sun, M.; Tsang, B. K. Role of X-linked inhibitor of apoptosis protein in chemoresistance in ovarian cancer:

- possible involvement of the phosphoinositide-3 kinase/Akt pathway. *Drug Resist Updat.* **2002**, *5*, 131–146.
- [44] Jung, Y.; Lippard, S. J. Direct cellular responses to platinum-induced DNA damage. *Chem Rev.* **2007**, *107*, 1387–1407.
- [45] Krieger, M. L.; Eckstein, N.; Schneider, V.; Koch, M.; Royer, H.-D.; Jaehde, U.; et al., Overcoming cisplatin resistance of ovarian cancer cells by targeted liposomes in vitro. *Int J Pharm* **2010**, *389*, 10–17.
- [46] Stathopoulos, G. P.; Antoniou, D.; Dimitroulis, J.; Stathopoulos, J.; Marosis, K.; Michalopoulou, P. Comparison of liposomal cisplatin versus cisplatin in non-squamous cell non-small-cell lung cancer. *Cancer Chemother Pharmacol.* **2011**, *68*, 945–950.
- [47] Jemal, A.; Bray, F.; Ferlay, J. Global Cancer Statistics. *Cancer J Clin.* **2011**, *61*, 69–90.
- [48] Lowy, D.; Schiller, J. Prophylactic human papillomavirus vaccines. *J Clin Invest.* **2006**, *116*, 1167–1173.
- [49] Rhodes, D. R.; Kalyana-sundaram, S.; Mahavisno, V.; Barrette, T. R.; Ghosh, D.; et al., Mining for regulatory programs in the cancer transcriptome. *Nat Genet.* **2005**, *37*, 579–583.
- [50] Yang, K.; Cai, Z.; Li, J.; Lin, G. A stable gene selection in microarray data analysis. *BMC Bioinformatics* **2006**, *7*, 228.
- [51] Schachtner, R.; Lutter, D.; Knollmüller, P.; Tomé, a. M.; Theis, F. J.; Schmitz, G. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics* **2008**, *24*, 1688–1697.
- [52] Shah, S.; Kusiak, A. Cancer gene search with data-mining and genetic algorithms. *Lancet* **2007**, *37*, 251–261.
- [53] Pal, N.; Aguan, K.; Sharma, A.; Amari, S.-i. Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC Bioinformatics* **2007**, *18*, 1–18.
- [54] Ancona, N.; Maglietta, R.; Piepoli, A.; D’Addabbo, A.; Cotugno, R.; Savino, M. On the statistical assessment of classifiers using DNA microarray data. *BMC Bioinformatics* **2006**, *7*, 387.

- [55] Zhang, S.; Cao, J. A close examination of double filtering with fold change and T test in microarray analysis. *BMC Bioinformatics* **2009**, *10*, 402.
- [56] Minn, A. J.; Gupta, G. P.; Siegel, P. M.; Bos, P. D.; Shu, W.; Giri, D. D.; et al., Genes that mediate breast cancer metastasis to lung. *Nature* **2005**, *436*, 518–524.
- [57] Györfy, B.; Lanczky, A.; Eklund, A. C.; Denkert, C.; Budczies, J.; Li, Q.; et al., An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat.* **2010**, *123*, 725–731.
- [58] Finak, G.; Sadekova, S.; Pepin, F.; Hallett, M.; Meterissian, S.; Halwani, F.; et al., Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res.* **2006**, *8*, R58.
- [59] Culhane, A. C.; Quackenbush, J. Confounding effects in „A six-gene signature predicting breast cancer lung metastasis“. *Cancer Res.* **2009**, *69*, 7480–7485.
- [60] Miller, L. D.; Coffman, L. G.; Chou, J. W. An Iron Regulatory Gene Signature Predicts Outcome in Breast Cancer. *Cancer Res.* **2011**, *71*, 6728–6737.
- [61] Buffa, F. M.; Harris, a. L.; West, C. M.; Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br J Cancer.* **2010**, *102*, 428–435.
- [62] Pawitan, Y.; Bjöhle, J.; Amler, L.; Borg, A.-L.; Egyhazi, S.; Hall, P.; et al., Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* **2005**, *7*, R953–R964.
- [63] Higgins, M.; Baselaga, J. Targeted therapy in breast cancer. *J Clin Invest.* **2011**, *121*, 3797–3803.
- [64] Staudacher, L.; Cottu, P. H.; Dieras, V.; Vincent-Salomon, A.; Guilhaume, M. N.; Escalup, L. Platinum-based chemotherapy in metastatic negative breast cancer : the Institut Curie experience. *Ann Oncol.* **2011**, 848–856.
- [65] Kapp, A. V.; Jeffrey, S. S.; Langerød, A.; Børresen Dale, A.-L.; Han, W.; Noh, D.-Y.; et al., Discovery and validation of breast cancer subtypes. *BMC Genomics* **2006**, *7*, 231.

- [66] Kreike, B.; van Kouwenhove, M.; Horlings, H.; Weigelt, B.; Peterse, H.; Bartelink, H.; et al., Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Res.* **2007**, *9*, R65.
- [67] Lehmann, B. D.; Bauer, J. A.; Chen, X.; Sanders, M. E.; Chakravarthy, A. B.; Shyr, Y.; et al., Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* **2011**, 2750–2767.
- [68] Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C. H.; Angelo, M.; et al., Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS* **2001**, *98*, 15149–15154.
- [69] Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **2002**, *99*, 6567–6572.
- [70] Dudoit, S.; Fridlyand, J. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J Am Stat Assoc.* **2002**, *97*, 37–41.
- [71] Li, T.; Zhang, C.; Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **2004**, *20*, 2429–2437.
- [72] Mackay, A.; Weigelt, B.; Grigoriadis, A.; Kreike, B.; Natrajan, R.; A'Hern, R.; et al., Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J Natl Cancer Inst.* **2011**, *103*, 662–673.
- [73] Loi, S.; Haibe-Kains, B.; Desmedt, C.; Wirapati, P.; Lallemand, F.; Tutt, A. M.; et al., Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **2008**, *9*, 239.
- [74] Haibe-Kains, B.; Desmedt, C.; Loi, S.; Culhane, A. C.; Bontempi, G.; Quackenbush, J.; et al., A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst.* **2012**, *104*, 311–325.
- [75] Geyer, F. C.; Rodrigues, D. N.; Weigelt, B.; Reis-Filho, J. S. Molecular classification of estrogen receptor-positive/luminal breast cancers. *Adv Anat Pathol.* **2012**, *19*, 39–53.
- [76] Burstein, H.; Griggs, J. Deep Time: The Long and the Short of Adjuvant Endocrine Therapy for Breast Cancer. *J Clin Oncol.* **2012**, *30*, 683–684.

- [77] Koch, M.; Royer, H.-D.; Wiese, M. A Microarray Tool Provides Pathway and GO Term Analysis. *Mol. Inf.* **2011**, *30*, 918–921.
- [78] Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; et al., Circos: an information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645.
- [79] Koch, M.; Wiese, M. Quality Visualization of Microarray Datasets Using Circos. *Microarrays* **2012**, *1*, 84–94.
- [80] Koch, M.; Wiese, M. Gene expression signatures of angiocidin and darapladib treatment connect to therapy options in cervical cancer. *J Cancer Res Clin Oncol.* **2012**.
- [81] Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **2005**, *102*, 15545–15550.
- [82] Koch, M.; Hanl, M.; Wiese, M. Feature extraction via composite scoring and voting in breast cancer. *Breast Cancer Res Treat.* **2012**, *135*, 307–318.
- [83] Hassan, M. R.; Hossain, M. M.; Bailey, J.; Macintyre, G.; Ho, J. W.; Ramamohanarao, K.; et al., A voting approach to identify a small number of highly predictive genes using multiple classifiers. *BMC Bioinformatics* **2009**, *12*, 1–12.
- [84] Popovici, V.; Chen, W.; Gallas, B. G.; Hatzis, C.; Shi, W.; Samuelson, F. W. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* **2010**, *12*, R5.
- [85] Shi, P.; Ray, S.; Zhu, Q.; Kon, M. a. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics* **2011**, *12*, 375.
- [86] Slawski, M.; Daumer, M.; Boulesteix, A. l. CMA, a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics* **2008**, *9*, 439.
- [87] Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus Clustering. *Broad Institute/MIT* **2003**, *16*, 1–34.
- [88] Wilkerson, M. D.; Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **2010**, *26*, 1572–1573.

- [89] Wilkerson, M. D.; Yin, X.; Hoadley, K. a.; Liu, Y.; Hayward, M. C.; Cabanski, C. R.; et al., Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res.* **2010**, *16*, 4864–4875.
- [90] Dalgaard, P. *Statistics and Computing*; Springer: New York, 2002; p 267.
- [91] Vogel, L. *Eclipse IDE 3.7*; Amazon Digital Services, 2011; p 120.
- [92] Boudreau, T.; Glick, J.; Greene, S.; Woehr, J.; Spurlin, V. *NetBeans: The Definitive Guide*; O'Reilly, 2002; p 672.
- [93] Boeck, H. *NetBeans Platform 6 Rich-Client-Entwicklung mit Java*; Galileo Computing, 2007.
- [94] Steinberg, D.; Budinsky, F.; Paternostro, M.; Merks, E. *EMF: Eclipse Modeling Framework (2nd Edition)*; O'Reilly, 2008; p 744.
- [95] Beck, K. *Test Driven Development by Example*; Addison-Wesley Verlag, 2003; p 220.
- [96] Meffert, K. *JUnit Profi Tipps Software erfolgreich testen*; Entwickler Press, 2006; p 200.
- [97] Wiest, S. *Continuous Integration mit Hudson*; dpunkt Verlag, 2011.
- [98] Zyl van, J.; Fox, B.; Casey, J.; Snyder, B.; O'Brien, T.; Redmond, E. *Maven: The Definitive Guide*; O'Reilly, 2008.
- [99] Collins-Sussman, B.; Fitzpatrick, B. W.; Pilato, C. M. *Version Control with Subversion*; O'Reilly, 2004; p 320.
- [100] Chacon, S. *Git*; <http://git-scm.com/book>, 2009.
- [101] Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *PNAS* **2003**, *100*, 9440–9445.
- [102] Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Proctical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol.* **1995**, *57*, 289–300.
- [103] Vacha, S. Ten Pitfalls of Microarray Analysis. *Agilent Technologies* **2003**, 1–12.
- [104] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2001; p 527.

- [105] Brier, G. W. Verfication of Forecasts Expressed in Terms of Probability. *Monthly Weather Rev.* **1950**, *78*, 1–3.
- [106] Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol.* **2004**, *3*, 1–25.
- [107] Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans Neural Netw.* **1999**, *10*, 988–999.
- [108] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine learning* **2002**, *46*, 389–422.
- [109] Breiman, L. Random Forests. *Machine learning* **2001**, *45*, 5–32.
- [110] Wold, S.; Ruhe, A.; Wold, H. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J Sci Comput.* **1984**, *5*, 735–743.
- [111] Ramírez, J.; Górriz, J. M.; Segovia, F.; Chaves, R.; Salas-Gonzalez, D.; López, M.; et al., Computer aided diagnosis system for the Alzheimer’s disease based on partial least squares and random forest SPECT image classification. *Neurosci Lett.* **2010**, *472*, 99–103.
- [112] Friedman, J. Greedy Function Approximation: A Gradient Boosting Machine. *An Stat.* **2001**, 1–39.
- [113] Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
- [114] Nishimura, D. A View From the Web. *Biotech Software & Internet Report* **2001**, *2*, 117–120.
- [115] Croft, D.; O’Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; et al., Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **2011**, *39*, D691–D697.
- [116] Ringnér, M.; Fredlund, E.; Häkkinen, J.; Borg, A. k.; Staaf, J. GOBO: gene expression-based outcome for breast cancer online. *PLoS One* **2011**, *6*, e17911.
- [117] Ashburner, M.; A. Ball, C.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; et al., Gene Ontology: tool for the unification of biology. *Nat Genet.* **2000**, *25*, 25–29.

- [118] Grossmann, S.; Bauer, S.; Robinson, P. N.; Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **2007**, *23*, 3024–3031.
- [119] Khatri, P.; Voichita, C.; Kattan, K. Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.* **2007**, *35*, 206–211.
- [120] Draghici, S.; Khatri, P.; Tarca, A. L.; Amin, K.; Done, A.; Voichita, C.; et al., A systems biology approach for pathway level analysis. *Genome Res.* **2007**, *17*, 1537–1345.
- [121] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; et al., Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.
- [122] Nikolsky, Y.; Ekins, S.; Nikolskaya, T.; Bugrim, A. A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol Lett.* **2005**, *158*, 20–29.
- [123] Ekins, S.; Bugrim, A.; Brovold, L.; Kirillov, E.; Nikolsky, Y.; Rakhmatulin, E.; et al., Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. *Xenobiotica* **2006**, *36*, 877–901.
- [124] Quackenbush, J. Microarray data normalization and transformation. *Nat Genet.* **2002**, *32*, 496–501.
- [125] Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **2003**, *19*, 185–193.
- [126] Smyth, G. K.; Speed, T. Normalization of cDNA microarray data. *Methods* **2003**, *31*, 265–273.
- [127] Carvalho, B.; Bengtsson, H.; Speed, T. P.; Irizarry, R. a. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **2007**, *8*, 485–499.
- [128] Almeida, A.; Rosa, P. L.; Rigaiil, G.; Hupe, P.; Meyniel, J.-p.; Decraene, C.; et al., ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. *Bioinformatics* **2008**, *24*, 768–774.
- [129] Zeller, G.; Henz, S.; Laubinger, S.; Weigel, D. Transcript normalization and segmentation of tiling array data. *Pac Symp Biocomput* **2008**, *12*, 527–538.

- [130] Autio, R.; Kilpinen, S.; Saarela, M.; Kallioniemi, O.; Hautaniemi, S.; Astola, J.; et al., Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics* **2009**, *10 Suppl 1*, S24.
- [131] Lutter, D.; Langmann, T.; Ugocsai, P.; Moehle, C.; Seibold, E.; Splettstoesser, W. D.; et al., Analyzing time-dependent microarray data using independent component analysis derived expression modes from human macrophages infected with *F. tularensis holartica*. *J Biomed Inform.* **2009**, *42*, 605–611.
- [132] Aoyagi, K.; Minashi, K.; Igaki, H.; Tachimori, Y.; Nishimura, T.; Hokamura, N.; et al., Artificially induced epithelial-mesenchymal transition in surgical subjects: its implications in clinical and basic cancer research. *PLoS One* **2011**, *6*, e18196.
- [133] Chang, E. C.; Charn, T. H.; Park, S.-H.; Helferich, W. G.; Komm, B.; Katzenellenbogen, J.; et al., Estrogen Receptors alpha and beta as determinants of gene expression: influence of ligand, dose, and chromatin binding. *Mol Endocrinol.* **2008**, *22*, 1032–1043.
- [134] Team, R. D. C. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **2008**, *1*, 2673.
- [135] Sean, D.; Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847.
- [136] Gautier, L.; Cope, L.; Bolstad, B. M.; Irizarry, R. a. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20*, 307–315.
- [137] Irizarry, R. A.; Hobbs, B.; Beazer-barcly, Y. D.; Antonellis, K. J.; Scherf, U. W. E.; Speed, T. P.; et al., Exploration , normalization , and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2003**, *4*, 249–264.
- [138] Stacklies, W.; Redestig, H.; Scholz, M.; Walther, D.; Selbig, J. pcaMethods, A bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **2007**, *23*, 1164–1167.
- [139] Schaefer, C. F.; Anthony, K.; Krupa, S.; Buchoff, J.; Day, M.; Hannay, T. PID: the Pathway Interaction Database. *Nucleic Acids Res.* **2009**, *37*, D674–679.
- [140] Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J. *Design Patterns: Elements of Reusable Object-Oriented Software*; Addison-Wesley Professional, 1995.

- [141] Durinck, S.; Moreau, Y.; Kasprzyk, A.; Davis, S.; De Moor, B.; Brazma, A.; et al., BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **2005**, *21*, 3439–3440.
- [142] Suzuki, R.; Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22*, 1540–1542.
- [143] Wickham, H. ggplot2. *Wiley Interdiscip Rev Comput Stat.* **2011**, *3*, 180–185.
- [144] Ross, M.; Zhou, X.; Song, G. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **2003**, *102*, 2951–2959.
- [145] Theocharidis, A.; Dongen, S. V.; Enright, A. J.; Freeman, T. C.; van Dongen, S. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc.* **2009**, *4*, 1535–1550.
- [146] Bendas, G.; Krause, A.; Bakowsky, U.; Vogel, J.; Rothe, U. Targetability of novel immunoliposomes prepared by a new antibody conjugation technique. *Int J Pharm* **1999**, *181*, 79–93.
- [147] Muscolini, M.; Cianfrocca, R.; Sajeve, A.; Mozzetti, S.; Ferrandina, G.; Costanzo, A.; et al., Trichostatin A up-regulates p73 and induces Bax-dependent apoptosis in cisplatin-resistant ovarian cancer cells. *Mol Cancer Ther.* **2008**, *7*, 1410–1419.
- [148] Beijnen, J. H. The application of inductively coupled plasma mass spectrometry in clinical pharmacological oncology research. *Mass Spectrom Rev.* **2007**, *27*, 67–100.
- [149] Pyeon, D.; Newton, M. A.; Lambert, P. F.; den Boon, J. a.; Sengupta, S.; Marsit, C. J.; et al., Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.* **2007**, *67*, 4605–4619.
- [150] Scotto, L.; Narayan, G.; Nandula, S. Identification of Copy Number Gain and Overexpressed Genes on Chromosome Arm 20q by an Integrative Genomic Approach in Cervical Cancer : Potential Role in Progression. *Genes Chromosomes Cancer* **2008**, *47*, 755–765.
- [151] Noordhuis, M. G.; Fehrmann, R. S. N.; Wisman, G. B. A.; Nijhuis, E. R.; Zanden, J. J. V.; Moerland, P. D.; et al., Involvement of the TGF- $\beta$  and b-Catenin Pathways in Pelvic Lymph Node Metastasis in Early-Stage Cervical Cancer. *Clin Cancer Res.* **2011**, *17*, 1317–1330.

- [152] Wilson, C. L.; Miller, C. J. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* **2005**, *21*, 3683–3685.
- [153] Bolstad, B. M. Comparing the effects of background , normalization and summarization on gene expression estimates. *Methods* **2002**, 1–10.
- [154] Rudy, J.; Valafar, F. Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinformatics* **2011**, *12*, 467.
- [155] Li, Y.; Zou, L.; Li, Q.; Haibe-Kains, B.; Tian, R.; Li, Y.; et al., Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat Med.* **2010**, *16*, 214–218.
- [156] Neve, R.; Chin, K.; Fridlyand, J.; Yeh, J.; Baehner, F.; et al., A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **2006**, *10*, 515–527.
- [157] Watanabe, T.; Komuro, Y.; Kiyomatsu, T.; Kanazawa, T.; Kazama, Y.; Tanaka, J.; et al., Prediction of sensitivity of rectal cancer cells in response to pre-operative radiotherapy by DNA microarray analysis of gene expression profiles. *Cancer Res.* **2006**, *66*, 3370–3374.
- [158] Desmedt, C.; Piette, F.; Loi, S.; Wang, Y.; Lallemand, F.; Haibe-Kains, B.; et al., Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res.* **2007**, *13*, 3207–3214.
- [159] Carew, J. S.; Giles, F. J.; Nawrocki, S. T. Histone deacetylase inhibitors: mechanisms of cell death and promise in combination cancer therapy. *Cancer Lett.* **2008**, *269*, 7–17.
- [160] Vakifahmetoglu, H.; Olsson, M.; Tamm, C.; Heidari, N.; Orrenius, S.; B, Z. DNA damage induces two distinct modes of cell death in ovarian carcinomas. *Cell Death Differ.* **2008**, *3*, 555–566.
- [161] Brown, R.; Clugston, C.; Burns, P.; Edlin, A.; Vasey, P.; Vojtesek, B.; et al., Increased accumulation of p53 protein in cisplatin-resistant ovarian cell lines. *Int J Cancer.* **1993**, *4*, 678–684.
- [162] Bragado, P.; Armesilla, A.; Silva, A.; A, P. Apoptosis by cisplatin requires p53 mediated p38alpha MAPK activation through ROS generation. *Apoptosis.* **2007**, *12*, 1733–1742.

- [163] Dolado, I.; Swat, A.; Ajenjo, N.; De Vita, G.; Cuadrado, A.; AR, N. p38alpha MAP kinase as a sensor of reactive oxygen species in tumorigenesis. *Cancer Cell*. **2007**, *2*, 191–205.
- [164] Ishida, O.; Maruyama, K.; Sasaki, K.; M, I. Size-dependent extravasation and interstitial localization of polyethyleneglycol liposomes in solid tumor-bearing mice. *Int J Pharm*. **1999**, *190*, 49–56.
- [165] Sancho-Martínez, S.; Prieto-García, L.; Prieto, M.; López-Novoa, J.; FJ, L.-H. Subcellular targets of cisplatin cytotoxicity: an integrated view. *Pharmacol Ther*. **2012**, *1*, 33–35.
- [166] Rosty, C.; Sheffer, M.; Tsafrir, D.; Stransky, N.; Tsafrir, I.; Peter, M.; et al., Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. *Oncogene* **2005**, *24*, 7094–7104.
- [167] Bidus, M. a. Prediction of Lymph Node Metastasis in Patients with Endometrioid Endometrial Cancer Using Expression Microarray. *Clin Cancer Res*. **2006**, *12*, 83–88.
- [168] Wilensky, R. L. R.; Shi, Y.; Mohler, E. R. E.; Hamamdzic, D.; Burgert, M. E.; Li, J. Inhibition of lipoprotein-associated phospholipase A2 reduces complex coronary atherosclerotic plaque development. *Nat Med*. **2008**, *14*, 1059–1066.
- [169] Gaurnier-Hausser, A.; Rothman, V. L.; Dimitrov, S.; Tuszynski, G. P. The novel angiogenic inhibitor, angiocidin, induces differentiation of monocytes to macrophages. *Cancer Res*. **2008**, *68*, 5905–5914.
- [170] Browne, E.; Wing, B.; Coleman, D. Altered Cellular mRNA Levels in Human Cytomegalovirus-Infected Fibroblasts : Viral Block to the Accumulation of Antiviral mRNAs. *J Virol*. **2001**, *75*, 12319–12330.
- [171] Flechner, S. M.; Kurian, S. M.; Head, S. R.; Sharp, S. M.; Whisenant, T. C.; Zhang, J.; et al., Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. *Am J Transplant*. **2004**, *4*, 1475–1489.
- [172] Wieland, S.; Thimme, R.; Purcell, R. H.; Chisari, F. V. Genomic analysis of the host response to hepatitis B virus infection. *PNAS* **2004**, *101*, 6669–6674.

- [173] Rozanov, D. V.; Savinov, A. Y.; Williams, R.; Liu, K.; Golubkov, V. S.; Krajewski, S.; et al., Molecular signature of MT1-MMP: transactivation of the downstream universal gene network in cancer. *Cancer Res.* **2008**, *68*, 4086–4096.
- [174] Slebos, R. J. C.; Yi, Y.; Ely, K.; Carter, J.; Evjen, A.; Zhang, X. Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma. *Clin Cancer Res.* **2006**, *12*, 701–709.
- [175] Ueno, M.; Itoh, M.; Kong, L.; Sugihara, K.; Asano, M.; Takakura, N.; et al., PSF1 Is Essential for Early Embryogenesis in Mice. *Mol Cell Biol.* **2005**, *25*, 10528–10532.
- [176] Laneve, P.; Gioia, U.; Ragno, R.; Altieri, F.; Di Franco, C.; Santini, T. The tumor marker human placental protein 11 is an endoribonuclease. *J Biol Chem.* **2008**, *283*, 34712–34719.
- [177] Zhang, J.; Cao, W.; Xu, Q.; Chen, W.-t. The expression of EMP1 is downregulated in oral squamous cell carcinoma and possibly associated with tumour metastasis. *J Clin Pathol.* **2011**, *64*, 25–29.
- [178] Insinga, A.; Monestiroli, S.; Ronzoni, S.; Gelmetti, V.; Marchesi, F.; Viale, A.; et al., Inhibitors of histone deacetylases induce tumor-selective apoptosis through activation of the death receptor pathway. *Nat Med.* **2005**, *11*, 71–76.
- [179] Mayer, E. L.; Baurain, J.-F.; Sparano, J.; Strauss, L.; Campone, M.; Fumoleau, P.; et al., A phase 2 trial of dasatinib in patients with advanced HER2-positive and/or hormone receptor-positive breast cancer. *Clin Cancer Res.* **2011**, *17*, 6897–6904.
- [180] Silver, D. P.; Richardson, A. L.; Eklund, A. C.; Wang, Z. C.; Szallasi, Z.; Li, Q.; et al., Efficacy of neoadjuvant Cisplatin in triple-negative breast cancer. *J Clin Oncol.* **2010**, *28*, 1145–1153.
- [181] Li, J.; Wang, C.-Y. TBL1-TBLR1 and beta-catenin recruit each other to Wnt target-gene promoter for transcription activation and oncogenesis. *Nat Cell Biol.* **2008**, *10*, 160–169.
- [182] Wang, H.; Shao, N.; Ding, Q. M.; Cui, J.; Reddy, E. S.; Rao, V. N. BRCA1 proteins are transported to the nucleus in the absence of serum and splice variants BRCA1a, BRCA1b are tyrosine phosphoproteins that associate with E2F, cyclins and cyclin dependent kinases. *Oncogene* **1997**, *15*, 143–157.

- [183] Bindra, R. S.; Glazer, P. M. Repression of RAD51 gene expression by E2F4/p130 complexes in hypoxia. *Oncogene* **2007**, *26*, 2048–2057.
- [184] Byrski, T.; Huzarski, T.; Dent, R.; Gronwald, J.; Zuziak, D.; Cybulski, C. Response to neoadjuvant therapy with cisplatin in BRCA1-positive breast cancer patients. *Breast Cancer Res Treat.* **2009**, *115*, 359–363.



# Danksagung

An dieser Stelle möchte ich allen danken, die zu dieser Arbeit beigetragen haben. Sehr herzlich möchte ich zuerst meinen persönlichen Dank richten an Prof. Dr. Wiese. Er vertraute mir diese Arbeit an und war stets zu einem fachlichen Dialog bereit. Er bot mir die Chance meinen Interessen freien Lauf zu lassen, sowie mich auf internationalen Kongressen fortzubilden um so auch professionelle Kontakte zu knüpfen.

Zudem danke ich allen Mitarbeitern der Arbeitsgruppe Wiese sowie der Arbeitsgruppe Bendas für das nette Arbeitsklima und das freundliche Miteinander. Außerdem gilt mein besonderer Dank folgenden Kollegen, die Arbeit wäre ohne ihre Hilfe nicht möglich gewesen:

**Markus Hanl**, Arbeitsgruppe Prof. Dr. Wiese.

**Gerd Bendas**, Arbeitsgruppe Prof. Dr. Bendas.

**Dr. Michaela Krieger**, Arbeitsgruppe Prof. Dr. Bendas.

**Brad Sherman**, Fredrick Sanger Institute, NIH.

**Melania Pintillie**, Cancer Research Center, Ontario.

**Barbara Bachtary**, Medizinische Universität, Wien.

**Martin Krzywinski**, Genome Sciences Centre, Vancouver BC.

**Dr. Judith Schenk** und **Christian van den Bos**, Lonza.

**Scott Preiss**, Glaxo-Smith-Kline. Belgien.

Mein ganz besonderer Dank gilt meiner Mutter Ursula, meinem Vater Johann und meinem Bruder Thomas, zu jeder Zeit fand ich bei Ihnen Unterstützung.



# Publikationen

## Wissenschaftliche Originalarbeiten

Koch M, Hanl M and Wiese M. Feature extraction via composite scoring and voting in breast cancer. **Breast Cancer Research and Treatment** 2012;135(1):307-318.

Koch M, Wiese M. Angiocidin and darapladib may present novel therapy options in cervical cancer. **Cancer Research and Clinical Oncology** 2013;139(2):259-267

Koch M, Krieger ML, Brenner N, Beier M, Jaehde U, Wiese M, Royer H-D and Bendas G. Efficient elimination of chemotherapy-resistant ovarian cancer cells by liposomal cisplatin. (in Präparation)

Koch M, Wiese M. Accessing cancer metabolic pathways by the use of microarray technology. **Current Pharmaceutical Design** 2012; [Epub ahead of print]

Koch M, Wiese M. Quality visualization of microarray datasets using Circos. **Microarrays** 2012; 1(2):84-94.

Koch M, Royer H-D, Wiese M. A Microarray Tool Provides Pathway and GO Term Analysis. **Molecular Informatics** 2011;30:918-21.

Krieger ML, Eckstein N, Schneider V, Koch M, Royer H-D, Jaehde U, Bendas G. Overcoming cisplatin resistance of ovarian cancer cells by targeted liposomes in vitro. **Int J Pharm** 2010;389:10-17.

## Poster und Kongressbeiträge

Koch M, Krieger M, Brenner N, Royer HD, Bendas G and Wiese M, „Microarray analysis of liposomal cisplatin treated A2780 ovarian cancer cells reveals mechanisms of chemoresistance“, 10th International Conference on Systems Biology, Edinburgh, Schottland, (Oktober 2010). Poster.

---

**Koch M**, Topolsky I, Krieger M, Brenner N, Royer HD, Jaehde U, Bendas G, Wiese M, „Research fields in Gene Expression Analysis“, Bayer-Schering Conference, LIMES-Institute (Life and Medical Sciences), Bonn, Deutschland, (September 2010). Poster.

**Koch M**, Wiese M, Brenner N, Royer HD, and Eckstein N, „Time-resolved monitoring of the transcriptome of MCF-7 breast cancer cells during the emergence of cisplatin resistance“, 8th European Conference on Computational Biology, Stockholm, Schweden, (Juni 2009). Poster.

# Verfassererklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die den verwendeten Werken wörtlich oder inhaltlich entnommenen Stellen als solche gekennzeichnet.

Stuttgart,

Martin Koch