Robuste Datenauswertung und Anwendungen von Oligonukleotid-Arrays in der Genexpressionsanalyse

Dissertation

Zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
der Humboldt-Universität zu Berlin

von

Stefan Röpcke
geboren am 10. 8. 1971 in Rostock

Präsident der Humboldt-Universität
Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I
Prof. Dr. Michael Linscheid

Gutachter

Prof. Dr. Reinhard Heinrich

Prof. Dr. Hanspeter Herzel

Tag der mündlichen Prüfung: 30. 9. 2003

Abriss

Die Doktorarbeit wurde 1999 mit der Zielsetzung begonnen, die Auswertung von Oligonukleotid-Arrays zu überarbeiten und wenn möglich zu verbessern. Diese neue Technologie dient der Genexpressionsanalyse und erlaubt es, tausende von Genen parallel zu untersuchen. Im Zuge dieser Arbeit gelang die Entwicklung eines robusten Verfahrens zur Datenanalyse von Oligonukleotid-Arrays, das auch den Vergleich von Experimenten sehr unterschiedlicher Qualität erlaubt. Gerade für die Untersuchung humaner Proben ist die Robustheit von großem Interesse, da meist mit sehr begrenzten Mengen an Gewebematerial gearbeitet wird. Unterschiedliche Bedingung bei der Gewinnung und Asservierung des Materials können zur Beeinträchtigung der Vergleichbarkeit der Ergebnisse führen.

Der Standardweg bei der Evaluierung von Auswertemethoden, ausreichend Wiederholungen zu erzeugen, war wegen der begrenzten Materialresourcen und des hohen Array-Preises versperrt. Durch genaue Spezifikation der Ziele der Analyse und durch Annahmen an das System konnte auf der Grundlage eines eingeschränkten Sets an Kontrollversuchen gezeigt werden, dass die vorgeschlagene Methode besser die Erwartungen erfüllt als herkömmliche Verfahren. Ein weiterer Teil der Arbeit bestand im Aufbau einer relationalen Datenbank und in der schrittweisen Automatisierung der Auswertung. Die strukturierte Speicherung ermöglicht Analysen über sämtliche Datensätze, was sich unter anderem bei der Suche nach selten exprimierten Genen als nützlich erweist.

Stellvertretend für andere Krebserkrankungen wurde eine detaillierte Analyse zweier publizierter Expressionsdatensätze zum Bronchialkarzinom vorgenommen. Man findet in beiden Datensätzen zwischen Tumor- und Normalgewebe differenziell exprimierte Gene. Bei der Gegenüberstellung der Ergebnisse der auf unterschiedlichen Array-Plattformen durchgeführten Analysen zeigt sich der deutliche Einfluss der Technologie auf die Expressionssignale.

Der spezielle Aufbau des verwendeten Oligonukleotid-Arrays gestattete die Entdeckung putativer Antisense-Transkripte. Die Koexpression einiger Sense- und Antisense-Sonden lassen sich durch Northern-Blot-Experimente bestätigen. Hier zeigt sich das Anwendungspotenzial der Array-Technologie für die Charakterisierung des humanen Transkriptoms.

Gerade für Modellorganismen können bereits mit einem Array große Teile des gesamten Genoms untersuchen werden. Damit ist man in der Lage, Informationen über die Transkriptionsmaschinerie selbst zu gewinnen. Man findet zum Beispiel einen Zusammenhang zwischen den Längen von Introns und Exons und der mittleren Expression von Genen in Hefe und Fruchtfliege. Daraus ergeben sich interessante Fragen für den Energiehaushalt der Zelle und den resultierenden evolutionären Druck auf die Genstruktur.

Die Vielfalt der Anwendungen und die Ausbaumöglichkeiten verdeutlichen die Bedeutung und das Potenzial der Array-Technologie für die Genexpressionsanalyse. Eine wichtige Aufgabe bleibt deshalb die weitere Verbesserung der Qualitätskontrolle der Experimente und der Datenanalyse.

Danksagung

An dieser Stelle möchte ich mich für die umfangreiche Hilfe bedanken, die ich von vielen Seiten empfangen habe. In erster Linie gilt mein Dank natürlich den Kollegen bei der metaGen, ohne deren Hilfe diese Arbeit nicht hätte entstehen können.

Als erstes möchte ich den Professoren Reinhard Heinrich und Hanspeter Herzel meinen Dank dafür aussprechen, dass sie sich für die Begutachtung meiner Promotion bereit erklärt haben.

Herrn Professor Rosenthal möchte ich dafür danken, dass er das Promovieren in der Firma metaGen möglich gemacht und gefördert hat. Das offene und kritische Arbeitsklima halfen mir, meine Arbeit voranzubringen. Letztlich konnten die neuen Verfahren nicht nur auf relevanten Daten getestet sondern sogar in die tagtägliche Praxis bei metaGen integriert werden. Dr. Christian Pilarsky hat maßgeblich bei der Entwicklung und Evaluierung des Analyseverfahrens mitgewirkt. Vor allem sein Interesse an der Verbesserung der Auswertung, viele gute Ideen und seine aktive, kritische Beurteilung vorgeschlagener Methoden waren integraler Bestandteil des Entwicklungsprozesses.

Besonders der Zuspruch und die tatkräftige Unterstützung von Dr. Detlev Mennerich erlaubten mir, einige Laborversuche selbst durchzuführen. Unterstützt wurde ich hierbei außerdem von Eva Klopocki, Anke Vogel und Simone Kaiser. So konnte ich Einblicke in die molekularbiologische Praxis gewinnen, die sich als sehr wertvoll bei der Datenauswertung erwiesen. Durch reines Literaturstudium sind diese wohl nicht zu erlangen. Die Daten, auf denen der Großteil dieser Arbeit beruht, sind in den Labors von Dr. Christian Pilarsky und Dr. Thomas Brümmendorf (Chipexperimente) und Dr. Edgar Dahl (RNA-ISH) generiert worden. Nicole Creutzburg hat freundlicherweise die RNA-ISH (Kapitel 4) durchgeführt.

In der Abteilung Bioinformatik gilt mein besonderer Dank meinem Mitstreiter Eike Staub für unzählige gute Ideen, kritische Komentare und inhaltsreiche Diskussionen. Klaus Herrmann half mir bei statistischen Problemen und prüfte zahlreiche Verfahren und den methodischen Teil dieses Textes. Bernd Hinzmann brachte bereitwillig seine Expertise im Bereich der Sequenzanalyse und natürlich seine Zeit ein. Xinhong Li stand mir bei der Clusteranalyse zur Seite. Außerdem sei noch Martin Stei für das kritsche Lesen des Manuskripts gedankt.

Inhaltsverzeichnis

1. Einführung

Die Basis der vorliegenden Arbeit bildet die Technologie der Oligonukleotid-Arrays[1] für die Genexpressionsanalyse. Die Hauptergebnisse bestehen in der Entwicklung und Etablierung eines robusten Verfahrens zur Datenanalyse und in Anwendungen, die sich daraus ableiten lassen. Dieses Kapitel erläutert den Einsatz der Expressionsanalyse bei der Suche nach krebsrelevanten Genen. Hier motiviert sich der gewählte Ansatz für die Datenanalyse. Ausführlich ist der Aufbau der verwendeten Oligonukleotid-Arrays mit ihren Sequenzsets beschrieben. Abschnitt 1.5 gibt eine Übersicht über die Labormethoden, die Teil eines jeden Chip-Experiments sind. Abschließend sind in Abschnitt 1.6 zum Vergleich zwei alternative Hochdurchsatzverfahren zur Genexpressionsanalyse dargestellt.

1.1. Begriffe und Abkürzungen

BLAST	Basic Local Alignment Search Tool [1]
bp, kb	Basenpaare, Kilobasen
cDNA	revers-komplementäre DNA (zur mRNA)
cDNA-Array	Glasträger mit lokalisierbaren, aufgetropften DNA-Fragmenten
cDNA-Chip	Synonym zu cDNA-Array
cRNA	revers-komplementäre RNA (zur mRNA)
CDS	proteinkodierender Bereich der mRNA
DNase	Desoxyribonuklease
dNTP	Desoxynukleosidtriphosphat
EST	Expressed Sequence Tag
EST-Cluster	Eine Menge von EST's, deren Sequenzen überlappen und die damit potenziell vom gleichen Transkript stammen.
GAPDH	Glycerinaldehyd-3-phosphat-Dehydrogenase
LOH	Loss of Heterozygosity
mM	Millimolar
MM	Missmatch: 25-Basen langes Oligonukleotid, zum PM-Oligo

[1] Unter Oligonukleotid-Array ist in dieser Arbeit immer der von der Firma Affymetrix hergestellte GeneChip® zu verstehen. Synonym werden benutzt: Oligo-Array, Oligo-Chip, Chip

	identisch, außer Position 13 ist durch das reverse Komplement ausgetauscht
mRNA	Messenger-RNA
NCBI	National Center for Biotechnology Information; Internetseite: www.ncbi.nlm.nih.gov
OD	Optische Dichte
Oligo-Array	Oligonukleotid-Array, GeneChip® von Affymetrix
Oligo-dT	Oligo Desoxythymidin
PCR	Polymerase Chain Reaction (Polymerase-Kettenreaktion)
PM	Perfect Match: 25-Basen langes Oligonukleotid, das als Sonde auf einen Affymetrix-Chip synthetisiert wurde
PMQ	3. Quartil der PM-Intensitäten; auch Kurzbezeichnung für das in Abschnitt 2.3 eingeführte Verfahren
Poly-A-RNA	RNA, die am 3'-Ende eine längere Adeninfolge enthält
Probe	Das zu untersuchende Material (z. B. mRNA, Tumorzellen)
RefSeq	RefSeq [2]steht für Referenzsequenz- Datenbank und wird am NCBI aufgebaut und verwaltet. Zurzeit sind unter anderen 15170 humane Transkripte enthalten, wobei 4848 von Fachwissenschaftlern einzeln überprüft wurden und damit den höchsten Qualitätsstatus *reviewed* erreicht haben. (Stand vom 26. 9. 2002)
RNase	Ribonuklease (RNA spaltendes Enzym)
RNase H	RNase mit Substratspezifität für RNA-DNA-Hybridmoleküle
rRNA	ribosomale RNA
Rohintensitäten	Signalwerte eines Oligo-Arrays, wie sie von der Bildverarbeitung von Affymetrix (MAS 5.0) errechnet und in der *CEL*-Datei gespeichert werden.
SAGE	Serial Analysis of Gene Expression; siehe Abschnitt 1.6
T_m	Schmelztemperatur
Target	Zielprotein für eine Krebstherapie
UTR	Nicht translatierter Bereich der mRNA

1.2. Einbettung und Motivation der Arbeit

Die vorliegende Arbeit entstand in der Zeit von Mai 1999 bis Dezember 2002 bei der Firma metaGen Pharmaceuticals. metaGen arbeitet auf dem Gebiet der Krebsforschung und setzt die Genexpressionsanalyse zur Identifikation im Tumor überexprimierter Gene ein. Aus Gewebematerial von Krebspatienten wird die mRNA isoliert und mit Hilfe einer neuen Technologie, den Oligonukleotid-Arrays, untersucht. Ziel der Arbeit war es, die

Datenauswertung mit Methoden der Statistik und der Informatik zu verbessern. Um die im Kapitel 2 eingeführte Methode im Kontext zu motivieren, muss die Strategie des Einsatzes der Expressionsanalyse in der Krebsforschung erläutert werden. Dazu gehört eine eingehende Betrachtung der Grenzen der Technologie und der zu treffenden Annahmen über das Untersuchungsmaterial. Im Zentrum des Interesses der Firma steht die Untersuchung von Gewebeproben krebskranker Patienten mit bestimmten Indikationen, darunter die häufigsten Tumorerkrankungen Prostata-, Brust-, Kolon- und Bronchialkarzinome. Krebs zählt zu den häufigsten Todesursachen in der westlichen Welt. Körpereigene Zellen mutieren über mehrere Stufen zu malignen Tumorzellen (siehe Kapitel 3). Der Pathologe kann am Gewebeschnitt den veränderten Phänotyp der transformierten Zellen erkennen. Erklärtes Ziel ist die Entwicklung von Krebsmedikamenten, also Substanzen, die die malignen Zellen erkennen und zerstören oder zumindest in ihrem Wachstum behindern. Der erste Schritt auf diesem Weg ist die Auswahl therapeutischer Zielproteine oder **Targets**. Ein therapeutisches Target im Sinne von metaGen ist ein im Tumorgewebe überexprimiertes Protein, das durch Antikörper oder Small-Molecules inaktiviert werden kann. Für eine Antikörpertherapie kommen membranständige oder sezernierte Proteine in Frage. Small-Molecules hingegen wirken meistens als Inhibitoren für Kinasen oder Phosphatasen. Was macht nun die Suche nach solchen Zielmolekülen so schwierig? Krebs ist eine als vielfältige multifaktorielle Krankheit. Sogar innerhalb von Tumorsubtypen, die morphologisch nicht zu unterscheiden sind, findet man unterschiedliche Proteine dereguliert. In den seltensten Fällen lässt sich bislang die Ursache für die maligne Transformation der Zellen bestimmen. So scheinen selbst die Zellen innerhalb eines Tumors eine enorme Variabilität aufzuweisen. Beispielsweise lassen sich über gute molekulare Tumormarker[1] nur etwa 20-30% der histologisch als maligne identifizierbaren Zellen färben (Prof. Dietel, Pathologie, Charité, Berlin). Darüber hinaus sind unsere experimentellen Möglichkeiten zur Untersuchung des Proteoms[2] von Zellen noch

---

[1] Man koppelt einen Farbstoff an einen spezifischen Antikörper für ein Protein, das verstärkt in Tumorzellen zu finden ist. Ein mit diesem Antikörper inkubierter Gewebeschnitt liefert Farbsignale in den Bereichen, in denen das Protein vorhanden ist. Bezeichnung: Immunhistochemie

[2] Proteom ist ein Kunstbegriff und bezeichnet in Anlehnung an das Genom eine Gesamtheit von Proteinen.

vergleichsweise bescheiden. Das heißt, man ist bisher nicht in der Lage, in einem Experiment tausende von Proteinen auf ihre differenzielle Expression hin zu untersuchen. Die Array-Technologie bietet zum ersten Mal die Möglichkeit, einen großen Teil der Gene parallel auf ihre differenzielle Regulation in unterschiedlichen Zellpopulationen hin zu prüfen. Findet man zwischen Tumor- und Normalgewebeproben differenziell exprimierte Gene, so werden diese als potenziell tumorrelevant angesehen. Für einen Therapieansatz sucht man letztendlich jedoch nach differenziellen Proteinen. Dem Einsatz der Expressionsanalyse als Filter für tumorrelevante Proteine liegt also die Hypothese zugrunde, dass zumindest bedingt Gen- und Proteinexpression korrespondieren. Das folgende Schema (Abbildung 1-1) veranschaulicht die Grundannahme für die Strategie des Einsatzes der Expressionsanalyse in der Therapieentwicklung.

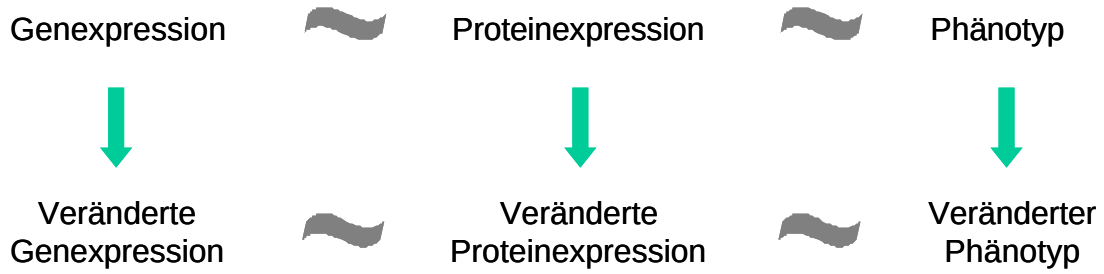


Abbildung 1-1 Schema für die Grundannahme des experimentellen Ansatzes
Die grünen Pfeile symbolisieren eine beobachtbare Veränderung oder Transformation. Ganz rechts ist die Transformation von Zellen als Veränderung des Phänotyps angedeutet (z.B. von benigne nach maligne). Die Transformation ist durch eine veränderte Zusammensetzung der Proteine bedingt. Diese Änderung der Proteinzusammensetzung muss durch eine veränderte Genexpression erzeugt worden sein. Die Schlangen symbolisieren kausale, funktionelle Zusammenhänge.

Inwieweit dieser Ansatz direkt zur Ursachenfindung der Krebserkrankungen beiträgt, lässt sich schwer abschätzen. So ist zum Beispiel nicht bekannt, was letztlich den malignen Phänotyp hervorruft, einige wenige mutierte beziehungsweise deregulierte Proteine oder die graduelle Veränderung vieler Bestandteile eines komplexen Netzwerkes von interagierenden Proteinen und Nukleinsäuren. Die Vielschichtigkeit des Problems soll das folgende Gedankenmodell illustrieren: Nehmen wir an, in einer Zelle gibt es ein Protein, dass normalerweise eine Kontrollfunktion bei der Replikation inne hat aber durch ein Mutation inaktiviert ist. Jetzt habe die Zelle noch einen Rückkopplungsmechanismus, der bei fehlender Funktion die Synthese dieses Proteins induziert. Die Genexpressionsanalyse detektiert also eine Hochregulation. Eine funktionelle Inhibition dieses Proteins hätte aber keine negative Wirkung auf den Tumor. Ist wiederum der Expressionsunterschied zwischen normalen und

Tumorzellen groß genug, ließe sich das trotzdem für einen therapeutischen Ansatz ausnutzen. Mit einem spezifisch an das deregulierte Target bindenden Antikörper kann man eine Anreicherung daran gebundener Zellgifte oder Radioisotope in den Tumorzellen erzielen.

Als Ergebnis der Genexpressionsanalyse erwartet man demzufolge Kandidatengene, die deutlich stärker im Tumor exprimiert sind. Was versteht man aber genau unter differenziell exprimiert? Sucht man eher nach konsistent oder nach stark differenziellen Genen? Was gilt als stark differenziell und welche Unterschiede können mit Hilfe der Array-Technologie aufgelöst werden? Im Folgenden sollen die Ziele für die Verfahrensentwicklung in der Datenanalyse und die damit verbundenen Grenzen der eingesetzten Technologie präzisiert werden. Nehmen wir an, man könnte die mRNA in Zusammensetzung und Menge einzelner Zellen zu jedem beliebigen Zeitpunkt bestimmen. Dann ließen sich mit der Genexpressionsanalyse unter anderem folgende Fragen beantworten:

1. Unterscheidet sich der Gesamt-RNA-Gehalt der Zellen, und falls ja, ist die Absolutanzahl an Transkripten pro Gen und Zelle oder die relative Menge bezüglich des Gesamt-RNA-Gehalts die funktionell entscheidende Größe, und gilt dies für alle Gene gleichermaßen?

2. Welche Gene sind zwischen zwei beliebigen Zellen differenziell exprimiert?

3. Wie groß ist der Unterschied für jedes einzelne Gen und ist dieser für die Zelle von Bedeutung?

4. Wie groß ist die Variabilität der Genexpression zwischen histologisch nicht zu unterscheidenden Zellen innerhalb eines Individuums und auch zwischen Individuen?

5. Was unterscheidet normale Epithelzellen und maligne Tumorzellen?

Aufgrund des hohen RNA-Bedarfs pro Experiment ist es nicht möglich, die RNA-Population von einzelnen Zellen zu messen. Da zusätzlich die Gesamtmenge der RNA, die in jedes Experiments eingesetzt wird, gleich ist und die Anzahl der Zellen, aus denen sie stammt, nicht erfasst ist, geschieht jede Messung relativ zum Gesamtpool. Man misst also immer relative Konzentrationen in Bezug auf den eingesetzten Pool an Messenger-Molekülen niemals Absolutkonzentrationen bezüglich einer festen Zellzahl. Behalten wir die Annahme bei, man könne einen mRNA-Pool in seiner Zusammensetzung genau ausmessen. Weiter setzen wir

voraus, dass der tumorrelevante Unterschied eines Gens in der relativen Expressionsänderung bezüglich des Gesamtpools besteht. Läge nun zum Beispiel die Gewebeprobe eines Lungenkrebspatienten vor mit zwei gut unterscheidbaren Zellphänotypen, normales Lungenepithel und Plattenepithelkarzinom (siehe Kapitel 3). Die Proben seien so beschaffen, dass man ausreichend viel mRNA isolieren kann. Dann lässt sich unter anderem folgendes untersuchen:

1. Welche Gene sind zwischen den Zelltypen differenziell relativ zum Gesamt-RNA-Pool?

2. Gibt es Gene, deren Expression praktisch Null in einem der Zelltypen ist, und sind vielleicht gerade solche relevant?

3. Wie groß ist der Unterschied für jedes einzelne Gen zwischen den Zellpools und ist dieser von Bedeutung?

4. Sind eventuell nur Gene interessant, deren Expressionsunterschiede zwischen den Zelltypen eine bestimmte Mindestgröße übersteigen?

5. Wie groß ist die Variabilität der Genexpression zwischen histologisch nicht zu unterscheidenden Zellpopulationen innerhalb eines Individuums und auch zwischen Individuen?

6. Gibt es Gene, die sehr konsistent exprimiert sind?

Die Bedeutung oder Relevanz bestimmter Eigenschaften der Proben lässt sich durch die Untersuchung großer Patientengruppen prüfen. Bei der Interpretation der Ergebnisse ist aber Vorsicht vor falschen Schlussfolgerungen geboten. Man analysiert tausende von Genen auf 50 oder 100 Gewebeproben. Es handelt sich also um ein hochgradig überbestimmtes System, und ein als differenziell exprimiert gefundenes Gen sollte deshalb immer als vorläufiges Resultat angesehen werden. Bei metaGen setzt man komplementäre Methoden der Expressionsanalyse zu dessen Bestätigung ein. Die verwendete Technologie weist eine bestimmte Sensitivität und Messfehler auf, was den Gültigkeitsbereich gewonnener Aussagen weiter einschränkt. So ist man zum Beispiel durch die Detektionsschwelle nicht in der Lage, das Nichtvorhandensein eines bestimmten Transkripts nachzuweisen.

Das Anwendungsfeld der Oligo-Arrays erschöpft sich aber nicht in der Suche nach zwischen Zellpopulationen differenziell exprimierten Genen. Die Expressionsanalyse vieler tausend Gene bietet zum Beispiel auch die Möglichkeit, neue, hypothetische Transkripte zu verifizieren. Nach der Sequenzierung des Genoms eines Organismus sind die Charakterisierung der Gene und deren Expression der nächste logische Schritt. Viele der Gene werden mit Hilfe von Computerprogrammen (*in silico*) auf der Basis von Sequenzeigenschaften vorhergesagt. Ob und wenn ja unter welchen Umständen sie tatsächlich transkribiert werden, ist dann nicht bekannt. Die Gene höherer Organismen weisen tendenziell wesentlich komplexere Strukturen auf. Sie sind länger, haben mehr Exons und werden oft alternativ gespleisst[1]. Oligo-Arrays erlauben die systematische Prüfung hypothetischer Transkripte und Transkriptvarianten und die Suche nach bisher unentdeckten Genen (Kapitel 4). Die Expressionsanalyse lässt sich auch zum Studium des zellulären Prozesses der Genexpression einsetzen. So ist man mit der Array-Technologie beispielsweise in der Lage, allgemeine Eigenschaften aller hoch exprimierten Gene aufzudecken (Kapitel 5). Sorgfältig konstruierte Zellsysteme sollten Aussagen über Synthesekapazitäten von mRNA's zulassen.

1.3. Die Beschreibung der verwendeten Arrays

Dieser Abschnitt beschreibt Oligonukleotid-Arrays, wie sie von der Firma Affymetrix (Santa Clara, USA) hergestellt werden. Man bezeichnet sie auch als GeneChips oder einfach als Chips. Oligo-Chips sind von einer Plastikkapsel umhüllte Glasträger, 12,8 mm^2. Auf die Glasoberfläche eines solchen Trägers können zurzeit bereits hunderttausende verschiedene Oligonukleotide[2], präzise lokalisiert, synthetisiert werden. Man bedient sich hierbei moderner Fertigungsverfahren unter Reinstraumbedingungen, wie sie für die Herstellung von Computer-Chips eingesetzt werden. Folglich kann man auf Know-How in der Fertigung bei der Präzision und Miniaturisierung zurückgegreifen. Man verliert allerdings an Flexibilität. Im Vergleich dazu werden bei cDNA-Arrays (siehe Abschnitt 1.6) PCR-Produkte mit einem

---

[1] Für eine ausführliche Betrachtung siehe Kapitel 5.

[2] Als Oligonukleotid bezeichnet man ein einzelsträngiges DNA-Molekül weniger Basen Länge. (oligo: griechisch einige) Synonym werden Oligomer oder kurz Oligo benutzt.

Spotting-Roboter auf eine Glasoberfläche gebracht. Sie lassen sich damit im molekularbiologischen Labor herstellen und direkt auf sich ändernde Bedürfnisse anpassen.

Die Glasoberfläche der Rohlinge der Oligo-Chips ist beschichtet, um bei der Synthese eine gute Kopplung der Desoxynukleotide zu ermöglichen. Vor das eigentliche Oligonukleotid muss noch ein Linker auf die Oberfläche synthetisiert werden. Der Linker dient als Abstandhalter, so dass das Oligonukleotid sich freier in der Hybridisierungslösung bewegen kann und damit die Bindungseffizienz erhöht. Die eigentliche Oligomer-Synthese erfolgt dann Base für Base oder Schicht für Schicht. In jeder Schicht werden lichtempfindliche 5'-Schutzgruppen selektiv von den wachsenden Oligomeren entfernt. Jeder gerade gekoppelte Baustein blockiert seinerseits durch Schutzgruppen das Weiterwachsen der Kette. Das genaue Positionieren der Lichtstrahlen erreicht man mit Hilfe photolithografischer Masken. Derzeit beträgt die Länge der Oligomere, die auf den Affymetrix-Chips synthetisiert werden 25 Nukleotide. Für jedes Oligonukleotid (***PM*-Oligo**, *perfect match*) ist auf dem Chip eine spezifische Negativkontrolle das *Missmatch*-Oligo (***MM*-Oligo**) synthetisiert, bei dem die 13-te Base durch ihr Komplement ausgetauscht ist. Bei metaGen Chip I erreichte man eine Zellengröße von 24x24 µm, bei dem neusten Chip von Affymetrix bereits 18x18 µm. Damit können auf Chip I 526x526 und auf dem neuen Chip 712x712 verschiedene Oligospezies synthetisiert und lokalisiert werden. Die Standardsonde für die Untersuchung einer bestimmten Zielsequenz besteht auf Chip I aus 20 spezifischen Oligonukleotiden plus die 20 entsprechenden *MM*-Oligos. Bei Affymetrix wird ein solcher Satz von Oligo-Sonden als *Probeset* bezeichnet. In dieser Arbeit ist aber ausschließlich die Bezeichnung **Sondenset** in Gebrauch, da das Wort Probe sonst irreführend für die untersuchte RNA oder für die Chip-Sonde stehen könnte. Das Prinzip der Anordnung der Paare ist auf den beiden Chips unterschiedlich. Liegen auf Chip I die zu einer Sonde gehörenden Oligos zusammen nebeneinander, so besteht auf den neuen Chips ein Standardset nur noch aus 11 Oligopaaren, die über den Chip verteilt liegen.

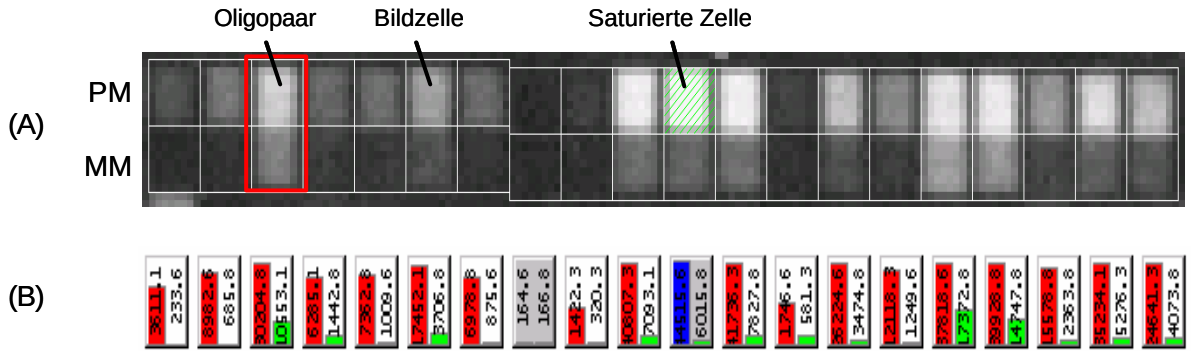


Abbildung 1-2 Beispiel für Scannerbild eines Sondensets auf Chip I
(A) *Dargestellt ist ein, vergrößerter Ausschnitt aus einem Chipbild, der alle Bildzellen eines Sondensets enthält. Die weißen Linien sind das Ergebnis der Bildanalyse, in der die Zellen akkurat lokalisiert werden müssen. Ein Oligopaar besteht aus einem PM-Oligo und einem MM-Oligo.* ***(B)*** *Das Ergebnis der Bildanalyse der Affymetrix-Software MAS 5.0 (siehe Abschnitt 2.1) Der rote Balken entspricht dem PM-Signal und der grüne dem MM-Signal. Ist das Paar grau hinterlegt (maskiert), so gehen die Signale nicht in die weitere Analyse ein.*

Zusätzlich zu der zu untersuchenden RNA bringt man noch markiertes B2-Oligo auf den Chip. Hier sind dafür bestimmte, wohldefinierte Kontrollregionen angelegt worden, wie beispielsweise ein Schachbrettmuster in jeder Ecke. Sie ermöglichen es der Bildanalysesoftware, ein Liniennetz über den Chip zu legen und damit die Oligospezies exakt zu lokalisieren. Zusätzlich vermitteln diese Regionen einen Eindruck davon, ob die Synthese der Oligos gut funktioniert hat. Die Sensitivität oder Detektionsschwelle der Chips gibt Affymetrix mit einem Molekül in hunderttausend an.

1.4. Die Sequenzsets der Arrays

Technisch ist man derzeit in der Lage, Chips mit etwa einer halben Million verschiedener, 25 Basen langer Oligonukleotide herzustellen. Für niedere Organismen, wie Bakterien, lässt sich bereits das komplette Genom auf einem Chip repräsentieren. Plant man die Genexpression höherer Organismen zu untersuchen, so konstruiert man Sonden für jedes potenzielle Transkript. Da die Oligos mit 25 Basen relativ kurz sind, und die Vorhersage ihres Hybridisierungsverhaltens bislang nicht möglich ist, wählt man immer ein Set von Oligos als Sonde. Für den aktuellen humanen Chip von Affymetrix (siehe Abschnitt 1.3) besteht das Sondenset für jedes Transkript aus 11 sequenzspezifischen Oligonukleotiden. Bei der Auswahl hat man darauf geachtet, dass es keine Transkripte gibt, die fälschlicherweise mit einem der Oligos kreuzhybridisieren können und das die Sequenz keine extreme Basenkomposition aufweist. Auf einem Chip dieses Typs lassen sich etwa 20000 Transkripte

repräsentieren. Für einige Eukaryonten, deren Transkriptome[1] bereits gut charakterisiert sind, lässt sich die Expression aller Gene mit einem sorgfältig konstruierten Chip messen. Will man humane Proben untersuchen, treten zwei Schwierigkeiten auf:

1. Das Transkriptom des Menschen besteht aus mehr als 20000 unterschiedlichen mRNA's. Aus diesem Grunde bietet Affymetrix zum Beispiel für die Untersuchung humaner Proben zwei komplementäre Chips an.

2. Das humane Transkriptom ist nur zum Teil gut charakterisiert und aufgrund der Grösse des Genoms ist man nicht in der Lage, für alle potenziellen Transkripte Sonden zu generieren.

Bei metaGen wurde 1998 ein Chip konstruiert, **metaGen Chip I**. Als Basisset dienten damals dbEST[2] und die nicht-öffentliche EST-Datenbank von Incyte Pharmaceuticals. Die EST's wurden zu Clustern zusammengefasst, und man generierte Konsensussequenzen. Als Konsensussequenz bezeichnet man das aus dem Sequenzcluster geschätzte mRNA-Fragment von dem die EST's ursprünglich sequenziert wurden. Zusätzlich konnte man sich noch die Information über den Gewebetyp, aus dem die EST's stammten, zunutze machen. Für die Chipkonstruktion wählte man besonders solche Konsensussequenzen aus, für die sich die relative Zahl der EST's zwischen Bibliotheken aus Tumor- beziehungsweise Normalgewebe unterschied [3]. Außerdem wählte man noch eine Reihe weiterer Sequenzen aus, die entweder schon als tumorrelevant in der Literatur beschrieben waren oder deren Proteinfamilien besonders interessierten. Letztlich bestand das Set für Chip I aus 3950 Sequenzen und bei Affymetrix konstruierte man dafür insgesamt 6117 Oligosets zu meist 20 Oligopaaren. Darunter waren 1066 cDNA-Fragmente, für die die 3'-5'-Orientierung des Transkripts nicht bekannt war. Für dieses Set konstruierte man pro Orientierung ein Oligoset. Zusätzlich konnten für 588 Sequenzen, von denen der kodierende Bereich und das 3'-UTR gut

---

[1]Transkriptom ist ein Kunstbegriff, der die Gesamtheit aller RNA-Moleküle eines Organismus bezeichnet, die jemals von irgendeinem Gen abgeschrieben werden können.

[2] dbEST ist der Teil von GenBank, der nur EST-Sequenzen enthält. GenBank ist eine Sammlung aller öffentlich verfügbaren DNA-Sequenzen. Zurzeit sind etwa 4,5 Millionen EST-Sequenzen in dbEST.

charakterisiert waren, jeweils ein Oligoset pro Sequenzregion konstruiert werden. Siehe hierzu auch Kapitel 4.

Seit Anfang 2002 sind die Affymetrix Arrays **HG-U133 A** und **B** bei metaGen im Einsatz. HG steht dabei für Humangenom und U133 für die zur Konstruktion verwendete Version von UniGene [4]. Auf den beiden Chips sind insgesamt über eine Million unterschiedliche Oligonukleotide synthetisiert, die in etwa 45000 Sondensätzen organisiert sind. Die verschiedenen Oligosets repräsentieren laut Aussage von Affymetrix über 39000 Transkriptvarianten von über 33000 humanen Genen. Als Basis für den Auswahlprozess dienten die Sequenzcluster vom UniGene Build 133[1] vom April 2001. Zusätzliche Datenquellen boten dbEST, die Datenbank von Sequenzrohdaten (trace repository) der Washingtoner Universität und die Genomdatenbank Golden Path der Universität von Kalifornien in Santa Cruz. Damit war man in der Lage, zusätzliche Filter und Qualitätskontrollen auf die Sequenzcluster anzuwenden. Sequenzen geringer Qualität oder solche die beispielsweise durch fehlerhaftes Clustern falsch zugeordnet waren, konnten identifiziert und eliminiert werden. Außerdem identifizierte man eine große Zahl Transkripte mit alternativen Spleißvarianten bzw. Polyadenylierungsstellen, die dann durch zusätzliche Sondensets auf den Chips repräsentiert wurden. Für Hybridisierungskontrollen sind auf den beiden Chips HG-U133 A und B Sondensets für die bakteriellen Gene *bioB, bioC, bioD* und *cre* vorhanden und für poly-A -Kontrollen die Gene *dap*, *lys*, *phe* und *thr*. Für den direkten Vergleich von Chip A und B sind auf beiden die gleichen 100 Sondensets für häufig exprimierte Gene synthetisiert worden. Außerdem findet man noch die Standardkontrollen *GAPDH, ß-Actin* und *ISGF-3*. Der Vorgänger von HG-U133 war der Chipsatz HG-U95, der auf Unigene Build 95 basierte. Etwa 12000 gut charakterisierte Gene sind auf dem Chip **HG-U95Av2** repräsentiert und weitere 50000 weniger gut bekannte Sequenzcluster verteilt auf den Chips HG-U95 B bis E. Die Sondensets auf diesen Chips bestehen aus 16 spezifischen Oligopaaren, und die Größe der Bildzellen beträgt 20 μm^2. Der Chip HG-U95Av2 wurde unter anderem in der Expressionsstudie der Bronchialkarzinome verwendet, die die Grundlage für Kapitel 3 bildet.

[1] http://www.ncbi.nlm.nih.gov/UniGene/build.shtml

1.5. Chipexperiment

Dieser Abschnitt gibt einen Überblick über die Chipexperimente, wie sie bei metaGen durchgeführt werden. Die entsprechende Laborarbeit war nicht Teil dieser Promotion und deshalb findet keine detaillierte Beschreibung der Protokolle statt[1]. Ziel der vorliegenden Darstellung ist ein prinzipielles Verständnis des Experiments und seiner Komplexität. Einige Effekte, die bei der Datenanalyse auftreten, können hier ihre Erklärung finden.

Der gemeinsame Ausgangspunkt für jedes Chipexperiment ist die Gesamtheit Messenger-RNA oder mRNA eines Zellpools, die entweder aus Gewebeproben oder aus Zellkulturen stammt. Im Gegensatz zu Zellkulturen, bei denen im Normalfall ausreichend viele Zellen mit ähnlichen Eigenschaften vorhanden sind, ist es bei Gewebeproben weitaus schwieriger, die gewünschten Zellpools zu bekommen. Bei metaGen untersucht man beispielsweise Gewebeproben solider Tumoren, die ein komplexes Gemisch verschiedener Zelltypen darstellen. Unterschiedliche Zusammensetzungen der Zellpools können Unterschiede in der mRNA erzeugen. Idealerweise möchte man einzelne Zellen untersuchen und durch den Vergleich der Charakteristika vieler Einzelzellen auf die für die maligne Transformation entscheidenden Faktoren schließen. Limitierend ist hier jedoch die große Menge an RNA, die die Verfahren verlangen, um verlässliche Signale zu liefern. Exemplarisch sei hier die Berechnung zum Mengenbedarf für ein cDNA-Array-Experiment von [5] sinngemäß zitiert: Die Menge an Gesamt-RNA, die pro Zellpool und Chip eingesetzt wird, beträgt etwa 50-200 µg oder 2-5µg poly-A-mRNA. Für ein Transkript, das nur als einzelne Kopie pro Zelle vorliegt, was etwa ein Molekül auf 100 000 entspricht, bedeutet das, dass etwa 300 dieses Typs für eine Hybridisierung der Sonde mhe genug kommen. Bei der Rechnung ging man von 100 µg Ausgangsmenge Gesamt-RNA, einer Arrayfläche von 800 mm^2 und einem Spotdurchmesser von 200 µm aus. Setzt man weiter voraus, dass die Transkripte durchschnittlich 600 bp lang sind und dass die Einbaurate etwa zwei fluoreszenzmarkierte Nukleotide auf 100 bp beträgt, so erwarten die Autoren etwa 12 Farbmoleküle pro 100 μm^2 gescannter Fläche, was an der unteren Detektionschwelle der Scanner liegt und damit kaum von Rauschsignalen unterscheidbar ist. metaGen setzt die Laser-gestützte Mikrodissektion ein, um hinreichend homogene Zellpools aus Gewebeschnitten zu erhalten. Da dieses

[1] Promotionsarbeit von Christoph Wissmann (Biochemie, FU-Berlin)

Verfahren sehr arbeitsintensiv ist, amplifiziert man die isolierte mRNA, um den Aufwand vertretbar zu halten und trotzdem die für Chipexperimente erforderlichen Mengen Ausgangsmaterial aus genau charakterisierten Zellen zu erreichen. Bei der Mikrodissektion schneidet man 10 µm dicke Scheiben von tiefgefrorenen Gewebeproben und bringt diese Schnitte auf Objektträger. Nach der Beurteilung durch einen Pathologen schneidet man Computer gestützt mit einem Laser die interessierenden Zellareale aus, kann diese dann vom Objektträger ablösen und in Reaktionsgefäßen zu den entsprechenden Zellpools sammeln. Pro Probe müssen 15 bis 40 Schnitte mit etwa 60 mm^2 Fläche disseziert werden. Nach groben Schätzungen entspricht das zwischen 100000 und 500000 Zellen. Das folgende Schema (Abbildung 1-3) veranschaulicht die einzelnen Schritte eines Chipexperiments.

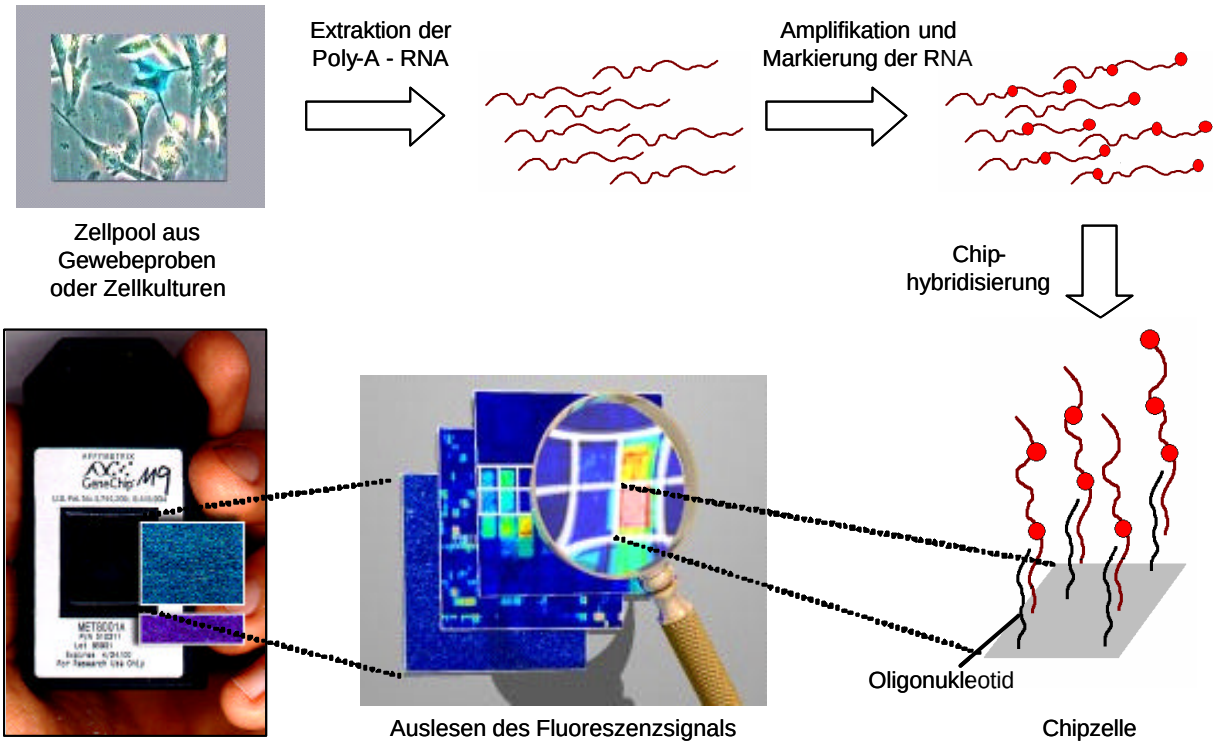


Abbildung 1-3 Schema eines Oligo-Chip-Experiments

Links unten sieht man einen Affymetrix Chip im Standardformat. Der eigentliche Oligoträger ist in einer Plastikhülle eingebettet, die mit einem Deckglas abgeschlossen ist und in die die Reagenzien über ein Ventil eingefüllt werden.

Die Zellen des Zellpools werden lysiert und anschließend homogenisiert. Zur Extraktion der Poly-A-RNA inkubiert man das Lysat mit biotinylierten Oligo-dT-Primern. Nach dem Abtrennen der Zelltrümmer gibt man an Streptavidin gekoppelte paramagnetische Partikel zu der Lösung. Streptavidin bindet kovalent an Biotin. Dadurch binden die paramagnetischen Partikel über das inkorporierte Biotin an die mRNA-Primer-Komplexe. Anschließend rührt man die Lösung auf und stellt das Reationsgefäß in einen magnetische Ständer, so dass sich

die Partikel zusammen mit den Poly-A-RNA-Molekülen an die Gefäßwand heften. Alle anderen Bestandteile des Zelllysats verbleiben in der Lösung, die dann abgenommen werden kann. Mit Hilfe von DEPC-Wasser trennt man die RNA-Moleküle wieder von den Partikeln. Viele Gruppen extrahieren Gesamt-RNA und nicht Poly-A-RNA und setzen die dann in die cDNA-Synthese ein.

Die aufgereinigte RNA wird anschließend in cDNA umgeschrieben. Dazu benutzt man Primer, die sich aus 24 Thyminbasen (Oligo-dT) am 3'-Ende und der T7- Promotersequenz am 5'-Ende zusammensetzen. Die Oligo-dT-Sequenz bildet ein Hybrid mit dem Poly-A-Ende der mRNA und dient somit als Primer für die reverse Transkriptase, die ausgehend von einem RNA-Molekül den revers-komplementären DNA-Strang erzeugt. Das verwendete Enzym hat keine RNase-H-Aktivität[1], so dass RNA-DNA-Hybridmoleküle entstehen, deren Gesamtheit als cDNA bezeichnet wird. Die cDNA ist ein Abbild des Transkriptomes eines Zellpools in einem bestimmten Zustand und ist wesentlich stabiler als die RNA. Deshalb ist sie meistens der Ausgangspunkt für Untersuchungen der Genexpression.

[1] RNase ist ein Enzym, das RNA-Moleküle spaltet. RNase-H ist eine RNase, die RNA verdaut, die Teil eines DNA-RNA-Hybridmolekül ist.

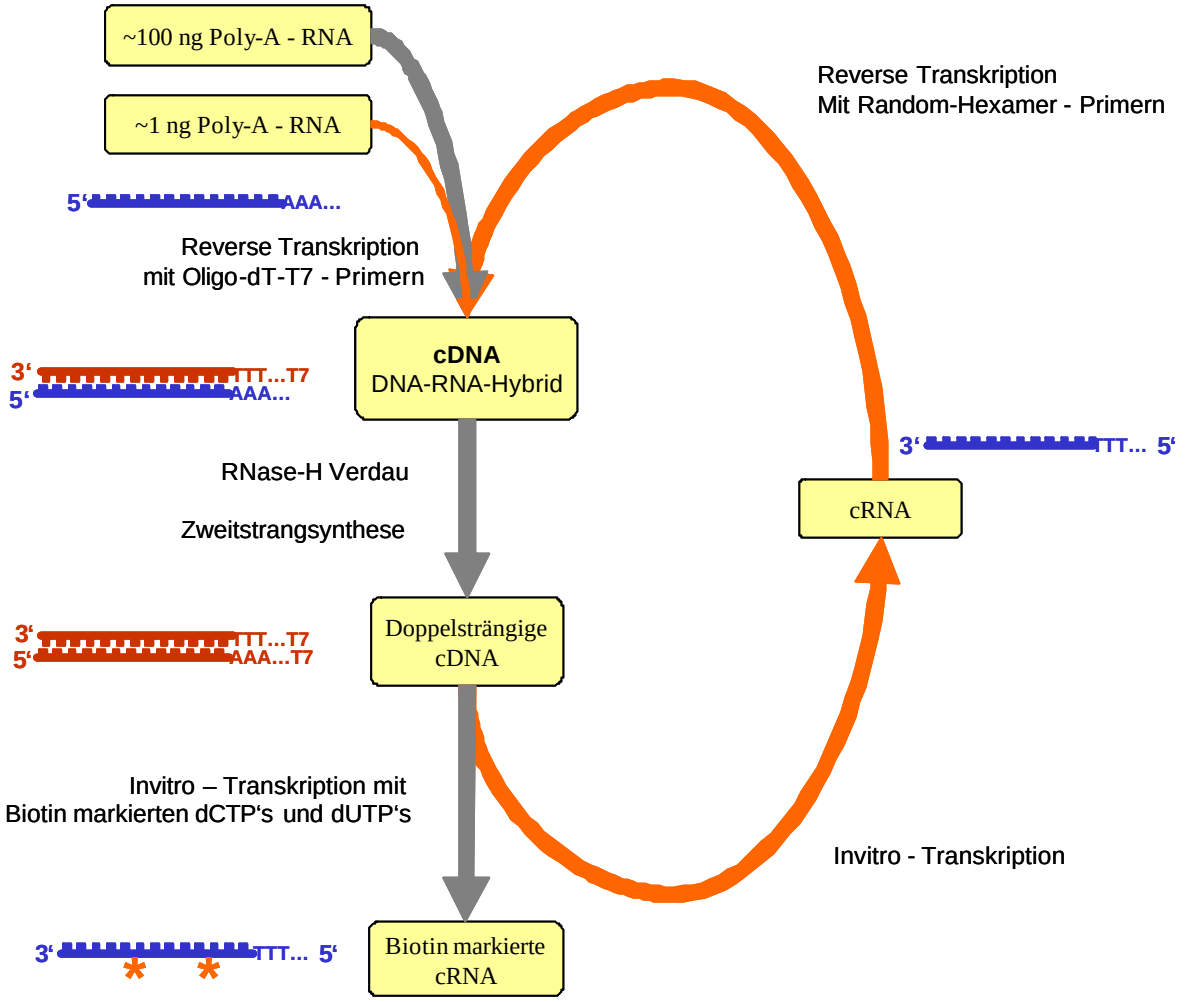


Abbildung 1-4 Schema der cRNA-Synthese für die Chiphybridisierung

Die grauen Pfeile markieren den Standardweg, der von Affymetrix vorgeschlagen wurde und bei uns für die Aufbereitung von RNA aus Zelllinien genutzt wird, da hier ausreichend Zellen vorhanden sind. Der orange Weg stellt die Voramplifikation dar, die zweimal durchlaufen wird. Die blauen und dunkelroten gezackten Streifen symbolisieren RNA- und DNA-Moleküle und die Orientierung der Stränge ist durch 3’ und 5’ markiert.T7 bezeichnet die Promotersequenz für die T7-RNA-Polymerase. Die orangen Sterne unten links deuten die Biotinmarkierung an.

Um die Qualität der cDNA zu prüfen, nimmt man 1µl ab und amplifiziert und quantifiziert das Kontrollgen Succinatdehydrogenase. Als Referenz benutzt man eine cDNA der Prostata-Zelllinie DU145, die man zu diesem Zweck vorher exakt auf ein Nanogramm je Mikroliter

konzentriert hat. Jetzt misst man per TaqMan-PCR[1] das Kontrollgen in beiden cDNA's. Erreicht neu synthetisierte cDNA vergleichbare Werte wie die Referenz, kann man davon ausgehen, dass die mRNA nicht degradiert und die cDNA-Synthese erfolgreich war.

Will man für die Chiphybridisierung komplementäre RNA (cRNA) gewinnen, benötigt man doppelsträngige DNA mit der entsprechenden Promotersequenz als Template für die T7-RNA-Polymerase. Dazu setzt man einen RNase-H-Verdau des DNA-RNA-Hybrids an, so dass nur noch Bruchstücken der RNA auf dem DNA-Strang zurückbleiben. Diese Bruchstücke dienen bei der anschließenden Zweitstrangsynthese als Primer. Dabei synthetisiert man mit Hilfe der *E. coli* DNA-Polymerase den revers-komplementären Strang, und es entsteht das doppelsträngige DNA-Template. Nach dem von Affymetrix empfohlenen Standardprotokoll folgt nun die In-Vitro-Transkription mit der T7-RNA-Polymerase unter Präsenz biotinylierter CTP's und UTP's. Das Ergebnis sind RNA-Moleküle, die revers-komplementär zur mRNA sind. Bevor diese dann auf den Chip hybridisiert werden, unterzieht man sie noch einer Fragmentierung in dem man sie für 10 min auf 97 °C erhitzt. Idealerweise hätten nun alle RNA-Fragmente die gleiche Länge, die gleiche Hybridisierungskinetik, die gleiche Anzahl inkorporierter Biotinmoleküle und die gleiche hohe Spezifität zu dem entsprechenden Oligonukleotid auf dem Chip.

Da der Prozess der Mikrodissektion, wie oben erläutert, sehr aufwendig ist und damit pro Gewebeprobe nur etwa ein Nanogramm mRNA gewonnen wird, muss diese vervielfältigt werden. Hierfür generiert man die cRNA ohne biotinylierte Nukleotide und erzeugt anschließend mit der reversen Transkriptase erneut RNA-DNA-Hybridmoleküle diesmal allerdings nicht mit Oligo-dT-Primern sondern mit Random-Hexamer-Primern. Damit hat man das Ausgangsprodukt für die Zweitstrangsynthese und die sich anschließende In-Vitro-Transkription vervielfältigt. Diese Amplifikationsrunde wird noch ein zweites Mal wiederholt bevor in der dritten die cRNA-Moleküle wie im Standardprotokoll eine Biotinmarkierung erhalten. Die eigentliche Vervielfältigung kommt dadurch zustande, dass die T7-Polymerase von einem DNA-Template nicht nur ein cRNA-Molekül erzeugt sondern zig.

---

[1] TaqMan-PCR ist eine PCR, die mit einem Gerät durchgeführt wird, das eine Fluoreszenzmessung während der Amplifikationszyklen erlaubt. Damit ist man in der Lage, einzelne Sequenzen relativ zueinander genau zu quantifizieren.

Zusammen mit der so präparierten RNA hybridisiert man dann den standardisierten Kontrollmix von Affymetrix für 16 Stunden bei 42 °C auf den Chip. Pro Experiment setzt man 5-40 µg Gesamt-RNA oder 0,2-5 µg poly-A-mRNA ein. Der Kontrollmix enthält bereits markierte und fragmentierte RNA-Moleküle zumeist bakteriellen Ursprungs. Nach der Abnahme der Hybridisierungslösung und einiger Waschschritte folgen zwei Färbeschritte mit biotinyliertem Streptavidin und Streptavidin-gekoppeltem Phycoeritrin. Nochmaliges Waschen schließt das Laborprotokoll ab, und der Chip kann nun mit dem Scannen beginnend ausgewertet werden.

1.6. Alternative Hochdurchsatzverfahren zur Genexpressionsanalyse

Außer den hier vorgestellten Oligonukleotid-Arrays finden noch andere Hochdurchsatzverfahren ihre Anwendung in der Genexpressionsanalyse. Zu den verbreitetsten Technologien zählen cDNA-Arrays und SAGE (Serial Analysis of Gene Expression), deren Grundprinzipien in diesem Abschnitt erläutert werden.

cDNA-Arrays (auch Microarrays genannt) stellen für viele Anwendungen eine direkte Alternative zu Oligo-Arrays dar. Die Entscheidung für eine der Plattformen fällt dann oft nach der Verfügbarkeit der Geräte oder aufgrund einer Kostenabschätzung. Ein solches Array besteht aus einem Glasträger, der derart beschichtet wird, dass die Oberfläche gut DNA bindet. Die Produkte, die als Sonden auf den Träger aufgebracht werden, sind meistens Klone bzw. PCR-Produkte aus cDNA-Banken oder auch Oligomere (50-60 bp). Sie liegen in Reaktionsgefäßen vor, die zur besseren Handhabe in Platten organisiert sind. Bevor man die DNA-Produkte auf das Array bringt, denaturiert man sie durch Erhitzen in einer speziellen Pufferlösung. Ziel ist es, später auf der Glasoberfläche in den Spots möglichst große einzelsträngige Sequenzbereiche zu fixieren, die dann bei der Hybridisierung die Probe gut binden. Mit dem Einsatz von Robotertechnik bringt man ein Tröpfchen aus jedem Reaktionsgefäß an einen exakt lokalisierten Platz auf dem Glasträger. Man ist bereits in der Lage 24000 Spots auf einem Objektträger unterzubringen. Die Tropfen werden auf dem Array eingetrocknet, und anschließend bindet man die DNA kovalent an die Glasoberfläche indem man das Array bäckt (bei etwa 80°C) oder mit UV-Licht bestrahlt (UV-Crosslinking).

Vergleichbar der Vorbereitung von Chip-Experimenten poolt man hier ebenfalls die Zellen und extrahiert die RNA. Ein Richtwert für die Ausgangsmenge ist 5µg mRNA. Zur cDNA-

Synthese dient eine Reverse-Transkriptase. Nach einem Protokoll von [6] benutzt man Oligomere bestehend aus 21 T-Nukleotiden als Primer und setzt zusätzlich zu den normalen dNTP's mit einem Fluoreszenzfabstoff markierte dCTP's ein. Als Farbstoffe sind Fluorescein-12, Cy3 und Cy5 gebräuchlich. Es entstehen mRNA-DNA-Hybridmoleküle. Die verwendete Transkriptase hat in diesem Falle eine RNase-H-Aktivität, so das die mRNA degradiert wird. Nach der Aufreinigung hybridisiert man die cDNA auf ein Array für sechs Stunden bei 62°C. Eine häufig verwendete Form ist die konkurrierende Hybridisierung von zwei verschieden markierten cDNA-Proben auf ein Array. Untersucht man beispielsweise Tumorzellen und normale Gewebezellen eines Patienten, so markiert man bei der cDNA-Synthese für die Zellen des einen Typs mit Cy3 und für die des anderen Typs mit Cy5. Jetzt hybridisiert man ein Gemisch der beiden markierten cDNA's auf ein Array. Misst man dann bei unterschiedlichen Wellenlängen die Intensitäten der beiden Farbstoffe auf einem Spot, so sollte sich in dem Signalverhältnis das Konzentrationsverhältnis der markierten Transkripte widerspiegeln. Der Hintergrund dieses Vorgehens ist, dass man Spotunregelmäßigkeiten zwischen Arrays von vornherein ausgleichen möchte. Ist beispielsweise in einem Spot die Sonde niedriger konzentriert und dadurch das Signal abgeschwächt, sollte trotzdem das Verhältnis der Signale zweier konkurrierend hybridisierter cDNA's auf diesem Spot gewahrt bleiben. Bei der Datenanalyse korrigiert man für jeden Farbkanal einzeln das Signal mit dem für den Spot spezifischen Hintergrund. Für weitere Analysen verwendet man den logarithmierten Quotienten der korrigierten Signale der beiden Kanäle. Dieser Wert beschreibt eine relative Expressionsänderung und wird im Folgenden mit *LogRatio* bezeichnet.

Eine andere verbreitete Technologie zur Genexpressionsanalyse nennt sich *Serial Analysis of Gene Expression* oder kurz **SAGE**[1] [7]. Ausgehend von einem mRNA-Pool führt unter Verwendung von 5'-biotinylierten Oligo-dT-Primern eine Erstrangsynthese durch. Es schließt sich ein RNase-H-Verdau und die Zweitstrangsynthese an. (Vergleiche 1.5) Als nächstes schneidet man die cDNA mit einem Restriktionsenzyms[2] (z.B. NlaIII). Durch die Kopplung

[1] Umfangreiche Informationen sind über die Seite http://www.sagenet.org verfügbar.

[2] Restriktionsenzyme erkennen eine bestimmte Basenfolge (Restriktionsstelle) und schneiden die DNA in definierter Art und Weise.

magnetischer Partikel über das Biotin der Primer lässt sich ein Teil der cDNA's isolieren. Man gewinnt einen Pool von Sequenzen, die gerade die 3'-Bereiche der mRNA-Moleküle vom Poly-A-Ende bis zur ersten Restriktionsstelle abdecken. Als nächstes ligiert man Linkersequenzen an die Fragmente und schneidet dann so, dass Sequenzen entstehen, die aus dem Linker und der 9 bp langen Tag-Sequenz bestehen. Diese Tags sind die spezifischen Signaturen, mit deren Hife später entschieden wird, welches Transkript hier vorlag. Aufgrund ihrer Kürze sind sie nicht immer eindeutig, und deshalb versucht man neuerdings Enzyme zu verwenden, die längere Tags produzieren. Im nächsten Schritt legiert man immer zwei Tags zu so genannten Ditags. Die Linker befinden sich jeweils außen und enthalten Primer-Bindungsstellen, so dass die Ditags vervielfältigt werden können. Anschließend entfernt man die Linkersequenzen und legiert die verbleibenden Ditags zu längeren Konkatameren. Die Produkte trennt man in einem Agarosegel der Länge nach auf und isoliert die Sequenzen zwischen 600 und 2500 bp. Nach der Aufreinigung werden diese schließlich kloniert und sequenziert. Im Ergebnis erhält man eine Liste von Tags und der entsprechenden Häufigkeit ihres Auftretens. Für eindeutig zuweisbare Tags schließt man aus ihrer Häufigkeit auf das Expressionsniveau des Transkripts.

2. Datenanalyse von Affymetrix GeneChips

Inhalt dieses Kapitels ist die Einführung und Validierung eines alternativen Verfahrens zur Auswertung von Oligo-Array-Experimente. Dieses Verfahren ist der zentrale Grundstein der vorliegenden Promotionsarbeit. Vor der Darstellung und Diskussion des Algorithmus wird ein Einblick in die Qualität der Rohdaten und die damit verbundenen Grenzen dieser Array-Technologie vermittelt. Für die effiziente Nutzung der Expressionsdaten sind darüber hinaus eine sorgfältige Proben- und Sequenzannotation und ein durchdachtes Datenmanagement unerlässlich.

2.1. Die Bildanalyse und die Beschreibung der Rohdaten

Die kleinste Einheit auf dem Bild des Chips, die der Scanner auflöst ist das Pixel von *3x3* µm Größe. Die kleinste Einheit, die auf dem Chip bei der Oligosynthese adressiert wird und die man dann für die weitere Datenanalyse benutzt, ist die Chip- oder Bildzelle. Sie ist auf dem metaGen Chip I *24x24* µm und auf den neuesten Chips *18x18* µm groß. Der Scanner tastet jede Chipzelle mit *8x8* beziehungsweise *6x6* Pixel ab. Die Bildanalysesoftware bietet die Möglichkeit einer visuellen Qualitätsprüfung jedes Chips. Artefakte, die beispielsweise durch Chips mit Produktionsfehlern oder durch verunreinigte Chemikalien auftreten, erkennt man durch einfache Betrachtung des Scannerbildes. Kleinere Artefakte können mit Hilfe von Qualitätstags markiert und für die weitere Analyse maskiert werden, so dass man nicht gezwungen ist, das ganze Chipexperiment zu verwerfen.

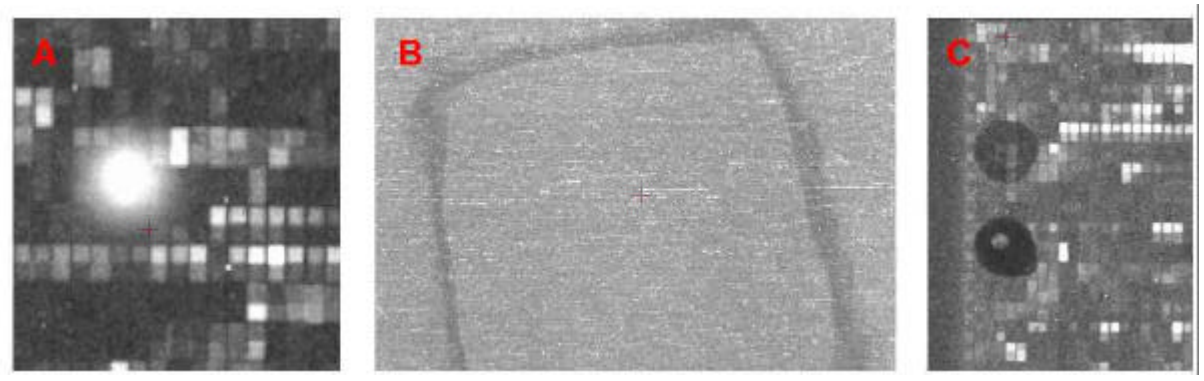


Abbildung 2-1 Typische Artefakte, die beim Scannen beobachtet werden
***Links**: Reflexion des Laserlichtes, **mitte**: Kratzspuren, **rechts**: vermutlich ein Synthesefehler*

Für die Prozessierung der Bilddaten wird von der Software über das Chipbild ein Netz aus Linien gelegt, so dass jede Zelle lokalisiert und dem richtigen Oligo zugeordnet werden kann.

Der äußere Rand jeder Bildzelle ist meistens dunkler und wird verworfen. Somit bleiben bei Chip I noch *6x6* der ursprünglichen 64 Pixel für die Schätzung des mittleren Signals, bei den neuen Chips sind es nur noch *4x4*. Die Software benutzt in der Standardeinstellung das dritte Quartil der Pixelwerte als Schätzer. Dieser bei der Bildanalyse errechnete Wert wird im Folgenden als **Rohintensitäten** bezeichnet. Er wird mit den *(x,y)*-Koordinaten der Bildzelle, der verwendeten Anzahl der Pixel und ihrer Standardabweichung in die *CEL*-Datei ausgegeben. Das gescannte Pixelbild wird in der *DAT*-Datei gespeichert und die Parameter des Chipexperiments in der *EXP*-Datei. Das Bild eines Experiments umfasst für beliebige Chips etwa 45 Megabyte (Mb), und die *CEL*-Dateien sind pro ausgewertetes Experiment etwa 7 Mb bei Chip I und 13 Mb bei dem neuen Affymetrix Chip groß.

Der Scanner oder die Software hat einen festen Wertebereich, der den Messbereich und die Genauigkeit von vornherein einschränken. Der Rundungsfehler erscheint deutlich kleiner als die Varianz der Messung und stellt damit keine limitierende Einschränkung dar. Bei unserer Scannereinstellung führt der eingeschränkte Messbereich wiederholt zu Sättigungswerten, die bei der Datenanalyse einer Sonderbehandlung bedürfen. Die Rohintensitäten haben eine asymmetrische eingipfelige Verteilung mit einem langen Schwanz, der für die meisten Experimente bis in den Sättigungsbereich reicht. Die Minimalwerte liegen meist jenseits von 300 und der Maximalwert ist quasi konstant bei etwa 46200. Abhängig von der Scannereinstellung hat man bis zu mehreren tausend gesättigten Werten. Für die Dichtefunktion der Verteilung ist keine parametrische Darstellung bekannt, insbesondere führt die Anwendung einer Logarithmusfunktion auch nicht zu normalverteilten Signalen.

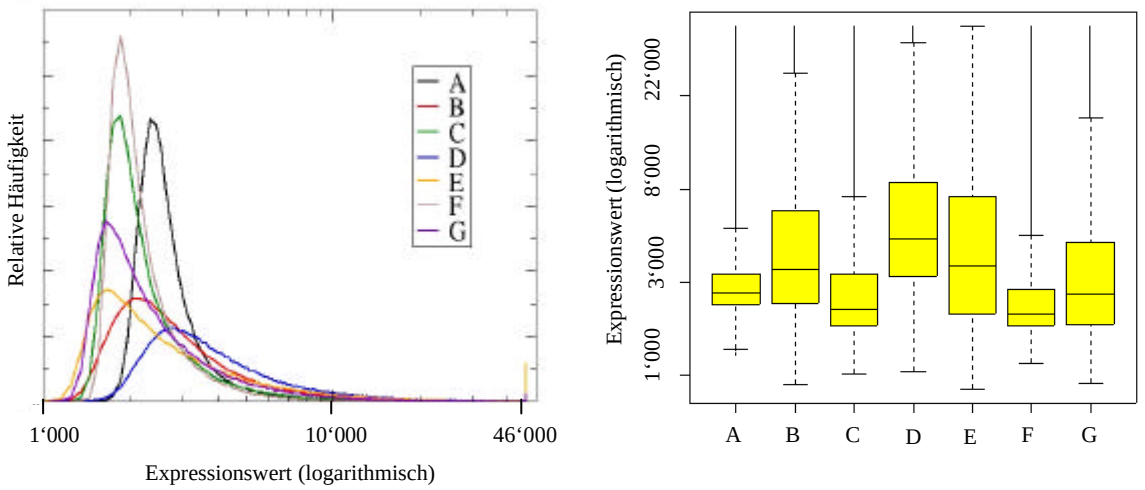


Abbildung 2-2 Die Verteilungen der Expressionswerte ausgewählter Experimente
***Links** sind die Histogramme der Expressionswerte von sieben willkürlich gewählten Chips. D bis G sind Beispiele für extrem unterschiedliche Verteilungen obwohl man bei diesen Versuchen RNA der gleichen Zelllinie hybridisierte. Die **rechte** Abbildung zeigt die gleichen Verteilungen als Box-Whisker-Plots (kurz **Box-Plot**). Die gelben Boxen enthalten jeweils alle Werte vom 25% bis zum 75% Quantil und die Linie in der Box markiert den Median. Die Punktlinie und der Querstrich markieren die Whisker. Der untere Whisker entspricht dem anderthalbfachen des Abstandes zwischen dem Median und dem 25%-Quantil und der obere Whisker dem zwischen Median und 75%-Quantil. Alle Punkte, die jenseits dieser Bereiche liegen sind einzeln eingetragen und aufgrund ihrer großen Zahl erscheinen sie oberhalb als dünne schwarze Linie.*

Es hat sich auch in der Routine als dienlich erwiesen, die Verteilungen in einem Histogramm und in einem Scatterplot mit einem verwandten Experiment darzustellen, um zusätzlich Hinweise auf Artefakte zu erhalten, die auf dem Scannerbild nicht einfach zu entdecken sind. Eine wichtige Beobachtung ist, dass der jeder Chip eine Topologie aufweist. Die Annahme scheint vernünftig, dass die niedrigsten 5% oder auch 10% der Rohintensitäten praktisch keine Signale von Transkripten darstellen, sondern die Vermessung eines leeren Chips repräsentieren. Da die Signale der gebunden Transkripte ungleichmäßig über den Chip verstreut sind, lassen sich Inhomogenitäten im Chiphintergrund nicht leicht mit bloßem Auge erkennen. Schaut man sich aber zum Beispiel die Verteilung des niedrigsten einen Prozents der Werte über den Chip an, so entdeckt man erhebliche Schwankungen. Der hier in die Signalwerte eingebrachte Fehler ist so groß, dass eine Korrektur zu einem merklichen Qualitätsgewinn führt.

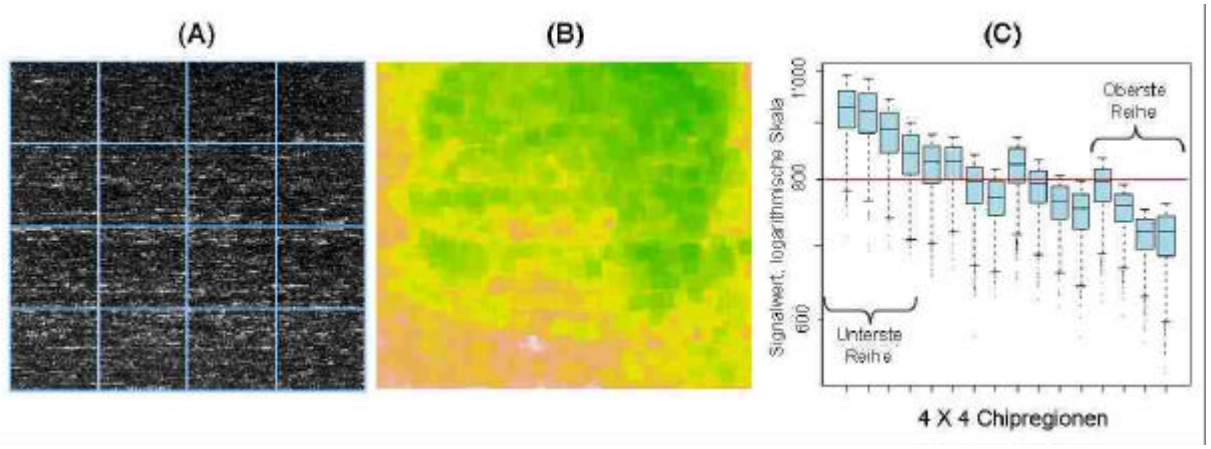


Abbildung 2-3 Inhomogenität des Chiphintergrundes

Die Hintergrundintensitäten des Scannerbildes eines Oligo-Chips lassen Tendenzen erkennen, die bei der Analyse nicht zu vernachlässigen sind. ***(A)*** *Scannerbild eines Chips; Die Intensitäten sind als Graustufen kodiert. Die Linien deuten die Aufteilung des Chipbildes in 4 X 4 Regionen an.* ***(B)*** *Das Scannerbild besteht aus etwa 500 X 500 Bildzellen. In einem gleitenden Fenster von 20 X 20 Zellen wird jeweils das 0,01-Quantil der Rohintensitäten gebildet und als Farbwert kodiert. Das 0,01-Quantil entspricht bei den 400 Intensitäten etwa dem fünft kleinsten Wert. Der Wertebereich erstreckt sich von 569,2 (dunkel grün) bis 922 (blass beige, weiß) für dieses Beispiel.* ***(C)*** *Der Chip wurde in 16 Regionen aufgeteilt (angedeutet durch die grauen Linien in A), und die Verteilungen der jeweils 10% niedrigsten Intensitäten sind als Box-Plots dargestellt. Die Reihenfolge in der die Regionen durchlaufen werden, ist von links nach rechts und von unten nach oben. Die rote Linie dient der Orientierung.*

Die Hintergrundwerte der Rohintensitäten variieren in den meisten Fällen pro Chip im Mittel um 500 bis 700, in extremen Fällen bis über tausend. Eine detailliertere Analyse zeigte, dass sich die Verteilungen der Intensitäten der Regionen (vergleiche Abbildung 2-3 C) nicht linear aufeinander transformieren lassen. Das Thema ist nicht weiter ausgeführt, da keine abschließende Lösung vorgeschlagen werden kann.

2.2. Die Aufgabenbeschreibung für die Chipanalyse

Die Technologie der Oligonukleotid-Arrays befindet sich seit etwa sieben Jahren im Routineeinsatz und ist damit noch verhältnismäßig jung. Bei metaGen werden seit April 1999 kontinuierlich Chipexperimente durchgeführt, was, wie auch in vielen anderen Gruppen weltweit, bereits zu einer stattlichen Menge an Expressionsdaten geführt hat. Eine zuverlässige Datenanalyse erwies sich als schwierig und konnte bisher nicht befriedigend gelöst werden. Zugespitzt bedeutet das: Es lassen sich zurzeit mit Array-Experimenten nur sehr eingeschränkt quantitative Aussagen über die Genexpression gewinnen. Das gilt übrigens für praktisch jedes Verfahren zur Expressionsanalyse. Das Unbehagen darüber ist in zahllosen

Publikationen belegt. Vielerorts arbeiten Forscher in Firmen wie auch an öffentlichen Einrichtungen an der Entwicklung besserer Verfahren zur Datenanalyse (Abschnitte 2.4, 2.6).

Ausgehend von den in der Einführung beschriebenen Zielen ergeben sich an eine gute Expressionsanalyse die folgende Erwartungen: Technische Artefakte müssen aufgedeckt und eliminiert werden. Die Hintergrundwerte sollten innerhalb eines Chips und natürlich auch zwischen verschiedenen Chips etwa im gleichen Bereich liegen. Man benötigt einen repräsentativen Expressionswert, der pro Gen und Experiment möglichst zur mRNA-Konzentration proportional ist. Die errechneten repräsentativen Expressionswerte sollten pro Gen über die Datensätze vergleichbar sein. Als Qualitätsmaß gilt hierbei, dass Wiederholungsexperimente möglichst gut die Werte reproduzieren und dass echte Unterschiede zwischen Experimenten sicher detektiert und gut geschätzt werden können. Eine weitere wichtige Aufgabe der Datenanalyse ist die Aufdeckung und Trennung der Hauptfehlerquellen, um hier schon möglichst beim Design der Experimente vorbeugen zu können. Entdeckt man beispielsweise eine Chipserie, die nachweislich mit Produktionsfehlern behaftet ist, so lassen sich von vornherein Qualitätseinbußen vermeiden.

Betrachtet man die vorliegenden Rohdaten und die zu erfüllenden Erwartungen, so besteht offensichtlich Bedarf an einer Bereinigung und Normierung. Zur Rechtfertigung der dafür notwendigen Datentransformationen müssen gewisse Annahmen getroffen werden:

1. Die Konzentration der an eine Oligo-Sonde gebundenen RNA-Moleküle ist eine monoton steigende Funktion des gemessenen Signals der Bildzelle.

2. Wählt man zufällig und unabhängig zwei Sets von Bildzellen eines Chips, so müssen die Verteilungen der jeweils kleinsten 2% der messbaren Signale übereinstimmen, falls die Sets insgesamt groß genug[1] sind. Das heißt, jede gemessene Verschiebung ist ein technisches Artefakt und darf durch Datentransformation behoben werden.

3. Die Gesamtmenge der in jeden Versuch eingesetzten RNA ist konstant.

4. Alle Gene auf dem Chip sollten aufgrund ihrer großen Anzahl eine repräsentative Teilmenge aller Gene des Genoms darstellen. Das legt nahe, auch die den gemessenen

[1] Chip I besteht aus circa 250 tausend Bildzellen.

Signalen zugrund liegende Verteilung für jedes Chipexperiment als identisch anzunehmen. Das erlaubt, die Datensätze zweier Chipexperimente vor dem Vergleich so zu transformieren, dass sich die Verteilungen ihrer Expressionswerte ähneln (Normierung).

5. Die Verteilung der Expressionsunterschiede zwischen zwei beliebigen RNA-Proben ist annähernd symmetrisch, dass heißt, man erwartet etwa gleich viele ähnlich hoch- wie runterregulierte Gene beim Vergleich zweier Chipexperimente.

Annahme 4 ist besonders schwerwiegend und wird nur aufgrund mangelnder Alternativen getroffen. So ist es eher unwahrscheinlich, dass die Verteilung der Expressionswerte hoch-proliferativer Tumorzellen der von seneszenten Zellen gleicht. Wären für Zelltypen invariant exprimierte Gene bekannt, so ließen sich deren Signale zur Normierung verwenden oder zumindest die Annahme prüfen.

Nach eingehender Analyse der Rohdaten und publizierter sowie verfügbarer Verfahren und nach Abschätzung der vorhandenen Möglichkeiten, erschien die Entwicklung einer alternativen Auswertung ausgehend von den in 2.1 als Rohintensitäten bezeichneten Werten, also den *CEL*-Dateien, als aussichtsreich. Die Qualität und die Konsistenz der Daten konnte wesentlich verbessert werden. Zusätzlich ergab sich die Notwendigkeit, eine Datenbank aufzubauen, mit der die Verwaltung hunderter Chipdatensätze und der dazu gehörigen Metainformationen[1] zu bewältigen ist.

2.3. Das Auswerteverfahren

Inhalt dieses Abschnittes ist das im Rahmen dieser Arbeit entwickelte und implementierte Verfahren zur Auswertung von Oligonukleptid-Arrays, wie sie in der Einleitung beschrieben sind. Ausgangspunkt sind Rohintensitäten, für jeden Chip also einem Intensitätswert pro Bildzelle beziehungsweise pro Oligospezies. Die Bezeichnung Rohintensitäten ist vielleicht etwas irreführend, da die *CEL*-Daten ja bereits durch Verdichtung aus den Pixelbilddateien entstanden sind. Sie hat sich aber als zweckmäßig für die Abgrenzung zu errechneten Werten

[1] Unter dem Begriff Metainformation seien alle Daten zusammengefasst, die neben den eigentlichen Signalen pro Chipexperiment erhoben werden.

späterer Bearbeitungsschritte erwiesen. Für die Hintergrundkorrektur teilt man den Chip in vier mal vier Regionen auf und berechnet in jeder dieser 16 Regionen den Mittelwert der 2% niedrigsten Zellsignale. Für jede einzelne Bildzelle ergibt die Rohintensität abzüglich des regionalen Hintergrundwertes das korrigierte Signal. Mit der Division jedes hintergrund-korrigierten Signals mit dem Median dieser Werte erreicht man eine Skalierung der Intensitäten bezüglich ihres Zentrums. Nach diesen beiden Korrekturschritten sind die Verteilungen der Signale einer großen Gruppe von Chips schon sehr gut angeglichen. Man könnte die normalisierten Rohdaten so belassen und jede Bildzelle, die zu einem Transkript gehört als unabhängige Messung für dasselbe Gen ansehen. Es hat sich aber aufgrund der schieren Datenfülle und weil dadurch die Aussagekraft nicht wesentlich gemindert wird als zweckmäßig erwiesen, die Daten weiter zu verdichten. Für jedes Oligoset berechnet man einen repräsentativen Expressionswert und einen Detektionsscore. Als repräsentativen Expressionswert hat sich das dritte Quartil der *PM*-Signale als guter Schätzwert behauptet. Er wird im Weiteren mit ***PMQ*** (*PM*-Quartil) abgekürzt. Um zu bewerten, ob ein bestimmtes Transkript auf dem Chip detektiert wird, lässt sich prüfen, ob in dem Sondenset das *PM*-Signal fast immer stärker als das entsprechende *MM*-Signal ist. Dazu eignet sich der Wilcoxon-Test für Paardifferenzen [8]. Besteht das Sondenset wie auf Chip I aus 20 (*PM, MM*)-Oligopaaren, so berechnet man die 20 Differenzen der *PM*-, und *MM*-Signale und sortiert diese nach der Größe ihrer Absolutbeträge. Entsprechend dieser Sortierung bekommt jetzt jede Differenz eine Rangzahl zugeordnet. Anschließend schaut man sich die Vorzeichen der Differenzen an und bildet die negative und die positive Rangsumme. Die negative Rangsumme berechnet sich als die Summe der Rangzahlen, die den Differenzen mit negativem Vorzeichen zugeordnet wurden und die positive Rangsumme entsprechend. Es gilt, dass die Summe der beiden Rangsummen für eine feste Anzahl Wertepaare konstant die Summe aller Rangzahlen ist. Für ein auf dem Chip nicht detektiertes Transkript erwartet man, dass im Mittel für die Hälfte aller Paare das *MM*-Signal größer ist als das *PM*-Signal. Das würde sich dann darin widerspiegeln, dass die positive und die negative Rangsumme annähernd gleich große Werte annehmen. Auf Grund der Konstruktion der Oligos kann die Erwartung noch genauer spezifiziert und damit der Test verbessert werden. Da die Missmatch-Oligos durch den Austausch der mittleren Base aus den *PM*-Oligos hervorgegangen sind, kann eine systematische Tendenz zu größeren *MM*-Signalen ($PM<MM$) von vornherein ausgeschlossen werden. Damit lässt sich für ein Sondenset die einseitige Fragestellung prüfen: $H_0 : PM \leq MM$ (Nullhypothese). Gegeben das beobachtete

Expressionsmuster, dann gibt der P-Wert des Wilcoxon-Tests die Wahrscheinlichkeit dafür an, dass das Oligo-Sondenset kein Transkript detektiert hat. Dieser P-Wert wird als **Detektionsscore** verwendet. Als Schwelle zum Filtern der Daten dient fast ausschließlich 0,05 oder 5% Irrtumswahrscheinlichkeit.

Formale Beschreibung des Algorithmus

Notation: Der Pfeil „←“ symbolisiert den Zuweisungsoperator. In der Programmiersprache *C* entspricht dem beispielsweise das Gleichheitszeichen.

Eingabe: Rohintensitäten für jede Bildzelle (I)

Bemerkung: Der Einfachheit halber nehmen wir an, dass I eindeutig ist und damit eine umkehrbar-eindeutige Abbildung zwischen Bildzelle und Rohintensität existiert. Trifft man diese Annahme nicht, muss man für I einen Index der Bildzelle mitführen.

(x,y) - Lokalisation einer Bildzelle oder Intensität auf dem Chip

Sondenset – Name des Oligosets (meist ein-eindeutig für ein bestimmtes Gen)

Oligopaar – Nummer des Oligopaares innerhalb eines Sondensets

(für metaGen Chip I meist eins bis zwanzig)

Typ – Typ des Oligos also *PM* (perfect match) oder *MM* (missmatch)

Ausgabe: pro Sondenset (verdichtete Daten)

Sondenset – Name des Oligosets

PMQ – repräsentativer Expressionswert, 3. Quartil der korrigierten *PM*-Signale

P_Wert – Detektionsscore, P-Wert des einseitigen Wilcoxon-Tests

Folgende Indexfunktionen liefern die *x,y*-Position zu einer Intensität

$$x_I : I \to \{1, \mathrm{K}\ , x_{\max}\}$$

$$y_I : I \to \{1, \mathrm{K}\ , y_{\max}\}$$

In diesem Fall symbolisieren die Pfeile eindeutige Abbildungen.

Die folgende Funktion liefert für die (x,y)-Koordinaten einer beliebigen Bildzelle eine Zahl zwischen eins und sechzehn. Dadurch wird der Chip in viermal vier durchnummerierte Regionen unterteilt.

$$region(x, y) \leftarrow funktion(x, y)\{$$
$$a \leftarrow x \; DIV \; \frac{x_{\max}}{4} + 1$$
$$b \leftarrow y \; DIV \; \frac{y_{\max}}{4} + 1$$
$$return \; (a + (b-1) \cdot 4)$$
$$\}$$

Bemerkung: *DIV* bezeichnet die ganzzahlige Division. Für eine Intensität I geben die Funktionen *sondenset(I), oligopaar(I) und typ(I)* jeweils den Namen des Sondensets, die Nummer des Oligopaares und den Typ des Oligos zu dem sie gehört zurück.

Hauptprogramm

1. Hintergrundkorrektur

   für i=1,…,16

   $$\Im \leftarrow \{I \mid region(x_I, y_I) = i\}$$
   $$hg_i \leftarrow mittel(\{I \mid w \in \Im \quad und \quad w < quantile(\Im, 0.02)\})$$

   für alle I

   $$I \leftarrow I - hg_i, \quad wobei \; i = region(x_I, y_I)$$

2. Skalierung

   für alle I

   $$I \leftarrow I / median$$

Bemerkung: *median* bezeichnet den medianen Wert aller hintergrund-korrigierten Intensitäten.

3. Berechnung des Detektionsscores

$$\text{für alle Sondensets } S$$

$$\Im^S \leftarrow \{ I \mid sondenset(I) = S \}$$

$$paare \leftarrow \{ (I_{PM}, I_{MM}) \mid I_{PM}, I_{MM} \in \Im^S \text{ und } oligopaar(I_{PM}) = oligopaar(I_{MM}) \text{ und } typ(I_{PM}) = PM \text{ und } typ(I_{MM}) = MM \}$$

$$P\_Wert \leftarrow wilcoxon.p\_wert(\ paare,\ H_0 : I_{PM} \le I_{MM}\ )$$

4. Berechnung des repräsentativen Expressionswertes

$$\text{für alle Sondensets } S$$

$$\Im^S \leftarrow \{ I \mid sondenset(I) = S \text{ und } typ(I) = PM \}$$

$$PMQ \leftarrow quantile(\ \Im^S, 0.75)$$

Bemerkung: Die Berechnung der Quantile und des Wilcoxon-Tests ist im Anhang ausgeführt.

Beispiel

Als nächstes wird zur Veranschaulichung ein konstruiertes Sondenset durchgerechnet. Das Sondenset bestehe aus 11 Oligopaaren (*PM, MM*), und die Werte seien bereits hintergrund-korrigiert. Die Oligopaare sind durchnummeriert und die entsprechenden Intensitäten in der folgenden Tabelle dargestellt.

Oligo-paare	1	2	3	4	5	6	7	8	9	10	11
PM	32	34	45	33	16	15	19	37	39	24	33
MM	22	14	24	11	18	27	12	23	31	13	14
PM-MM	10	20	21	22	-2	-12	7	14	8	11	19

Für die Berechnung des Detektionsscores findet der Wilcoxon-Test für Paardifferenzen seine Anwendung. Von den tatsächlichen Intensitätswerten zieht man sich auf die Rangzahlen zurück, da die Verteilungsfunktion nicht bekannt ist. Man berechnet zuerst die Differenzen *PM-MM*) und ordnet diese dann nach ihrem Absolutbetrag. Jede der Differenzen erhält damit eine Rangzahl. Anschließend summiert man alle Rangzahlen von Differenzen mit negativem Vorzeichen zur negativen Rangsumme und alle Rangzahlen von Differenzen mit positivem Vorzeichen zur positiven Rangsumme auf.

***PM-MM* geordnet**	-2	7	8	10	11	-12	14	19	20	21	22
Rangzahl	1	2	3	4	5	6	7	8	9	10	11

Das ergibt für die negative Rangsumme $R^- = 1+6 = 7$ und für die positive Rangsumme $R^+ = 2+3+4+5+7+...+11 = 59$. Als P-Wert bei der Anwendung des einseitigen Wilcoxon-Tests[1] auf die gepaarte Stichprobe vom Umfang 11 erhält man 0,009277. Die Nullhypothese $H_0 : PM = MM$ kann damit auf dem 1%-Niveau abgelehnt werden. Als Detektionsschwelle wurde meistens 0,05 (5%-Niveau) benutzt. Für das Beispiel geht man also davon aus, dass das Transkript spezifisch detektiert wird. Der repräsentative Expressionswert ergibt sich als 3. Quartil der *PM*-Werte, für das Beispiel 35,5.

Während des gesamten Prozesses können Qualitätsparameter zum Filtern der Daten generiert werden. Beispiele hierfür sind die Standardabweichung der Pixelwerte aus der Bildanalyse oder die Information, ob ein bestimmter Signalwert im Sättigungsbereich liegt. Die Homogenität des Hintergrundes oder die Zahl der Oligopaare, bei denen das *PM*-Signal stärker als das *MM*-Signal ist, liefern Auskunft über die Qualität des Chipexperiments als ganzes. Die bislang beschriebenen Algorithmen lassen sich auf jeden Chip separat anwenden. Für die Chip-übergreifende Analyse von Datensätzen, kann eine zusätzliche Normierung der Daten ihre Vergleichbarkeit signifikant verbessern. Aufgabe der Normierung ist es, die Daten der Einzelexperimente so zu transformieren, dass die von der Technologie stammende Streuung möglichst minimiert wird. Dabei müssen die in den Proben vorhandenen interessanten Unterschiede aber erhalten bleiben. Für eine Reihe von Datensätzen erzielt man durch Logarithmierung und anschließende lineare Anpassung befriedigende Ergebnisse. Die bislang beste Normierung im oben beschriebenen Sinne konnte mit stückweiser polynomialer Regression[2] erreicht werden. Siehe Anhang 1 für eine detailliertere Beschreibung. Das Verfahren basiert auf der Form der Punktwolke. Man legt also kein mathematisches Modell für systematische und zufällige Effekte zugrunde, durch das sich die Beobachtungen und

---

[1] Der Aufruf der R-Funktion zur Berechnung des Tests mit allen Parametern lautet *wilcox.test*(pm,mm, alternative="greater", paired=T, exact=T), wobei *pm* und *mm* die Vektoren der Intensitäten der *PM*- und *MM*-Oligos bezeichnen.

[2] Berechnet wurde die Anpassung mit Hilfe der R-Funktion *loess*.

deren Abweichungen vom Erwartungswert interpretieren ließen. Wäre die Expression bestimmter Gene[1] vorhersagbar, so könnte man daran die Normierungsverfahren eichen und so die Datenqualität wesentlich verbessern. Die bakteriellen Spike-Kontrollen sind dafür nicht geeignet, da die RNA nicht die gleiche Aufarbeitung erfahren hat wie die Probe.

2.4. Das Verfahren von Affymetrix

Affymetrix hat die Verfahren für die Bild- und Datenanalyse in der *Microarray Analysis Suite* (**MAS 5.0**) zusammengefasst. In der aktuellen Version sind grundlegend veränderte Algorithmen der Datenanalyse implementiert, die im Folgenden beschrieben werden. Ausgangspunkt für die Diskussion die in den *CEL*-Dateien gespeicherten Rohintensitäten. Grundsätzlich können zwei Analysewege beschritten werden, die Einzelchipauswertung und die paarweise Chip-Auswertung.

Für die Einzelchipauswertung lässt sich zusammenfassend feststellen, dass vergleichbar unserem Algorithmus ein Detektionsscore und ein repräsentativer Expressionswert berechnet werden. Der Detektionsscore ergibt sich ebenfalls aus dem P-Wert des einseitige Wilcoxon-Rangsummentest. (siehe vorigen Abschnitt) Allerdings wird hier zuerst für jedes Oligopaar ein Diskriminanzscore R berechnet: $R=(PM-MM)/(PM+MM)$ Die Differenz wird durch die Summe relativiert. Für ein Probeset, dass kein Transkript detektiert erwartet man eine gleichmäßige Verteilung der Diskriminanzscores um den Median Null. In der Software MAS 5.0 lässt sich der Test mit einer Konstante $t > 0$ parametrisieren. Die Nullhypothese, gegen die man dann testet, lautet: $median(R - t) = 0$ und die Alternativhypothese: $median(R - t) > 0$. Der Anwender erreicht mit der Erhöhung von Tau, dass nur noch Transkripte als detektiert gelten, in deren Sondensets fast alle *PM*-Signale um mindestens $t \cdot (PM+MM)$ größer als die *MM*-Signale sind. Auch für den Detektionsscore (P-Wert des Wilcoxon-Tests) kann der Anwender zwei Schwellwerte a_1 und a_2 festlegen, womit sich die Expressionswerte in drei Klassen wie verlässliche (0 = P-Wert < a_1), zweifelhafte (a_1 = P-Wert < a_2) und nicht zu trauende (P-Wert > a_2) einteilen lassen. Die Originalbezeichnung in MAS 5.0 lauten *present, marginal, absent*. Für die Berechnung eines repräsentativen Expressionswertes, wird nach der

[1] Das perfekte Kontrollgen ist in jeder untersuchten Zelle mit der gleichen Anzahl mRNA-Moleküle vertreten.

Hintergrundkorrektur ein robustes gewichtetes Mittel der logarithmierten Differenzen (*PM-MM*) nach dem Einschritt-Biweight-Tukey – Verfahren gebildet. (Siehe Anhang 1 zu Kapitel 2) Der entscheidende Unterschied zu unserem Verfahren ist die Annahme, dass die *MM*-Signale eine gute Schätzung für das lokale, oligospezifische Hintergrundrauschsignal darstellen. Folgt man der Annahme, so ist die Differenz (*PM-MM*) ein guter Schätzer für die absolute, spezifische Expression. Im Vergleich zu früheren Versionen wird darauf geachtet, dass es bei der Berechnung nicht zu negativen Expressionswerten kommt. Für die Ausgabe werden die Werte rücktransformiert und skaliert. (Siehe Anhang 1) Zur Unterscheidung von anderen errechneten Expressionswerten wird dieser Schätzer im Folgenden mit **AvgDiff** (*average difference*) bezeichnet.

Für die paarweise Auswertung von Chipexperimenten bietet die Microarray Suite einen gesonderten Algorithmus an. Ein Chip wird als Referenz definiert. Die Intensitäten der beiden Experimente durchlaufen die Hintergrundkorrektur und werden anschließend angepasst, so dass sie zwischen Experimenten vergleichbar sind. Mit Hilfe des Wilcoxon-Tests kann geprüft werden, ob die Signale eines Sondensets auf dem einen Chip signifikant höher oder niedriger sind als auf dem Referenzchip. Schwellwerte, ab wann eine Änderung der Expression als signifikant gilt, kann der Anwender festlegen. Die Normierung der beiden Chip-Datensätze lässt sich wahlweise auf Basis einer ausgezeichneten Menge von Sondensätzen (invariant exprimierte Kontrollgene) oder basierend auf allen Intensitäten durchführen. Ein Wert für die Stärke der Änderung der Expression (*LogRatio*) kann wiederum mit Hilfe des Mittelwertschätzers von Tukey (Siehe Anhang zu Kapitel 2) bestimmt werden.

2.5. Evaluierung des in 2.3 vorgestellten Verfahrens

Dieser Abschnitt befasst sich mit der Performanz des Verfahrens hinsichtlich der unter 2.2 beschriebenen Ziele und den entsprechenden Annahmen. Generell hat man mit der Schwierigkeit zu kämpfen, dass das Probenmaterial nur in sehr begrenzten Mengen vorhanden ist, man misst also tausende von Transkripten parallel in vielleicht 30 Tumor-, Normalgewebe-Paaren pro Tumorentität. Abschließende, statistisch abgesicherte Aussagen sind folglich nicht zu erzielen. Zum Zwecke der Leistungsabschätzung der Technologie und der Analyseverfahren konnten drei verschiedene Datensätze generiert werden:

Set 1 Zwei Zelllinien LnCAP, DU145, je sechs Wiederholungen (12 Chipexperimente)

Set 2 Acht Kulturen der Zelllinie RT4 auf Chips aus unterschiedlichen Produktionschargen über einen längeren Zeitraum

Set 3 Eine Chipreihe mit fünf unterschiedlichen Konzentrationen der Spike-RNA

Mit Hilfe von Set 1 lässt sich demonstrieren, dass das vorgestellte Verfahren für hoch qualitative Daten gut reproduzierbare Signale errechnet und sich auch Unterschiede in der Genexpression gut reproduzieren lassen. Die LnCAP-Zelllinie stammt ursprünglich aus einer Lymphknotenmetastase eines Prostatakarzinoms, und wurde beim ATCC unter der Nummer CRL-1740 beschafft. Die Zelllinie DU145 (ATCC: HTB-81) etablierte man aus Zellen einer Hirnmetastase eines Prostatakarzinoms. Nachdem die Chipdaten der zwei Zelllinien die Standardauswertung durchlaufen hatten, wurden noch alle normalisierten Expressionswerte an die Daten eines ausgewählten Experiments mit Hilfe der *Loess*-Korrektur angepasst. Der Korrelationskoeffizient liegt danach für die sechs Wiederholungen der LnCAP-Zelllinie im Mittel bei 0,99 (Minimum: 0,97) und für die Wiederholungen der DU145-Zelllinie im Mittel bei 0,98 (Minimum: 0,96). Paarweise Darstellungen sind im Anhang 1 beigefügt. Die Korrelationskoeffizienten zwischen den Zelllinien reichen von etwa 0,87 bis 0,90. Wendet man den T-Test an, um für jedes Gen einen Score für die differenzielle Expression zwischen den beiden Zelllinien zu erhalten, so kann man anschließend zählen, wie viele der Gene eine bestimmte Schwelle unterschreiten. Es zeigt sich jedoch, dass die Anwendung der T-Statistik allein zu viele falsch positive liefert. Das gleiche gilt übrigens auch für das nicht-parametrische Pendant, den U-Test von Mann, Whitney und Wilcoxon (Anhang A). Beispielsweise erreichen 1433 Gene (~1/4 aller Gene) den kleinsten P-Wert, also den besten Score, des U-Tests. Legt man die gleiche Schwelle für die T-Statistik an, so erreichen 1722 Gene diesen Score.

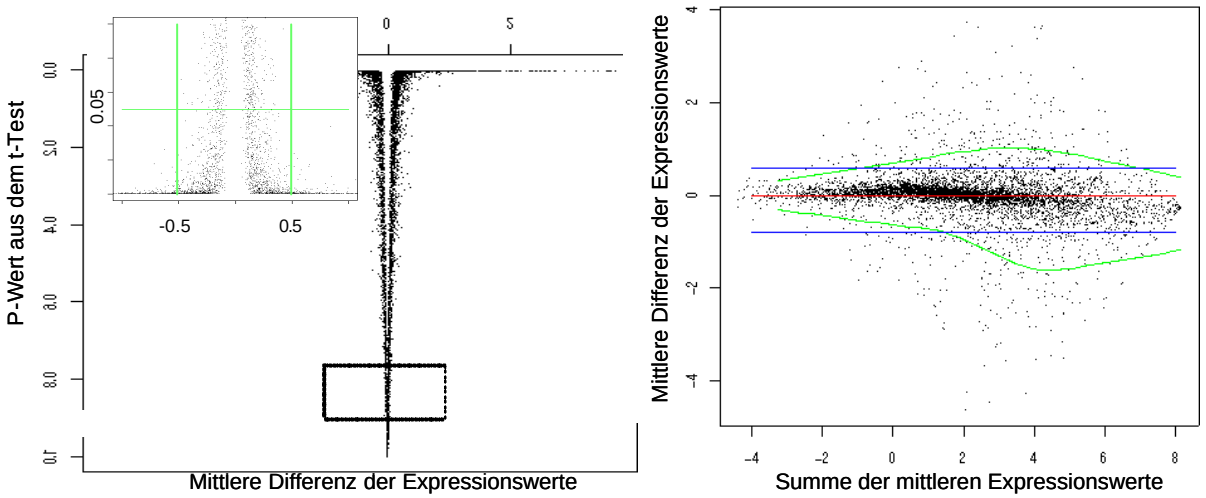


Abbildung 2-4 Differenzielle Expression zwischen den Zelllinien DU145 und LnCAP
Links *ist für jedes Sondenset der P-Wert des t-Tests gegen mittlere Differenz der logarithmierten und normalisierten PM-Quartile (Expressionswerte)zwischen den Experimenten der beiden Zelllinien aufgetragen. Die links oben eingesetzte Graphik ist ein vergrößerter Ausschnitt (Rahmen).* ***Rechts*** *ist die mittlere Abweichung (Differenz) der Expressionswerte in Abhängigkeit von ihrer Summe dargestellt. Die rote Linie liegt bei y=0, und die blauen bei y=-0,8 und y=0,6. Zur Erzeugung der grünen Linien wurde in einem entlang der x-Achse gleitenden Fenster das 5%- und das 95%-Quantil der y-Werte bestimmt und diese dann mit Hilfe der Loess-Korrektur geglättet.*

Die P-Wert-Schwelle für die T-Statistik scheint schwierig festzulegen zu sein. Um nicht minimale Unterschiede als signifikante differenzielle Expression zu werten, sollte man auch für mittlere Differenzen[1] eine untere Schwelle festlegen. Abbildung 2-4 suggeriert, dass alle Gene, die mittlere Differenzen außerhalb des Intervalls (-0,5; 0,5) aufweisen, als differenziell exprimiert gelten könnten. Es sind allerdings für nur 381 die Differenzen kleiner -0,5 aber für 575 Gene größer als 0.5. Ähnliche Unterschiede findet man auch für andere Schwellen, was eine leichte Asymmetrie in der Verteilung aufdeckt (leichte S-Kurve in Abbildung 2-4 rechts). Um etwa 300 (~5%) hoch- wie runterregulierte Gene zu erhalten, müsste man die Schwellen auf -0,8 und 0,6 legen.

Der zweite Datensatz ist über etwa zwei Jahre entstanden. Besonders in der ersten Zeit wiesen die Chips erhebliche qualitative Unterschiede auf. Zum Teil mussten sogar ganze Serien ausgetauscht werden. Um gravierende Produktionsfehler früh zu erkennen, hybridisierte man

[1] Da jeweils gleich viele (6) Wiederholungen analysiert wurden, gilt dass die mittlere Differenz gleich der Differenz der Mittelwerte ist.

bei metaGen einen Chip jeder Produktionsserie mit RNA der Zelllinie RT4 (ATCC: HTB-2). Diese Zelllinie stammt ursprünglich aus einem Übergangszellpapilom der Harnblase und wurde uns von der Arbeitsgruppe Knüchel an der Uni Regensburg zur Verfügung gestellt. Aus diesen Experimenten fielen sieben auf, deren Verteilungen ihrer Rohintensitäten besonders deutlich voneinander abwichen (dargestellt in Abbildung 2-2). Das Interesse an diesem Datensatz rührt daher, dass er in etwa die Variabilität aller jemals bei metaGen durchgeführten Chipexperimente widerspiegelt. Die komplizierte Aufbereitung der RNA und hier vor allem die Mikrodissektion und die RNA-Amplifikation stellen zusätzliche Fehlerquellen für die Messungen dar, die nicht leicht zu beherrschen sind. Über die für diese Technologie lange Zeit der Nutzung haben sich die Daten zu hunderten Versuchen in unserer Datenbank angesammelt, und man möchte natürlich in der Lage sein, falls gewünscht, jedes Experiment mit jedem vergleichen können. Außerdem wächst das Volumen an frei verfügbaren Daten aus öffentlich geförderten Projekten sprunghaft an. Will man solche Datensätze, die von unterschiedlichen Gruppen erzeugt wurden, für übergreifende Analysen nutzen, kommen zusätzliche Anforderungen an die Datennormierung hinzu. Die Auswertung der Daten im Set 1 war in diesem Sinne eine besondere Situation, in der man wichtige Einflussgrößen, die die Daten verzerren, bewusst konstant gehalten hat. Ähnliches gilt für die von Affymetrix und GeneLogic (USA) publizierten Kontrollexperimente. Die Daten weisen allesamt eine sehr geringe Streuung auf, und die Ergebnisse sind nicht einfach auf Experimente mit aus Gewebeproben gewonnener RNA verallgemeinerbar. Für planbare komplexe Experimente, wie beispielsweise die Untersuchung des Zellzyklus der Hefe, ist es aufgrund der produktionsbedingten Schwankungen ratsam, die benötigten Arrays en block zu bestellen und zu verarbeiten.

Die Datensätze der RT4-Versuche wurden mit dem in Abschnitt 2.3 vorgestellten Verfahren ausgewertet und normiert. Die folgende Abbildung fast die Resultate zusammen.

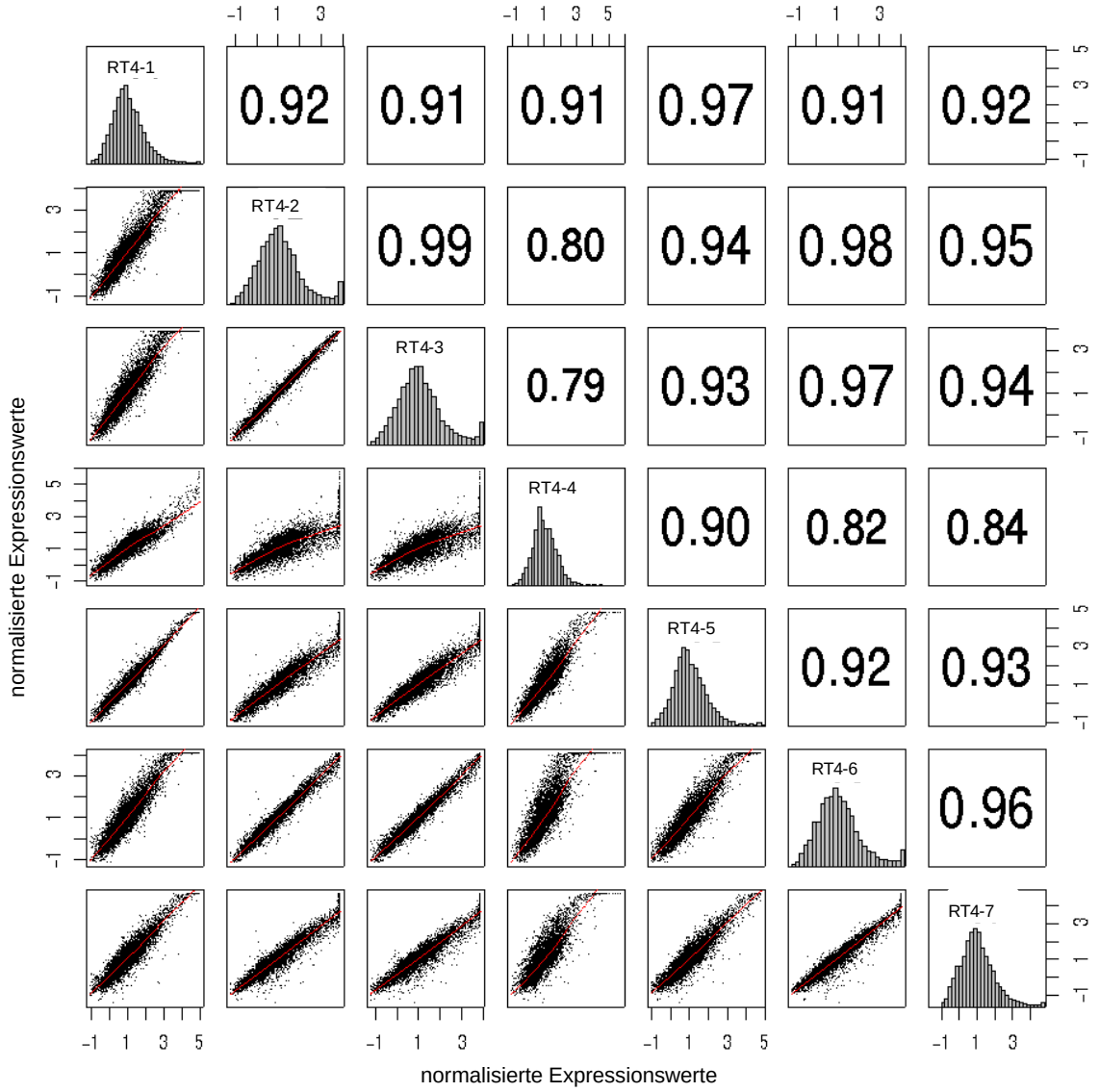


Abbildung 2-5 Reproduzierbarkeit der *PM*-Quartile

Dargestellt sind die Verteilungen und die Korrelation der logarithmierten und normierten Expressionswerte (PM-Quartile) von sieben Chipexperimenten mit der Zelllinie RT4. In der Diagonalen sind die Verteilungen der Expressionswerte als Histogramme dargestellt. Die Felder links unterhalb der Diagonalen veranschaulichen die Korrelation im Scatter-Plot. Die Loess-Regression ist jeweils als graue Linie angedeutet. Rechts oben sind für jeden Vergleich die Korrelationskoeffizienten aufgeführt.

Nach der *Loess*-Korrektur ist die Korrelation der meisten Datensätze recht gut. RT4-4 fällt jedoch heraus. Auffällig ist auch die unterschiedliche Anzahl gesättigter Werte, die zum Teil die Anpassung verfälschen können. Der gleiche Datensatz wurde auch mit der Analysesoftware MAS 5.0 von Affymetrix ausgewertet, um die Algorithmen direkt vergleichen zu können. Dazu sei vorausgeschickt, dass für qualitativ hochwertige Daten wie bei den Replikationsexperimenten des Datensatzes 1 kaum Unterschiede in den Resultaten

festzustellen sind. Beide Verfahren arbeiten erwartungstreu. Der fundamentale Unterschied zwischen beiden Ansätzen besteht in der Frage, welcher gemessene Signalwert den besten Schätzer für die absolute spezifische Expression eines Transkripts liefert. Die PMQ-Methode beruht nur auf den Signalen der *PM*-Oligos (*perfect match*). Bei der Affymetrix-Methode geht man davon aus, dass die *MM*-Signale (*missmatch*) gute Schätzwerte für das oligospezifische Hintergrundsignal darstellen. Folglich wäre die Differenz *PM-MM* ein besserer Schätzwert für die absolute spezifische Expression pro Oligosonde und der *AvgDiff*-Wert ein besserer Schätzwert pro Sondenset.

Für die Suche nach differenziell exprimierten Genen nutzt man in den meisten Fällen als Maß einfach die Stärke der Änderung der Expression pro Gen beispielsweise zwischen einer Tumor- und einer Normalprobe. Für PMQ-Signale wie auch für die Signalwerte von Affymetrix (*AvgDiff*) entspricht das gerade dem Quotienten der Einzelwerte. Arbeitet man mit logarithmierten und um die Null zentrierten Daten, so eignet sich die Differenz der Einzelwerte (*LogRatio*) als Maß für die Expressionsänderung. Da es sich bei dem RT4-Datensatz um RNA-Proben derselben Zelllinie handelt, können die Experimente als Wiederholungen angesehen werden. Man erwartet demzufolge keine Änderungen in der Genexpression (*LogRatio* = 0). Da dieser Datensatz über etwa zwei Jahren parallel zu den anderen Chipexperimenten erhoben wurde, erlaubt er die Abschätzung von Fehlerraten bei der Charakterisierung differenziell exprimierter Gene. Wie der Vergleich in Abbildung 2-5 zeigt, ist das gerade für Experimente, die auf unterschiedlichen Chipserien durchgeführt wurden, nicht einfach zu lösen. Um die Ergebnisse der Auswertungen dahingehend zu prüfen, ließen sich die Differenzen der Expressionswerte der sieben oben bereits verwendeten Experimente zu einem achten Referenzexperiment bilden. Die Referenzprobe ist ebenfalls mRNA der Zelllinie RT4.

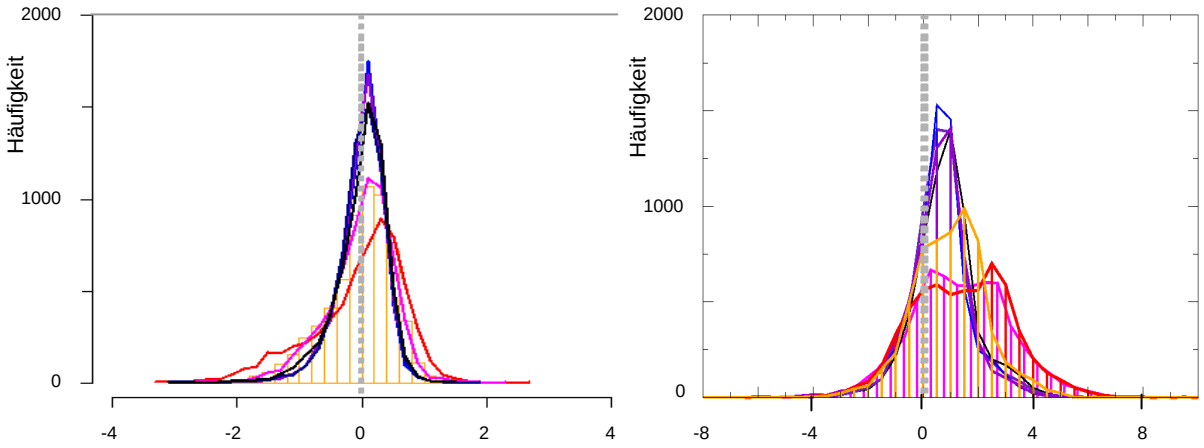


Abbildung 2-6 Vergleich der Expressionsunterschiede mit der PMQ- und der Affymetrix-Methode

In beiden Graphiken sind die überlagerten Histogramme der errechneten Expressionsunterschiede für sieben Chipexperimente im Vergleich zu einer Referenz. Als Probe für die Versuche diente ausschließlich mRNA aus einer RT4-Zelllinie. Nach der Berechnung der Expressionswerte wurden die Experimente gegeneinander normiert und anschließend die Expressionsänderungen pro Sondenset bestimmt. (siehe Text) Die grauen senkrechten Strichellinien markieren die Erwartung, dass sich die Expressionswerte bei Wiederholungsexperimenten nicht ändern. ***Links:*** *Die Expressionswerte wurden mit der PMQ-Methode errechnet.* ***Rechts:*** *MAS 5.0 von Affymetrix, vergleichende Analyse; Als Referenzexperiment (Baseline) diente wiederum das ausgewählte (achte) Experiment. Es ist zu beachten, dass die Form der Histogramme der beiden Methoden nicht direkt vergleichbar ist, da der Wertebereich und damit die Skalierung der x-Achse nicht festliegen.*

Für beide Methoden gilt, dass die Histogramme erheblich voneinander abweichen. Das zeigt, dass unabhängig vom Normierungsverfahren beim Vergleich beliebiger Chipexperimente mit relativ großen Streuungswerten zu rechnen ist, die höchstwahrscheinlich Technologie bedingt sind. Die Abbildung spricht für die Überlegenheit der PMQ-Methode gegenüber der von Affymetrix. Die entscheidende Beobachtung hier ist die deutliche Verschiebung der Verteilungen aus der Null bei der Affymetrix-Auswertung, was in krassem Gegensatz zu der Annahme steht, dass etwa gleich viele Gene hoch- wie runterreguliert sind.

Ein weiterer wichtiger Punkt bei der Evaluierung ist die Reproduzierbarkeit von Konzentrationsunterschieden, deren Werte in den RNA-Pools bekannt sind. Andernfalls könnte man fälschlicher Weise annehmen, dass eine Methode, die die Varianz und damit auch die echten Expressionsunterschiede staucht, besser reproduzierbare Ergebnisse liefert. Um das zu untersuchen, wurde der dritte Datensatz generiert. Zu jedem Hybridisierungs-Cocktail gibt man drei Mikroliter Kontroll-RNA-Mix (Abschnitt 1.5). Dieser enthält im Standardfall Spike-RNA bestehend aus fragmentierter baktertieller RNA der Gene *BioB, BioC, BioD, CreX*, die

so dosiert sind, dass sie letztlich in festen Mengen 15 nM, 50 nM, 250 nM und 1 µM im Hybridisierungscocktail vorliegen. Gibt man den Kontrollmix in unterschiedlichen Mengen in den Hybridisierungs-Cocktail, so erhält man eine Verdünnungsreihe der bakteriellen Transkripte. Dazu wurde in je drei Chipexperimenten ½, 1, 2 und 6 µl Kontrollmix zugegeben. Die eigentliche Probe, in diesem Falle RNA aus Prostata- und Blasentumoren sollte davon nicht beeinflusst sein. Das gesamte experimentelle Setting ist Tabelle 2-1 zu entnehmen.

Tabelle 2-1 Experimentelles Setup für die Variation der Spike-Kontrollen

		Errechnete mRNA-Mengen im Kontrollmix				
Genname	**Organismus**	**in 0,5 µl**	**in 1µl**	**in 2µl**	**in 3µl**	**in 6µl**
BioB	E. coli	2,5	5	10	15 nM	30
BioC	E. coli	8,3	16,7	33,3	50 nM	100
BioD	E. coli	41,7	83,3	166,7	250 nM	500
CreX	P1 Bakteriophage	166,7	333,3	666,7	1 µM	2 µM

Orange markiert sind die Mengen im Standardfall (3µl). Ausgenommen diese Versuchsreihe werden jedem Hybridisierungscocktail 3 µl vom Kontrollmix beigegeben. Die Werte sind, wenn keine Einheit angegeben ist, in Nanomol (nM) oder in Micromol (µM) aufgeführt. Die Farbkodierungen der Gennamen gilt für die Abbildung 2-7.

Die Experimente wurden mit dem PMQ- und mit dem Affymetrix-Verfahren ausgewertet und normiert. Zieht man jetzt die repräsentativen Expressionswerte der bakteriellen Spike-Transkripte heraus und setzt sie zu den eingesetzten Mengen in Beziehung, so würde man im optimalen Fall folgendes erwarten:

1. Die jeweils drei Wiederholungen liegen dicht beieinander.

2. Die Werte steigen linear mit der eingesetzten Molaren Menge.

3. Die Werte steigen mit der Molaren Menge für jedes Transkript in gleicher Weise. Die Regressionsgeraden der einzelnen Spike-Kontrollen sind vergleichbar. Das heißt, der Anstieg der Regressionsgeraden ist gleich.

Fände man ein Gesetz über den Zusammenhang zwischen Signal und Transkriptmenge, so könnte man es auf alle untersuchten Transkripte anwenden. Dafür müsste aber zusätzlich gelten, dass die Spike-Kontrollen repräsentativ für die RNA-Probe sind. Die folgenden Graphiken veranschaulichen die Expressionswerte der Spike-Transkripte in Abhängigkeit vom eingesetzten Volumen des Kontrollmix und in Abhängigkeit von der Molaren Menge. Ausgewählt wurden jeweils die Sondensets der 3'-Enden der vier bakteriellen Gene. Pro Konzentration lagen drei Chipexperimente als Wiederholungen vor. Die Regressionsgeraden sind mit der Methode der kleinsten Quadrate berechnet.

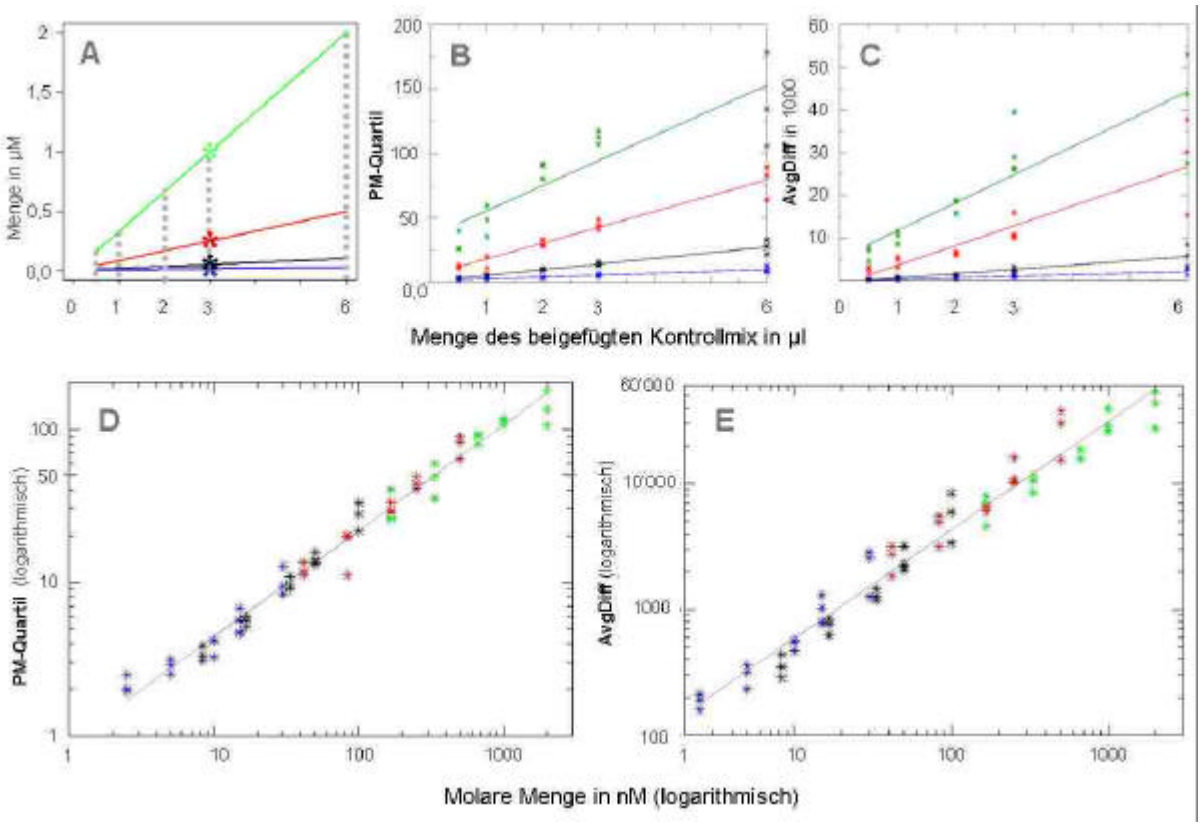


Abbildung 2-7 Reproduktion definiert eingesetzter Molarer Mengen bakterieller Transkripte

Dargestellt sind die errechneten Expressionswerte in Abhängigkeit vom Volumen des eingesetzten Kontrollmix und von der errechneten Molaren Menge der Transkripte. Die Farbkodierung ist wie in Tabelle 2-1: grün CreX, rot BioD, schwarz BioC, blau BioB. In der oberen Leiste sind die Expressionswerte in Abhängigkeit vom Volumen des Kontrollmixes dargestellt. Graphik ***A*** *zeigt die errechnete Molare Menge in Abhängigkeit vom Volumen des Kontrollmixes laut Tabelle 2-1. Die großen Sterne markieren den Standardfall, bei dem 3 µl Kontrollmix eingesetzt werden. In Graphik* ***B*** *sind die Werte der PMQ-Methode dargestellt. Die Gleichungen der Regressionsgeraden lauten von unten (blau BioB) nach oben (grün CreX): y=1,15+4,4x; y=6+12,2x; y=36+20x. Graphik* ***C*** *zeigt die Werte der Affymetrix-Auswertung. Hier lauten die Gleichungen der Regressionsgeraden von unten (blau BioB) nach oben (grün CreX): y = -105+377x; y=-411+1016x; y=-1006+4624x; y=5189+6524x. In den unteren zwei Graphiken sind die Expressionswerte in Abhängigkeit von den Molaren Mengen aufgetragen (2,5 nM bis 2 µM, Tabelle 2-1). Die logarithmische Darstellung erscheint informativer.* ***D*** *zeigt die Ergebnisse der PMQ-Methode. Regression (graue Linie): $y=0,91x^{0,69}$.* ***E*** *zeigt das Ergebnis des Affymetrix-Verfahrens. Regression: $y=80,7x^{0,86}$.*

Die Graphiken *A* bis *C* zeigen für beide Auswertemethoden einen linearen Zusammenhang, also Proportionalität von Transkriptmenge und Expressionswert. Allerdings ist der Nulldurchgang nicht gegeben. Die Werte für *CreX* erreichen den Sättigungsbereich, was durch das Abknicken der Steigung bei 3 µl und 6 µl für beide Methoden zu erkennen ist. Errechnet man die eingesetzten Molaren Mengen aus dem Volumen des Kontrollmixes und Konzentrationen der einzelnen bekteriellen Transkripte, so zeigt sich eine Abhängigkeit des

Proportionalitätsfaktors vom Transkript oder zumindest vom Expressionsniveau. Logarithmiert man die Skalen, so zeigt sich ein annähernd linearer Zusammenhang (Abbildung 2-7 D, E), was für eine Gesetzmäßigkeit spricht. Das bedeutet, vermutlich gibt es zwar keinen linearen aber einen funktionalen Zusammenhang zwischen der Molaren Menge der Messenger-Moleküle und dem Expressionssignal.

Zusammenfassend lässt sich im Vergleich zu dem von Affymetrix bereitgestellten Verfahren feststellen, dass die Reproduzierbarkeit der Werte von Experimenten hoher Qualität ähnlich gut ist. Auch die Expressionsunterschiede lassen sich gut reproduzieren, unabhängig davon mit welcher Methode man die Experimente auswertet. Das PMQ-Verfahren wurde für Experimente entwickelt, die auf geringe Mengen RNA unterschiedlicher Qualität beruhen. Es zeigte sich, dass die PMQ-Methode robuster ist bei großen Signal- und Streuungsschwankungen. Das ermöglicht die Auswertung von Experimentserien, die sonst nicht vergleichbar wären. Die Variantion der Spike-Kontrollen lässt einen gesetzmäßigen Zusammenhang zwischen der Molaren Menge einer mRNA-Spezies und dem gemessenen Expressionssignal vermuten. Die Spike-Transkripte sind jedoch bakteriellen Ursprungs und haben nicht die gleiche Aufarbeitung durchlaufen, wie die RNA-Probe, so dass sie nicht als repräsentativ für Transkripte in humanen Zellen angesehen werden können. Eine Kontrollstrategie mit vergleichbaren Spike-Transkripten, die möglichst früh bei der RNA-Aufbereitung eingebracht werden, ist denkbar. Die gute Korrelation in Abbildung 2-7 D, E spricht dafür, dass sich so die Normierung wesentlich verbessern ließe. Aufgrund der großen Anzahl an Experimenten weltweit und der Erfolg versprechenden Modellierungsansätze [9], besteht Hoffnung, dass das Hybridisierungsverhalten der einzelnen Oligosonden bald gut vorhergesagt werden kann.

2.6. Literatur zu Verfahren der Datenanalyse

Parallel zur Entstehung und Verbesserung der Array-Technologie selbst entwickeln sich die Verfahren zur Datenanalyse. Es stellte sich heraus, dass sich die Qualität der Ergebnisse durch neue Algorithmen deutlich verbessern ließ. Anfänglich wurden diese Entwicklungen hauptsächlich innerhalb von Firmen vorangetrieben. In den letzten Jahren sind aber die Preise der Arrays rapide gesunken und damit auch für akademische Gruppen erschwinglich geworden. Im Zuge dessen findet man jetzt umfangreiche Datensätze frei zugänglich, und auch in der Datenanalyse konnten wesentliche Fortschritte publiziert werden. Gerade in den

letzten zwei Jahren ist eine Flut von Arbeiten erschienen, die sich aber zu großen Teilen auch überschneiden. Die relevantesten Publikationen in Bezug auf dieses Kapitel seien hier kurz vorgestellt. Eine Reihe von Autoren beschäftigt sich direkt mit der Normierung und Datentransformation. Ziel ist dabei immer die Verbesserung der Vergleichbarkeit von Chipexperimenten. Man minimiert die technisch bedingte Varianz durch Datentransformationen ohne die biologisch interessanten Unterschiede zu verfälschen. In [10] wird ähnlich wie bei den hier vorgestellten Analysen eine stückweise polynomiale Regression mit Hilfe der R-Funktion *loess* eingesetzt. [11] verwendet kubische Splines. Kürzlich wurde auch gezeigt, dass durch geeignete Verfahren die Varianz der Expressionsdaten über den gesamten Messbereich stabilisiert werden kann [12], [13]. In den weitreichendsten Ansätzen wird versucht die Expressionssignale direkt zu modellieren. Laut den Autoren sind die Ansätze sehr vielversprechend [9], [14]. Außerdem wird in den Arbeiten auch darauf hingewiesen, dass Schätzer die nur auf den PM-Signalen beruhen, informativer sind, was sich mit unseren Erfahrungen deckt (siehe oben).

Kommerziell erhältliche Softwarepakete, wie beispielsweise der Expressionist von GeneData (Basel, Schweiz)[1], konzentrieren sich meist stärker auf Qualitätskontrolle, Datenmanagement und Endanwenderanalysen. Unter letzteren seien Cluster- und Hauptkomponentenanalyse, Korrelationsanalyse und eine Reihe von Verfahren zur graphischen Aufbereitung der hochkomplexen Daten erwähnt. Für die Rohdatenanalyse beschränkt man sich im Wesentlichen auf die Methode von Affymetrix.

2.7. Datenmanagement

Parallel zur Entwicklung der Analysealgorithmen wurde begonnen, eine Datenbank aufzubauen und die Ein- und Ausgabe der Daten zu organisieren und zu automatisieren. Dieser Abschnitt umfasst eine kurze Erläuterung der Vorteile des Einsatzes von Datenbanksystemen (DBS) für die Verwaltung von Massendaten. Anschließend folgen die Beschreibung der Datenmodellierung und dessen Ergebnis, das Datenbankschema. Abschließend folgen einige Aspekte der Implementierung.

[1] www.genedata.com

Spricht man von einer Datenbank, meint man in erster Linie eine große Menge strukturierter Daten, die in elektronischer Form abgespeichert sind. Im Bereich der Bioinformatik kennt man hier unter anderem die Sequenzdatenbanken. Das sind mittlerweile Gigabyte große Dateien mit speziellen Suchindexen. In der Informatik finden relationale Datenbanken die weiteste Verbreitung, und sie sind das Mittel der Wahl, wenn es gilt, große Datenvolumen zu verwalten. Es ist deshalb auch verwunderlich, dass sich bei den Sequenzdatenbanken immer noch die flache Dateistruktur hält. Die Basisstruktur zur Speicherung von Daten in jedem relationalen Datenbanksystem ist die Relation oder die Tabelle. Jedes Datenobjekt der realen Welt wird als Eigenschaftstupel repräsentiert. Beispielsweise könnte man Personen durch die Eigenschaften Name, Vorname, Geburtsdatum und Adresse charakterisieren. In der Datenbank entstünde dann eine Tabelle mit den entsprechenden Spalten für jede Eigenschaft. Beziehungen zwischen Objekten können ebenfalls in Form von Relationen also Tabellen ausgedrückt werden. Für das Beispiel ließe sich ein Relation „ist Tochter von“ erstellen, wobei dann bestimmte Spalten die Tochter spezifizieren und andere Spalten die Mutter oder den Vater. Oft belegt man die Tabellen mit Primärschlüsseln[1], um das Konsistenthalten zu vereinfachen und das Suchen zu beschleunigen. Ein wichtiger praktischer Vorteil relationaler Datenbanken ist, dass es verschiedene umfangreiche Softwaresysteme gibt, die zumindest in ihrer Kernfunktionalität ausgereift und für die Handhabung großer Datenvolumina ausgetestet sind. Jedes Datenbanksystem bietet auch eine Variante der standardisierten Abfrage- und Datenmanipulationssprache SQL.

Das Buch von [15] erwies sich als ausgesprochen hilfreich bei der Planung und Erstellung unseres Systems. Allerdings völlig im Gegensatz zur Lehrbuchvorstellung ist der Datenbankentwicklung bei uns nicht eine Analyse- und Planungsphase vorangegangen. Weder war anfänglich klar, welche Daten in die neue Form überführt werden sollen, noch war abzusehen, welche Struktur am besten die im Entstehen begriffenen Auswerteverfahren unterstützt. Die Datenbank wuchs anfänglich eher wie ein Geschwür. Waren neue Datentypen für die Auswertung erforderlich, so entstanden zusätzliche Spalten oder neue Tabellen. Roeing: *„Datenbanken aufzubauen bedeutet den Weg zwischen zwei sehr unterschiedlichen*

[1] Ein Primärschlüssel ist ein eindeutiger Bezeichner für jedes Tupel in einer Relation, also für jede Zeile einer Tabelle. Oft wird einfach eine laufende Nummer verwendet, die bei jedem neuen Tabelleneintrag (INSERT) hochgezählt wird.

Welten zu beschreiten. Auf der einen Seite steht de reale Welt in ihrer vielschichtigen Komplexität, auf der anderen der EDV-technische Ausschnitt, der durch Hard- und Software in der Machbarkeit stark beschränkt ist.“

Mitte 2001 konnte dann die Datenbankgruppe die gesamte Datenbank neu strukturieren und schließlich Anfang 2002 wurden die Daten in das neue Schema übernommen. Um das volle Potenzial der Datenbanktechnologie nutzen zu können, reicht es nicht aus, die schieren Expressionsdaten zu speichern und bei Bedarf auszulesen. Zusätzlich müssen auch die Sequenz- und Probendaten modelliert und eingepflegt werden. Gehen die Probenzahlen in die hunderte, wird der Nutzen einer durchdachten DB-Struktur bald deutlich. Gerade bei Patientendaten von klinischen Partnern gestaltet es sich oft schwierig, die notwendigen Informationen lückenlos zu beschaffen und in automatisch auswertbare Form zu bringen. Beispielsweise gehören die Verlaufsdaten bei Krebspatienten zu den wichtigsten klinischen Parametern, und sie müssen sorgfältig über Jahre erfasst werden. Die Schwierigkeit bei der Planung einer neuen Datenbank liegt nun darin abzusehen, welche Daten notwendig sind, um eine anfangs oft nicht einmal klar formulierbare Fragestellung zu bearbeiten. Versucht man dabei zu viel zu erfassen und zu verwalten, droht man leicht von der Datenflut überwältigt zu werden. Ein anschauliches Beispiel bieten hier Prozesse in einem molekularbiologischen Labor. Jeder Prozess erfordert eine Vielzahl an Schritten und Bezeichnerwechseln, die alle einen gewissen Einfluss auf die Qualität des Ergebnisses haben. Eine akribische Erfassung böte die Möglichkeit, Fehler zurückzuverfolgen oder auch detaillierte Kostenrechnungen aufzustellen. Der Aufwand für den Aufbau, die Pflege und die Nutzung einer solchen allumfassenden Datenbank steht dazu allerdings in keinem Verhältnis. Es gibt eine Reihe von Systemen zur Auswertung von Array-Experimenten, in denen Datenbanken zum Einsatz kommen. Beispielhaft seien hier ArrayExpress[1] und der Expressionist von GeneData[2] erwähnt. Standardisierungsbemühungen waren bislang nicht erfolgreich.

[1] http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html

[2] http://www.genedata.com

Datenstruktur und -modellierung

Die mit der Expressionsanalyse verwobenen Daten lassen sich inhaltlich in drei Gruppen teilen, die Probendaten, die Sequenzdaten und die Experimentdaten zusammen mit den Rohintensitäten und den verdichteten Werten.

Zu den Probendaten zählt alles, was zu der RNA bekannt ist, die letztlich auf den Chip hybridisiert wird. Die Art und Struktur dieser Daten unterscheiden sich von Arbeitsgruppe zu Arbeitsgruppe erheblich. Welche Information zu den einzelnen Proben verfügbar und welche für die Auswertung wichtig ist, hängt stark von der spezifischen Aufgabenstellung und vom Untersuchungsmaterial ab. Eine allgemeingültige Struktur für beliebige Fragestellungen entwickeln zu wollen, erscheint daher utopisch. Für die Analyse von Krebsgewebe sind besonders Daten relevant, die den Patienten und den Tumor näher charakterisieren. Beispielsweise interessiert man sich für Unterschiede der Genexpression zwischen Invasionsfront und Resttumor innerhalb eines Kolonkarzinoms. Dazu muss aus den Probendaten abzuleiten sein, welche Chipexperimente zum gleichen Patienten mit dieser Diagnose durchgeführt wurden und zusätzlich muss die Lokalisation innerhalb des Tumors, woher die RNA stammt, gespeichert sein.

Unter Experimentdaten sind alle Informationen zu verstehen, die direkt mit der Chiphybridisierung und dem Aufbau der Arrays zu tun haben. In Art und Struktur der gespeicherten Daten spiegelt sich das Vorgehen beim Chipexperiment und bei der Datenanalyse wider. Beispielsweise erhebt man sicherlich in anderen Arbeitsgruppen andere Qualitätsparameter, die sich nicht in dem vorliegenden Datenbankmodell abbilden lassen. Für diesen Bereich scheinen allerdings Standardisierungen sinnvoll. Wäre man in der Lage, sich auf einheitliche oder wenigstens ähnliche Protokolle und wenige Auswerteverfahren zu einigen, so ließen sich Datensätze verschiedener Arbeitsgruppen besser vergleichen, und der Wert der erzeugten Daten würde sich vervielfachen.

Die Sequenzdaten umfassen alle Informationen, die zu den Sonden und den entsprechenden Transkripten in der Datenbank zusammengetragen worden sind. Will man die Expressionsdaten effizient nutzen, ist es unerlässlich, umfangreiche Information zu den Sonden und damit zu den untersuchten Transkripten einer automatischen Analyse zugänglich zu machen. Anfänglich begnügte man sich für jeden Chip mit einer Liste bestehend aus einer Spalte für die Sondensets und einer für die Sequenzbezeichner. Zum Teil fügte man noch eine

Kurzbeschreibung aus dem entsprechenden Eintrag in der Sequenzdatenbank bei. Für Einzelgenanalysen ist dieses Vorgehen für den Anwender in den meisten Fällen ausreichend. Findet man ein interessantes Expressionsmuster, so sucht man aus der Liste den Sequenzbezeichner heraus und hat damit über die bekannten Datenbanken Zugang zur Sequenz selbst und zu weiteren öffentlich verfügbaren Informationen. Meistens möchte man jedoch die Daten nicht nur einmal auswerten und die zehn besten Gene auswählen. Hat man umfangreiche Datensätze vorliegen, so ergeben sich zahllose Analysevarianten und interessante Fragestellungen, so dass immer andere Gene in den Fokus rücken. Es ist auch denkbar, die Analyse von vornherein mit einer Klassifizierung des Sequenzsets zu beginnen und anschließend die Expressionsdaten einzubeziehen. Beispielsweise könnte man sich für die Koexpression von Rezeptor-, Ligandenpaaren interessieren. Bei metaGen wird sukzessive die SequenzDB ausgebaut, die zwar nicht die Sequenzen selbst aber Informationen über die Sequenzen enthält. Im Fokus der Firma liegen besonders solche Eigenschaften, die das entsprechende Protein als Target für die Krebstherapie interessant machen können. Dazu zählen vor allem funktionelle Charakteristika und die zelluläre Lokalisation. Affymetrix hat Anfang 2002 den Nutzern eine umfangreiche Sequenzannotation über das Internet[1] zur Verfügung gestellt. Unter Verwendung des SRS-Systems[2] sind Datenbanken wie beispielsweise *RefSeq*, Geneontology und OMIM mit den Sondensets verknüpft. Zusätzlich wird eine Proteinannotation basierend auf einer HMM-Analyse[3] für die Transkripte geboten, für die der kodierende Bereich bekannt ist. Über das SRS-System sind die einzelnen Datenquellen indiziert und verknüpft, so dass der Anwender die Möglichkeit hat, für ein im Expressionsverhalten auffälliges Sondenset direkt abzurufen, welches Gen sich dahinter verbirgt und ob beispielsweise ein genetisches Syndrom dafür bekannt ist.

Die detaillierte Darstellung der Tabellen und des Schemas sind in Anhang A zu finden.

---

[1] http://www.affymetrix.com

[2] SRS-Sequence Retrieval System; http://srs.ebi.ac.uk

[3] HMM-Hidden Markov Model: weit verbreitetes Verfahren zur Suche nach Proteinmodulen, die strukturelle Eigenschaften finden helfen

Aspekte der Implementierung

Wie oben bereits beschrieben wuchs unsere Datenbank parallel zu den Experimenten. Sukzessive konnten Prozesse strukturiert und automatisiert werden. Als Datenbanksoftware wird bei metaGen Oracle (Oracle Corporation, Redwood Shores, USA) eingesetzt. Zum Laden großer, wohlstrukturierter Datenmengen dient unter Oracle ein spezielles Programm namens SQL*Loader.

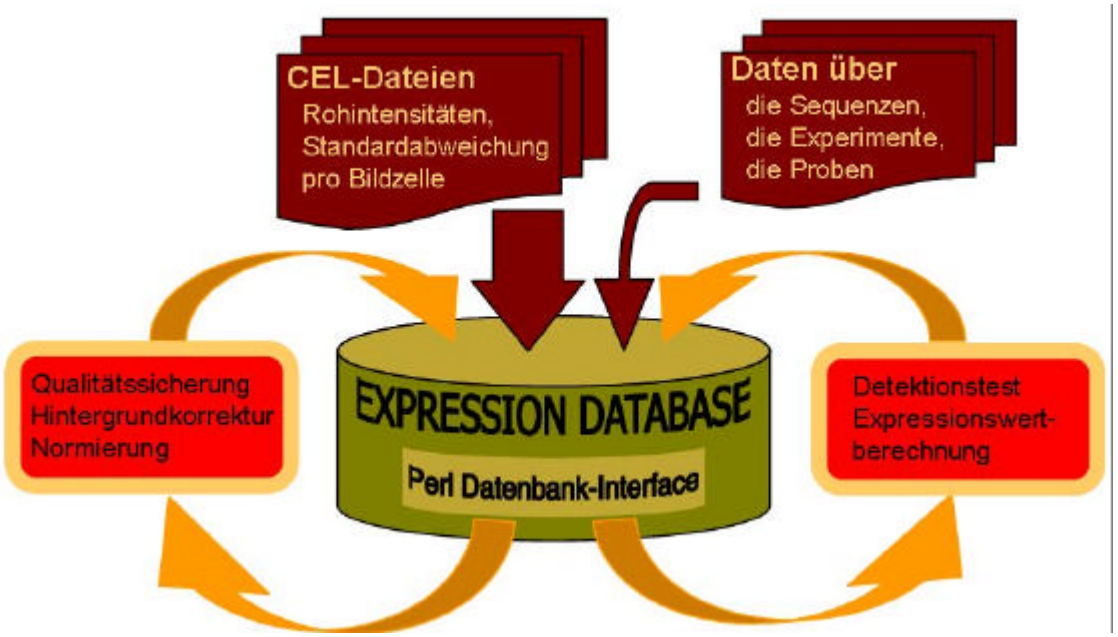


Abbildung 2-8 Schema der Datenflüsse

Zentral in grüner Farbe ist die in Oracle implementierte relationale Datenbank symbolisiert. Die eingehenden Daten stammen aus verschiedensten Quellen (siehe Text). Die äußeren roten Felder symbolisieren aufwendige Berechnungen, die effiziente Speicherzugriffe verlangen. Die Pfeile veranschaulichen Datenflüsse. Aufgrund dessen, dass zur Programmierung fast ausschließlich die Sprache Perl zum Einsatz kam, diente meistens auch das Perl-Modul DBI (database interface) für das Auslesen der Daten.

Die Erfassung der Probendaten erfolgt zum großen Teil über Masken und Erfassungsroutinen, die von der Datenbankgruppe erarbeitet wurden. Ein wichtiger Bestandteil sind hierbei Konsistenzchecks und die Verwendung kontrollierten Vokabulars. Zur Manipulation der Daten und zur Implementierung der in Abschnitt 2.3 beschriebenen Verfahren wurde die

Programmiersprache Perl[1] und das Statistikpaket R[2] eingesetzt. Am stärksten strukturiert sind die Probendaten. Bei der Erfassung ist größte Sorgfalt geboten, da von verschiedenen Personen zu verschiedenen Zeiten Daten erhoben werden. Zusätzlich unterscheidet sich die Art der Information zwischen Tumorentität. Das größte Datenvolumen enthält bei weitem die Rohdatentabelle der Oligoarrays (CEL_Raw, 16,3 GB). Hier muss darauf geachtet werden, dass die Speicherstrukturen einen effizienten Zugriff auf Teilmengen für verschiedenste Anwendungen und Analysewerkzeuge erlauben. Das erreicht man durch das Anlegen von Datenbankindexen (Volumen des Indexbereichs etwa 13,6 GB). Die gesamte Datenbank umfasst zurzeit etwa 800 Chipexperimente und benötigt einen Speicherplatz von 50 GB.

Wünschenswert wäre die Etablierung von Standards, so weit es möglich ist. Große Teile der Experiment- und der Sequenzdaten sollten sich vereinheitlichen lassen. Für etliche Teilprozesse der Array-Experimente gibt es ja bereits Quasistandards. Daten aus verschiedenen Labors ließen sich dann wesentlich leichter vergleichen. Die einzelnen Forschergruppen könnten sich dann auf die aufgabengerechte Modellierung ihrer Probendaten konzentrieren. Bisher sind jedoch befriedigende Lösungen nicht in Sicht[3].

---

[1] www.perl.org

[2] www.R-project.org

[3] Ein Konsortium, dass sich um Standardisierungen bemüht: www.mged.org

3. Genexpressionsanalyse in der Lungenkrebsforschung

Im November 2001 sind in PNAS[1] zwei umfangreiche Studien zur Genexpressionsanalyse in Lungentumoren erschienen ([16], [17]). Für die eine Arbeit bediente man sich der Oligo-Array-Technologie und für die andere der cDNA-Array-Technologie. Die Datensätze eignen sich, die Anwendbarkeit der im zweiten Kapitel vorgestellten Verfahren zu demonstrieren, und die unterschiedlichen Plattformen für die Genexpressionsanalyse zu vergleichen. Außerdem lassen sich anhand dieser Daten ausgezeichnet typische Fragestellungen der Krebsforschung diskutieren. In diesem Kapitel wird zuerst ein Überblick über das Bronchialkarzinom gegeben und anschließend der Ansatz und die Ergebnisse der beiden Publikationen erläutert. Die Forschungsziele der Arbeiten von Bhattacharjee et al und Garber et al unterscheiden sich kaum, sogar der Fokus liegt bei beiden auf den Adenokarzinomen. Der Hauptteil des Kapitels beschreibt eine vergleichende Reanalyse der beiden Datensätze. Dafür bedarf es einer Vorverarbeitung der Daten, die Inhalt von Abschnitt 3.6 ist. Allgemein wird der Frage nachgegangen, ob die Art und Weise des Einsatzes der Array-Technologien den angestrebten Versuchszielen angemessen sind. Unter anderem wird untersucht, ob die Daten Aussagen über die absolute Genexpression zulassen oder nur für relative Messungen eingesetzt werden sollten. Ein oberflächlicher Vergleich der veröffentlichten Daten zeigt, dass die Ergebnisse erheblich von einander abweichen und folglich die Analysen nicht redundant sind. Geht man davon aus, dass beide Technologien verwertbare Resultate liefern, so kann es zwei wesentliche Gründe für die großen beobachteten Unterschiede geben, die jeweils unterschiedliche Konsequenzen nach sich ziehen:

I. Die biologische Vielfalt der Genexpression der untersuchten Gewebe ist erheblich größer als die beobachtete. Das hieße, dass die Probenzahl bei weitem nicht ausreicht, um alle differenziell expremierten Gene finden zu können. In der Konsequenz wäre die *Summe* der differenziellen Gene aus beiden Versuchsreihen als potenziell tumorrelevant weiter zu verfolgen.

II. Mit beiden Verfahren misst man zwar die richtige Tendenz, verfälscht aber teilweise stark die tatsächlichen Häufigkeiten der mRNA-Moleküle. Geht man davon aus, dass es relativ wenige Schlüsselfaktoren gibt, die die Zellen transformieren und diese für

[1] PNAS: Procedings of the National Academy of Science

einen Tumortyp immer wieder ähnlich sind, so sollte man weitere Analysen auf die *Schnittmenge* der differenziellen Gene der beiden Versuchsreihen beschränken.

Im Folgenden wird auf die oben beschriebenen Gründe als *Arbeitshypothese I und II* verwiesen. Die detaillierte Analyse in Abschnitt 3.7 ergibt starke Hinweise für Arbeitshypothese II. Nachfolgend können Listen von Genen generiert werden, die in beiden Datensets konsistent differenziell zwischen Tumor- und Normalproben exprimiert sind. Zur Validierung der Ergebnisse dienten die in einem Übersichtsartikel als tumorrelevant beschriebene Gene. Die Trennung aller Adenokarzinompatienten mit guter beziehungsweise schlechter Prognose aufgrund der Genexpressionswerte mit Hilfe von Clusterverfahren gelingt nicht.

3.1. Lungenkrebsstatistiken [18]

Lungenkrebs ist mit 12,3% die häufigste Krebsform weltweit. Im Jahr 2000 gab es schätzungsweise 1,2 Millionen Neuerkrankungen, wobei Tabakrauchen mit 80-90% definitiv die häufigste Ursache ist. Tabak wurde von Seefahrern zur Zeit der großen Entdeckungen aus Amerika nach Europa gebracht. Die rasante weltweite Verbreitung der Zigarette begann Ende des 19. Jahrhunderts mit der Erfindung der Bonsack Zigarettendrehmaschine. Bis zu diesem Zeitpunkt wurde Lungenkrebs ausgesprochen selten diagnostiziert. In der medizinischen Literatur waren bis 1898 weltweit insgesamt nur 140 Fälle beschrieben. Durch massive Studien in den fünfziger und sechziger Jahren konnte das Tabakrauchen unanfechtbar als ursächlich für Lungenkrebs nachgewiesen werden [19]. Bronchialkarzinome stellen in dieser Beziehung eine Ausnahme dar, da für keine andere Krebsform ein so deutlicher kausaler Zusammenhang zu einem einzigen Risikofaktor gezeigt werden konnte. Damit ist auch der Weg zur Vermeidung beziehungsweise Minimierung des Auftretens von Lungenkrebs klar vorgezeichnet. Er besteht schlicht aus der Eindämmung des Rauchens, vor allem dadurch, dass man versucht, das Anfangen mit dem Rauchen von Kindern und Jugendlichen zu verhindern und dass man das Aufhören unterstützt und fördert. Jemand, der zeitlebens raucht, trägt ein 20-30 fach höheres Risiko, Lungenkrebs zu entwickeln als ein Nichtraucher. Einmal erkrankt sterben trotz verbesserter Therapien etwa 90% der Patienten. Für 2000 schätzt man weltweit 1,1 Millionen Todesopfer, das entspricht 17,8% aller Krebsopfer. Etwa 11% der starken Raucher entwickeln in ihrem Leben Krebs. Seltene familiäre genetische Prädispositionen können zu einem bis zu 2,5 fach höheren Krebsrisiko führen. Als weitere

Risikofaktoren gelten das Passivrauchen, Asbest und Radon. Die karzinogene Wirkung von Radon wurde vor allem bei Bergarbeitern untersucht.

3.2. Pathogenese und Verlauf der Erkrankung

Als Quellen für diesen Abschnitt dienten die Lehrbücher über Histologie [20], über Pathologie [21] und über die molekulare Grundlagen von Krebs [22].

Das proximale[1] Lungenepithel besteht aus mehrreihigem, hochprismatischem Epithel mit Zilien an seiner luminalen Oberfläche. Wichtiger Bestandteil der Epithelschicht sind die schleimbildenden Becherzellen. Eingestreut in die Basalschicht des Epithels sind neuroendokrine Zellen. Weiter distal werden die Epithelzellen eher würfelförmig und verlieren ihre Zilien bis sie in den Alveolen (Lungenbläschen) schließlich plattenepithelförmig sind [22]. Die schleimbildenden Becherzellen werden auch von den zentralen Atemwegen zur Peripherie hin seltener. Bronchialkarzinome entspringen in den meisten Fällen (~70%) dem mehrreihigen, hochprismatischen Epithel der proximalen Atemwege. Kleinzellige Tumoren entwickeln sich sehr wahrscheinlich aus neuroendokrinen Vorläuferzellen. Adenokarzinome sind meist peripher lokalisiert und entstehen aus den schleimbildenden Epithelien der Bronchien, aus den Clara-Zellen oder aus den Pneumozyten vom Typ II. Großzellige Lungentumoren sind entdifferenzierte Formen der anderen nichtkleinzelligen Subtypen [21]. Die folgende Abbildung zeigt einen Schnitt durch bronchiales Epithelgewebe. Darstellungen der Bronchialwand, eines Adenokarzinoms und eines Plattenepithelkarzinoms sind im Anhang B zu finden.

[1] proximal: zur Körpermitte hin gelegen, im Gegensatz zu distal

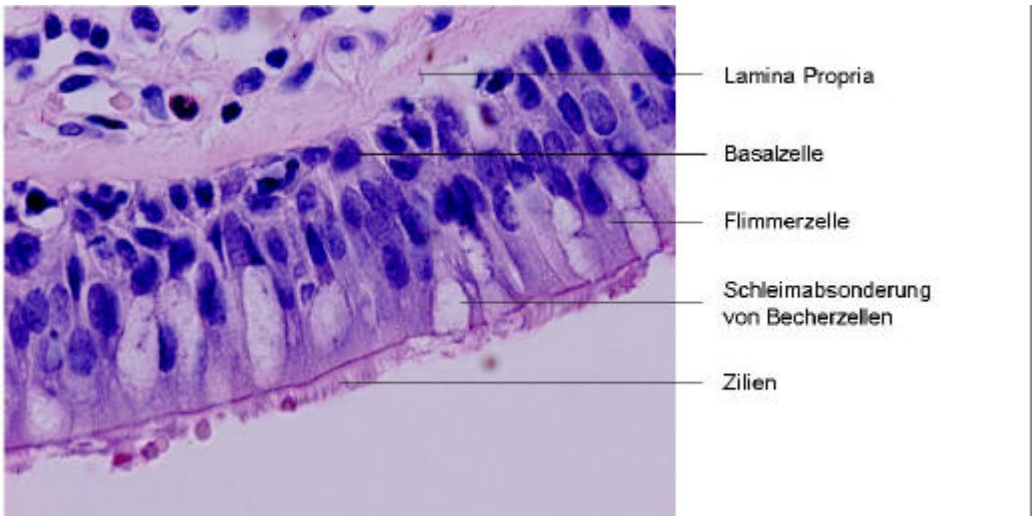


Abbildung 3-1 Histologie des normalen Lungengewebes
Mit einem Lichtmikroskop aufgenommene Schnitte (Hämatoxylin/Eosin-Färbung) von histologisch normalem Lungenepithel der Bronchialwand (40X). Das Epithel wird als mehrreihig und hochprismatisch beschrieben. Weitere Aufnahmen sind im Anhang zu finden.

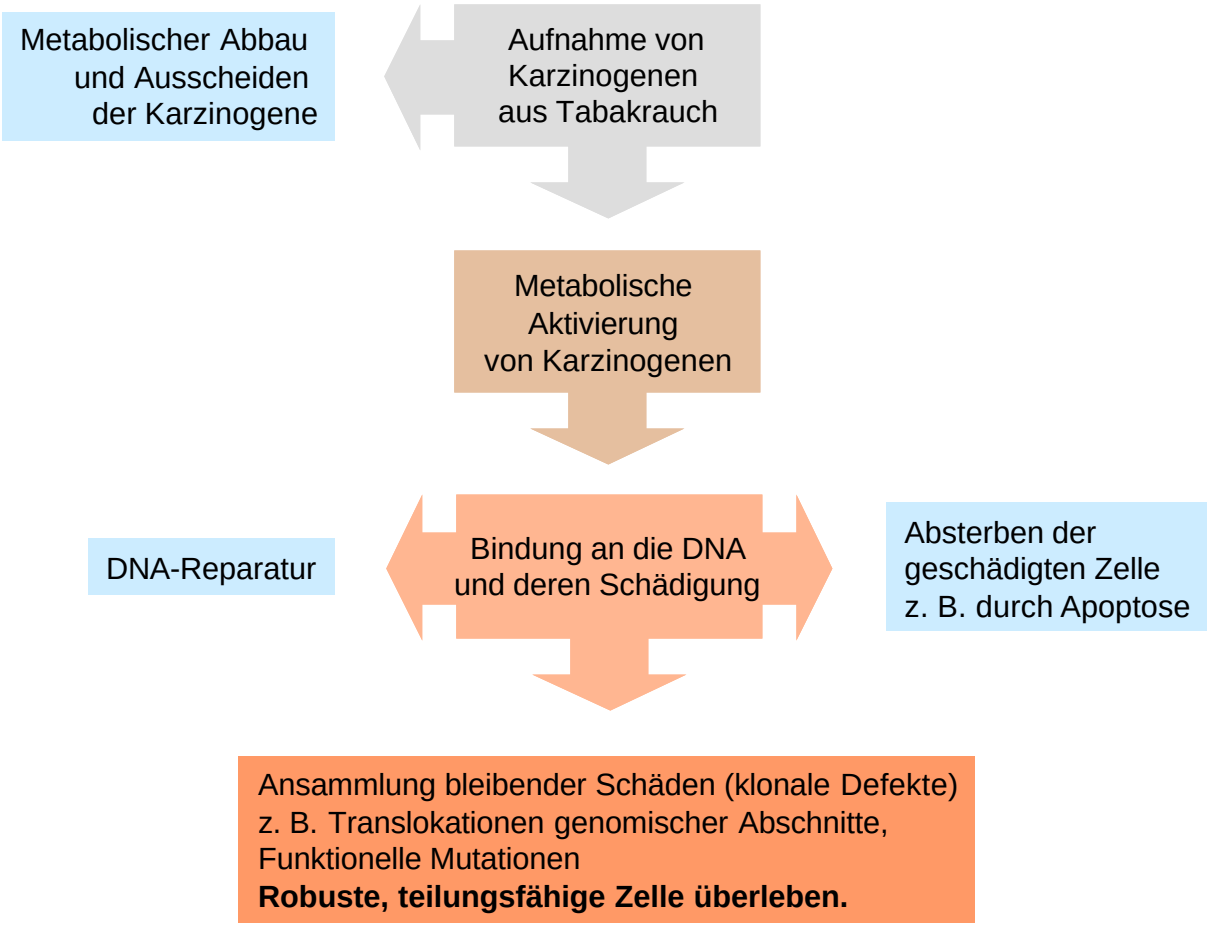
Epithelzellen teilen sich permanent und sind ständig dem Tabakrauch und damit den im Rauch enthaltenen Karzinogenen ausgesetzt. Das motiviert die Vorstellung der Karzinogenese als Evolution Darwinschen Typs von einzelnen Körperzellen. Durch sukzessive Akkumulation somatischer Mutationen[1] erreichen einige Zellen Wachstumsvorteile und schließlich die Fähigkeit zur Überwindung zellulärer Kontrollmechanismen wie Kontakthemmung oder Anoikes[2]. Das führt erst zu Zellen, die in der Lage sind, mehrschichtig zu wachsen (Hyperplasien) und später aufgrund von Nährstoff- und Platzmangel zu immer aggressiveren, invadierenden und metastasierenden Tumorzellen. Bei Untersuchungen des Epithels von Rauchern ohne Lungenkrebs fand man zelluläre Atypien wie beispielsweise den Verlust von Zilien und eine verstärkte Zellproliferation. Außerdem ergaben molekularbiologische Analysen hunderttausende bis zu 90000 Zellen umfassende Gewebsläsionen, die

[1] Somatische Mutationen sind Mutationen in Zellen eines Organismus, die nicht die Keimbahn betreffen und damit nicht vererbt werden.

[2] Anoikes bedeutet auf Griechisch das Fallen der Blätter im Herbst. In der Zellbiologie bezeichnet man damit einen Mechanismus bei Vielzellern, der zum Absterben von Zellen führt, die sich aus dem natürlichen Verband herauslösen.

jeweils Allelverluste als klonale Defekte[1] aufwiesen. Bei der Krebsentstehung kann eine bevorzugte Ordnung der Allelverluste festgestellt werden. Zuerst gehen unterschiedliche Regionen von Chromosom 3p verloren und anschließend von Chromosom 9p, dem Genort des bekannten Tumorsupressorgens *p16*. Außerdem konnte eine spezifische Änderung von methylierten Promotorbereichen in diesen chromosomalen Loci nachgewiesen werden [18].

[1] In jeder gesunden diploiden Zelle gibt es pro Gen zwei Allele. In Tumorzellen kommt es häufig zu chromosomalen Abberationen, die zum Gewinn oder Verlust von Allelen bestimmter Gene führen. Ist die Zelle trotz der Abberation lebens- und teilungsfähig und vererbt dadurch den Defekt an ihre Tochterzellen, so spricht man von einem klonalen Defekt.



Schema 3-1 Hypothetischer Verlauf der Karzinogenese[1]

Zigarettenrauch enthält über 20 bekannte Lungenkrebsspezifische Karzinogene, darunter Polyzyklische Kohlenwasserstoffe und tabakspezifische Nitrosamine, wobei wohl NNK: 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone das am besten untersuchte ist. Die aus dem Tabakrauch von den Zellen aufgenommenen Verbindungen werden von körpereigenen Enzymen verstoffwechselt. Man hat zum Beispiel herausgefunden, dass Proteine der Glutathione-S Transferase Familie an der Detoxifizierung polizyklischer aromatischer Kohlenwasserstoffe beteiligt sind und dass Proteine der Zytochrom *P450* Familie für die Aktivierung von NNK wichtig sind. Aktivierte Karzinogene binden an die DNA und schädigen diese in dem sie beispielsweise eine Transversion von Guanin zu Thymin bewirken [18]. Man schätzt, dass jede Zelle etwa 1000 bis 100000 derartige Schädigungen pro Tag

[1] Das Schema ist abgewandelt aus [18] Minna, J. D.; Roth, J. A. und Gazdar, A. F. (2002): Focus on lung cancer, Cancer Cell (Band 1), Nr. 1, Seite 49-52. übernommen.

erfährt. Sie werden im besten Fall repariert, treiben die Zellen in die Apoptose oder persistieren im schlimmsten Fall als somatische Mutationen. Letztere können wiederum zur Störung der Regulation oder Fehlfunktion von Proteinen oder zu genomischen Instabilitäten führen. Die beobachtbare Tumorprogression von der Hyperplasie über das Karzinom in situ bis zum invasiven beziehungsweise metastasierenden Karzinom unterstützt die Vorstellung, dass erst eine Reihe spezifischer genetischer Veränderung zur Entstehung eines Tumors führen.

3.3. Standarddiagnose und –therapie

Die in diesem Abschnitt dargestellten Daten sind aus [18] und [23] entnommen oder aus dem Internet[1].

Als häufigste Symptome des Bronchialkarzinoms sind Husten und später Schmerzen und Gewichtsverlust beschrieben. Im Allgemeinen gilt, dass die Erkrankung erst relativ spät erkannt wird, da sie, in der Peripherie lokalisiert, in den Frühstadien kaum Beschwerden verursachen. Die Bewertung eines Lungenkrebspatienten erfolgt oft durch histopathologische Charakterisierung von Gewebematerial, dass meist durch Biopsien gewonnen wird. Häufig angewendete Verfahren sind außerdem Computertomographie und Lungenfunktionstests. Neuere Verfahren sind die flexible Fibreoptik-Bronchoskopie und die CT-geführte Feinnadelbiopsie. Der Befund wird nach der internationalen TNM-Klassifikation[2] bewertet und eingeordnet. Die große Mehrzahl der Bronchialkarzinome lassen sich histologisch in vier Subtypen unterteilen, und zwar den kleinzelligen Lungenkrebs und die drei nichtkleinzelligen Formen das Plattenepithelkarzinom, das Adenokarzinom inklusive dem bronchiolo-alveolären Karzinom und dem großzelligen Bronchialkarzinom. Die wichtigste und therapierelevante Unterscheidung ist die zwischen kleinzelligen und nichtkleinzelligen Tumoren und die Einstufung des Patienten nach dem Grad der Erkrankung. Bei nichtkleinzelligen Tumorpatienten werden meistens noch Lymphknoten aus dem Mediastinum untersucht, um frühzeitig Hinweise auf eine mögliche Metastasierung zu erhalten. Bei den meisten Patienten tritt die Krankheit erneut auf und entwickelt Fernmetastasen. Das liegt zum einen an der

[1] Informationsnetzwerk: http://www.alcase.org

[2] http://www.hc-sc.gc.ca/hpb/lcdc/bc/ccocs/lungpoumon/lcg_i_e.html

Aggressivität der häufigsten Formen, die früh anderes Gewebe invadieren und metastasieren und zum anderen daran, dass die Krankheit nicht rechtzeitig erkannt wird, was sich dann negativ auf die Prognose bei chirurgischen Eingriffen auswirkt. Allgemein gilt, dass Patienten mit kleinzelligen Bronchialkarzinomen eine schlechtere Prognose haben.

Die derzeitigen Standardtherapien für die verschiedenen Tumortypen sind in Tabelle 3-1 zusammengefasst. Nichtkleinzellige Bronchialkarzinome werden, soweit es irgendwie möglich ist, chirurgisch entfernt. Generell führt eine Therapie, auch wenn keine Heilung erreicht werden kann, zur Linderung der Symptome und meistens zu einer Verlängerung des Lebens des Patienten.

Tabelle 3-1 Standardtherapieformen beim Bronchialkarzinom

	Kleinzelliges Bronchialkarzinom		Nichtkleinzelliges Bronchialkarzinom	
Progression	Lokal beschränkt	fortgeschritten	Kein mediastinaler Lymphknotenbefall	Metastasen in den Lymphknoten
Strahlen-therapie	Brustkorb mit Chemo-therapie		Bei Inoperabilität	Hohe Dosis oder mit Chemotherapie und/oder Resektion
Chemo-therapie (Cisplatin, Etoposide, Irinotecan)	In Kombination mit Strahlen-therapie	Mehrere Therapeutika in Kombination		Neoadjuvant oder in Kombination mit Strahlentherapie
Resektion			Lobe-, oder Pneumektomie	Resektion in Kombination mit Chemotherapie
Therapieziel	Heilung, Linderung	Linderung, Lebens-verlängerung	Heilung, Linderung	Linderung, Lebensverlängerung
5-Jahres-Überleben	10% - 15%	< 5%	50-60% Stadium I 15-20% Stadium II	< 12%
Erreichte mittlere Überlebens-verlängerung		10 – 16 Monate		1 Jahr 35%, 2 Jahre 15%

Quellen: [23], http://www.alcase.org

Auf Grund der vielen Neuerkrankungen und der eher bescheidenen Therapieerfolge wird auf breiter Front nach neuen Behandlungsformen geforscht. In der Entwicklung befinden sich unter anderem neue bildgebende Verfahren wie Autofluoreszenz – Bronchoskopie, effektivere

Strahlentherapieformen und neue Chemotherapeutika mit geringeren Nebenwirkungen. Einige Medikamente aus der molekularbiologischen Forschung befinden sich auch mittlerweile in der klinischen Entwicklung, darunter sind Inhibitoren für Angiogenese, EGF und Her2/neu Rezeptoren und weiteren Tyrosinkinasen, RAS, Farnesyltransferasen, CDK. In einem anderen Ansatz der Chemoprävention wird versucht schon die Entstehung der Krankheit beispielsweise durch Vitamin E zu verhindern oder wenigstens zu verzögern.

3.4. Die Arbeit von Garber et al [17]

In diese Studie gingen insgesamt 67 Tumorproben aus insgesamt 56 Patienten mit unterschiedlichen histo-pathologischen Befunden ein. Die größte Kohorte bestehend aus 41 Proben stammt von Adenokarzinomen. Die restlichen Proben setzen sich wie folgt zusammen: 16 Plattenepithelkarzinome, jeweils fünf Proben großzelliger (LCLC), kleinzelliger Bronchialkarzinome (SCLC) und normaler Bronchialgewebe und eine fötale Gewebeprobe. Von 11 Patienten konnten jeweils zwei Tumorproben gewonnen werden. Die Experimente sind im Labor von Brown und Botstein in Stanford nach deren Standardverfahren [24] durchgeführt worden. Den Autoren zu Folge wurden auf die 24000-Spot-Arrays 23100 cDNA-Klone aufgebracht, die wiederum 17108 Gene repräsentieren. Die Gewebeproben wurden grob zurechtgeschnitten, anschließend die mRNA extrahiert und unter Zugabe von Cy5-kojugierten dUTP's revers transkribiert. Für die Hybridisierung wurde ein Referenzpool aus 11 Zelllinien erzeugt, ebenso die RNA extrahiert und unter Prominenz von Cy3-conjugierten dUTP's revers transkribiert. Eine detailiertere Beschreibung der Technologie befindet sich in Abschnitt 1.6 in der Einführung. Der Datenanalyse hat man eine Filterkaskade vorgeschaltet, die zur Auswahl von 835 Genen, repräsentiert durch 918 Spots auf dem cDNA-Array, führte. Damit beschränkt man alle folgenden Analysen auf nur noch etwa 4% der Originaldaten.

Die Filterkaskade:

1. Signale, die im Referenzpool nur wenig stärker als der Hintergrund sind, wurden als schwach markiert. Schwellwert: Signal / Hintergrund = 1.5

2. Nur Spots die in höchstens 5 der 73 Experimente als schwach erscheinen werden weiterbetrachtet.

3. Anschließend wird eine Heuristik angewendet, die darauf abzielt, sich bei der weiteren Analyse möglichst auf Gene zu beschränken, die möglichst reproduzierbare Expressionswerte liefern und gleichzeitig zwischen den Proben variieren (siehe Originalarbeit).

Als repräsentativen Expressionswert benutzten die Autoren den LogRatio, der sich pro Spot aus dem Logarithmus des Quotienten von hintergrundkorrigierten Cy3-Signal und Cy5-Signal ergibt. Die Daten analysierte man mit der bekannten Implementierung eines hierarchischen Cluster-Verfahrens von Eisen und Spellmann. Dadurch ließen sich Gene identifizieren, die spezifisch in histologischen Tumorsubtypen exprimiert sind. Durch eine detaillierte Analyse der Adenokarzinomproben konnte man drei Gruppen identifizieren, die sich in ihren Expressionsmustern unterscheiden. Im Vergleich der Gruppen ergab eine Kaplan-Meier-Analyse der Überlebensdaten der Patienten einen deutlichen Einfluss der Gruppenzugehörigkeit auf die Lebenserwartung. Mit Hilfe des T-Tests generierte man Listen von Genen, die spezifisch zwischen den Tumorsubtypen differenziell exprimiert sind.

3.5. Die Arbeit von Bhattacharjee et al [16]

In diese Studie sind 186 Gewebeproben von Lungenkrebspatienten und 17 Normalgewebeproben eingegangen. Auch hier ist für die Adenokarzinome die größte Probenzahl erfasst worden, nämlich 127. Darüber hinaus analysierte man 21 Plattenepithelkarzinome, 20 pulmonary Carcinoid, 6 kleinzellige Lungenkarzinome und 12 Adenokarzinome aus Metastasen (siehe Tabelle 3-2 im folgenden Abschnitt). Von 36 Patienten gingen zu Validierungszwecken mehrfach Proben in die Analyse ein. Bei der Asservierung des Gewebes achtete man darauf, dass es bis spätestens 30 Minuten nach der operativen Entnahme bei -140 °C eingefroren war. Unter den 125 Patienten mit öffentlich verfügbaren klinischen und histologischen Daten befinden sich 53 Männer und 72 Frauen. Dabei sind 17 als Nichtraucher, 51 als durchschnittliche und 54 als starke Raucher klassifiziert. Die RNA wurde im Unterschied zu dem bei metaGen verwendeten Verfahren durch Zugabe von Trizol und anschließender Aufreinigung über eine Säule aus den Zellen extrahiert. Um zu prüfen ob die mRNA degradiert ist, generierte man mit einem Teil der RNA einen Northern-Blot und hybridisierte anschließend eine Sonde für *ß-Aktin*. Ist die mRNA intakt, so beobachtet man eine saubere Bande auf der richtigen Höhe (siehe auch Kapitel 5). Dem von Affymetrix empfohlenen Protokoll folgend, wurde ausgehend von 15-20

Mikrogramm Total-RNA unter Zugabe von T7-Oligo-dT-Primern eine Erststrangsynthese und anschließend eine Zweitstrangsynthese durchgeführt. Die Invitrotranskription, die Fragmentierung und die Hybridisierung liefen analog dem in der Einleitung beschriebenen Standardverfahren. Die Chips wurden mit 10 Mikrogramm aufbereiteter RNA bei 45°C für 16 Stunden hybridisiert.

Für die Datenanalyse wurden die aggregierten Werte der Affymetrix Standardauswertung benutzt, also pro Sondenset (*Probeset*) der Average-Differenz-Wert und der Detektionsscore[1]. Auch in dieser Arbeit wurde eine Filterkaskade zwischengeschaltet, um von möglichen Artefakten freizukommen und die Analyse auf eine möglichst informative Teilmenge der Gene einzuschränken. Als erstes sortierte man Experimente aus, die schon beim Scannen visuell Artefakte erkennen ließen oder bei denen nur 30% der Sondensets als *„present"* von dem Programm akzeptiert wurden. Dann wählte man ein „typisches" Experiment als Referenz und berechnete für jedes andere Experiment eine lineare Regressionsgerade bezüglich dieser Referenz. Eine Steigung der Geraden größer vier betrachtete man als Indiz für schlechte Qualität und eliminierte den Datensatz und wiederholte das Experiment. Für die Normierung der Datensätze bediente man sich allerdings eines nicht-linearen Verfahrens (siehe folgenden Abschnitt). Insgesamt hybridisierte man 52 Gewebeproben doppelt, wobei nach der Auswertung die Expressionswerte von 45 Replikaten einen Korrelationskoeffizienten größer 0,9 aufwiesen. Vergleichbar dem Ansatz von Garber et al schränkte man die Analyse auf die Gene ein, deren Expression sich einerseits verlässlich messen lässt und andererseits zwischen den Patienten variiert, also informativ ist. Genauer heißt das, für die Cluster-Analysen benutzte man nur Sondensets, die erstens in den 45 Wiederholungen einen Korrelationskoeffizient größer 0,8 aufwiesen und für die zweitens die Standardabweichung ihrer Expressionswerte über alle Experimente 50 Einheiten übersteigt. Diese Filteranwendung ergab 675 Gene. Auch in dieser Arbeit ließen sich für histologische Subtypen spezifische Gene identifizieren. Einige sind bereits als Marker beschrieben und werden von den Autoren diskutiert. Wiederum konnte man drei Subgruppen in den Adenokarzinomproben aufgrund

---

[1] Der in der Affymetrix-Software errechnete Detektionsscore leitet sich in der neusten Version aus dem Wilcoxon-Test ab (siehe Kapitel 2). In der für die vorliegende Auswertung eingesetzten Vorgängerversion wurde der Detektionsscore über eine Heuristik bestimmt und konnte die Werte *present, marginal, absent* annehmen.

ihrer Expressionsmuster unterscheiden. Und eine Kaplan-Meier-Analyse ergab einen Einfluss der Gruppenzugehörigkeit auf die Lebenserwartung.

3.6. Die eigene Analyse - Vorverarbeitung

Will man einen öffentlich verfügbaren Datensatz analysieren, so besteht der erste Schritt immer darin, sich die Datenquelle zu erschließen. Dazu gehören, das Verstehen von Struktur und Inhalt der Rohdaten, das Einbetten in die eigene Datenbank und die Aufarbeitung der Information über die untersuchten Proben, über die Laborexperimente und über die Gene. Es hat sich als erstrebenswert erwiesen, möglichst alles, was für eine Auswertung von Relevanz sein könnte, in die Datenbank einzupflegen. Das ist ein erheblicher Mehraufwand, der sich nur für Datensätze hinreichender Qualität und Vollständigkeit lohnt. Die Bearbeitung komplexer medizinischer und biologischer Fragestellungen kann durch die wohl strukturierte Speicherung der Information wesentlich unterstützt werden. (Siehe dazu auch Abschnitt 2.6)

Tabelle 3-2 Subklassifikation des Bronchialkarzinoms und Probenzahlen der Publikationen

Zelltyp	Häufigkeit[1]		Herkunft[2]	Probenzahlen	
	Frauen	Männer		Garber	Bhattacharjee
Kleinzelliges Bronchialkarzinom	20%	22%	Zentraler Abschnitt des Bronchialbaumes	5	6
Karzinoid	2%	2%	Diffuses neuroendokrines System	0	20
Adenokarzinom	46%	24%	Schleimbildende Epithelien der Lungenperipherie	41	127 und 12 Metastasen
Plattenepithel-karzinom	20%	40%	Schleimbildende Epithelien der Segment- und Subsegmentbronchien	16	21
Großzelliges Bronchialkarzinom	8%	7%	Entdifferenzierte[3] Plattenepithel- und Adenokarzinome	5	0
Normales Lungengewebe				5	17

Als nächster Schritt folgt die für die Vergleichbarkeit der Experimente obligatorische Vorverarbeitung. Das Filtern und die Normierung der **cDNA-Daten** erfolgten entlang der in der Publikation vorgeschlagenen Methoden. Leider gibt es bei diesen Daten keine spezifischen Negativkontrollen, wie das Missmatch-Oligo bei den Affymetrix-Chips. Verlangt man, dass für einen Spot die gemessenen Signale in wenigstens 67 der 72 Experimente das Anderthalbfache des Hintergrundes erreichen (Filterregel aus der Publikation von Garber et al), so bleiben für die weiteren Analysen noch 9097 der anfänglich

[1] Quelle: [23] Seeber, Siegfried und Schütte, Jochen (1998): Therapiekonzepte Onkologie, Springer-Verlag.

[2] Entnommen aus [21] Riede, Ursus-Nikolaus und Schaefer, Hans-Eckart (1999): Allgemeine und spezielle Pathologie, Thieme Verlag, Stuttgart, New York.

[3] In einem intakten Gewebe sind die Zellen ausdifferenziert, so dass sie optimal Funktion erfüllen können. Im Zuge der Karzinogenese bildet sich die Differenzierung mehr und mehr zurück (Entdifferenzierung).

circa 24000 Messpunkte. Für die cDNA-Daten wurden zwei Berechnungen durchgeführt, eine mit den gemessenen Absolutsignalen und eine mit den relativen Änderungen bezüglich des Referenz-RNA (Siehe hierzu auch Abschnitt 1.6 und Kapitel 4). Die mittlere absolute Expression eines Gens lässt sich als Mittelwert der hintergrundkorrigierten und logarithmierten Werte über alle Experimente schätzen (*LogSignal*), wobei die Signale der Referenz-RNA unberücksichtigt bleiben. Für die Untersuchung von relativen Expressionsänderungen hingegen dienten die logarithmierten Quotienten aus Cy5 und Cy3 Signal (*LogRatio = log(Cy5/Cy3)*) als Ausgangspunkt. In den standardisierten Rohdaten-tabellen, wie sie auf der Internetseite http://www.genome.stanford.edu bereitgestellt sind, wird das hintergrundkorrigierte Signal mit *CH2D_MEAN* und der logarithmierte Quotient mit *LOGRAT2N* bezeichnet. Die Cy3-Signale rühren in jedem Experiment von der RNA des gemeinsamen Referenzpools her. Durch die Quotientenbildung verliert man Information über das absolute Expressionssignal. Korreliert beispielsweise die Absolutmenge der RNA der Zellen gut mit dem gemessenen Signal, so wird diese Abhängigkeit durch die Quotientenbildung zerstört. Dieses interne Normierungsverfahren beruht auf den Annahmen, dass die RNA des Referenzpools für jedes Experiment identisch ist und dass der Hauptfehler der Signalwerte von der Inhomogenität der Spots herrührt. Zur Normierung wurden die Ausgangsdaten (*LogRatios*, wie auch *LogSignals*) pro Experiment um die Null zentriert und auf Standardabweichung eins skaliert. Die Verteilungen können damit recht gut angeglichen werden (Siehe Anhang B: Darstellung der Verteilungen als Box-Plots). Allerdings bleibt die Korrelation der Expressionswerte zwischen den Experimenten mit Korrelationskoeffizienten im Bereich von etwa 0,5 bis 0,8 unter den Erwartungen. Die Datensätze von drei Experimenten fallen völlig heraus, was auf Fehler bei der Durchführung der Laborexperimente oder bei der Datenverarbeitung hindeutet. Zur Validierung der Normierung wurden die wenigen Proben im Datensatz von Garber et al verwendet, von denen mehrere vom gleichen Patienten stammten. Da es keine echten Wiederholungsexperimente gibt, lässt sich nur prüfen, ob die transformierten Daten völlig den Erwartungen zuwider laufen. Detailliert wurden dazu die zwei Plattenepithelkarzinomproben des Patienten 246 und die zwei Adenokarzinomproben des Patienten 320 aus der Publikation untersucht. Die Änderung der Expression eines Gens zwischen zwei Proben *A* und *B* wird definiert als die Differenz der *LogRatios* dieses Gens in Bezug auf die Referenz in den Proben *A* und *B*.

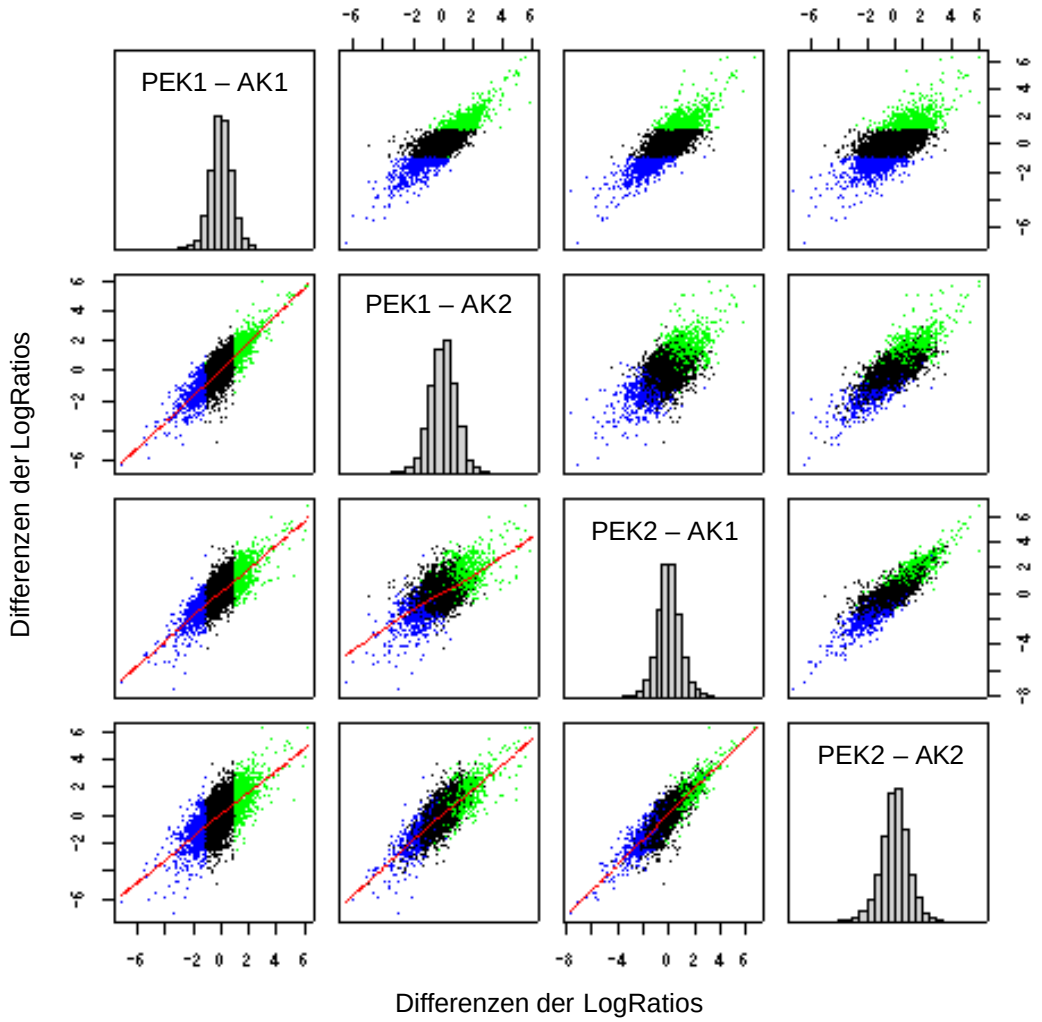


Abbildung 3-2 Korrelation der Expressionsänderung (LogRatio)

Für jeweils zwei replizierte Plattenepithelkarzinomproben (PEK1, PEK2) und Adenokarzinomproben (AK1, AK2) auf den cDNA-Arrays wurden als erstes die Differenzen der normierten LogRatio-Werte in den Kombinationen (PEK1,AK1), (PEK1,AK2), (PEK2,AK1) und (PEK2,AK2) gebildet. Die Verteilungen der Differenzen sind als Histogramme in den diagonalen Feldern abgebildet. In den Feldern außerhalb der Diagonalen sind jeweils die Differenzen der Einzelvergleiche gegeneinander aufgetragen. In dem Feld links unten (und rechts oben) beispielsweise sind die Werte von (PEK1 – AK1) und von (PEK2 – AK2) dargestellt. Die Regressionslinie ist rot markiert. Die Scatterplots der rechten oberen Dreiecksmatrix entsprechen gespiegelt gerade denen der unteren. Die im ersten Vergleich (PEK1 – AK1) höchsten 10% der Werte (PEK1 > AK1) sind hellgrün markiert und die niedrigsten 10% (PEK1 < AK1) blau.

Es ist augenfällig, dass sich die Änderungen der Expressionswerte in den vier möglichen Kombinationen zumindest qualitativ reproduzieren lassen. Die Korrelation der absoluten Expressionswerte ist der der Quotienten vergleichbar. Die folgende Abbildung 3-3 ist das Pendant für die logarithmierten und normierten Cy5-Signale (*LogSignal*) zu Abbildung 3-2

für die *LogRatios*. Liefert das Referenzsignal wenig spotspezifische Information, so gilt annähernd folgende Äquivalenz bei der Berechnung der Expressionsänderung:

$$\frac{A}{B} \approx \frac{A/C_A}{B/C_B} = \frac{A}{B} \cdot \frac{C_B}{C_A} \text{ mit } \frac{C_B}{C_A} \approx 1$$

oder für die logarithmierten Werte $\log(A) - \log(B) \approx \log(A/C_A) - \log(B/C_B)$.

Dabei stellt *A* das Cy5-Signal für ein bestimmtes Gen auf einem Array dar. *B* bezeichnet das Cy5 Signal auf einem anderen Array und *C* das entsprechende Referenzsignal (Cy3). Die Division des Signalwertes durch die Referenz ist nur hilfreich, wenn sich damit systematische spotspezifische Fehler korrigieren lassen. Andernfalls fügt man nur die Streuung des Referenzsignals der Streuung des Cy5-Signals hinzu und verliert damit Genauigkeit.

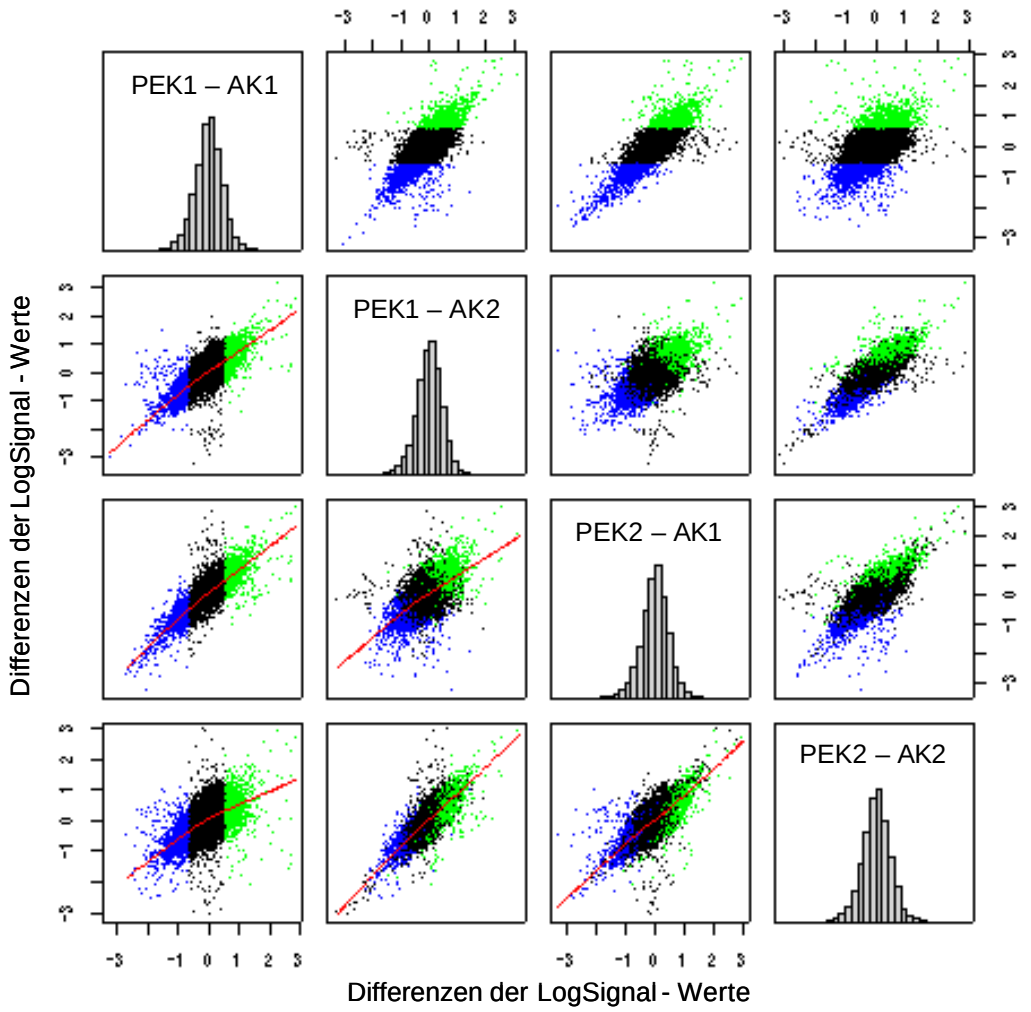


Abbildung 3-3 Korrelation der Expressionsänderungen (*LogSignal*)

Für jeweils zwei replizierte Plattenepithelkarzinomproben (PEK1, PEK2) und Adenokarzinomproben (AK1, AK2) auf den cDNA-Arrays wurden als erstes die Differenzen der normierten LogSignal-Werte gebildet. Die Verteilungen der Differenzen sind als Histogramme in den diagonalen Feldern dargestellt. In den Feldern außerhalb der Diagonalen sind jeweils die Differenzen der Einzelvergleiche gegeneinander aufgetragen. Eine stückweise Regression (Loess) ist durch die rote Linie angedeutet. Die Scatterplots der rechten oberen Dreiecksmatrix entsprechen gespiegelt gerade denen der unteren. Die im ersten Vergleich (PEK1 – AK1) höchsten 10% der Werte (PEK1 > AK1) sind hellgrün markiert und die niedrigsten 10% (PEK1 < AK1) blau.

Aufgrund des Mangels an echten Wiederholungsversuchen lässt sich nicht bewerten, welchen qualitativen Gewinn die konkurrierende Hybridisierung der Proben gegen den Referenz-Pool bringt. Der erheblich höhere experimentelle Aufwand für das Zweifarbsystem und die vorliegenden Daten lassen es lohnend erscheinen, diese Frage genauer zu untersuchen. Für die vergleichende Analyse des folgenden Abschnitts ist nur der Fakt von Bedeutung, dass die Absolutsignale (*LogSignal*) reproduzierbar sind. Das bedeutet, dass sich aus den cDNA-Daten Informationen über die absolute mRNA-Konzentration gewinnen lassen (dazu Kapitel 5).

Die **Oligo-Array-Daten** sind von den Autoren (Bhattacharjee et al) als Rohdaten bereitgestellt worden und konnten deshalb mit dem in Kapitel 2 beschriebenen Verfahren ausgewertet werden. Das Ergebnis besteht aus einem repräsentativen Expressionswert (*PM-Quartil*) und einem P-Wert (Detektionsscore). Für jedes Sondenset mit P-Wert kleiner 0,05 nimmt man an, dass es ein Transkript detektiert hat. Zur Normalisierung wurden die logarithmierten Expressionswerte pro Experiment um den Mittelwert zentriert und auf Standardabweichung eins skaliert. Der Vergleich einer Stichprobe von Datensätzen ergab, dass etliche zwar gut jedoch nicht linear korreliert waren. Für eine nichtlineare Anpassung diente die Funktion *loess* aus dem Statistikpaket R. Die genaue Beschreibung der Methoden und deren Validierung sind Teil von Kapitel 2. Die Verteilungen der normierten Werte in Box-Plot-Darstellung sind in Anhang B beigefügt. Zur Veranschaulichung sind die Expressionswerte von zwei Normalgewebe- und zwei Adenokarzinomproben aufgearbeitet worden (Abbildung 3-4). Der Eindruck der Abbildungen 3-2, 3-3, 3-4, dass sich die Oligo-Array-Daten besser reproduzieren lassen als die cDNA-Array-Daten wird auch durch die Korrelationskoeffizienten bestätigt: Die Korrelationskoeffizienten der normierten Expressionswerte liegen im Mittel bei 0,934 (1.- 3. Quartil: 0,924- 0,946) für die Oligo-Chip-Daten und bei 0,47 (1.- 3. Quartil: 0,35- 0,59) für die *LogRatios* und bei 0,62 (1.- 3. Quartil: 0,54- 0,71) für die *LogSignal*-Werte der cDNA-Arrays.

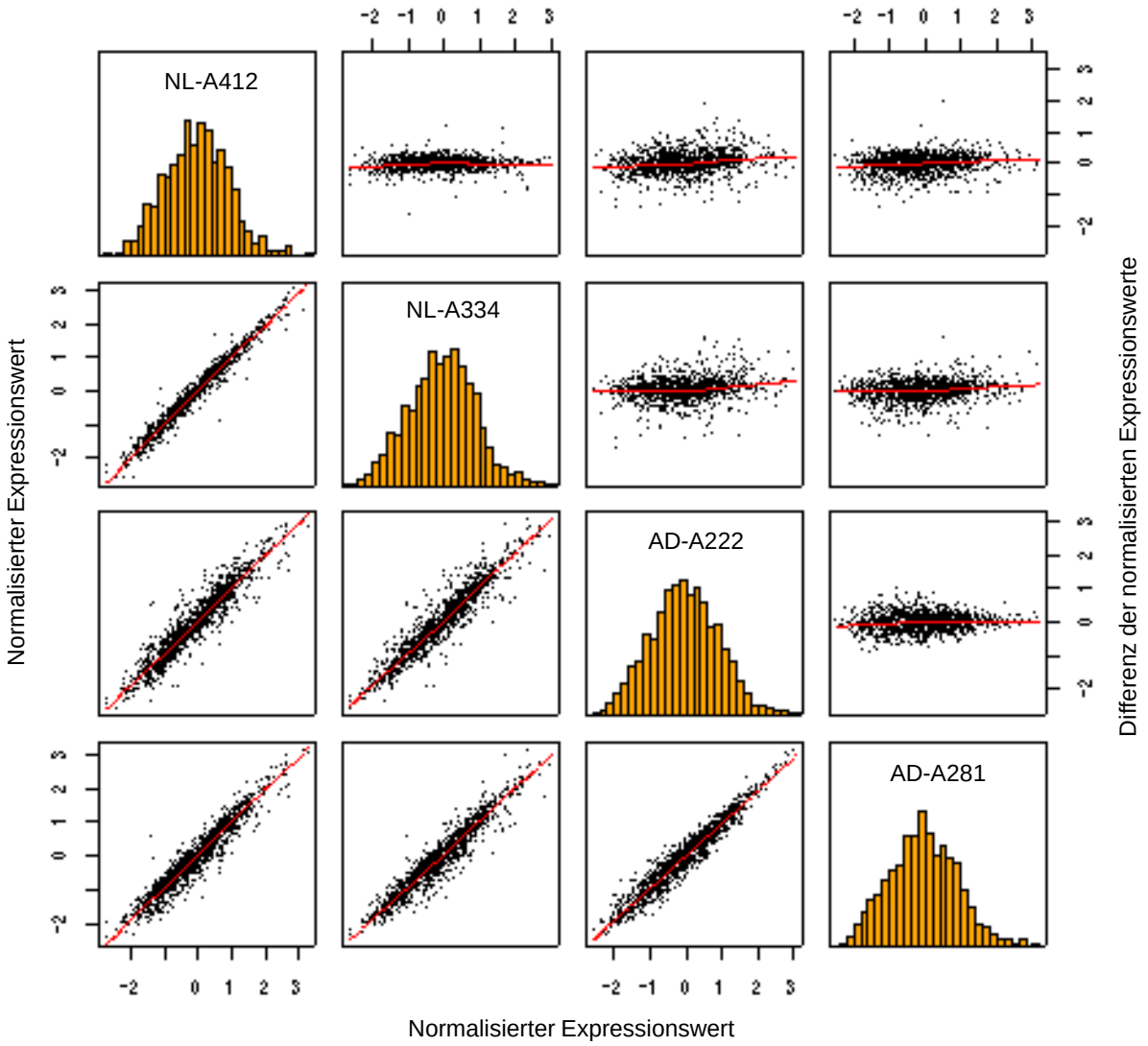


Abbildung 3-4 Die Korrelation und Streuung der Werte von vier Chipexperimenten
Dargestellt sind die Verteilungen und die Korrelation der logarithmierten und normierten Expressionswerte (PM-Quartile) von zwei Normalgeweben (NL-A412, NL-A334) und zwei Adenokarzinomproben (AD-A222, AD-A281). Die Bezeichner wurden aus der Publikation von Bhattacharjee et al übernommen. In der ***Diagonalen*** *sind die Verteilungen der Expressionswerte als Histogramme dargestellt. Die Felder* ***links*** *unterhalb der Diagonalen veranschaulichen die Korrelation im Scatter-Plot.* ***Rechts*** *oben sind für jeden Vergleich pro Gen die Differenzen der Expressionswerte gebildet worden und in Abhängigkeit vom Expressionswert dargestellt.*

Die Expressionsanalyse setzt man in der Krebsforschung meistens dazu ein, zwischen normalen Zellen und Tumorzellen differenziell exprimierte Gene zu identifizieren. Man interessiert sich also für Expressionsunterschiede. Bei den gut korrelierten Daten der Oligo-Arrays besteht nun die Gefahr, dass die Signalstärke von der Qualität der Oligosonden dominiert wird. Das hieße für eine sehr gute Sonde, sind die passenden Transkripte im Hybridisierungscocktail vorhanden, so detektiert man ein konstant hohes Signal auf dem

Chip unabhängig von der tatsächlichen Konzentration. Die unterschiedlichen Signale, der Sonden auf dem Chip gäben dann lediglich darüber Auskunft, ob die Sonde ein hohes Bindungspotenzial hat. Um das zu prüfen und die Ergebnisse der unterschiedlichen Technologien vergleichen zu können, müssen daher die Expressionsänderungen zwischen Proben analysiert werden. Arbeitet man beispielsweise zweimal Normalgewebe und zweimal Tumorgewebe aus jeweils den selben Patienten auf, dann erwartet man, dass die tumorrelevanten differenziellen Gene konsistent als hoch- beziehungsweise runterreguliert gefunden werden unabhängig davon, welche der Normal- und Tumorproben man vergleicht.

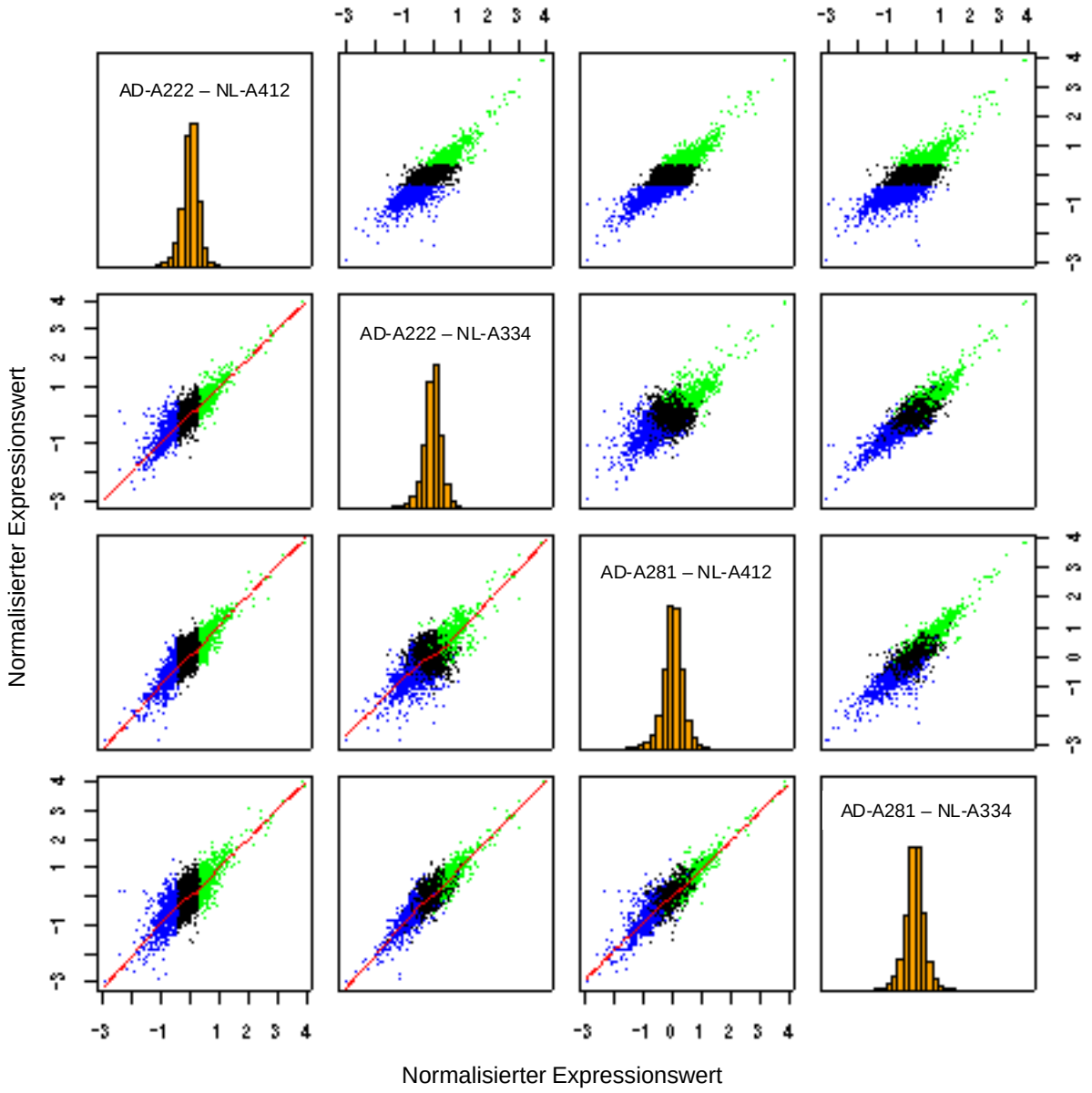


Abbildung 3-5 Korrelation der Expressionsänderungen von Oligo-Array-Daten

Für jeweils zwei Normalgewebeproben (NL-A412, NL-A334) und zwei Adenokarzinomproben (AD-A222, AD-A281) vom gleichen Patienten wurden als erstes die Differenzen der normierten Expressionswerte (PM-Quartile) gebildet. Die Verteilungen der Differenzen sind als Histogramme in den diagonalen Feldern dargestellt. In den Feldern außerhalb der Diagonalen sind jeweils die Differenzen der Einzelvergleiche gegeneinander aufgetragen. Eine stückweise Regression (Loess) ist durch die rote Linie angedeutet. Die Scatterplots der rechten oberen Dreiecksmatrix entsprechen gespiegelt gerade denen der unteren. Die im ersten Vergleich (AD-A222 – NL-A412) höchsten 10% der Werte (AD-A222 > NL-A412) sind hellgrün markiert und die niedrigsten 10% (AD-A222 < NL-A412) blau.

Hier zeigt sich, dass bei den Oligo-Array-Daten durchaus vergleichbare Streuungen auftreten wie bei den cDNA-Daten (Siehe Abbildung 3-2).

Die letzte vorbereitende Maßnahme bestand aus dem Abgleich der Array-Sequenzsets. Für den Affymetrix-Chip HG-U95Av2 dienten die so genannten Targetsequenzen als Ausgangsbasis[1]. Das ist eine Datei von Sequenzen, die aus dem am 3'-Ende gelegenen Bereich der zu untersuchenden mRNA's ausgewählt und auf bakterielle Verunreinigungen und repetetive Sequenzen getestet wurden. Aus den Targetsequenzen sind von Affymetrix beim Chipdesign die Oligos ausgewählt worden. Für das cDNA-Array sind in den Dateien für jeden Spot Sequenzbezeichner der auf die Glasträger aufgebrachten EST-Klone angegeben. Die entsprechenden Sequenzen sind in öffentlichen Datenbanken verfügbar und bildeten damit für das cDNA-Array das Basisset. Wie bereits in Abschnitt 1.4 näher erläutert, dient zurzeit Unigene Built 148 als aktueller Sequenzpool über den andere Sequenzsets verknüpft werden. Als ausreichend strenges und zuverlässiges Verfahren für den Sequenzvergleich hat sich Programm *BLAST* mit folgenden Parametern erwiesen: Schwellwert für den E-Wert (E-Value) 10^{-50}, mindestens ein *HSP*[2] der Länge 120 bp und 97% Identität, bester Treffer. Ein Sequenzabgleich dieser Art lässt sich immer kritisieren, und es ist sehr wahrscheinlich, dass die relativ strenge Wahl der Parameter dazu führt, dass etliche Target- beziehungsweise EST-Sequenzen die das gleiche Transkript auf den Arrays detektieren, nicht als äquivalent erkannt werden. Eine Abschwächung der Stringenz führte jedoch zu vielen falsch-positiven Treffern, und das Ziel der vorliegenden Analyse ist ja nicht die Suche nach neuen, noch wenig charakterisierten Transkripten. Das beschriebene Vorgehen führte zu einem Set von 3644 Unigene-Clusterrepräsentanten, die auf dem Affymetrix-Chip U95Av2 durch 4257 Sondensets und auf dem cDNA-Array durch 4121 Spots repräsentiert sind. Für die Annotation der Sequenzen ließen sich über die Verknüpfung von UniGene andere Datenquellen wie zum Beispiel die RefSeq-Datenbank am NCBI nutzen. Welche Daten für die Sequenzen genau verfügbar sind und wie diese organisiert sind, ist in Kapitel 2 beschrieben.

---

[1] Für eine genaue Beschreibung des Chips und der Sequenzsets siehe Abschnitt 1.4 und die Affymetrix-Internetseite http://www.affymetrix.com

[2] HSP (*High Scoring Pair*): Lokaler Treffer maximaler Ausdehnung beim Vergleich zweier Sequenzen. Vergleicht man beispielsweise eine mRNA-Sequenz mit der entsprechenden genomischen Sequenz, so sollte jedes Exon ein HSP ergeben.

3.7. Vergleichende Analyse - Resultate

Die den Publikationen zugrunde liegenden Oligo- beziehungsweise cDNA-Array-Typen gehören zu den am häufigsten für die Genexpressionsanalyse verwendeten Technologien. Ein Vergleich der mit den unterschiedlichen Technologien gewonnenen Ergebnisse lässt sich auf Basis der absoluten oder der relativen Expressionswerte durchführen. Jedes Oligonukleotide-Array hybridisierte man mit einer markierten RNA-Probe. Und damit misst man pro Chipexperiment die absoluten mRNA-Konzentrationen in dieser Probe. Relative Änderungen der Genexpression erhält man durch die Verrechnung mehrerer Experimente. Jedes cDNA-Array hingegen wurde mit einer Cy5-markierten Probe und einer Cy3-markierten Referenz parallel hybridisiert. Die beiden Signale können auf jedem Spot mit unterschiedlichen Wellenlängen vom Laser detektiert und quantifiziert werden. Aus den Intensitätswerten der einzelnen Kanäle (*LogSignal*) lässt sich potenziell auf die absolute mRNA-Konzentration schließen. Verrechnet man die Signale der Probe und der Referenz, so erhält man relative Expressionswerte (*LogRatio*), die sich für die Bestimmung von Expressionsänderungen eignen aber nicht für Aussagen über die mRNA-Konzentrationen. Als erstes sollte geklärt werden, inwieweit die LogSignal-Werte der cDNA-Arrays und repräsentativen Expressionswerte der Oligo-Arrays vergleichbare Ergebnisse liefern. Eine gute Übereinstimmung wäre ein Hinweis auf eine Proportionalität zwischen Molekülkonzentration und Expressionswert. Um das zu prüfen, lassen sich die Mittel der Expressionswerte innerhalb der beiden Datensets für jedes Gen, das auf beiden Arrays repräsentiert ist, bilden. Trägt man diese Werte in einer Graphik gegeneinander auf, kann man keinen systematischen Zusammenhang erkennen (Abbildung 3-6 links). Für eine detailliertere Analyse lässt sich eine Kontingenztafel berechnen, indem erst die beiden Wertebereiche jeweils in Dezile[1] zerlegt werden. Es entstehen zehn mal zehn mögliche Dezilkombinationen (Felder), wobei auf jedes der Gene in eine Kombination zutrifft. Liegt beispielsweise ein Gen im dritten Dezil der

[1] Zur Berechnung der Dezile einer Menge von Zahlen (alle Expressionswerte eines Datensatzes) bringt man diese in aufsteigende Reihenfolge und bestimmt anschließend zehn Zahlen, so dass 10% aller Werte kleiner als die erste Zahl sind und 90% aller Werte größer als die erste Zahl sind. Die zweite Zahl ist so gewählt, das 20% kleiner und 80% größer sind und so weiter. Dadurch unterteilt man den Wertebereich in Intervalle unterschiedlicher Größe, wobei jeder gleich viele Werte enthält.

cDNA- Array-Daten und im fünften Dezil der Oligo-Array-Daten, so bedeutet dass: Erstens 20% aller LogSignal-Expressionswerte des cDNA-Array-Datensatzes sind kleiner als der Wert dieses Gens und 70% sind größer. Und zweitens 40% aller Expressionswerte des Oligo-Array-Datensatzes sind kleiner als der entsprechende Wert dieses Gens und 50% sind größer.

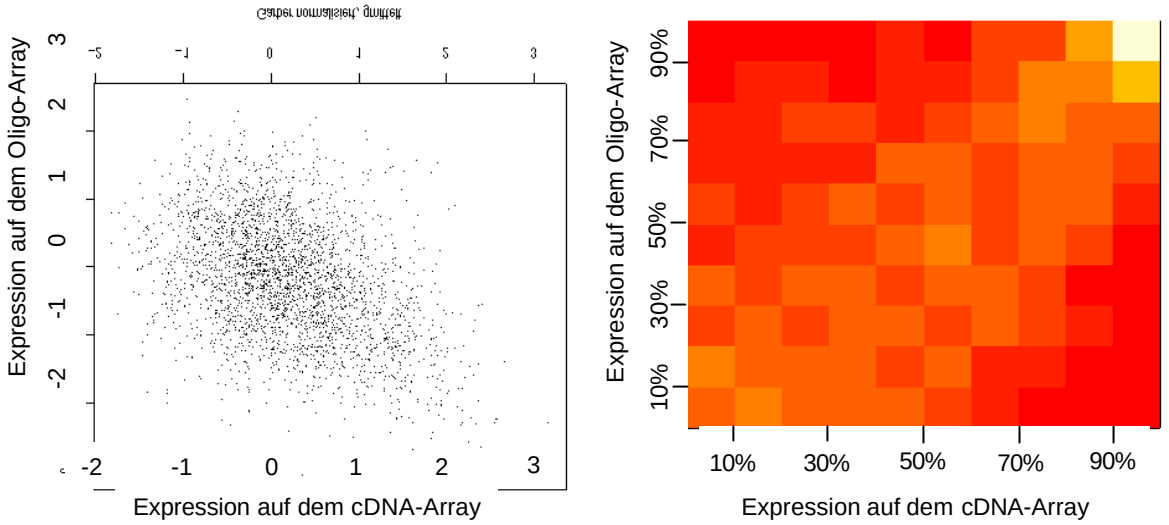


Abbildung 3-6 Korrelation mit unterschiedlichen Technologien gemessener Expressionssignale

Links *sind die mittleren Expressionssignale der beiden Datensätze für die Gene, die auf beiden Arrays repräsentiert sind (3644), gegeneinander aufgetragen. Die Expressionssignale wurden dafür wie im vorigen Abschnitt beschrieben logarithmiert und normiert.* ***Rechts*** *ist die Kontingenztafel farbkodiert dargestellt. Dabei gilt: je heller desto höher die Inzidenz (Anzahl der Elemente).*

Die gemessenen und normalisierten Expressionswerte der beiden Sets scheinen nicht direkt quantitativ vergleichbar. Bei der Analyse der Kontingenztafeln zeigen die jeweils 10% höchsten Signale die beste Übereinstimmung. Für die anderen Bereiche sieht man einen breiten Rücken entlang der Diagonalen, der von einer schwachen Korrelation herrührt. Dem Anschein nach lassen sich aus den cDNA-Arraydaten auch Aussagen über die absolute Genexpression ziehen. Die große Streuung der Werte kann verschiedene Ursachen haben. Zum einen liegen den Versuchsreihen unterschiedliche Gewebeproben zu Grunde. An Gewebeschnitten (z. B. Anhang B) lässt sich erahnen, wie komplex und vielseitig Zellverbände und deren molekularbiologischen Charakteristika sind (Arbeitshypothese I). Zum anderen unterscheiden sich die Methoden von der Gewebeasservierung über die RNA-Gewinnung bis zur Hybridisierung erheblich (Arbeitshypothese II).

Als nächstes sollte geprüft werden, ob die als differenziell exprimiert gefundenen Gene in beiden Datensätzen die gleichen sind. Wie in Tabelle 3-2 aufgelistet, stehen in Garbers Set 41

Adenokarzinom-, 16 Plattenepithelkarzinom- und 5 Normalproben zur Verfügung und in Bhattacharjees Set sind es der Reihenfolge entsprechend 139, 21 und 17. Diese drei Gewbetypen eigneten sich auf Grund der höheren Fallzahlen für eine tieferführende Untersuchung. Während der Analyse fiel auf, dass die Genexpression zwischen den Tumor- und Normalproben innerhalb eines Datensatzes (gleiche Technologie) stärker korreliert als zwischen Tumorproben auf unterschiedlichen Array-Typen (Siehe Abbildung 3-7). Zum Beispiel zeigen die mittleren Expressionswerte der mit den Oligo-Arrays untersuchten Plattenepithelkarzinomproben eine bessere Inzidenz mit den auf Oligo-Arrays untersuchten Normalproben als mit den Plattenepithelkarzinomproben des cDNA-Array-Datensatzes.

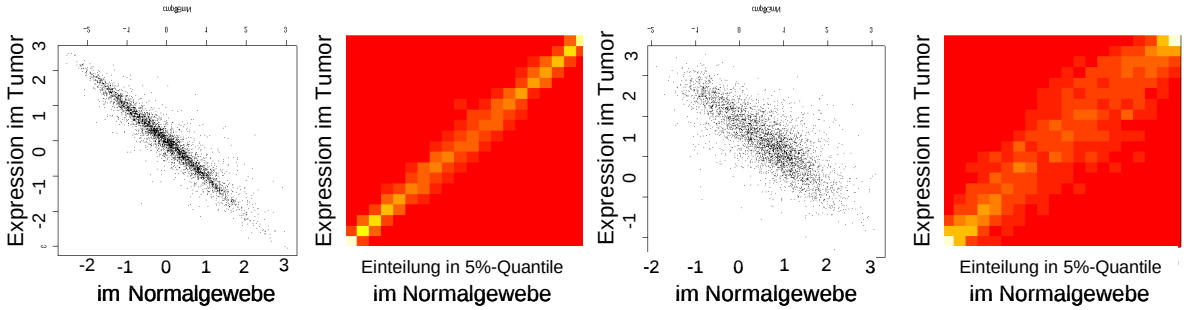


Abbildung 3-7 Korrelation der mittleren Expression im Tumor versus Normalgewebe
Dargestellt ist die mittlere Expression der Gene in Plattenepithelkarzinomproben gegen die mittlere Expression in den Normalproben, jeweils als Scatterplot und als farbkodierte Kontingenztafel. Für die Berechnung der Kontingenztafeln wurden die Wertebereiche in alle 5%-Quantile (5%-Perzentile) unterteilt. Den beiden Abbildungen ***links*** *liegt der Oligo-Array-Datensatz zugrunde und den beiden Abbildungen* ***rechts*** *der cDNA-Array-Datensatz.*

Die Scatterplots für den Vergleich der mit den unterschiedlichen Array-Technologien erzeugten Werte, eingeschränkt auf die einzelnen Gewebetypen, ähneln dem in Abbildung 3-6 links und sind deshalb nicht aufgeführt. Die Daten favorisieren Arbeitshypothese II: Geht man von zwei Einflussfaktoren auf das Signal aus, einen für die Charakteristik des Gewebetyps und einen für die Technologie, so überwiegt der Einfluss der Technologie. Genauer lässt sich das mit Hilfe der Varianzanalyse untersuchen: Man definiert zwei unabhängige Faktoren oder Merkmale, den Gewebetyp mit den Ausprägungen Adenokarzinom, Plattenepithelkarzinom und Normalgewebe und die Technologie mit den Ausprägungen cDNA-Array und Oligo-

Array. Um eine ANOVA[1] (*Analysis of Variance*) zu rechnen wurden die Datensätze vorher eingeschränkt, da man sonst rechen– wie auswertetechnisch an Grenzen stößt, und der Informationszuwachs minimal wäre. Anders ausgedrückt: Man schränkt den Datensatz so ein, dass, falls die Technologie einen signifikanten Einfluss auf die gemessenen Expressionsunterschiede in dem reduzierten Set hat, so gilt dies auch für den kompletten Datensatz. Die Proben wurden für diese Analyse auf die Plattenepithelkarzinome und die Normalgewebe reduziert. Die Menge der zu untersuchenden Gene lässt sich auf 799 einschränken, indem man nur noch solche weiter betrachtete, die in einem der beiden Datensätze zu den 10% im Normalgewebe im Mittel höchsten Signalen gehören. Das Ergebnis der Analyse zeigt einen hochsignifikanten[2] Einfluss der Technologie. Es ist allerdings bei der Interpretation Vorsicht geboten, da trotz gravierender Transformationen, die Daten von der Normalverteilung abweichen.

Die Schwierigkeit liegt hier darin, dass sich die mit unterschiedlichen Technologien erhobenen Expressionsdaten nicht direkt quantitativ vergleichen lassen. Zeigt beispielsweise eine bestimmtes Oligoset für ein Gen immer sehr hohe Signale, obwohl das Gen nur relativ moderat exprimiert ist, so kann das daran liegen, dass die gewählten Oligos einen hohen GC-Gehalt haben und damit im Schnitt mehr RNA-Moleküle binden. Die Signale bleiben in diesem Fall zwischen Arrays der gleichen Technologie immer noch gut reproduzierbar, aber Vergleiche über Technologiegrenzen hinweg können zu Fehlinterpretationen führen. Oft interessiert man sich besonders für Expressionsänderungen oder differenzielle Expression zwischen Zellen in unterschiedlichen physiologischen Zuständen, in diesem Fall zwischen Zellen des normalen Lungengewebes und Tumorzellen. Für die Bewertung der publizierten Ergebnisse stellt sich unter anderem die Frage, ob man ein Gen, dass auf den cDNA-Arrays als stark oder konsistent hochreguliert erscheint, auf den Oligo-Arrays dem entsprechend wieder findet. Geht man nun von Arbeitshypothese I aus, so sollten die relativen

---

[1] Siehe unter anderem [8] Sachs, Lothar (1997): Angewandte Statistik, Springer, Berlin Heidelberg..

[2] Als Modell diente *e ~ tech + type*, wobei *e* für den repräsentativen Expressionswert, *tech* für die Technologie und *type* für den Gewebetyp steht. Die Berechnung wurde mit der R-Funktion *aov* durchgeführt und ergab für den Einfluss der Technologie den P-Wert: $2{,}2*10^{-16}$.

Expressionsänderungen jedes Transkripts unabhängig von der Technologie einer Zufallsstichprobe aus einer einzigen Verteilung ähneln. Gilt jedoch für viele untersuchte Gene, dass die Streuung der Expressionsänderungen innerhalb einer Technologie wesentlich kleiner ist als die Gesamtvarianz, so würde man Arbeitshypothese II bevorzugen und folgern, dass der durch die Technologie eingebrachte Fehler die quantitativ gemessenen Expressionsunterschiede zwischen den Gewebetypen überwiegt. Da keine gepaarten Tumor-, Normalproben vorlagen, wurde pro Sonde innerhalb der Datensätze jede Differenz[1] zwischen einem Tumorwert und einem Normalwert gebildet. Zum Beispiel ergeben sich damit bei dem Vergleich von Adenokarzinomproben (41) und Normalgewebeproben (5) für die cDNA-Daten 205 (41*5) Differenzen pro Spot und für die Oligo-Chip-Daten entsprechend 2363 (139*17) Differenzen pro Sondenset. Diese Werte wurden pro Datensatz auf Standardabweichung eins skaliert, um eine bessere Vergleichbarkeit zu erreichen. Mit Hilfe der Varianzanalyse lässt sich auch für die Expressionsänderungen ein signifikanter Einfluss der Technologie feststellen. (Siehe oben.) Etwas plastischer wird das Ergebnis, betrachtet man die Verteilungen detailliert pro Gen. Für jedes Gen lässt sich der Interquartilsbereich der Expressionsunterschiede bestimmen, also das Intervall zwischen erstem und drittem Quartil aller Differenzen. Es stellte sich heraus, dass für 134 der 799 Gene (16,8%) die Intervalle nicht überlappen und damit die Verteilungen klar separierbar sind. Für einzelne Gene führt das letztlich zu unterschiedlichen Aussagen abhängig davon, welchen Datensatz man analysiert hat. Zum Beispiel könnte man auf Basis der Oligo-Array-Daten annehmen, dass das Gen *SPARCL1* fast immer im Tumor geringer exprimiert ist als im Normalgewebe (Abbildung 3-8). Im cDNA-Datensatz hingegen zeigt es sich als praktisch unverändert und nur in seltenen Fällen als im Tumor runterreguliert.

[1] Da die Signalintensitäten logarithmiert und zentriert sind, stellen die Differenzen ein vernünftiges Maß für Expressionsunterschiede dar. Der Erwartungswert der Differenzen ist Null. Siehe auch Kapitel 2.

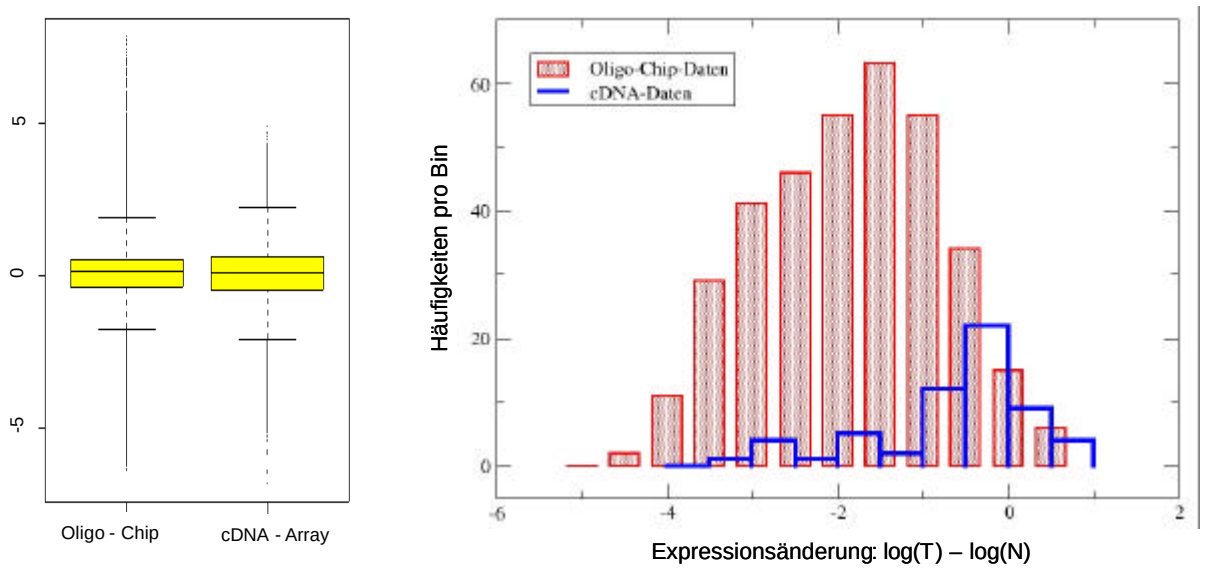


Abbildung 3-8 Analyse der Verteilungen der Expressionsänderungen
Links *sind die Boxplots der Verteilungen aller mittleren Expressionsunterschiede zwischen den Plattenepithelkarzinom- (PEK) und den Normalgewebeproben (NL) der 799 ausgewählten Gene pro Datensatz nach der Zentrierung und der Skalierung mit dem Reziproken der Standardabweichung dargestellt. Die* ***rechte*** *Abbildung zeigt die Histogramme der Werte für ein Beispielgen, und zwar SPARCL1 (RefSeq: NM_004684). Die Binweite beträgt 0,5. Die Verteilungen enthalten jeweils unterschiedlich viele Messwerte, da vom Oligo-Chip 357 (21 PEK mal 17 NL) und vom cDNA-Array 80 (16 PEK mal 5 NL) paarweise Vergleiche berechnet werden konnten.*

Die beste Übereinstimmung der mittleren Expressionsänderungen erreicht man in den jeweils extremsten 5-10% der Daten. Für die Auswahl von differenziell exprimierten Gene wurden innerhalb jeder Technologie folgende zwei Regeln angewendet (Siehe auch Kapitel 2):

1. Der P-Wert des T-Tests auf differenzielle Expression ist kleiner 0,05. Die Nullhypothese lautet dabei: Die Expressionsunterschiede zwischen Normalgewebe und der Tumorentität sind symmetrisch um die Null verteilt.

2. Der errechnete mittlere Expressionsunterschied muss zu den 10% größten oder zu den 10% kleinsten Werten gehören.

Im Folgenden werden ausschließlich Gene als differenziell bezeichnet, die diese beiden Regeln erfüllen.

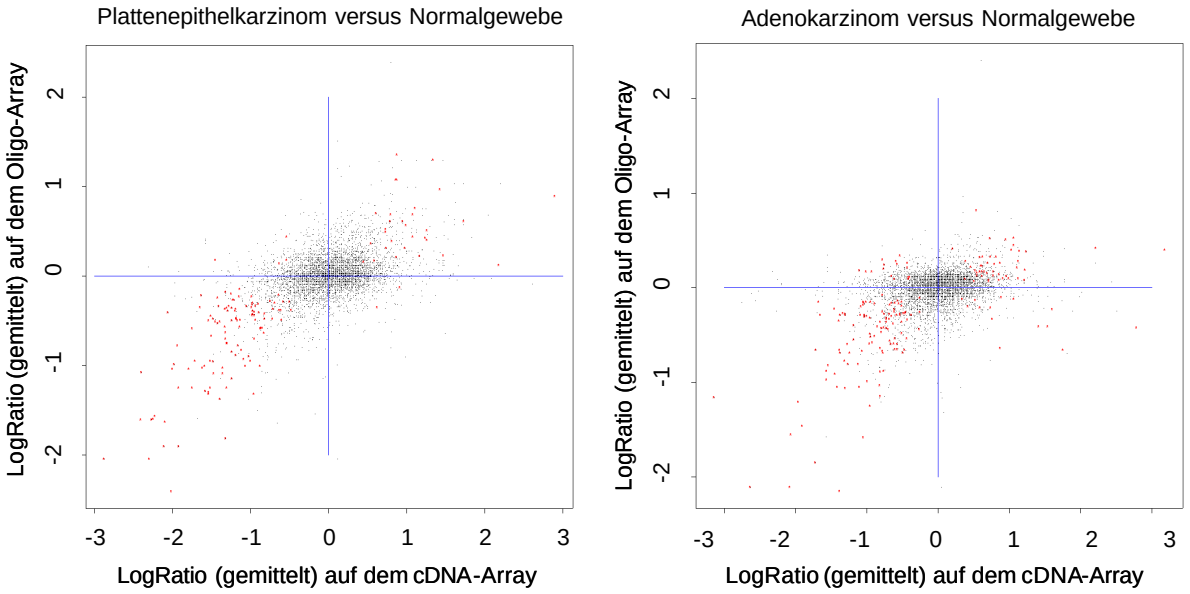


Abbildung 3-9 Korrelation der Expressionsunterschiede

Errechnet wurden die mittleren Expressionsunterschiede (LogRatios) zwischen Tumor- und Normalproben innerhalb der beiden Technologien. Für alle auf beiden Arrays repräsentierten Gene sind diese Werte gegeneinander aufgetragen. ***Links*** *ist der Vergleich für das Plattenepithelkarzinom und* ***rechts*** *für das Adenokarzinom dargestellt.* ***Rot*** *markiert sind alle Gene, in einem der Sets zu den extremsten 10% der Werte zählen und die der t-Test als differenziell detektiert. Die* ***blauen*** *Linien kennzeichnen die Nulllinien und sollen dem Betrachter eine Hilfe sein.*

Es gilt, dass sich die errechnete differenzielle Expression für die meisten nach obigen Regeln ausgewählten Gene mit beiden Array-Technologien bestätigen lässt. Trotz der höheren Patientenzahlen zeigen die Adenokarzinome eine schlechtere Konkordanz als die Plattenepithelkarzinome (Abbildung 3-9). Das stimmt mit der Beobachtung der beiden Autorenkollektive der Originalpublikationen überein, dass die Adenokarzinome in sich inhomogener sind. Aufgrund des großen Einflusses der Technologie empfiehlt es sich, weitere Analysen vorrangig auf der Schnittmenge der in beiden Datensets als differenziell exprimiert identifizierten Gene durchzuführen.

Tabelle 3-3 Zahlen der als differenziell exprimiert gefundenen Gene

Richtung der Expressionsänderung		**Hoch im Tumor: Gene (Sonden/Spots)**	**Runter im Tumor: Gene (Sonden/Spots)**
Oligo-Array	Plattenepithelk.	401 (422)	387 (410)
	Adenokarzinom	417 (431)	404 (422)
cDNA-Array	Plattenepithelk.	287 (303)	355 (378)
	Adenokarzinom	352 (366)	385 (404)
Plattenepithelk auf Oligo- **UND** cDNA-Array		99	162
Adenok auf Oligo- **UND** cDNA-Array		81	136
Plattenepithelk auf Oligo- **ODER** cDNA-Array		490	418
Adenok auf Oligo- **ODER** cDNA-Array		607	517

Entschieden, ob ein Gen als differenziell exprimiert gilt, wurde nach den weiter oben beschriebenen Regeln. „Plattenepithelk." Steht als Abkürzung für Plattenepithelkarzinom und „Adenok." Adenokarzinom.

Wie stimmen nun die Chipdaten mit bereits bekanntem Wissen überein? Um ein möglichst objektives Bild von der Abdeckung und der Konkordanz der Chipdaten mit der Literatur zu erhalten, beziehen sich die folgenden Betrachtungen auf alle Gene, die in einen aktuellen Übersichtsartikel [25] als in die Progression von Lungentumoren involviert beziehungsweise differenziell exprimiert beschrieben sind. Dabei sind Gene, über die in dem Artikel nur spekuliert wird und andere Gene die in der umfangreichen Literatur zum Bronchialkarzinom jedoch nicht in diesem Artikel beschrieben sind, nicht berücksichtigt. Damit kommt man auf 52 Gene, von denen sind 25 auf beiden Chips, jeweils 9 nur auf dem cDNA-Array beziehungsweise auf dem Oligo-Chip und die restlichen 9 überhaupt nicht repräsentiert.

Tabelle 3-4 Übersicht über einige in der Literatur als tumorrelevant beschriebener Gene[1]

Gen	Endogene Funktion	Mechanismus	Gen-expression	Protein-expression	PEK		Adenok.	
					cDNA-Array	Oligo-chip	cDNA-Array	Oligo-chip
c-erbB-1 (EGFR)	EGF-Rezeptor, Signal-transduktion	Genamplifikation, Punktmutation	hochreguliert, 13% (NSCLC)	korreliert mit schlechter Prognose	hoch	**hoch**	**hoch**	hoch
c-erbB-2 (Her2/neu)	Rezeptor-Tyrosinkinase	Genamplifikation	hochreguliert	korreliert mit schlechter Prognose	neutral	runter	hoch	hoch
MYCN, v-MYC, LMYC	Transkrip-tionsfaktor, Onkogen	Genamplifikation, deregulation, Promoter-translokation	hochreguliert	hochreguliert, häufiger in SCLC	hoch, hoch, n.r.	runter, hoch, n.r.	**hoch, hoch,** n.r.	hoch, neutral, n.r.
Cyclin D1	Zellzyklus-kontrolle	Genamplifikation, Promoter-translokation		hochreguliert	**runter**	runter	runter	**runter**
APC	Signal-transduktion, Tumor-supressor	fehlerhafte Methylierung, Punktmutation + LOH oder 2. Mutation	runterreguliert	runterreguliert	neutral	neutral	**runter**	neutral
factor VIII (F8)	Gerinnungs-faktor, Marker für Blut-kapillare	Angiogenese am Tumor		wider-sprüchlich	**runter**	neutral	**runter**	runter
RAR-beta	Kernrezeptor	fehlerhafte Methylierung	runterreguliert in 50% NSCLC	runterreguliert in 50% NSCLC, jedoch wider-sprüchlich	hoch	*?hoch*	**hoch**	neutral

PEK steht als Abkürzung für Plattenepithelkarzinom und „Adenok" für Adenokarzinom, n.r. bedeutet das Gen ist nicht repräsentiert. Für die Charakterisierung der Gene als hoch- oder runterreguliert wurden jeweils die 20% differenziellsten in Betracht gezogen. Der Eintrag erscheint fett gedruckt, falls das Gene zu den 10% differenziellsten zählt und rot, falls das Ergebnis der Literatur widerspricht. RAR-beta ist gesondert markiert, da hier selbst in der Literatur gegensätzliche Resultate publiziert sind.

In Tabelle 3-4 ist einer Auswahl der Gene gelistet. Die vollständige Tabelle findet man in Anhang B. Der EGF-Rezeptor ist ein bekanntes Onkogen, dessen erhöhte Expression zu vermehrter Zellproliferation führt. Her2/Neu ist eine strukturell verwandte Rezeptor-Tyrosinkinase, die Einfluss auf die Mitogenese, das Überleben, die Invasion und die Angiogenese hat [26]. Das Protein fungiert als Zielmolekül für die wohl zurzeit erfolgreichste

[1] Die RefSeq – Bezeichner der Gene dieser Tabelle: c-erbB-1 – NM_005228, c-erbB-2 – NM_004448, MYCN – NM_005378, v-MYC – NM_002467, LMYC – NM_005376, APC – NM_000038, RAR-beta – NM_000965

Antikörpertherapie gegen Brustkrebs. Das entsprechende Medikament heißt Herceptin®[1] oder Trastuzumab. Man hatte herausgefunden, dass eine Überexpression von Her2/Neu mit einer schlechten Prognose korreliert und dem oft eine Amplifikation des Genortes zugrunde liegt. In der Klinik untersucht man die Brustkrebspatientinnen auf solche Amplifikationen mit FISH[2] und auf starke Proteinexpression von Her2/Neu durch einen immunhistochemischen Test[3]. Bei positivem Befund behandelt man die Patientin mit Trastuzumab oft auch in Kombination mit Chemotherapeutika [27]. Einen ähnlichen Therapieansatz untersucht man zurzeit auch für nichtkleinzellige Bronchialkarzinome in klinischen Studien, und für eine kleine Gruppe von Patienten sind die Ergebnisse viel versprechend [26].

Die Analyse der Arraydaten ergibt eine Vielzahl weiterer, nicht in dem Artikel erwähnter Gene, die konsistent differenziell exprimiert sind. Ein interessantes Beispiel ist das Gen *stefin A*, dass einen Inhibitor der lysosomalen Cystein-Proteinasen Cathepsin B, H, L und S [28] kodiert. Die Expression des Gens ist konsistent in den beiden oben disskutierten Expressionsstudien im Plattenepithelkarzinom hochreguliert und in den Adenokarzinomproben runterreguliert. Es ist bereits beschrieben, dass das Gen im Plattenepithelkarzinom höher exprimiert ist als im Adenokarzinom der Lunge [29] und dass das Protein ebenfalls in Plattenepithelkarzinomen des Kopfes und des Halses erhöht exprimiert ist [30]. Die Autoren fanden Hinweise darauf, dass eher ein geringeres Vorhandensein des Proteins mit einer schlechten Prognose bezüglich Wiederkehr der Krankheit und Überlebenszeit einhergeht.

Bei der Analyse und Aufbereitung der Daten fiel auf, dass einige Gene gemeinsam dereguliert sind, die alle zu einem Komplex gehören, der an der Aktivierung von Plasminogen zu Plasmin beteiligt ist. Plasmin ist eine Serin-Proteinase, die unter anderem Proteine der extrazellulären Matrix spaltet. Eine wichtige Aufgabe von Plasmin ist die Spaltung von Fibrin

---

[1] Genentech Inc., South San Francisco, California

[2] FISH – fluorescence in-situ hybridisation

[3] Ein immunhistochemischer Test besteht darin, dass ein Gewebeschnitt mit einem für das Protein spezifischen und markierten Antikörper inkubiert wird. Die Stärke des Signals gibt Aufschluss über Höhe der Proteinexpression.

in Blutgerinnseln. Es akkumuliert vor allem an Wunden, entzündetem Gewebe und an Tumoren und ist an Um- und Neuformungsprozessen des Gewebes beteiligt. Plasminogen wird seinerseits proteolytisch gespalten und zu Plasmin prozessiert. Das kann durch die zwei Plasminogenaktivatoren, PLAU oder PLAT geschehen. PLAU kann über den Rezeptor PLAUR an die Zellmembran binden und erreicht durch die Bindung eine erhöhte Aktivität. Rezeptorgebundes PLAU findet man an den Spitzen auswachsender Nerven, an der Vorderseite migrierender Leukozyten und metastasierender Tumorzellen. Außerdem konnte gezeigt werden, dass PLAU das Zellwachstum stimuliert. Ein Inhibitor von PLAU, an dem bereits viel geforscht wurde, ist PAI-1. Bindet PAI-1 an das rezeptorgebundene PLAU, werden die drei Proteine von der Zelle internalisiert und dort PLAU und PAI-1 degradiert und PLAUR wieder zur Zelloberfläche gebracht. PLAU ist bereits als in verschiedenen Tumoren überexprimiert beschrieben, und es gibt gegen dieses Protein Antikörpertherapien in der klinischen Prüfung. Tetranectin ist ein sezerniertes Protein, das an Plasminogen bindet.

Auf dem Oligo-Chip sind folgende Gene repräsentiert: *plasminogen, PLAU, PLAT, PLAUR, PAI-1, tetranectin*. Sowohl *plasminogen* wie auch *PLAU* und *PLAT* sind häufig stark hochreguliert, *PAI-1* und *tetranectin* sind hingegen vermindert exprimiert. Aufgrund der Expression von *tetranectin* lassen sich Plattenepithelkarzinom- und Normalgewebeproben des Oligo-Array-Datensets perfekt diskriminieren (Abbildung 3-10). Auf dem cDNA-Array ist *tetranectin* bedauerlicherweise nicht repräsentiert.

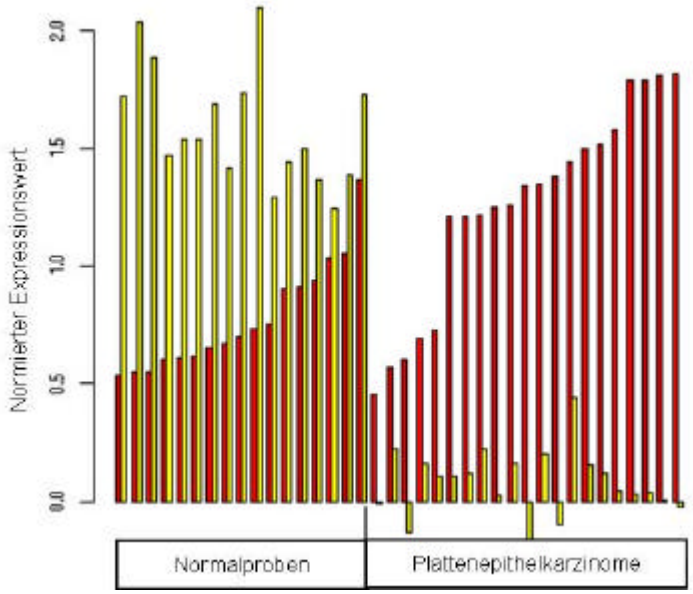


Abbildung 3-10 Die Expression von *PLAU* und *tetranectin*
Es sind die Genexpressionswerte vom Oligo-Chip des Plasminogenaktivators PLAU (rot) und des Plasminogen- bindenden Proteins Tetranectin (gelb) in den Normalproben und Plattenepithelkarzinom-proben dargestellt.

Eine Liste von Beispielgenen, die auf beiden Technologieplattformen als differenziell exprimiert auffallen, ist in Anhang B beigefügt.

Abschließend wurde mit einem eigenen Ansatz versucht, Gene zu identifizieren, deren Expression sich mit der Prognose der Adenokarzinompatienten unterscheidet. Dahinter verbirgt sich die Hoffnung, besonders aggressive Tumoren mit schlechter Prognose bei Anwendung herkömmlicher Standardtherapien molekularbiologisch zu charakterisieren und damit neue Therapieansätze zu eröffnen. Teil der publizierten Arbeiten war eine Clusteranalyse der Adenokarzinomproben. Für einige der gefundenen Patientengruppen mit jeweils ähnlichen Expressionsmustern konnte eine Tendenz zu besserer beziehungsweise schlechterer Prognose beobachtet werden. Wie von den Autoren selbst angemerkt, ist bei der Interpretation dieser Ergebnisse aufgrund der eingeschränkten Datenlage Vorsicht geboten. Eine weitere Arbeit, die eine Korrelation von Expressionsprofile mit der Überlebenserwartung auf Grundlage des Oligo-Array-Datensatz herstellt, wurde auf dem ISMB-Kongress[1] präsentiert [31]. Bei der Suche nach Genen, die eine differenzielle Expression bezüglich unterschiedlicher Prognose zeigen, lässt sich auch anders vorgehen: Laut Probenbeschreibung

[1] Intelligent Systems for Molecular Biology, 2002, Edmonton, Kanada

können in beiden Datensätzen jeweils zwei Gruppen von Patienten gebildet werden, eine mit relativ langen und eine mit sehr kurzen Überlebenszeiten. Für die Analyse dienten mehr als 32 Monate und weniger als 12 Monate als entsprechende Schwellen. Eine noch schärfere Trennung der Prognose ist nicht möglich, da sonst vom cDNA-Datensatz zu wenige Proben nutzbar sind. Der erste Schritt der Analyse bestand daraus, Listen zu generieren, in denen sich die Gene nach unterschiedlichen Kriterien für das Differenziellsein sortiert lassen. Genau wie für den Vergleich der Tumor- und Normalgewebe wurde der T-Test und die mittlere Expressionsänderung berechnet. Zusammenfassend muss konstatiert werden, dass es keine Gene gibt, die konsistent über beide Technologien in Adenokarzinomproben mit schlechterer Prognose höher beziehungsweise niedriger exprimiert sind.

Um sich über diese Beobachtung ein klareres Bild zu verschaffen, wurde als nächstes eine Clusteranalyse der Adenokarzinomproben des Oligo-Chips durchgeführt. Die Clusteranalyse gehört zum Gebiet der explorativen Methoden, deren Ziel es ist, mehrdimensionale Daten in unserem Fall also Experimente beziehungsweise Gene so zu gruppieren, dass in einem bestimmten Sinne ähnliche Daten in eine Gruppe und die in diesem Sinne verschiedenen Daten in unterschiedliche Gruppen fallen. Die Ähnlichkeit von Daten wird dabei über den Abstand definiert, was vor allem auch für die Expressionsanalyse wohl der kritischste Schritt ist. Um nach koexprimierten Genen zu suchen, kann man dafür beispielsweise den Korrelationskoeffizient benutzen. Da die Clusterverfahren der Vorsortierung der Daten und der Hypothesengenerierung dienen soll, ist es legitim auch einfach alle verfügbaren und einigermaßen begründbaren Abstandsmaße und Rechenverfahren auf die Daten anzuwenden, um auf diese Weise Gruppen zu finden, die gut mit klinischen Parametern korrespondieren und sich damit eventuell interpretieren lassen. Das Verfahren läuft wie folgt ab: Zuerst berechnet man pro Gen Indikatoren für eine differenzielle Expression und sortiert die entstehende Genliste danach. Als Indikatoren wurden die Differenz der mittleren Expression innerhalb der beiden Probengruppen und der P-Wert des T-Tests benutzt. Als nächstes wählt man die besten Gene der Liste und schränkt das Datenset darauf ein. Wie viele Gene das sind, ist der Willkür des Anwenders unterworfen und kann obiger Argumentation folgend an das Ziel der Analyse angepasst werden. Eine Fragestellung könnte beispielsweise lauten: Wie schafft man mit möglichst wenigen untersuchten Genen eine gute Trennung der Tumor- und Normalproben? Mit der Genauswahl lässt sich dann eine Matrix der Expressionswerte erstellen, die als Ausgangspunkt für verschiedenste Clusteranalyse dient. Die Rechnungen und Visualisierungen ließen sich einfach und schnell mit dem Computerprogramm GeneMath

(Applied Maths, Belgien) durchführen. Zur Demonstration ist das Ergebnis für ein leichtes Problem gezeigt: Die Trennung von Plattenepithelkarzinom- und Normalgewebeproben im Oligo-Chip-Datensatz. Die Aufgabe ist leicht, da ja schon ein einzelnes Gen, *tetranectin*, die zwei Gruppen perfekt trennt (Abbildung 3-10). Mit dem Ziel, möglichst nur stark differenzielle darzustellen, wurden 64 Gene ausgewählt.

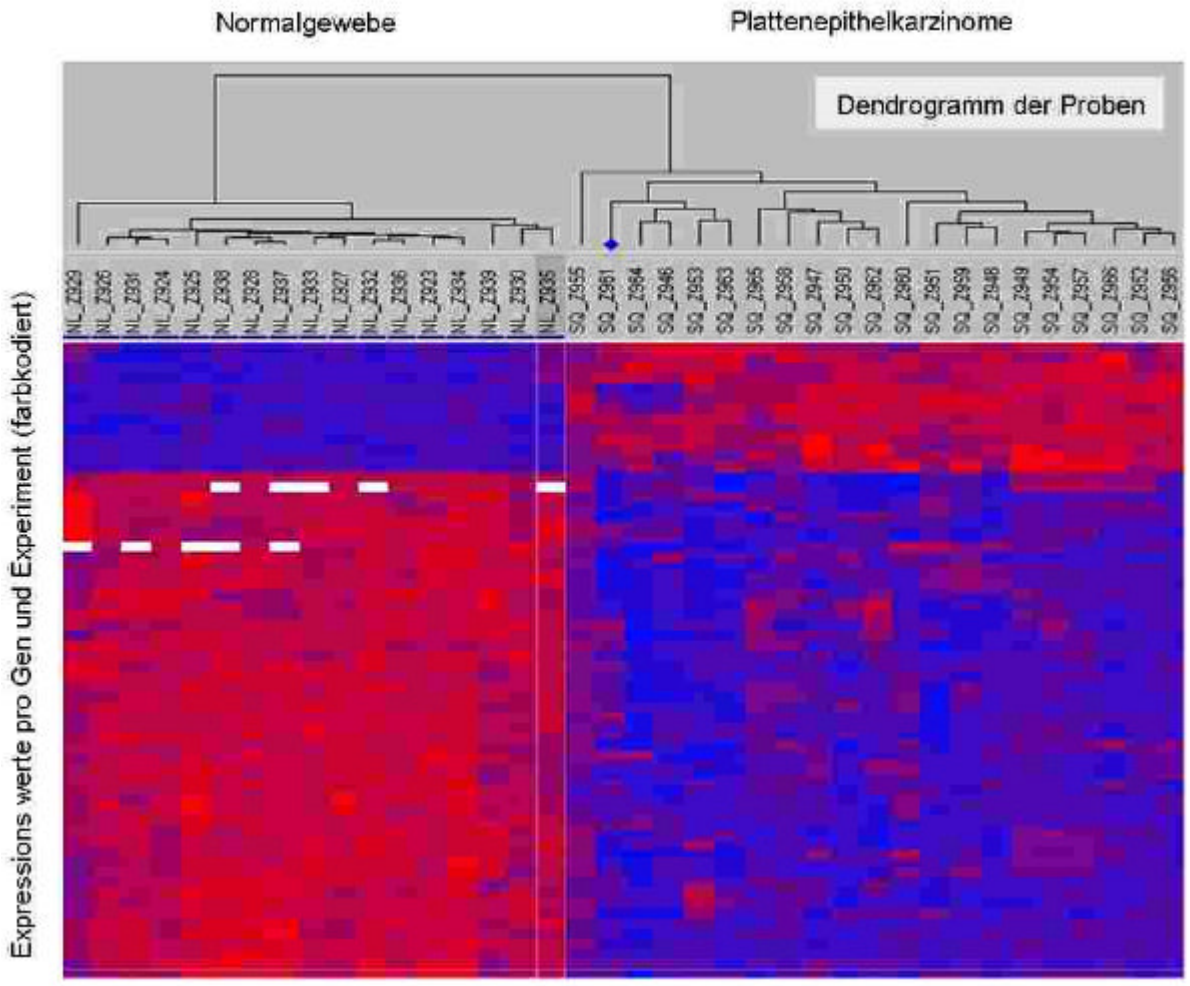


Abbildung 3-11 Ergebnis der Clusteranalyse für 64 Gene

Es wurde die 64 am stärksten differenziellen Gene aus der Liste ausgewählt, die Expressionswerte in einer Matrix aufbereitet. Die Proben konnten mit dem Programm GeneMath einem hierarchischen Clustering (complete linkage) unterzogen werden. Der Abstand zwischen zwei Proben ist definiert als die Euklidische Distanz ihrer Expressionswerte. Zur Darstellung wurden die Werte mit einer blau-rot-Skala farbkodiert, je roter desto höher und je blauer desto geringer. Gilt der Wert nach dem Detektionsscore als nicht verlässlich, so ist das Feld weis.

Die Normalgewebe und die Plattenepithelkarzinome lassen sich perfekt trennen. Innerhalb dieses Datensatzes gilt daher, dass sich zwei beliebige Plattenepithelproben einander in der Expression dieser 64 Gene ähnlicher sind als eine Karzinomprobe und eine Normalprobe.

Darüber hinaus fiel bei der Analyse auf, dass deutlich mehr Gene im Tumor vermindert exprimiert sind als erhöht.

Als nächstes sollten die Adenokarzinomproben, für die Überlebenszeiten der Patienten bekannt waren, geclustert werden. Ausgehend von den oben beschriebenen Listen ließen sich die 747 differenziellsten Gene auswählen und deren Expressionswerte in einer Zahlenmatrix zusammenfassen. Mit dem Programmpaket GeneMath lassen sich unterschiedliche Clusterverfahren basierend auf unterschiedlichen Abstandsmaßen zwischen Proben testen. Ziel ist es, eine Parameterkombination zu finden, bei der bezüglich der Überlebenszeit möglichst homogene Cluster entstehen. Das heißt, Proben mit ähnlichem Expressionsprofil sollten auch eine vergleichbare Prognose haben. Es wird nicht verlangt, dass zum Beispiel alle Patienten mit guter Prognose in ein Cluster fallen. Orthogonal dazu lassen sich die Expressionsprofile der Gene clustern. Als Abstandsmaß diente dabei der Korrelationskoeffizient von Pearson.

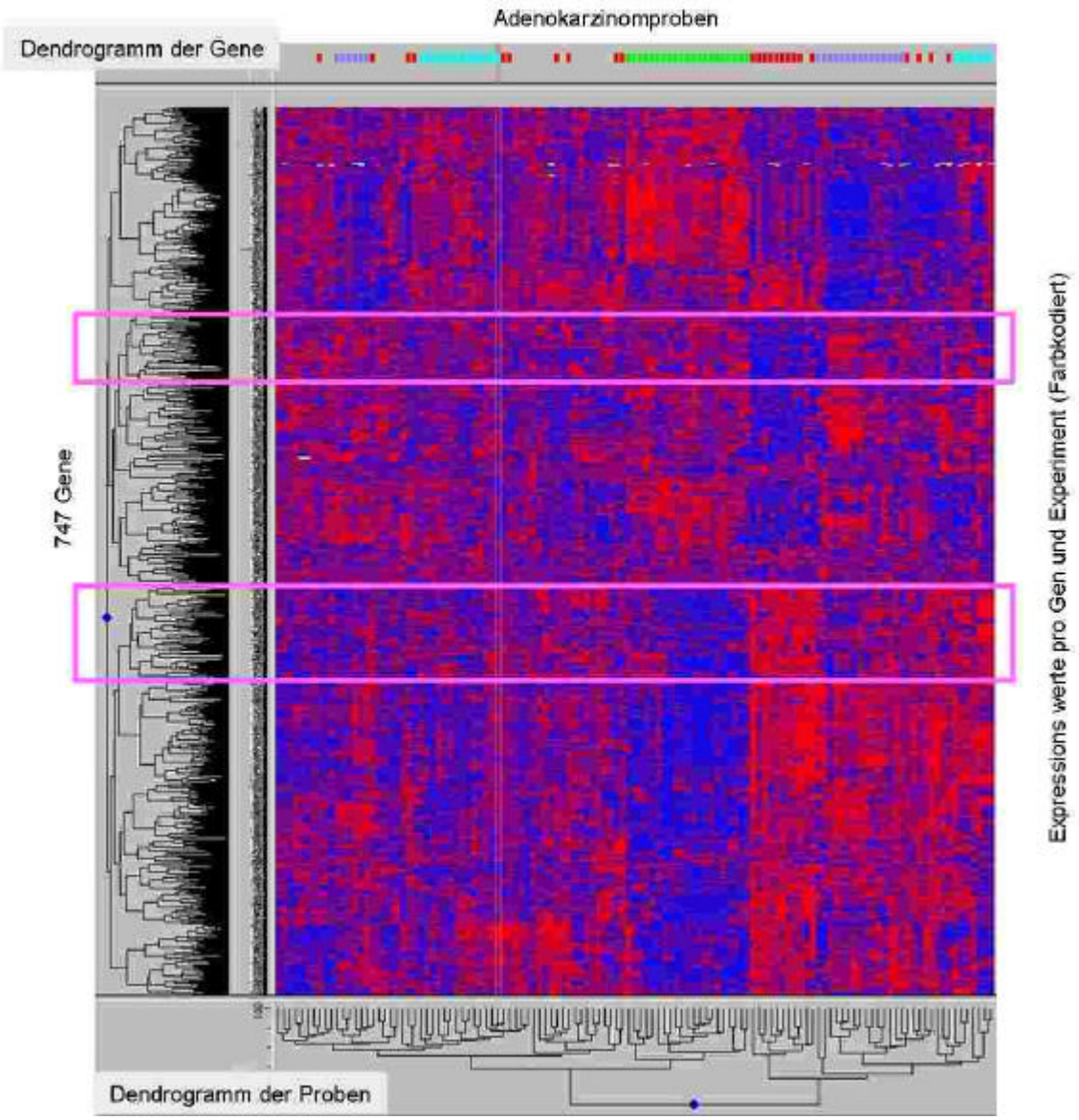


Abbildung 3-12 Clusteranalyse der Adenokarzinomproben der Oligo-Array-Daten
Es wurde die 747 am stärksten differenziellen Gene aus der Liste ausgewählt und die Expressionswerte in einer Matrix aufbereitet. Die Proben und die Gene konnten mit dem Programm GeneMath einem hierarchischen Clustering (complete linkage) unterzogen werden. Der Abstand zwischen Genen ist definiert als der Korrelationskoeffizient der Expressionswerte und der Abstand zwischen zwei Proben als die Euklidische Distanz. Die Dendrogramme veranschaulichen das Ergebnis des Clusterings. Die Werte wurden mit einer blau-rot-Skala farbkodiert, je roter desto höher und je blauer desto geringer. Gilt der Wert nach dem Detektionsscore als nicht verlässlich, so ist das Feld weis. Die pinkfarbenen Rahmen markieren zwei Gencluster, auf die im Text verwiesen wird. Zusätzlich sind einige Probengruppen mit farbigen Rechtecken über der Matrix markiert.

Beim Versuch, mit Hilfe verschiedener Clusterverfahren die Daten bezüglich der Prognose zu gruppieren, bestätigt sich die Beobachtung aus der Analyse der Listen zur differenziellen Genexpression zwischen Patienten mit langer und kurzer Überlebenszeit: Es lassen sich keine typischen Expressionsmuster für Patienten mit guter beziehungsweise schlechter Prognose

identifizieren. Das beste Ergebnis, was sich durch Kombination unterschiedlicher Verfahren und Abstände erreichen ließ, ist in Abbildung 3-12 dargestellt. Die Proben von Patienten mit schlechter Prognose sind mit rot über der Matrix markiert. Zehn der insgesamt 24 Proben mit schlechter Prognose und eine mit guter Prognose fallen in eine Gruppe, wobei sich die anderen 14 Proben mit schlechter Prognose auf andere Cluster verteilen. Diese Gruppe wird im Folgenden das rote Cluster genannt. Eine weitere Gruppe mit 21 Proben (hell grüne Markierung) zeigt ein homogenes Expressionsprofil und enthält ausschließlich solche mit guter Prognose. Die mit Blautönen gekennzeichneten Probengruppen sind zwar kleiner, zeigen aber ebenfalls homogene Expressionsmuster. Zwei auffällige Gruppen von Genen sind in der Abbildung mit pinkfarbenen Linien umrahmt. Die meisten Gene der oberen Box sind im hellgrünen Cluster relativ hoch und im roten Cluster gering exprimiert. Die Gene der unteren Box verhalten sich umgekehrt, im hellgrünen Cluster sind sie gering und im roten Cluster hoch exprimiert.

3.8. Zusammenfassung und Diskussion

Die beiden Forschergruppen auf deren Datensätzen dieser Abschnitt beruht, untersuchten Expressionsprofile einer Reihe von Subtypen des Bronchialkarzinoms. Besonderes Augenmerk lag dabei auf den Adenokarzinomen, was sich in der Anzahl und der mitgelieferten Beschreibung der Proben und in der Sorgfalt der Analyse widerspiegelt. Jede der Gruppen konnte Subpopulationen der Adenokarzinomproben identifizieren und diese teilweise sogar mit Überlebensdaten assoziieren. Hauptunterschiede der beiden Studien bestehen in den höheren Fallzahlen, in der ausgiebigeren Annotation der Proben der Gruppe aus Massachusetts und in der Technologieplattform. Die Resultate der in diesem Kapitel vorgestellten Reanalyse der Datensätze lassen sich wie folgt zusammenfassen:

1. Das in Kapitel 2 eingeführte Verfahren zur Arrayanalyse liefert reproduzierbar zwischen Experimenten vergleichbare Expressionswerte.

2. cDNA-Array-Experimente können auf Basis der hintergrundkorrigierten Signalwerte Aufschluss über die absolute Konzentration der Transkripte geben.

3. Die Mittelwerte der normierten absoluten Expressionssignale der auf beiden Arrays repräsentierten Gene korrelieren insgesamt nur schwach zwischen den

Technologieplattformen. Dabei gilt, dass sich die höchsten 10% der Signale am besten bestätigen.

4. Untersucht man differenziell exprimierte Gene zwischen Tumor- und Normalgewebe in beiden Datensätzen, so findet man wiederum nur eine schwache Korrelation in den errechneten Expressionsunterschieden. Eine gute Übereinstimmung erreicht man in den jeweils differenziellsten 10% der Gene für die zusätzlich der T-Test einen signifikanten Unterschied anzeigt.

5. Die als differenziell in der Literatur beschriebenen und auf den Arrays repräsentierten Gene können zum Teil mit den Expressionsdaten bestätigt werden.

6. Die Plasminogenaktivierung und deren Regulation scheint ein wichtiger Prozess in der Progression des Bronchialkarzinoms zu sein.

7. Die Plattenepithelkarzinomproben sind mit Hilfe der Expressionsdaten gut von Normalgeweben zu trennen, was sich durch eine Clusteranalyse demonstrieren lässt. Adenokarzinomenpatienten mit ähnlicher Prognose zeigen im Gegensatz dazu keine spezifischen Expressionsmuster.

Eine Reihe von Problemen bleibt ungelöst. Ist eine der Technologien überlegen? Um hier Klarheit zu schaffen, könnte eine Anzahl von Proben parallel auf beiden Technologie-plattformen gemessen werden. Interessant wäre es auch zu prüfen, ob die Adenokarzinome in ihren Expressionsprofilen tatsächlich inhomogener sind als die Plattenepithelkarzinome. Es ist aber schwer abzuschätzen, wie groß die Probenzahlen für eine definitive Aussage sein müssten. Außerdem können die Regeln für das Differenziellsein überprüft und gegebenenfalls angepasst werden, sobald es mehr Gene gibt, deren Expression im Bronchialkarzinom qualitativ wie quantitativ gut untersucht ist.

4. Die Entdeckung neuer Transkripte mit Hilfe von Oligonukleotid-Arrays

Nach den Veröffentlichungen über die Sequenzierung des humanen Genoms [32], schließt sich als nächste Aufgabe die genaue Charakterisierung des Transkriptoms an. Es gilt also möglichst alle Gene zu finden, die in einem spezifischen Zelltypen unter ganz bestimmten physiologischen Bedingungen in RNA umgeschrieben werden. Einen wichtigen Beitrag haben hier bis Dato die EST-Datenbanken geliefert. Kartiert man die entstehenden Sequenzen zurück auf das Genom, kann man eine große Anzahl der transkribierten Gene und oft sogar Spleißstellen und kleine Exons finden. Eine andere weit verbreitete Technik ist die automatische Annotation genomischer Sequenz. Man nimmt an, dass die Gene etwa 3% des humanen Genoms ausmachen. Außerdem sind die Transkripte in zum Teil sehr kleine Exons untergliedert und die Länge der dazwischen liegenden Introns übersteigt die der Exons bei weitem, so dass man für Genvorhersagen im humanen Genom eine wesentliche schlechtere Performanz der Programme erwartet als beispielsweise im Hefegenom. Man extrahiert hierzu Eigenschaften der Sequenz bekannter Transkripte, die diese von nichtkodierenden Bereichen unterscheiden. Die wohl bekannteste Eigenschaft ist die unterschiedliche Häufigkeit der vier Basen an den drei Positionen der Kodons in proteinkodierenden Regionen (CDS). Außerdem macht man sich konservierte Sequenzmotive zu nutze, um beispielsweise Exon-Intron – Übergänge oder Polyadenylierungsstellen zu identifizieren. Funktionelle RNA's, die nicht für Proteine kodieren und aus sehr kurzen Exons bestehen, sind mit den derzeit gebräuchlichen Computerprogrammen praktisch nicht zu entdecken. Die Chiptechnologie eignet sich ebenfalls hervorragend zur Suche nach neuen Genen. Theoretisch könnte man das gesamte Genom mit Oligos abrastern und dann die entsprechenden Sonden auf Oligo-Chips synthetisieren. Damit wäre man in der Lage, alle Transkripte, die in den untersuchten Zellen in ausreichender Zahl vorhanden sind, zu detektieren. Die Oligodichte ist heutzutage allerdings im Bereich von einer halben Million pro Chip und für das humane Genom bräuchte man vielleicht ein halbe Milliarde.

Anhand der auf dem Chip repräsentierten Transkripte lassen sich interessante Analysen durchführen. Wie in Abschnitt 1.3 bereits erwähnt, sind auf metaGen Chip I zwei Sets von Sequenzpaaren, die als Basis für dieses Kapitel dienen:

Set 1. 1066 Transkriptfragmente[1] für die Sondensets in beiden Orientierungen konstruiert wurden.

Set 2. 588 Gene mit jeweils einem Sondenset für die 3'-UTR und einem für die kodierende Region.

Bei der Analyse einzelner Gene fand man wider Erwarten etliche, für die sowohl die Sense- wie auch Antisense-Sonde relative konsistent Expressionssignale liefern. In Set 2 fallen einige Sequenzen auf, bei denen die CDS-Sonde ein höheres Signal zeigt als die 3'-UTR-Sonde. Auf Grund der reversen Transkription beginnend am Poly-A-Ende der mRNA kann man erwarten, dass die 3'-UTR der untersuchten Sequenzen die stärksten Signale liefert. Es könnte sich hierbei um alternative Spleißvarianten beziehungsweise alternativ polyadenylierte Transkripte handeln. Ebenso gut können den Beobachtungen technische Artefakte zugrunde liegen, auf die weiter unten eingegangen wird.

4.1. Insilico-Analyse

Den Beobachtungen sollte in einer detaillierten Analyse nachgegangen werden. In Kapitel 2 ist der P-Wert des Wilcoxon-Tests als Detektionsscore beschrieben, mit dem die Bewertung, ob ein Transkript in einem RNA-Pool vorhanden ist, möglich sein soll. Bei der großen Anzahl parallel untersuchter Sequenzen kommt es sicherlich vor, dass die Signale eines Oligosets suggerieren, ein Transkript sei vorhanden, obwohl nur rein zufällig die *PM*-Signale für fast alle Oligopaare höher sind als die *MM*-Signale. Die große Anzahl der bereits durchgeführten Experimente versetzt uns jedoch in die Lage, mit einiger Sicherheit zufällige und echte Effekte zu diskriminieren.

Durch die Beschränkung auf eine gut charakterisierte Teilmenge der beiden Sequenzsets konnte die Analyse wesentlich vereinfacht werden. In die Auswahl gelangten nur Sondensets, für die mindestens 18 der 20 Oligos genau eine Sequenz in der *RefSeq*-Datenbank treffen.

---

[1] Zum Zeitpunkt der Chipkonstruktion existierte für dieses Transkript nur ein EST-Cluster, dessen Orientierung nicht bekannt war. Für den Chip konstruierte man daher jeweils Sense- und Antisense-Sonden.

RefSeq [2] ist eine Datenbank von Referenzsequenzen am NCBI, die sich dadurch auszeichnet, dass die Einträge von Experten kuriert sind. Darüber hinaus arbeitet man zurzeit intensiv an ihrem Ausbau. Da man sich auf gut charakerisierte Sequenzen beschränkt, steht für die meistens einiges an Zusatzinformation online zur Verfügung. Außerdem schließt man damit bekannte Genfamilien aus der Analyse aus, für die man oft nicht einwandfrei entscheiden kann, welches Transkript gerade detektiert wird. Der Nachteil liegt im Verlust von Sequenzen, die noch nicht in diese Datenbank übernommen wurden und solche, von denen mehrere Varianten vorhanden sind. Zum Beispiel sind in der Datenbank einige gut charakterisierte genetische Varianten des Gens *BRCA1* zu finden, die alle von ein und demselben Sondenset getroffen werden. Pro Genom und damit pro Organismus existiert aber normalerweise nur eine Variante dieses Gens. Die von einer Gewebeprobe erhobenen Expressionsdaten beziehen sich auf nur ein bestimmtes Gen und sind damit valide. Von den 588 CDS-UTR-Sondenpaaren (Set 2) treffen 213 eindeutig Referenzsequenzen unter den beschriebenen Bedingungen. Von den 1066 Sense-Antisense-Paaren aus Set 1 treffen 102. Das von Set 1 wesentlich weniger die Bedingungen erfüllen, lässt sich mit der Auswahl der Sequenzen erklären. Zur Zeit der Konstruktion von metaGen-Chip I, 1998, waren insgesamt viel weniger Transkripte gut charakterisiert. Fand man EST-Cluster in Datenbanken, für die weder der proteinkodierende Bereich noch die 3'-5'-Richtung des Transkripts bekannt war, brachte man zwei Sondensets für dieses Cluster auf den Chip, eins für jede Richtung. Im Gegensatz zu den CDS-UTR-Paaren war zu der Zeit also das entsprechende Gen noch nicht bekannt beziehungsweise nicht hinreichend charakterisiert.

Es wurden Datensätze von 310 Chipexperimenten ausgewählt, die bereits mit dem Verfahren der Einzelchipanalyse aus Kapitel 2 ausgewertet worden waren. Für die paarweise Betrachtung schien eine zusätzliche Normierung zwischen den Chipdatensätzen nicht notwendig. Die RNA für die Chiphybridisierungen stammte aus 63 Zelllinienpräparaten und aus 247 Tumor- oder Normalgewebeproben von Krebspatienten. Mit den beiden Sequenzsets lassen sich inhaltlich völlig unterschiedliche Fragen bearbeiten. Der Grund dafür, dass sie hier parallel dargestellt sind, liegt in der Anwendbarkeit desselben Verfahrens für ihre Auswertung: Sei eine Menge von Sequenzen gegeben, die durch je zwei Oligosets auf dem Chip repräsentiert sind. Weiterhin habe man die Detektionsscores, also die P-Werte aus dem Wilcoxon-Test, aus den oben beschriebenen 310 Experimenten vorliegen. Dann lässt sich eine Zählstatistik anfertigen, in dem man für jede Sequenz erfasst, wie oft eins der Sondensets

beziehungsweise beide gleichzeitig ein P-Wert kleiner 0,05 erreichen und damit die Detektion eines Transkripts anzeigen.

Genauer: Bezeichne *E* die Menge der 310 Experimente, dann bildet man für jede Referenzsequenz *r*:

$$NANB_r \leftarrow |\{e \mid e \in E \text{ und } p\_Wert(a_r, e) \geq 0{,}05 \text{ und } p\_Wert(b_r, e) \geq 0{,}05\}|$$
$$ANB_r \leftarrow |\{e \mid e \in E \text{ und } p\_Wert(a_r, e) < 0{,}05 \text{ und } p\_Wert(b_r, e) \geq 0{,}05\}|$$
$$NAB_r \leftarrow |\{e \mid e \in E \text{ und } p\_Wert(a_r, e) \geq 0{,}05 \text{ und } p\_Wert(b_r, e) < 0{,}05\}|$$
$$AB_r \leftarrow |\{e \mid e \in E \text{ und } p\_Wert(a_r, e) < 0{,}05 \text{ und } p\_Wert(b_r, e) < 0{,}05\}|$$

Dabei bezeichnen a_r und b_r die Sondensets zur Referenzsequenz r und $p\_Wert(x,e)$ den P-Wert aus dem Wilcoxon-Test angewendet auf die Daten des Sondensets *x* aus dem Experiment *e*. Bei korrekter Zählung muss die Summe der vier Zahlen wieder 310 ergeben. Für quantitative Aussagen bildet man die Differenzen der PM-Quartile der beiden Sondensets pro Referenzsequenz und mittelt anschließend über die 310 Experimente.

4.2. Resultate der Insilico-Analyse

Für die *RefSeq*-Sequenzen sind Länge und die Orientierung in der Datenbank angegeben. In den meisten der untersuchten Fälle konnte eine Übereinstimmung der Chipdaten mit den Annotationen gefunden werden, dass heißt von den Sense-Antisense-Sequenzpaaren (Set 1) wird nur die in der annotierten Orientierung auf dem Chip detektiert, und man findet eine höheres Signal für das weiter 3' gelegene Sondenset. In Set 2 der CDS-UTR-Paare (213 insgesamt) wurden aber auch 27 entdeckt, deren Sondenset aus der CDS in mehr als 10 Versuchen überhaupt und häufiger als das aus der 3'-UTR etwas detektiert hat. In 13 weitere Paaren weist die Differenz der mittleren Expressionswerte von 3'-UTR und CDS einen deutlichen negativen Wert auf (Differenz der *PM*-Quartile < -2). Aufgrund des hohen Anteils an Fällen, die den Erwartungen widersprechen, kann man mit einiger Sicherheit ausschließen, dass es sich um rein zufällige Effekte handelt.

Tabelle 4-1 Ergebnis der Zählstatistik für 13 widersprüchliche Beispiele

RefSeq Bezeichner	Wie oft detektiert				Differenz 3'-UTR - CDS	Beschreibung der Sequenz
	-/-	U/-	-/C	U/C		
NM_002428	5	4	275	26	3,0178	matrix metalloproteinase 15 (membrane-inserted) (MMP15)
NM_002205	32	0	267	11	-0,2919	integrin, alpha 5 (fibronectin receptor, alpha polypeptide) (ITGA5)
NM_000074	69	0	238	3	0,7015	TNF (ligand) superfamily, member 5 (hyper-IgM syndrome) (TNFSF5)
NM_004148	65	2	199	44	-2,0445	ninjurin 1 (NINJ1)
NM_000629	140	1	167	2	2,2297	interferon (alpha, beta and omega) receptor 1 (IFNAR1)
NM_015865	2	1	162	145	-5,1664	solute carrier family 14 (urea transp.), member 1 (Kidd blood group) (SLC14A1)
NM_003839	174	2	130	4	-2,3329	TNF receptor superfamily, member 11a, activator of NFKB (TNFRSF11A)
NM_006286	11	1	126	172	0,4282	transcription factor Dp-2 (E2F dimerization partner 2) (TFDP2)
NM_002009	197	7	93	13	-0,2663	fibroblast growth factor 7 (keratinocyte growth factor) (FGF7)
NM_003239	29	55	76	150	0,8726	transforming growth factor, beta 3 (TGFB3)
NM_001559	228	0	76	6	-5,0271	interleukin 12 receptor, beta 2 (IL12RB2)
NM_003015	309	0	1	0	-5,3675	secreted frizzled-related protein 5 (SFRP5)
NM_001153	5	4	6	295	-5,5452	annexin A4 (ANXA4)

Die mRNA-Sequenzen sind über ihren Bezeichner eindeutig in der Referenzsequenzdatenbank (RefSeq) identifizierbar. Die vier folgenden Spalten enthalten die Zählstatistik, wie im vorherigen Abschnitt beschrieben. -/- symbolisiert die Anzahl der Experimente (aus 310) in denen keines der Oligosets ein Signal liefert. U/- zählt, in wie vielen Experimenten die 3'-UTR-Sonde etwas detektiert hat aber nicht die CDS-Sonde, -/C entsprechend anders herum. U/C zählt die Experimente, in denen beide Sonden gleichzeitig ein Signal liefern. In der Differenzspalte ist die mittlere Differenz der PM-Quartile von 3'-UTR-Sonde und CDS-Sonde eingetragen. Farblich unterlegt sind gerade die Werte, die den Widerspruch zur Annotation belegen.

Während der Analyse viel auf, dass es keine deutliche Korrelation zwischen dem Abstand der Sonde zum annotierten 3'-Ende des Transkripts und der mittleren Signalstärke gibt. Bei der Chipkonstruktion rät Affymetrix die Sonde aus den letzten 600 Basen zum 3'-Ende der mRNA zu wählen, da die reverse Transkription am Poly-A-Schwanz beginnt und das Enzym nicht beliebig lange Moleküle umschreiben kann. Genauer gesagt, verbirgt sich dahinter die Modellvorstellung, dass das Enzym eine bestimmte Effizienz bei der Kettenverlängerung hat. Folglich bricht die Synthese an jeder Base mit einer gewissen kleinen aber positiven Wahrscheinlichkeit ab. Diese Wahrscheinlichkeiten summieren sich über die Länge des Moleküls auf, so dass nach dem Modell das Vorkommen von Transkripten negativ mit ihrer Länge korreliert.

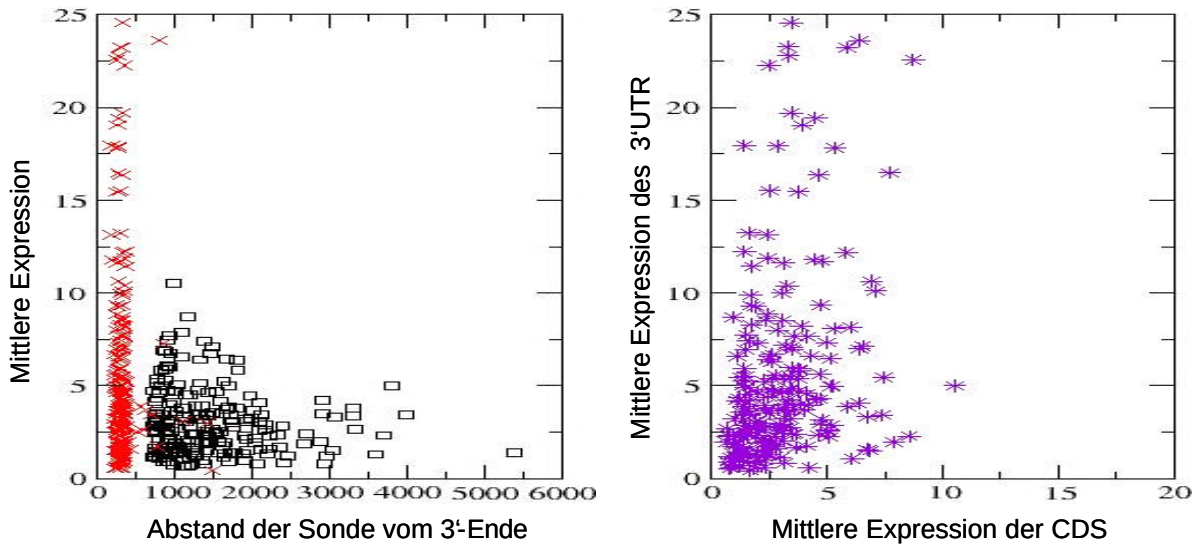


Abbildung 4-1 Die Beziehung zwischen mittlerer Expression und Abstand der Sonde vom 3'-Ende

Links *ist für das CDS-UTR-Sequenzset das mittlere Expressionssignal in Abhängigkeit vom Abstand des Oligosets vom 3'-Ende dargestellt. Mit roten Kreuzen sind die Sonden für die 3'-UTR's markiert und mit schwarzen Rechtecken die für die kodierenden Bereiche. Man kann hier ablesen, dass die CDS-Sonden selten hohe Expressionssignale liefern, aber eine deutliche negative Korrelation ist nicht erkennbar.* ***Rechts*** *sind die mittleren Expressionssignale der 213 CDS-UTR-Paare direkt gegeneinander aufgetragen. Wieder erkennt man, dass die höheren Signale tendenziell eher von den 3'-UTR-Sonden erreicht werden.*

Im Folgenden sind einige Gründe aufgelistet, die zu den beobachteten Fällen führen können:

1. Der Einfluss, den der Abstand der Sonde vom 3'-Ende auf das Expressionssignal hat, könnte wesentlich geringer sein als erwartet und damit von der Molekülkonzentration überlagert sein. Das hieße, die Enzyme erzeugten doch häufiger als vorausgesetzt Produkte, die weit über 1000 Basen lang sind.

2. Andere sequenzabhängige Einflussgrößen, wie zum Beispiel die Basenkomposition, fallen viel stärker ins Gewicht und verhindern damit ein klares Bild der Abhängigkeit des Signals von der Position der Sonde.

3. Bindet der Oligo-dT-Primer nicht nur am poly-A-Ende sondern auch innerhalb des Transkripts, so entstehen bei der Amplifikation kürzere Produkte, die eventuell nicht mehr den Bereich der 3'-Sonde überdecken. Die Synthese des kurzen Produkts kann dabei die des eigentlichen Transkripts verhindern.

4. Es gibt für viele untersuchte Gene noch Transkriptvarianten, die nicht in der Datenbank vermerkt sind. Besonders alternative 3'-Enden sind hier zu beachten.

5. Die Annotation der 3'-Enden in der Datenbank ist falsch. Viele Sequenzierungsprojekte von cDNA-Banken konzentrieren sich auf den proteinkodierenden Bereich und legen weniger Augenmerk auf die UTR's.

6. Der Einfluss, den die Sonde auf das Signal hat, könnte so stark sein, dass er andere Effekte überlagert. Für stark bindende Oligo-Sonden könnten wenige Zielmoleküle ausreichen, um ein detektierbares Signal zu erreichen, wohingegen schwach bindende Sonde eventuell nie ein Signal liefert.

Vermutlich liegt den Beobachtungen ein schwer zu entflechtendes Gemisch von Einflüssen zu Grunde. Auf jeden Fall ist man mit derlei Analysen in der Lage, Genvorhersagen zu überprüfen, neue Varianten bekannter Gene und deren Gewebespezifität zu entdeckten. Ein sorgfältiges auf diesen Zweck abgestimmtes Chipdesign würde uns in die Lage versetzen, das Transkriptom im großen Maßstab genauer zu charakterisieren. Die Untersuchungen am CDS-UTR-Datenset wurden an dieser Stelle abgebrochen, da inhaltlich das Interesse an dem Sense-Antisense-Set wesentlich stärker war.

Tabelle 4-2 Sense-Antisense-Koexpression

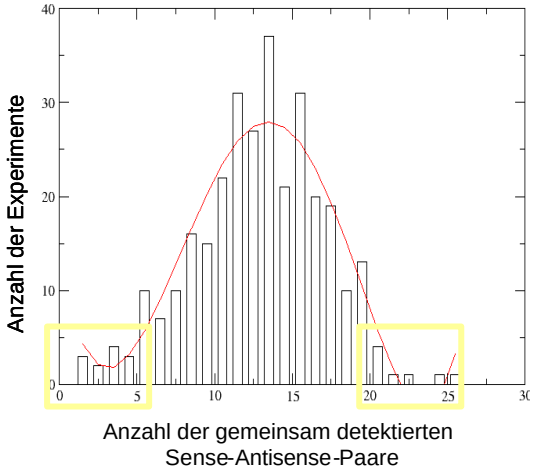
RefSeq	Wie oft detektiert				5'-EST's		3'-EST's		ETS's	PMQ-Mittel		Beschreibung der Sequenz
Bezeichner	S,A	S,A	S,A	S,A	A	S	S	A	gesamt	S	A	
NM_018011	10	2	24	274	4	31	118	17	184	8.9632	2.642	hypothetical protein FLJ10154 (FLJ10154)
NM_016127	3	3	34	270	56	1	2	117	229	10.775	1.0367	HSPC035 protein (LOC51669)
NM_018509	9	3	32	266	0	35	93	11	155	10.571	1.2585	hypothetical protein PRO1855 (PRO1855)
NM_006526	26	19	37	228	3	6	19	1	39	1.938	0.9468	zinc finger protein 217 (ZNF217)
NM_018975	4	0	97	209	5	28	113	5	177	6.0006	1.0738	TRF2-interacting telomeric RAP1 protein (RAP1)
NM_016629	19	2	98	191	14	1	0	37	61	8.7378	1.1901	hypothetical protein (LOC51323)
NM_024026	35	65	32	178	16	7	35	4	64	2.333	0.8784	mitochondrial ribosomal protein 63 (MRP63)
NM_016617	8	2	125	175	2	6	1	67	78	3.9674	0.3897	hypothetical protein (BM-002)
NM_020188	7	2	139	162	53	2	5	83	176	3.9187	2.9169	DC13 protein (DC13)
NM_021238	5	1	146	158	1	16	110	8	143	7.893	1.0875	TERA protein (TERA)
NM_030912	7	1	147	155	12	2	2	34	57	7.953	1.6552	tripartite motif protein TRIM8 (TRIM8)
NM_022349	38	41	81	150	5	1	0	17	25	4.2719	1.4491	CD20-like precusor (LOC64166)
NM_015070	39	93	31	147	0	3	6	5	47	1.0976	1.7779	KIAA0853 protein (KIAA0853)
NM_006283	14	2	156	138	0	29	106	0	153	12.063	2.2953	transforming, acidic coiled-coil cont. 1 (TACC1)
NM_015385	101	34	43	132	1	11	30	0	53	2.4284	0.7084	SH3-domain protein 5 (ponsin) (SH3D5)
NM_014050	67	55	64	124	0	95	30	2	153	2.9389	0.6677	mitochondrial ribosomal protein L42 (MRPL42)
NM_017689	62	121	16	111	0	9	45	2	75	9.2354	1.9437	hypothetical protein FLJ20151 (FLJ20151)
NM_023037	86	11	122	91	0	9	14	1	29	2.0853	0.5399	putative gene product (13CDNA73)
NM_022781	25	6	201	78	0	7	54	3	70	2.2709	0.6197	hypothetical protein FLJ21343 (FLJ21343)
NM_007106	36	2	194	78	1	26	35	2	72	2.738	0.6532	ubiquitin-like 3 (UBL3)
NM_016052	1	232	0	77	0	18	2	1	22	1.411	3.0785	CGI-115 protein (LOC51018)

Die mRNA-Sequenzen sind über ihren Bezeichner eindeutig in der Referenzsequenzdatenbank (RefSeq) identifizierbar. Die vier folgenden Spalten enthalten die oben beschriebene Zählstatistik. Die Spalte S,A (nicht Sense, nicht Antisense) enthält die Anzahl der Experimente (aus 310) in denen keines der Oligosets ein Signal liefert. S,A (nicht Sense, Antisense) zählt, in wie vielen Experimenten die Sense detektierende Sonde kein Signal liefert, aber die die Antisense-Transkripte detektiert , S,A entsprechend anders herum. S,A zählt die Experimente, in denen beide Sonden also für Sense und Antisense gleichzeitig ein Signal liefern. In den EST-Spalten ist die Datenbanksuche zusammengefasst (Referenzsequenz gegen dbEST_human1). In den mittleren mit S überschriebenen Spalten sind die ETS's gezählt, die das Sense-Transkript bestätigen. In den äußeren, mit A überschriebenen sind die gezählt, die von einem potenziellen Antisense-Transkript stammen. In der PMQ-Mittel-Spalte sind die mittlere PM-Quartile von Sense detektierender Sonde(S) und Antisense(A) detektierender Sonde eingetragen. Die farblichen Markierungen weisen besondere Felder aus und dienen damit der besseren Lesbarkeit der Tabelle. Außer gelb markeiren alle anderen Farben Felder, die das Vorhandensein von Antisense-Transkripten belegen.

Unter den 102 ausgewählten Sense-Antisense-Paaren waren 21, die in einem Viertel der 310 Experimente für beide Sondensets ein positives Signal lieferten. Ein weiteres Paar zeigte in 13 Experimenten nur für die Antisense-Sonde ein Signal und in 15 anderen nur für die Sense-

[1] Mit der Sequenz der RefSeq-Datenbank für das jeweilige Gen wurde mit dem Programm BLAST in der humanen Fraktion der EST-Datenbank dbEST gesucht.

Sonde. In 17 der 21 Fälle ist das Signal für das annotierte Transkript wesentlich stärker und variabler. Es scheint, als ob die meisten der untersuchten potenziellen Antisense-Transkripte nur ein geringes Expressionsniveau erreichen. Für acht weitere Sequenzpaare widersprechen die Chipergebnisse der annotierten Orientierung. Als Kriterium galt hier, dass eine der beiden Sonden in mindestens 10 Experimenten ein Transkript detektiert und dass das insgesamt für den Gegenstrang häufiger geschieht. Die Vermutung liegt natürlich nahe, dass es sich hier um technische Artefakte handeln könnte. Aus diesem Grunde ist die Prüfung einiger Hypothesen ratsam, die es erlauben würden, zwischen technischen beziehungsweise biologischen Faktoren zu diskriminieren. Dazu kann man zuerst die allgemeine Qualität der Chips und der resultierenden Datensätze betrachten. Die koexprimierten Transkriptpaare könnten sich in einigen Chipexperimenten häufen, weil deren RNA-Aufbereitung nicht richtig funktioniert hat oder deren Hybridisierungslösungen verunreinigt waren.



Experiment	Anzahl			Mittlerer Hintergrund	Q3 – Q1
	PM < MM	gesättigt	detektiert		
No_69_1_3	41804	1546	2394	837.07	2755.5
No_138_1_1	41927	1928	2266	845.55	2449
No_184_1_1	44132	328	2217	1068.6	1932.5
No_185_1_1	44918	330	2120	965.22	1603.3
No_242_1_1	42214	1043	2389	904.21	2404
No_254_1_1	41582	1202	2487	850.38	2465.7
mdamb231_p5	40594	4449	2480	1938.55	5719.7
B61_1_1	41547	1478	2469	567.92	2214.8
No7_3rd_15ug	53027	2	1297	875.83	340.2
No16_3rd_15ug	55847	4	983	1151.68	592
No_27_1_1	50842	530	1526	2235.63	3417.5
No_78_1_1	55829	190	776	1134.03	2099
no_203x_1_2	50574	1	1526	844.81	575
C32_1_1	51371	124	1414	569.65	637.8
Cal6_14rere	51750	32	1433	1441.33	789
Cal6_14__06_10	52238	20	1326	1301.98	662
Cal6_7rere_06_10	51885	1	1346	894.78	392.5

Abbildung 4-2 Die Sense-Antisense-Koexpression ist kein technisches Artefakt.
Links *Pro Experiment wurde gezählt, wie viele der 102 Sense-Antisense-Paare gemeinsam die Detektionsschwelle erreicht haben. Anschließend kann in einem Histogramm darstellen, in wie vielen von den 310 Experimenten die entsprechende Anzahl koexpremierter Sondensets etwas detektieren konnte. Die rote Kurve ist reine Ästhetik. In der Tabelle* ***rechts*** *sind einige Eckdaten der laut der Verteilung extremsten Experimente (gelbe Rahmen im Histogramm links) zusammengefasst. In der oberen Hälfte der Tabelle sind die Werte für die Experimente der rechte gelbe Box aufgeführt und in der unteren Hälfte die für die linke gelbe Box. Die Spalte Q3-Q1 enthält den Interquartilsabstand also die Differenz zwischen 75%- und 25%-Quantil der Rohintensitäten. Die Spalte „detektiert" enthält die Anzahl der Sondensets aus 6117, deren Detektionsscore kleiner 0,05 ist. In „PM<MM" ist die Anzahl der Oligopaare (gesamt etwa 120'000) aufgeführt, in denen das MM-Signal einen höheren Wert als das PM-Signal aufweist. Die Daten zeigen erhebliche Differenzen, was dafür spricht, dass die Unterschiede in der Sense-Antisense-Koexpression allgemeine Qualitätsunterschiede in den Datensätzen widerspiegeln. Orange unterlegt sind jeweils die Werte, die für eine schlechtere RNA-Qualität der Probe hindeuten.*

Das in Abbildung 4-2 dargestellte Ergebnis der Untersuchung zeigt keinen Hinweis darauf, dass es sich hierbei um einfache Verunreinigungen handelt. Erreichen beispielsweise für den gesamten Chip mehr Signale die Detektionsschwelle, so gibt es auch mehr Sense-Antisense-Koexpression. Als nächstes wurde der Einfluss der Voramplifikation, des GC-Gehalts der Oligosets und deren Abstand vom 3'-Ende der mRNA geprüft. Beim Abstand der Oligos vom 3'-Ende ließ sich auch für dieses Sequenzset kein wesentlicher Effekt auf die mittleren Expressionssignale feststellen (Siehe Abbildung 4-1). Zur Untersuchung des GC-Gehalts wurde pro Sequenz der mittlere GC-Gehalt aller Oligos bestimmt und dieser zur Anzahl der Experimente, in denen das zugehörige Sense-Antisense-Paar koexprimiert ist, in Beziehung gesetzt (Abbildung 4-3 links). Man erkennt, dass tendenziell die GC-reichen Sequenzen keine

Antisense-Expression aufweisen. Allerdings ist der Effekt zu schwach bei so wenigen Sequenzen, als dass man eine gesicherte Aussage treffen könnte.

Wie in der Einleitung beschrieben, muss man die RNA aus Gewebeproben voramplifizieren, um ausreichend Material für eine Chiphybridisierung zu erhalten. Ist die beobachtete Koexpression ein Artefakt der RNA-Aufbereitung, so liegt die Vermutung nahe, dass der Effekt, sich durch die Voramplifikation verstärkt. Um diese Hypothese zu testen, gruppiert man am besten alle Experimente, in solche, die mit Gewebeproben und solche, die mit Zelllinien durchgeführt wurden. Die RNA aus Zelllinien bedurfte keiner Voramplifikation, da sie immer in ausreichender Menge verfügbar war. Erwartet man einen verstärkenden Effekt der Voramplifikation, so sollten die Paare tendenziell häufiger in den Gewebeproben gemeinsam ein Signal liefern. Das ließe sich daran erkennen, dass die meisten Punkte in der rechten Abbildung 4-3 unterhalb der Identitätsgeraden liegen. Offensichtlich ist das nicht der Fall, und damit ist es unwahrscheinlich, dass die Voramplifikation für die Detektion von Antisense-Transkripten verantwortlich ist.

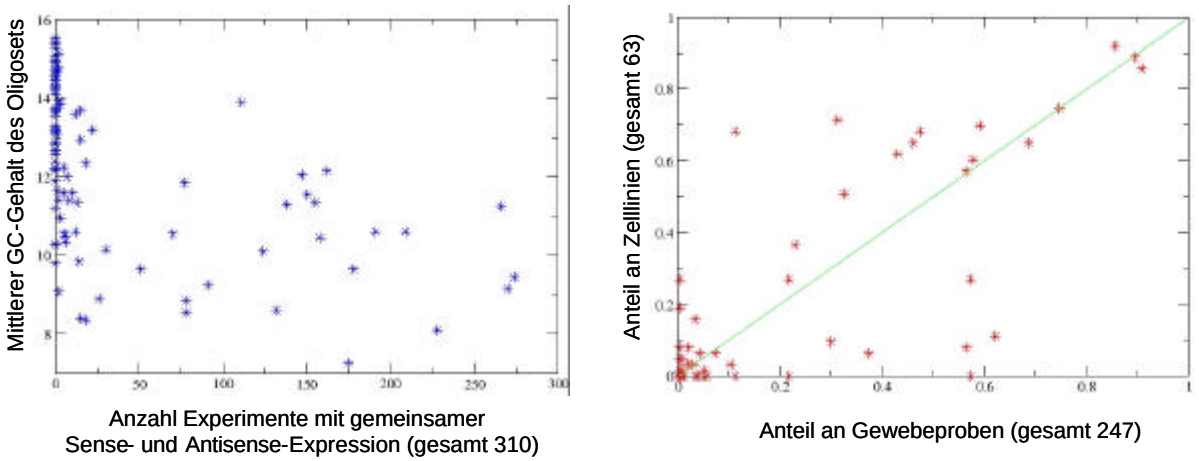


Abbildung 4-3 Potentielle Einflussfaktoren auf die Sense-Antisense Koexpression

*Die **linke** Darstellung zeigt den Zusammenhang zwischen dem mittleren GC-Gehalt der Oligomere in diesen Sondensets und der Anzahl der Experimente mit Sense-Antisense-Koexpression. In dem **rechten** Plot sind die Ergebnisse nach Gewebeproben und Zelllinien getrennt dargestellt. Auf der y-Achse ist für jedes Sondenpaar die Größe des Anteils an allen Zelllinienexperimenten aufgetragen, in denen das Paar koexprimiert ist. Die x-Koordinate entspricht dem Anteil an allen Experimenten mit Gewebeproben. Zeigen beispielsweise für eine Sequenz die beiden Oligosets ein Signal in 21 Zelllinienexperimenten und in 124 Experimente mit Gewebeproben, dann entspräche das gerade dem Punkt x=1/2 und y=1/3 im Koordinatensystem.*

Die anschaulichste Hypothese für die mögliche Entstehung von Artefakten ist das interne Binden des Oligo-dT-T7-Primers. Beispielsweise kann man in EST-Datenbanken erkennen, wenn eine zu hohe Konzentration an Oligo-dT-Primer eingesetzt wurde. Dann binden die Primer an interne imperfekte Adenin-Folgen und initiierten an diesen Stellen die reverse Transkription. Im Ergebnis findet man in den Datenbanken Stapel von EST's, die jeweils an diesen Adenin-Folgen enden. Beim Chipexperiment sind die Primer die gesamte Zeit im Reaktionsmix vorhanden, so dass sie nach dem RNase-H-Verdau an Adenin-Folgen des Erststranges binden können. (Abschnitt 1.5) Die T7-Promotersequenz wird dabei nicht doppelsträngig, aber laut [33] reicht es für die Funktion der T7-Polymerase, wenn er einzelsträngig vorliegt.

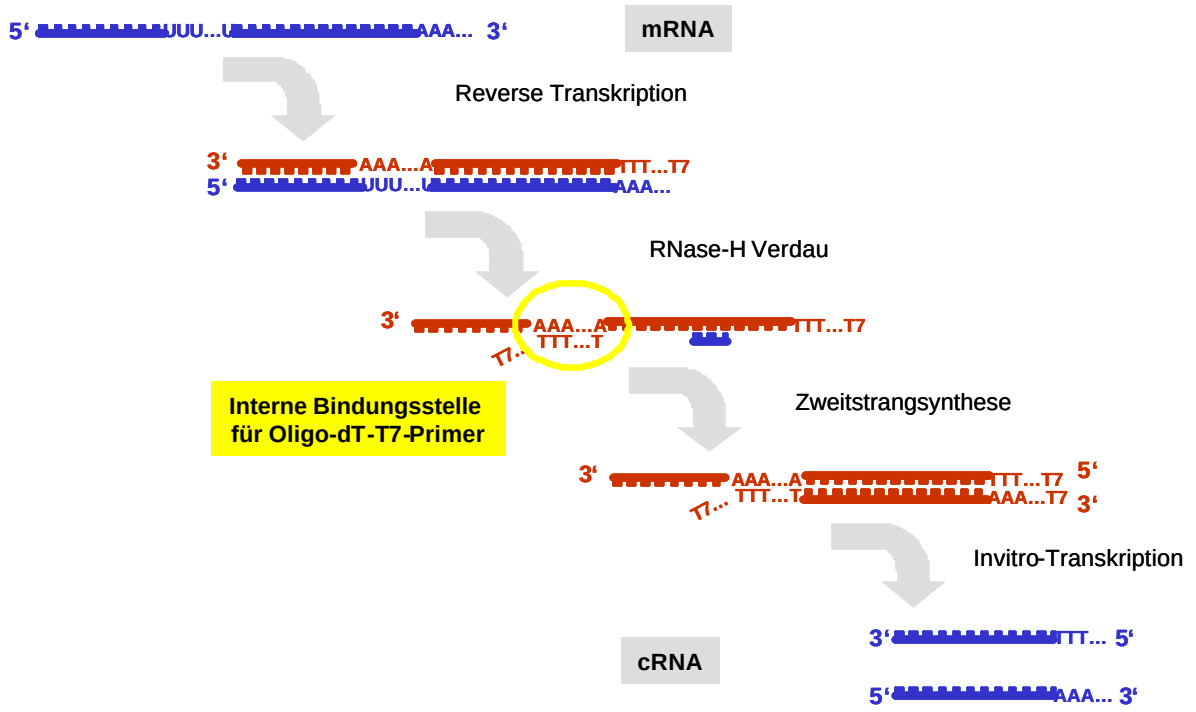


Abbildung 4-4 Hypothese zur Synthese von artifiziellen Antisenseprodukten
Dargestellt ist die RNA-Aufbereitung von der mRNA zur cRNA. Ist in der mRNA eine Urazil-Folge, so besteht die Möglichkeit, das der Oligo-dT-T7-Primer nach dem RNase-H-Verdau an den Erststrang bindet und für die Zweitstrangsynthese als Primer fungiert. Letztlich können cRNA-Kopien von der Sense- und Anitsense-Sequenz entstehen.

Es ist schwierig abzuschätzen, wie lang und exakt die Adeninfolge sein muss, um bei einer bestimmten Primerkonzentration unter unseren spezifischen Reaktionsbedingungen einen Effekt erwarten zu können. Der direkteste Ansatz schien der Vergleich der Primersequenz mit der mRNA-Sequenz und ihrem reversen Komplement zu sein. Zur T7-Promotersequenz konnten keine ähnlichen Sequenzbereiche gefunden werden, so dass es ausreicht, nach Bindungsstellen für das Oligo-dT zu suchen. Nach der Hypothese (Abbildung 4-4) sind deshalb Urazil-Folgen in der mRNA-Sequenz zu suchen, die in Richtung 5’ vom Oligoset liegen. Der Liste der 102 Sense-Antisensepaare zeigt beispielsweise die Referenzsequenz NM_018011 für beide Oligosets ein Signal in 274 der 310 Experimente. Die Oligos liegen auf der Datenbanksequenz zwischen den Basen 1183 und 1620 (in 5’-3’ Orientierung) und an Position 1136 findet man die Urazil-Folge UUUUUUGUUUUUGUUUUUGUUUUUUU. Damit passt diese Sequenz ausgezeichnet zu der Theorie. Zur Bewertung aller 102 ausgewählten Sequenzen

wurde mit dem Programm FASTA[1] Urazil-Folgen gesucht und nach ihrer Lage in Bezug auf die Oligos klassifiziert.

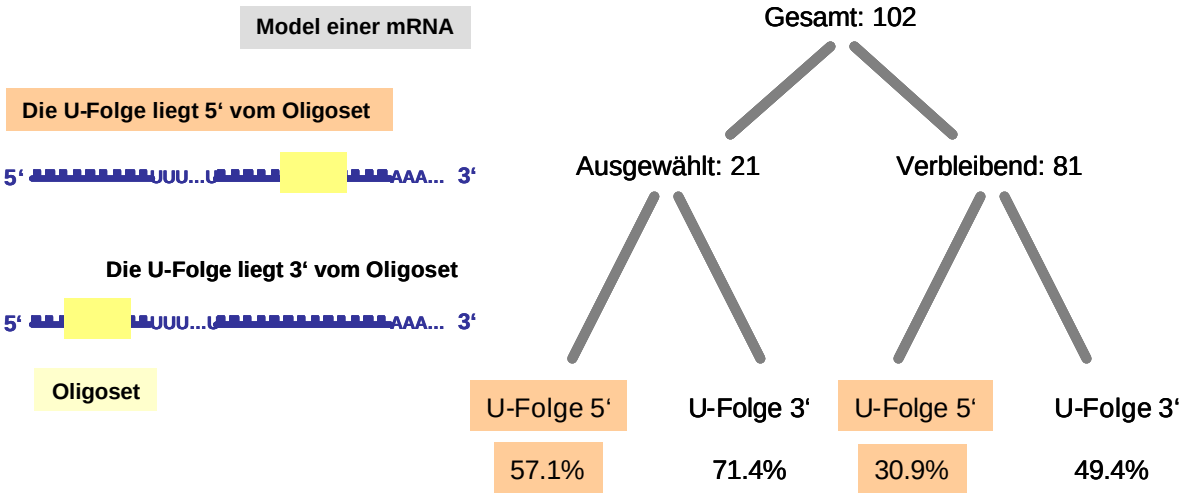


Abbildung 4-5 Schema zur Auswertung der Urazil-Folgen
Links *sind die Lagemöglichkeiten der Urazil-Folgen in der mRNA bezüglich den Oligos des Sondensets veranschaulicht.* ***Rechts*** *ist die Auswertung schematisch dargestellt. Von den 102 wurden 21 ausgewählt, für die in etwa ¼ der Experimente die Sense- und die Antisensesonde ein Signal liefern (Tabelle 4-2). Anschließend wurden in jedem dieser Sequenzsets die Urazil-Folgen und deren Lage erfasst. Angegeben ist der prozentuale Anteil an Sequenzen jedes Sets, die eine solche Folge (FASTA-Score > 25) in dem entsprechenden Sequenzbereich besitzen.*

Die Auswertung zeigt, dass man damit über die Hälfte der Fälle (12 von 21, 57,1%) erklären könnte. Es ist aber eben kein generelles Sequenzmotif aller, die eine Sense-Antisense-Koexpression zeigen. Ob der Unterschied, den man im Vergleich zu den verbleibenden 81 Sequenzen feststellt, signifikant ist, kann aufgrund der vielen weiteren Einflüsse und der geringen Umfänge der Sets nicht befriedigend beantwortet werden.

4.3. Literatur über Antisense-Transkripte

Im Gegensatz zu den Prokaryonten sind in Eukaryonten erst relativ wenige überlappende Sense-Antisense-Transkriptpaare in der Literatur beschrieben. Darunter befinden sich die für

---

[1] FASTA ist ein ursprünglich von Pearson und Lipman entwickeltes Computerprogram zum Vergleich von Sequenzen. Es eignet sich besonders für lokale Alignments und damit für kurze Sequenzen. Das Program wird am EBI gepflegt: www.ebi.ac.uk/fasta33.

die Tumorentwicklung wichtigen Gene *c-myc*, *p53* und *bcl-2* [34]. Das Transkriptpaar *Igf2r/Air* wird im Folgenden exemplarisch etwas genauer erläutert, da kürzlich in einem Experiment in der Maus deutliche Hinweise auf den funktionellen Mechanismus gewonnen werden konnten [35]. *Air* ist ein 108 kb langes Gen, das spezifisch nur vom paternalen Allel transkribiert und nicht gespleißt wird. Es liegt in einer Cluster von Genen, die durch genomisches Imprinting[1] allelspezifisch transkribiert werden. Man findet immer wieder nicht-kodierende Transkripte in derartigen Imprinting Clustern, und es gibt deutliche Hinweise darauf, dass sie an der Repression benachbarter Gene beteiligt sind [36]. Der Promoter des vom paternalen Allel exprimierten *Air*-Gens liegt in einer differenziell methylierten Region innerhalb von Intron 2 des *Igf2r*-Gens. Diese genomische Region enthält ein Methylierungsmuster, das sie von der Oozyte erbt und durch die Embryogenese erhält. Die Deletion der gesamten Region hebt das Imprinting auf, und die drei normalerweise maternalen Gene *Igf2r, Slc22a2* und *Slc22a3* können exprimiert werden. Sleutels und andere konnte nun zeigen, dass nicht die Deletion des genomischen Bereichs per se den Effekt verursacht sondern die fehlende Expression der *Air*-RNA. Sie fügten eine Polyadenylierungssignal 4,2 kb stromab vom Transkripttionsstart des *Air*-Gens ein, so dass die Zellen nur eine verkürzte RNA bilden, die nicht mehr mit *Igf2r* überlappt. Im Ergebnis ist das Methylierungsmuster nicht beeinträchtigt und die Expression der verkürzten Variante von Air ist die gleiche wie die des normalen Gens. Die Repression der drei Gene also der Effekt des Imprinting ist vollständig aufgehoben. Wohl gemerkt überlappt *Air* nur mit *Igf2r* und nicht mit den anderen beiden Genen. Der genaue Mechanismus ist damit noch nicht aufgeklärt, aber man hat zum ersten Mal zeigen können, dass ein Antisensetranskript direkt in die Genregulation involviert ist.

Da die meisten Methoden zur Untersuchung der Genexpression nicht spezifisch für den Einzelstrang sind, ist die Unterscheidung, ob ein Signal von einem Sense- oder Antisensetranskript stammt, im Allgemeinen nicht möglich. Das könnte der Grund dafür sein, dass man sehr wenige Sense-Antisense-Transkriptpaare entdeckt hat, falls das Phänomen

---

[1] Als Imprinting bezeichnet man einen Prozess, der bewirkt, dass bestimmte Gene nur von dem von der Mutter ererbten (maternalen) Allel und andere nur von dem vom Vater ererbten (paternalen) Allel transkribiert werden.

tatsächlich viel häufiger ist als bisher angenommen. Kürzlich sind zwei Publikationen erschienen, in denen systematisch nach überlappenden Transkripten in Sequenzdatenbanken gesucht wird. In der ersten Arbeit [37] benutzten die Autoren die mRNA-Sequenzen von *RefSeq* und ein kuriertes mRNA-Set vom EMBL und suchten mit Hilfe des BLAST-Programms nach Sense-Antisense-Paaren (NAT[1]). Sie fanden 174 Paare und sechs Sets von Sequenzen die mehr als ein Pendant haben. Dabei wurde unterschieden, zwischen Transkripten, die vom gleichen Genort stammen (cis-NAT), und solchen, die von nicht-überlappenden Genen abgeschrieben werden (trans-NAT). Von der einen Sorte identifizierte man 87 und von der anderen 80. Der Überlappungsbereich variierte zwischen 32 und 1606 bp Länge. In der Arbeit wird auf die technischen Probleme hingewiesen, die sich für cDNA-Array-Experimente ergeben, sollte sich bestätigen, dass überlappende Transkripte relativ häufig sind. Zusatzinformation zu den Ergebnissen findet man auf der Internetseite der Autoren: *http://www.hgmp.mrc.ac.uk/Research/Antisense*. In der zweiten Arbeit von Jay Shendure [38] wurde UniGene (siehe Einleitung) als Basis genommen. Im Vergleich zur Arbeit von Lehner et al hat man hier eine wesentlich größere Sequenzmenge als Datenbasis. Allerdings sind die Daten auch weniger gut kuriert, was mehr Vorsicht bei der Analyse verlangt. Die Autoren selektierten als erstes cDNA-Bibliotheken die gerichtet kloniert wurden und von guter Qualität sind. Anschließend suchten sie in UniGene nach Sequenzclustern, in denen eine im Vergleich zu allen signifikante Zahl von EST's in entgegengesetzter Orientierung vorkommt. Um sich rigoros vor genomischen Verunreinigungen zu schützen, mappten die Autoren die EST's zurück auf die genomische Sequenz und evaluierten, ob die Sense-Antisense-Transkriptpaare unterschiedliche Exon-Intron-Strukturen und Poly-A-Schwänze aufwiesen. Auf diese Weise konnten 144 humane und 73 murine solcher UniGene-Cluster entdeckt werden. In obiger Terminologie betrachtete man hier allerdings nur cis-NAT's. Schließlich konnten noch mit Hilfe einer einzelstrangspezifischen RT-PCR 33 von 39 ausgewählten Beispielen experimentell verifiziert werden. Auch hier sind auf einer Internetseite weitere Informationen veröffentlicht: http://arep.med.harvard.edu/antisense.html.

---

[1] Die Autoren bezeichnen Transkripte, die in revers-komplementärer Orientierung mit anderen identische Sequenzbereiche gemeinsam haben, als natürliche Antisensetranskripte, abgekürzt NAT's.

Insgesamt haben die in der Literatur bereits beschriebenen Fälle und die Ergebnisse der vorgestellten Analysen sehr wenig Überlapp. Lehner findet nur fünf in der Literatur bekannte Paare wieder und Shendure findet zehn aus der Literatur und dem Set von Lehner zusammen.

4.4. Laborversuche zur Prüfung der Expression putativer Antisense-Transkripte

Da sich also unsere Beobachtungen durch die beschriebenen Tests nicht als technische Artefakte entlarven ließen und man außerdem in der Literatur immer und mehr evaluierte Beispiele findet, führte der nächste logische Schritt ins Labor. Es sollte für einige ausgewählte Beispiele, die Existenz der Antisense-Transkripte in natura nachgewiesen werden. Der erste wichtige Schritt ist der Test, ob sich die Antisense-Transkripte auch in nicht-amplifizierter RNA nachweisen lassen. Hierzu sollte die Hybridisierung von Northern Blots[1] mit Oligos dienen. Da es sich dabei um eine etablierte Methode handelt, konnten gebrauchsfertige Membranen mit RNA aus unterschiedlichen humanen Geweben kommerziell bezogen werden. Nebenbei bemerkt, verbieten die hohen Auftragungsmenge von 2µg Poly-A-RNA oder 20µg Gesamt-RNA und der damit verbundene Aufwand die Verwendung mikrodissezierter Zellen. Für die meisten Untersuchungen benutzt man doppelsträngige DNA-Sonden, so zum Beispiel PCR-Produkte einer bestimmten cDNA. Gibt es in einem mRNA-Pool zu einem bestimmten genomischen Lokus nur Transkripte derselben Orientierung, so sollte der Einsatz von einzelsträngigen und doppelsträngigen Sonden zum gleichen Ergebnis führen. Bei der hier behandelten Fragestellung ist es jedoch entscheidend, dass die Sonden spezifisch für den Einzelstrang sind. Dazu wurden vier Sense-Antisense-Paare aus der Liste der 102 ausgewählt. Das Hauptkriterium für die Auswahl war ein möglichst hohes Signal der Sondensets auf dem Chip (mittlerer PMQ-Wert), vor allem für das Antisensetranskript. Außerdem sollte im 5'-Bereich der Oligos keine lange Urazil-Folge vorliegen (siehe oben). Die für die Weiterbearbeitung ausgewählten Gene lauten *CGI-115*, *ZNF217*, *Ponsin* und *DC13* und sind in Tabelle 4-2 hellblau markiert. Zufälligerweise war unter anderem bei metaGen gerade entdeckt worden, dass die Gene *OPN3* und *KMO* entgegengesetzt orientiert sind und am 3'-Ende überlappen. ([39], [40]) Ihr

[1] Für eine ausführliche Beschreibung des Verfahrens siehe 0.

Überlappungsbereich sollte deshalb bei den Laborversuchen als Positivkontrollen dienen. Leider war dieses Gen nicht durch ein Sense-Antisense-Sondenpaar auf dem Chip repräsentiert.

Für die ersten vier Gene wurden in jeder Orientierung bis zu neun nicht überlappende Oligomere aus den Chipsondensets herausgesucht und bei metabion (Martinsried, BRD) bestellt. Für *OPN3* und *KMO* wurden jeweils acht Oligos mittels des Primer-Design-Programms von DNA*Star (Madison, USA) bestimmt. Sämtliche verwendete Oligomere sind im Anhang C gelistet. Für jedes potenzielle Transkript konnte den in den Abschnitten 0 und 0 dargestellten Protokollen folgend der Oligomix radioaktiv markiert und auf die Blots hybridisiert werden. Das Gen *CGI-115* zeigte praktisch keine Expression auf dem Northern Blot und wurde deshalb nicht weiter bearbeitet. Es folgt die Beschreibung der anderen Gene und die Ergebnisse der Experimente.

Das Genpaar *OPN3-KMO* schien sich hervorragend als Positivkontrolle zu eignen, da hier bereits Analysen mit Northern Blots durchgeführt worden sind. Abbildung 4-6 zeigt eine Skizze der bisher aufgeklärten Genstruktur [39].

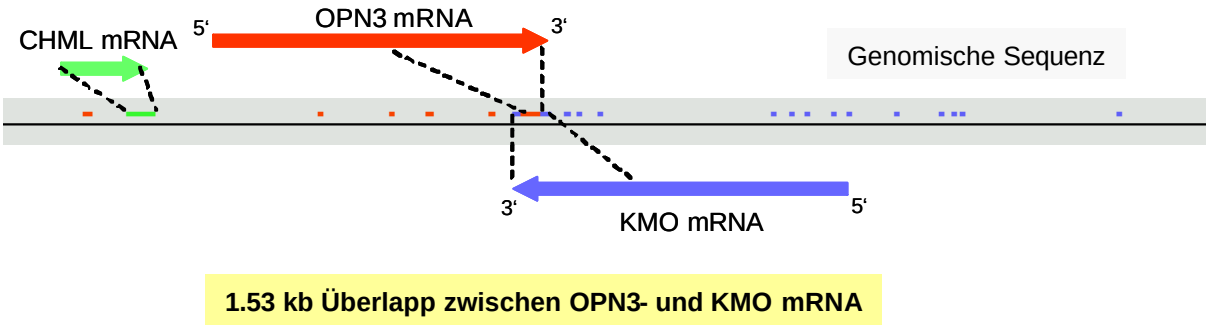


Abbildung 4-6 Überlappende Transkripte als Positivkontrolle

Die bunten Pünktchen symbolisieren die Exons, eine Farbe pro mRNA. CHML ist ein Gen, das nur aus einem Exon besteht und innerhalb des ersten Introns von OPN3 liegt. Für KMO wie auch für OPN3 existieren mehrere Spleißvarianten, wobei laut der Vorhersage nur die lange(etwa 5 kb) mit OPN3 überlappt.

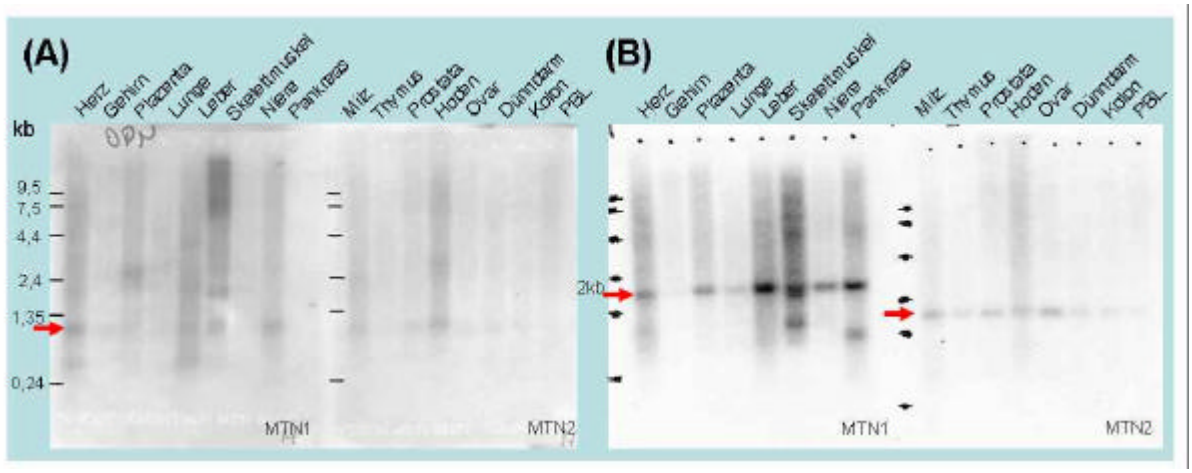


Abbildung 4-7 Northern Blot OPN3-KMO

Dargestellt ist das Ergebnis der Northern Blot Hybridisierung. Die Längenmarker an allen Blots sind gleich und mit schwarzen Strichen am Rand markiert. PBL steht für periphere Blutlymphozyten. ***(A)*** *Der OPN3-Blot zeigt nur ein schwaches Signal nach drei Tagen Exposition auf einem Röntgenfilm. Die stärksten Banden sollten eigentlich bei 2,1 und 2,5 kb liegen und zwar besonders in Plazenta und Leber. Bei etwa 1,2 kb wird eine schwache Bande detektiert, die allerdings mit keinem beschriebenen Transkript korrespondiert.* ***(B)*** *Der KMO-Blot zeigt nach Exposition übernacht eine eindeutige Bande bei etwa 2 kb. Allerdings sollte laut Publikation die Bande bei etwa 5 kb liegen.*

Das Ergebnis der Oligohybridisierung widerspricht den publizierten Daten. Für *OPN3* war insgesamt nur ein sehr schwaches Signal nach langer Exposition zu detektieren. Der Northern Blot für *KMO* zeigt eine klare Bande und sogar auf der Höhe von etwa 2 kb, was laut Publikation dem häufigsten Transkript entspricht. Was allerdings klar dagegenspricht, dass hier *KMO* detektiert wird, ist die Tatsache, dass laut Genmodell nur die lange Variante mit *OPN3* überlappt und die Oligos aus genau diesem Bereich gewählt wurden. In der Arbeit von Grit Kasper et al ist gezeigt, dass die Masse an EST's deutlich zwei Cluster bilden und damit das dargestellte Genmodell unterstützen und den Ergebnissen der Oligohybridisierung widersprechen.

Das zweite untersuchte Gen heißt ***ZNF217***. Es wurde über positionale Klonierung einer im Mammakarzinom amplifizierten genomischen Region auf Chromosom 20q13.2 entdeckt [41]. Aufgrund der Sequenzähnlichkeit konnten die Autoren das Gen den Transkriptionsfaktoren der Krüppel-Familie zuordnen, die sich durch spezifische Sequenzmotive auszeichnet. Und zwar besitzt das Gen eine DNA-bindende Domäne und acht Zinkfingermotive. Weiterhin gibt es Hinweise, dass die genomische Amplifikation eine erhöhte Genexpression nach sich zieht und dass es sich hierbei um ein Onkogen handeln könnte [42]. Interessant ist auch in diesem

Zusammenhang, dass für einen anderen Transkriptionsfaktor desselben Typs *WT1* bereits ein Antisense-Transkript *WIT1* nachgewiesen werden konnte [34].

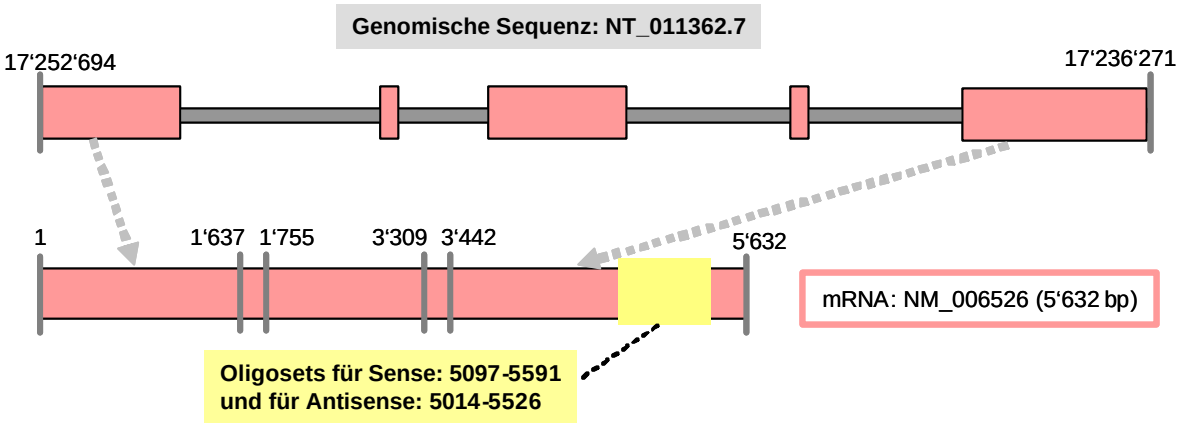


Abbildung 4-8 Genmodel von ZNF217

Die Referenzsequenz von ZNF217(NM_006526) überspannt einen genomischen Bereich von 16'023 bp und gliedert sich in fünf Exons. Die beiden äußeren Zahlen geben die Positionen des Gens auf der genomischen Referenzsequenz an. Die Lage der Oligos der Chipsonden ist durch die gelben Streifen angedeutet. Die Orientierung der Datenbanksequenz ist von links 5' nach rechts 3'. In einer beschriebenen Variante fehlt das vierte Exon von links [41].

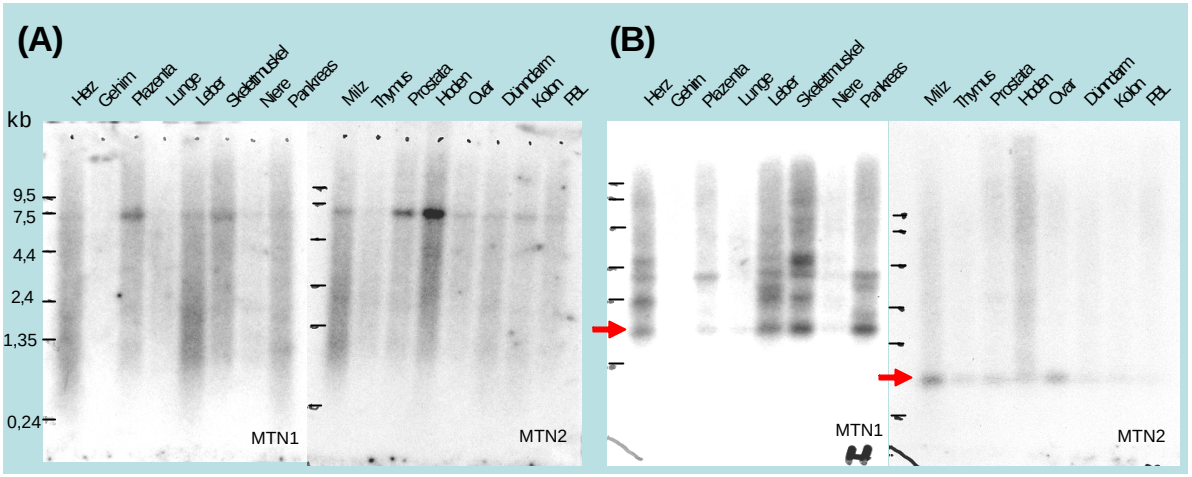


Abbildung 4-9 Northern Blot ZNF217

Dargestellt ist das Ergebnis der Northern Blot Hybridisierung.Die Längenmarker an allen Blots sind gleich und mit schwarzen Strichen am Rand markiert. PBL steht für periphere Blutlymphozyten. ***(A)*** *Das detektierte Transkript scheint das korrekte ZNF217 voller Länge zu sein. Die Bande liegt etwa bei den angegebenen 6 kb und die Gewebeverteilung entspricht auch soweit dem publizierten Ergebnis: insgesamt in vielen Geweben, am stärksten im Hoden und nicht exprimiert im Gehirn.* ***(B)*** *Der Blot für das Antisense-Transkript zeigt eine recht klare Bande bei etwa 500 Basen Länge (roter Pfeil). Die vielen Nebenbanden auf dem MTN1-Blot könnten von ungenügendem Waschen der Blots nach der Hybridisierung herrühren.*

Für *ZNF217* zeigen die Hybridisierungsexperimente hoffnungsvolle Ergebnisse. Das Sense-Transkript, für das ein Northern Blot bereits veröffentlicht ist, konnte mit den Oligos bestätigt werden [41]. Und auch die Oligos für das Antisense-Transkript zeigen ein Signal. In der Publikation ist außerdem ein 4 kb großes Transkript spezifisch im Hoden detektiert worden. Dass das bei der Oligohybridisierung nicht detektiert wird, ließe sich damit erklären, dass die kurze Variante auch eine verkürztes 3'-Ende hat und damit nicht mehr den Sondenbereich überdeckt. Für dieses Gen wurden der Genort und alle EST-Sequenzen, die von einem Antisense-Transkript stammen könnten, sorgfältig analysiert. Man findet gespleißte EST's[1] auf dem Gegenstrang, die sehr wahrscheinlich ein Transkript darstellen. Allerdings gibt es keine Evidenz, dass die Sequenzen der beiden Transkripte überlappen. Nummeriert man die Exons des Gens *ZNF217* von 1 bis 5 (5' –> 3') so ergibt sich folgendes Bild: Zwei der durch die gespleißten EST's vorhergesagten Exons liegen zwischen Exon 3 und 4 von *ZNF217* und

[1] Beispiele für gespleißte EST's, auf dem Gegenstrang des Gens ZNF217: AU118778, BE089097 (*GenBank Accession*)

eins liegt zwischen Exon 4 und 5. Bei einem weiteren gerichtet klonierten EST, dass in Antisense-Orientierung in der Datenbank annotiert ist und mit *ZNF217* überlappt, handelt es sich vermutlich um ein Klonierungsartefakt. Vergleichbar der Hypothese für Artefakte bei Chipexperimenten, könnte auch hier das interne Binden von Oligo-dT-Primern in der Zweitstrangsynthese zur fehlerhaften Detektion eines Antisense-Transkripts führen. Diese Vermutung liegt nahe, da das 5'-Ende dieses putativen Transkripts exakt mit dem 3'-Ende von *ZNF217* abschließt und direkt davor (3') auf dem *ZNF217*-Sequenz eine Urazilfolge zu finden ist.

Als nächstes wird das Gen ***Ponsin***[1] diskutiert. In der Publikation [43] ist ein Northern Blot für *Ponsin* in der Maus abgebildet, der auf viele gewebsspezifische Spleißvarianten hindeutet. Die stärkste Expression findet man demnach im Herzen und das Transkript hat eine Länge von etwa 7,5 kb. Die Autoren fanden eine Bindung des Proteins an *l-Afadin* und *Vinculin*. Aufgrund dessen und aufgrund der Lokalisation des Proteins im Gewebe legen sie eine regulatorische Funktion bei der Zell-Zell- und bei der Zell-Matrix-Adherens nahe. In einer weiteren Arbeit konnte gezeigt werden, dass Ponsin in die intrazellulälre Signalübertragung von Insulin involviert ist [44]. Laut Datenbankeintrag in *RefSeq* ist das Gen auf Chromosom 10q23.3-q24.1 lokalisiert und die Proteinsequenz enthält eine SORB-Domäne (pfam02208[2]) und drei SH3-Domänen (pfam00018[3]).

---

[1] Synonyme Namen für Ponsin sind SORBS1, SORB1, CAP (CBL assoziiertes Protein), SH3P12, SH3P5.

[2] Für weitere Informationen zu dieser Proteindomäne: http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF02208

[3] Für weitere Informationen zu dieser Proteindomäne: http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00018

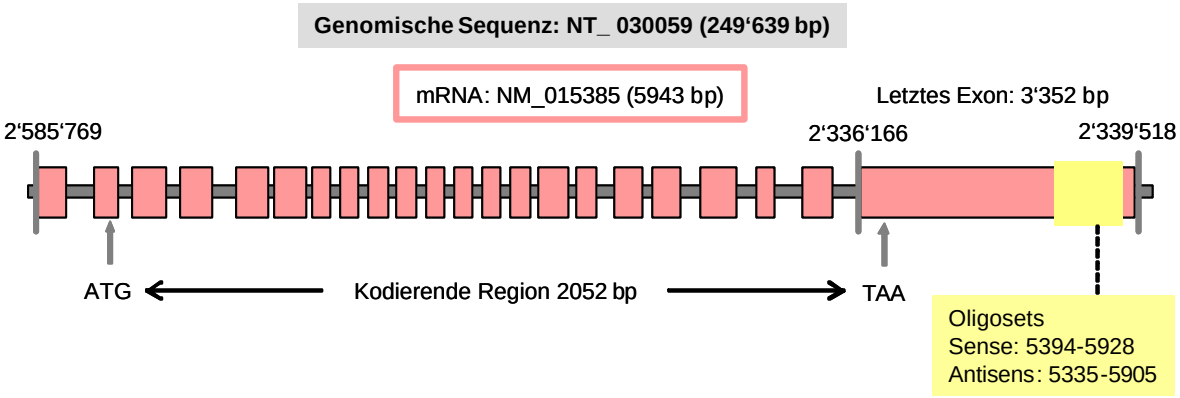


Abbildung 4-10 Genmodel einer Spleißvariante von Ponsin/SORBS1

Das Gen überspannt einen genomischen Bereich von etwa 250 kb. Die Zahlen über den grauen Markierungslinien geben die Positionen des Gens auf der genomischen Referenzsequenz (NT_030059) an. Die Referenzsequenz von Ponsin (NM_015385) gliedert sich in mindestens 22 Exons, wobei das letzte über 3 kb groß ist. Dieses Gen wird in unterschiedlichen Spleißformen transkribiert. Eine Variante ist in der Graphik mit dem proteinkodierenden Bereich eingetragen. ATG und TAA deuten die Positionen des Start- und des Stopkodons an. Die Lage der Oligos der Chipsonden ist durch die gelben Streifen angedeutet. Die Orientierung der Datenbanksequenz ist von links 5’ nach rechts 3’.

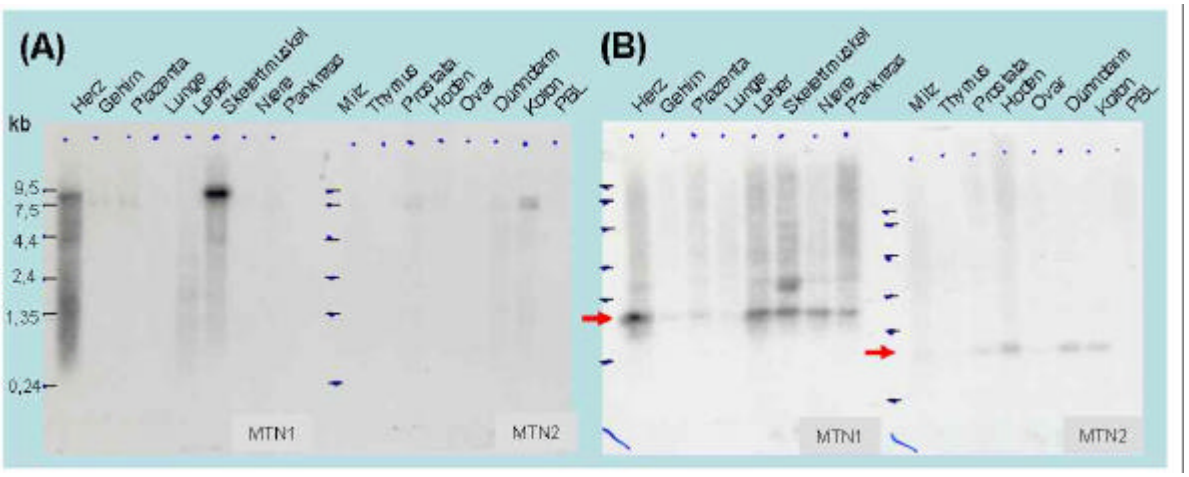


Abbildung 4-11 Northern Blot von Ponsin/SORBS1

Dargestellt ist das Ergebnis der Northern Blot Hybridisierung.Die Längenmarker an allen Blots sind gleich und mit schwarzen Strichen am Rand markiert. PBL steht für periphere Blutlymphozyten. ***(A)*** *Die Sonde detektiert Ponsin. Das Transkript hat eine Länge von über 7,5 kb, das am stärksten in der Skelettmuskulatu, im Herzen und im Kolon exprimiert ist. Auf dem Blot MTN2 läßt sich eine Doppelbande erkennen.* ***(B)*** *Der Blot für das Antisense-Transkript zeigt ebenfalls eine recht klare Bande etwas weniger als 1 kb Länge (roter Pfeil).*

Auch für dieses Gen sind die Ergebnisse der Hybridisierungsexperimente recht vielversprechend. Das Expressionsmuster von *Ponsin* ähnelt dem für die Maus publizierten.

Außerdem wurde im Rahmen der Doktorarbeit von Christoph Wissmann bei metaGen ein Northern Blot für das humane *Ponsin* mit einer doppelsträngigen Sonde angefertigt. Die ganze Vielfalt der Transkriptvarianten kann man natürlich durch den eingeschränkten Bereich, aus dem die Oligos gewählt wurden, auf diese Weise nicht untersuchen. Es ist allerdings bemerkenswert, dass das auf dem Blot detektierte Transkript um mindestens 1,5 kb länger ist als die in der Datenbank (*RefSeq*) angegebene Variante mit etwa 6 kb Länge. Die Antisensesonde detektiert ein kurzes Transkript in vielen Geweben und verstärkt im Herzen, in der Leber im Skelettmuskel, in der Niere und in der Pankreas. Für den Blot MTN2 zeigt die Sonde die stärkste Expression im Hoden, im Dünndarm und im Kolon. Auf EST-Ebene gibt es keinen Hinweis auf Antisense-Transkripte.

Das letzte untersuchte Gen ***DC13*** ist in der Literatur nicht beschrieben, zeigt aber mit die stärkste Expression der Antisensesonde auf dem Chip. Laut Datenbankeintrag liegt es auf Chromosom 16q23.3, und die mRNA ist vollständig aus dendritischen Zellen isoliert und sequenziert worden.

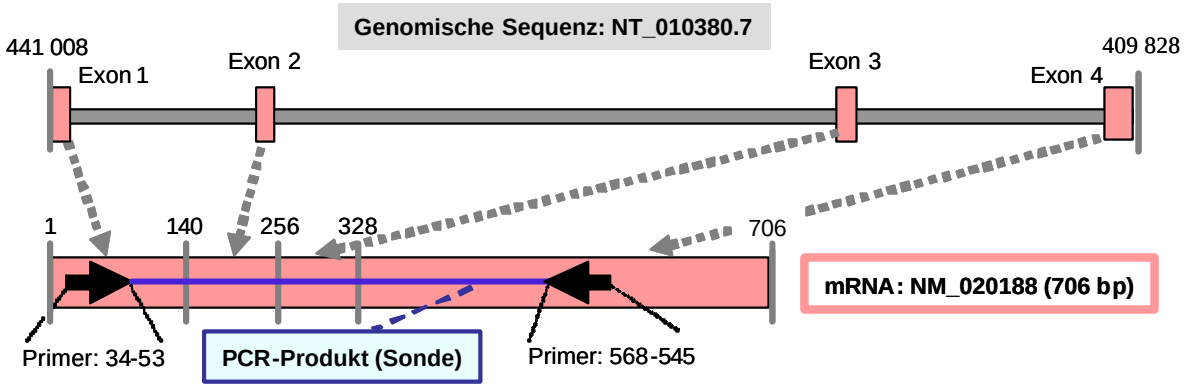


Abbildung 4-12 Genmodel von DC13

Die Referenzsequenz von DC13(NM_020188) überspannt einen genomischen Bereich von 31181 bp und gliedert sich in vier Exons. Die Zahlen an den grauen Markierungslinien oben geben die Lage des Gens auf der genomischen Referenzsequenz an. Unten markieren sie die Position der Exongrenzen auf der mRNA-Sequenz. Die dunkelblaue Linie verdeutlicht die Lage des PCR-Produkts, das als Sonde für die Insitu-Hybridisierung diente. Die Oligos der Chipsonde liegen für beide Orientierungen über die gesamte Sequenz verteilt und sind deshalb nicht extra eingetragen. Die Orientierung der Datenbanksequenz ist von links 5' nach rechts 3'.

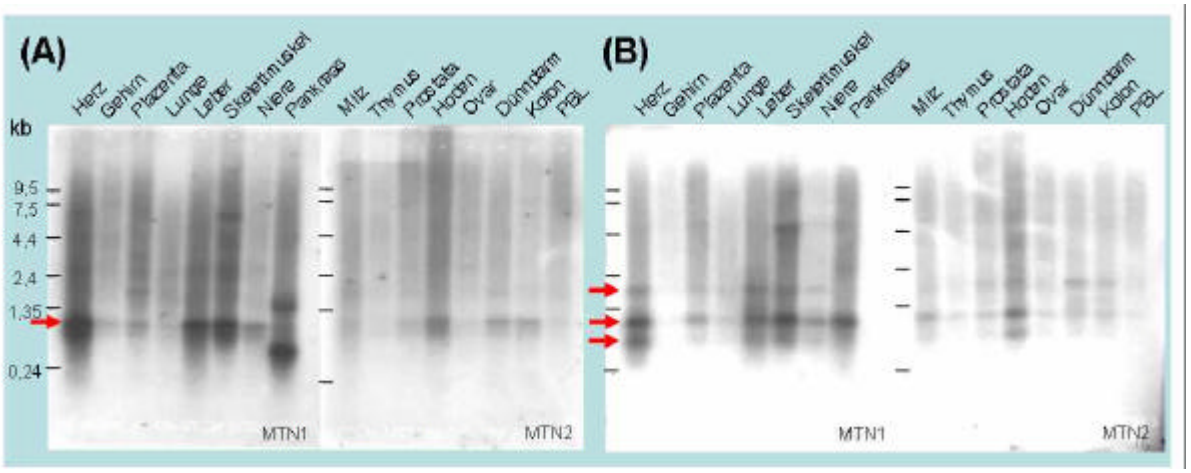


Abbildung 4-13 Northern Blot DC13

Dargestellt ist das Ergebnis der Northern Blot Hybridisierung. Die Längenmarker an allen Blots sind gleich und mit schwarzen Strichen am Rand markiert. PBL steht für periphere Blutlymphozyten. ***(A)*** *Die Sonde detektiert eine Doppelbande mit dem stärksten Signal bei circa 1 kb (roter Pfeil). Das Transkript ist am stärksten exprimiert im Herzen, in der Leber, in der Skelettmuskulatur und auf MTN2 im Hoden. In der Pankreas ist ein völlig anderes wahrscheinlich artifizielles Bandenmuster zu sehen.* ***(B)*** *Der Blot für das Antisense-Transkript zeigt drei recht klare Bande und die stärkste liegt bei etwa 1 kb Länge (rote Pfeile). Die Verteilung der Gewebsexpression ähnelt der in Abbildung (A).*

Für dieses Gen ist noch keine Expressionsstudie veröffentlicht. Der Blot der Sonde für das Sense-Transkript zeigt eine deutliche Bande bei etwa 1 kb und wenig darunter eine zweite. *DC13* erscheint ubiquitär exprimiert mit den stärksten Signalen im Herzen, in der Leber, in der Skelettmuskulatur und im Hoden. Sehr schwach, wenn überhaupt, ist es in der Lunge, Thymus, im Ovar und in den peripheren Blutlymphozyten exprimiert. Das abweichende Bandenmuster im Pankreas scheint ein technisches Artefakt zu sein, da es sich bei einer Wiederholung des Experiments auf einem anderen Blot (hier nicht gezeigt) nicht reproduzieren ließ. Die Sonde für das Antisense-Transkript zeigt drei deutliche Banden, wobei die stärkste etwa auf der gleichen Höhe liegt, wie die des Sense-Transkripts. Das längere Transkript ist etwa 2 kb und die kürzeren etwa 500 Basen lang. Die Gewebsverteilung der detektierten Anitsense-Transkripte entspricht ziemlich gut der der Sense-Transkripte mit der stärksten Expression im Herzen, in der Leber, im Skelettmuskel und im Hoden.

Auch für dieses Gen wurde eine detaillierte Analyse der verfügbaren EST-Sequenzen und der genomischen Region vorgenommen. Man findet eine große Zahl gespleißter EST-Sequenzen, größtenteils *DC13* bestätigen. Die Referenzsequenz weist eine Länge von 706 Basen auf. Die Sequenzrohdaten und eine Assemblierung der EST's sprechen dafür, dass das Haupttranskript im 3'-Bereich nicht vollständig in der Referenzsequenz enthalten ist. Es ergibt sich eine Konsensussequenz für *DC13* von 936 Basen Länge, was ziemlich exakt mit der beobachteten Bande im Northern-Blot übereinstimmt. Darüber hinaus findet man Evidenz für eine Reihe alternativ gespleißter Varianten. Laut Datenbankannotation (*RefSeq*) liegt das Gen *BM039*[1] auf dem Gegenstrang direkt neben *DC13*. Eine detaillierte Analyse der EST's zeigt jedoch, dass das 5'-Exon des *DC13*-Haupttranskripts vollständig von dem längeren *BM039*-5'-Exon überlappt wird. Eine putative Spleißvariante von *DC13* überlappt nicht mit dem 5'-Exon von *BM039* aber mit 6. Exon gezählt vom 5'-Ende. Die anderen Exons von *DC13* zeigen in der EST-Datenbank keine Hinweise auf Antisense-Transkripte. Zusammenfassend kann man

---

[1] *BM039 – RefSeq*: NM_018455 (vorhergesagt). Zur Lage der Sequenzen *BM039* und *DC13* eignet sich der Genome Browser Kalifornischen Universität in Santa Cruz, USA: „http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=13991209&hgt.right1=+%3E+&position=chr16%3A71911199-72003478"

konstatieren, dass die EST-Daten nicht ausreichend Information liefern, um die Sense-Antisense-Koexpression im Chipexperiment und auf dem Northern-Blot zu erklären. Der Überlapp mit dem Gen *BM039* bietet zwar einen Erklärungsansatz, aber es liegen nur zwei der zwanzig Oligosonden im gemeinsamen Bereich der beiden Gene.

Um die Genexpression von *DC13* weiter zu untersuchen, konnte eine RNA In Situ Hybridisierung (ISH) durchgeführt werden. Das Verfahren ist bei metaGen etabliert und wurde von Nicole Creutzburg im Labor von Dr. Edgar Dahl durchgeführt. Von einem doppelsträngigen DNA-Template stellt man einzelsträngige Fluoreszens-markierte Sonden her. Es entstehen eine Antisense-Sonde, die spezifisch für das Sense-Transkript sein sollte und eine Sense-Sonde, mit der potenzielle Antisense-Transkripte detektiert werden können. Dann hybridisiert man diese separat auf Gewebeschnitte, die auf Objektträgern fixiert sind. Ist das Experiment erfolgreich, so kann man anschließend die Verteilung der Transkripte in verschiedenen Zelltypen unter dem Mikroskop feststellen. Was das Experiment aus methodischer Sicht interessant macht, ist die Verwendung einer Sense-Sonde als Negativkontrolle im Standardverfahren. Dass heißt, normalerweise generiert man eine Sense- und eine Antisense-Sonde und hybridisiert beide auf ähnliche Schnitte. Findet man nun ein Signal der Sense-Sonde, das ja für die Detektion eines potenziellen Antisense-Transkripts steht, verwirft man das Ergebnis des Experiments. Die Grundannahme dabei ist, dass es kein Transkript gibt, welches die Sense-Sonde spezifisch bindet, also revers-komplementär zu ihr ist. Geben dann Sense- und Antisense-Sonde ein Signal, so geht man von einer unspezifischen Bindung beider Sonden aus. Für das hier untersuchte Gen *DC13* gab es also keine Negativkontrolle, da Sense- und Antisense-Sonde ein Signal liefern sollten. Die praktische Durchführung ist in Abschnitt genau beschrieben.

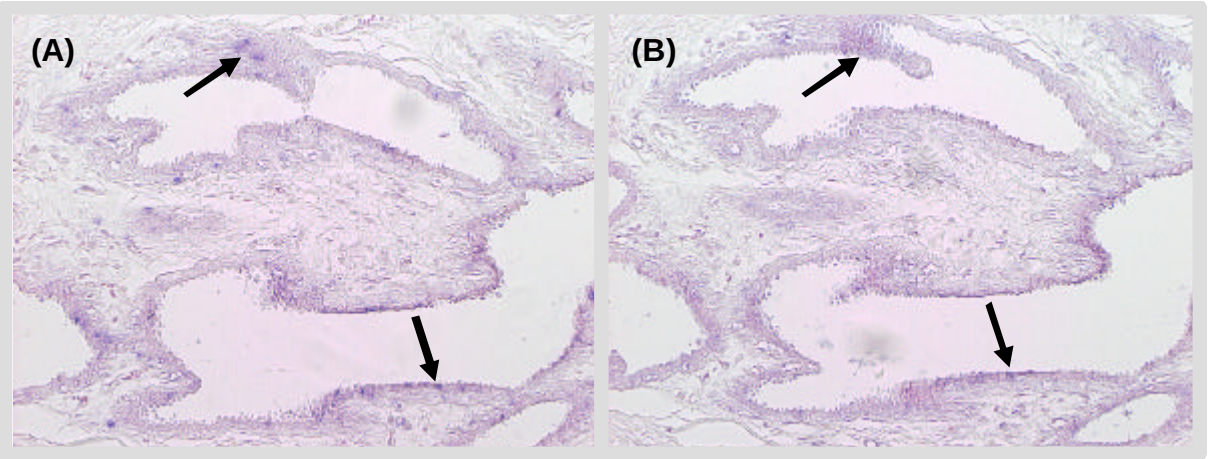


Abbildung 4-14 RNA In Situ Hybridisierung (ISH) für DC13
Dargestellt ist das Ergebnis der ISH für DC13. Die Schnitte stammen von einer Normalgewebeprobe einer Brustkrebspatientin. Das Gewebe umschließt Hohlräume der Drüsen und ist am äußeren Rand durch Epithelzellen begrenzt. ***(A)*** *Das Ergebnis der Hybridisierung der Antisense-Sonde, die die Referenzsequenz detektiert. Man erkennt ein schwaches Signal in einigen Epithelzellen (schwarzer Pfeil).* ***(B)*** *Das Ergebnis der Hybridisierung der Sensesonde, die ein potentielles Antisense-Transkript detektiert. Auch hier erkennt man ein sehr schwaches Signal in einigen Epithelzellen (schwarze Pfeile).*

Das vollständige Ergebnis der RNA In Situ Hybridisierung ist im Anhang C zu finden. Es wurden Gewebeschnitte von vier Brustkrebspatientinnen analysiert. Bei einer Probe zeigten Sense- und Antisense-Sonde ein starkes Signal, bei einer zweiten ein schwaches (Abbildung 4-14). In den anderen beiden Proben konnte nur eine schwache Expression der Referenzsequenz in einigen Tumorzellen aber keine Expression eines Antisense-Transkripts nachgewiesen werden. Im Anhang sind auch HE-gefärbte Schnitte beigefügt, auf denen die Gewebsstrukturen klarer hervortreten. Zusammenfassend lässt sich sagen, dass man in zwei der vier untersuchten Gewebe deutliche Expressionssignale der Sense-Sonde erhält. Das liefert einen zusätzlichen Beleg für die Existenz eines zu *DC13* revers-komplementären Transkripts. Die Sonde ist 534 Basen lang, wovon etwa 100 Basen im Überlappungsbereich mit *BM039* liegen. Daher ist auch für dieses Experiment nicht auszuschließen, dass das Signal der Sense-Sonde auf der Detektion von *BM039* beruht.

Fasst man die Ergebnisse der detaillierten Untersuchungen für die ausgewählten Gene zusammen, so lässt sich kein abschließendes Urteil über die Existenz und die Charakteristika putativer Transkripte des Gegenstranges fällen. Repetetive Elemente sowie Transkripte von anderen Genorten konnten auf Basis der aktuellen Datenbanken als Erklärung ausgeschlossen werden. Durch Northern-Blot-Hybridisierungen ließen sich die Chipdaten bestätigen.

Allerdings zeigen die Ergebnisse für das *OPN3/KMO* als Positivkontrolle, dass man zum Teil mit unspezifischen oder zumindest schwer zu interpretierenden Signalen rechnen muss. Der einzige Weg vorwärtszukommen liegt scheinbar in weiteren Laborversuchen. Als nächster Schritt sind für die drei Gene *DC13, Ponsin* und *ZNF217* die Durchführung von Ribonuklease-Protection-Assays (RPA) geplant. RPA ist ein sehr sensitives und bereits für diesen Zweck verwendetes Verfahren, mit dem der endgültige Nachweis der untersuchten Antisense-Transkripte in lebenden Zellen gelingen sollte, vorausgesetzt sie existieren. Anhand der Länge der entstehenden Fragmente lässt sich eine Hypothese über putative Transkripte generieren. Beispielsweise sollte für *DC13* erkennbar sein, ob in Antisense-Orientierung ausschließlich der Überlappungsbereich mit *BM039* nachgewiesen werden kann.

4.5. Labormethoden

Dieser Abschnitt enthält eine Beschreibung der Laborexperimente. Verfahren, die im Zuge der Anfertigung dieser Promotion selbst durchgeführt worden sind, erfahren eine genaue Darstellung mit Protokollen und verwendeten Lösungen, so dass danach eine Wiederholung des Experiments möglich sein sollte. Verfahren, die nicht selbst durchgeführt wurden, wie beispielsweise die RNA In Situ Hybridisierung, werden erläutert, und es wird für die Protokolle auf die entsprechenden Quellen verwiesen.

Verwendete Abkürzungen

ATP	Adenosin-Triphosphat
EDTA	Ethylendiamintetraacetat
DEPC	Diethylpyrocarbonat
dH_2O	deionisiertes Wasser
SDS	Natruimdodecylsulfat
SDS-PAGE	SDS-Polyacrylamid-Gelelektrophorese
SSC	Saline Sodium Citrate
TAE	Tris-Acetat-EDTA Puffer
TBS	Tris-gepufferte Salzlösung
TE	Tris-EDTA Puffer

Abkürzungen, die auch in anderen Teilen der Arbeit benutzt werden, sind in der Einführung aufgelistet.

Gelelektrophorese

Die Gelelektrophorese ist ein Verfahren, bei dem geladene Makromoleküle in einer gallertartigen Masse (Gel) mit Hilfe eines elektrischen Feldes aufgetrennt werden. In der vorliegenden Arbeit diente das Verfahren zur Aufreinigung von PCR-Produkten und zur Kontrolle der Ergebnisse anderer Experimente. Agarosegele wurden nach einem üblichen Verfahren aus 1xTAE-Puffer und Agarose (1 g pro 100 ml) hergestellt. Man versetzt die Gellösung mit Ethidiumbromid (0,5 µg/ml). Dieser Stoff lagert sich in die DNA-Moleküle ein und kann unter UV-Licht sichtbar gemacht werden. Das Gel wird in einer Gelkammer mit der Probe beladen und an die Kammer wird ein Spannungsfeld angelegt (80-120 Volt für Gele von etwa 10 cm Länge). Da DNA- und RNA-Moleküle negativ geladen sind, laufen die Produkte in Richtung des Pluspools. Die vernetzte Struktur des Gels bewirkt, dass kurze Produkte schneller und große Produkte langsamer laufen. Je nach Wahl der Bedingungen können große oder kleine Unterschiede zwischen Produkten nachgewiesen werden. In Sequenziergelen beispielsweise werden Differenzen von einer Base aufgelöst. Zur Abschätzung der Länge lässt man einen Größenstandard mitlaufen und vergleicht dann die Höhe der beobachteten Bande (Ethidiumbromid) damit.

PCR – Polymerase-Kettenreaktion

Das Verfahren der Polymerase-Kettenreaktion (Polymerase Chain Reaction, PCR) geht auf Arbeiten aus der Mitte der 80er Jahre zurück und erlaubt die spezifische Amplifikation von DNA Fragmenten. Der Einsatz der thermostabilen DNA-abhängigen DNA Polymerase aus dem thermophilen Bakterium *Thermus aquaticus* (*Taq* DNA Polymerase) ermöglicht die gezielte Amplifikation einzelner DNA-Sequenzen. Im Folgenden wird das Grundprotokoll der PCR beschrieben, wie es entsprechend angepasst verschiedentlich eingesetzt wurde. Für die Amplifikation der für diese Arbeit verwendeten DNA-Sequenzen war es nicht nötig, die PCR besonders zu optimieren, da die Produkte nicht besonders lang sind und immer ausreichend Template (cDNA) zur Verfügung stand. Der Einfachheit halber lassen sich daher so genannte „PCR-Beads“ einsetzen (Ready-To-Go™ PCR-Beads, Amersham Pharmacia Biotech). Das ist eine vorgefertigte Mischung, der in jeder Reaktion konstant bleibenden Bestandteile (Mastermix), aufgeteilt in Einheiten für einen 25µl-Reaktionsansatz.

Für einen 25µl-Reaktionsansatz gibt man folgendes hinzu:

1. 2,5 µl je Primer (10 pmol/µl)
2. 5 µl cDNA (1 ng/µl)
3. 15 µl Wasser

Laut Hersteller sind durch die Beads dann anderthalb Einheiten *Taq*-DNA-Polymerase, 10 mM Tris-HCl (pH 9,0), 50 mM KCl, 1,5 mM $MgCl_2$, je 200 mM dATP, dCTP, dGTP und dTTP und Stabilisatoren im Ansatz.

Die PCR-Reaktion führt man mit einem automatischen Termo-Cycler nach folgendem Programm durch:

1. Initiale Denaturierung 2 min bei 94 °C
2. Zyklus (40 Runden)
   2.1 Denaturierung 30 sec bei 94 °C
   2.2 Primer-Annealing 30 sec bei 58 °C
   2.3 DNA-Synthese 1,5 min bei 72 °C
3. Terminale Elongation 10 min bei 72 °C

Die Temperatur für das Primer-Annealing wird den spezifischen Primern angepasst, ebenso die Zeit für den DNA-Syntheseschritt an die Länge des Produkts.

Aufreinigung von DNA-Produkten

Um Produkte nach der PCR aufzureinigen, trennt man die Lösung in einem Agarosegel auf. Bestrahlt man das Gel von unten mit UV-Licht, so sollte das DNA-färbende Ethidiumbromid eine starke Bande in der erwarteten Laufhöhe zeigen. Das entsprechende Gelstückchen lässt sich dann mit Hilfe eines Skalpels ausschneiden und daraus das DNA-Produkt eluieren. Als Grundlage diente der QIAquick Gel Extraction Kit der Firma QiaGen (Hilden, BRD) und das vom Hersteller vorgeschlagene Protokoll. Dabei wird das Gelstück zuerst vollständig gelöst und über einer Säule die DNA aufgereinigt. Nach dem Eluieren quantifiziert man die DNA, in dem man die optische Dichte (OD) mit dem Photometer misst.

Northern-Blot-Hybridisierung

Northern Blots stellt man her, indem RNA in einem Agarosegel aufgetrennt und anschließend auf eine Zellulose- oder eine Nylonmembran transferiert wird. Konstruiert man jetzt eine radioaktiv markierte Sonde, die zu einer bestimmten Sequenz revers-komplementär ist, und hybridisiert diese auf den Blot, so lässt sich damit das Transkript in dem RNA-Pool nachweisen. Darüber hinaus ist das Verfahren in der Lage, neben der Stärke der Expression durch die Gelauftrennung auch die Größe des Transkripts und möglicher Varianten zu zeigen. Der Name Northern Blot entstand in Analogie zum Southern Blot von Edwin Southern, bei dem in ähnlicher Weise DNA in Gelen aufgetrennt und auf Membranen gebracht wird. Diese Blots können anschließend ebenfalls mit Sonden hybridisiert werden. Die Blot-Technologie ist etabliert, und es konnten kommerziell erhältlichen MTN-Blots (‘‘Multiple Tissue Northern‘‘-Blot 1+2, Clontech, Palo Alto, USA) eingesetzt werden. Die beiden Nylon-Membranen enthalten mRNA von 16 unterschiedlichen humanen Geweben (MTN1: Herz, Hirn, Plazenta, Lunge, Leber, Skelettmuskel, Niere, Pankreas; MTN2: Milz, Thymus, Prostata, Hoden, Ovar, Dünndarm, Dickdarm, periphere Blut-Lymphozyten). Die wichtigste Eigenschaft ist hierbei, dass die RNA aus den Geweben nicht amplifiziert ist. Es folgt das Protokoll für die Hybridisierung von Northernblots mit einem Mix aus Oligo-Sonden (nach Angaben des Herstellers):

1. Waschen der Blots in 1 X SSC, 0,1% SDS bei Raumtemperatur

2. Zweimal mit DEPC Wasser waschen

3. Prähybridisierung mit vorgewärmter ExpressHyb-Lösung (Clontech), 30 min bei 37°C
   Dafür legt man die beiden Blots (MTN1 und 2) Rücken an Rücken, schweißt sie in Plastikfolie ein und füllt 5 ml ExpressHyb-Lösung dazu. Anschließend muss sämtliche Luft aus der Folienhülle gedrückt werden, und sie wird luftdicht zugeschweisst. Die Blots sollten jetzt an den Außenseiten vollständig von Hybridisierungslösung umgeben sein.

4. Zugabe der markierten Oligosonde zu 5 ml frischer vorgewärmter ExpressHyb-Lösung; zur Vorbereitung der Sonde siehe Abschnitt 0

5. Austausch der Prähybridisierungslösung durch die markierte Sonde; Es sollten keine Luftblasen nach dem Verschweißen in der Folienhülle zurückbleiben.

6. Hybridisierung 2 h bei 37 °C

7. Waschen der Blots 2 X 15 min mit Waschlösung I bei Raumtemperatur
   Waschlösung I: 2 X SSC, 0,05% SDS in DEPC-Wasser

8. Waschen der Blots 2 X 15 min mit Waschlösung II bei 37 °C
   Waschlösung II: 0,1 X SSC, 0,1% SDS in DEPC-Wasser

9. Plazieren der Blots in einer Filmkassete und Auflegen eines Films

10. Exposition über Nacht

Die Restradioaktivität kann man während der Waschschritte mit dem Geigerzähler messen. Bei sehr schwachen Signalen (<40) sollte das Waschen abgebrochen werden und man belässt den Film bis zu vier Tagen zur Exposition in der Kassette.

Radioaktive Markierung von Oligonukleotiden

Die eingesetzten Oligos sollen 5'-endmarkiert werden mit [?-^{32}P]-ATP. Für die Markierungsreaktion wurden Ready-To-Go™ T4 Polynukleotidkinase (Amersham Pharmacia Biotech) eingesetzt. Laut Hersteller sind in jedem gelieferten Reaktionsgefäß 8-10 Einheiten FPLCpure™ T4 Polynukleotidkinase, 50mM Tris-HCl (pH 7,6), 5 mM DTT (Dithiothreitol), 0,1 mM Spermidine, 0,1 mM EDTA (pH 8,0), 0,2 mM ATP und Stabilisatoren. Die Markierung erfolgt nach folgendem Protokoll:

1. Zugabe von 25 µl Wasser zu Ready-To-Go T4-Polynukleotidkinase (PNK)

2. Mischen durch Hoch- und Runterpipettieren 2-5 min bei Raumtemperatur

3. Zugabe von 5-10 pmol Oligogemisch

4. Auffüllen auf 49 µl mit Wasser

5. Zugabe von 1 µl [?-^{32}P]-ATP (3'000 Ci/mmol, 10 µCi/µl)

6. Gut mischen
7. Zentrifugieren
8. Inkubation bei 37 °C, 30 min
9. Terminieren der Reaktion durch Zugabe von 2,5 µl EDTA (0,5 molar)

Damit ist ein Gemisch aus Oligo-Sonden radioaktiv markiert und bereit für den Einsatz in einer Northern-Blot-Hybridisierung.

RNA In Situ Hybridisierung (ISH)

In diesem Abschnitt wird das Verfahren der RNA In Situ Hybridisierung und die dafür nötigen Vorbereitungen am Beispiel des Gens *DC13* beschrieben. Als erstes musste ein doppelsträngiges DNA-Produkt hergestellt werden, von dem sich dann über In-Vitro-Transkription (IVT) spezifische Einzelstrangsonden herstellen lassen. Dazu wurden ausgehend von der Referenzsequenz mit dem Programm GeneQuest (DNA*Star Inc., Madison, USA) zwei Primer generiert:

Vorwärtsprimer **5'-GGGTAATACGACTCACTATAGGGATCCAGGGTTTTCATATTTCTCCA-3'**
Rückwärtsprimer **5'-GGGATTTAGGTGACACTATAGGGCGTCTGGCAAGCGGTTCA-3'**

Dabei ist die rot markierte Sequenz der für *DC13* spezifische Teil und die blaue Sequenz markiert den T7-Promoter beim Vorwärtsprimer und den SP6-Promoter beim Rückwärtsprimer. Bei der Auswahl der Primer muss man darauf achten, dass sie nicht in einer repetitiven Region liegen und in etwa gleiche Annealingtemperaturen aufweisen. Die Primer wurden von metabion (Martinsried) bezogen. Anschließend konnten die PCR und die Aufreinigung des Produkts den in 0 und 0 dargestellten Protokollen folgend durchgeführt werden (2 Ansätze à 25 µl). Als Template wurde kommerziell erworbene cDNA (Clontech, USA) aus Skelettmuskulatur eingesetzt, da das Signal auf dem Northern-Blot für dieses Gewebe vergleichsweise stark war. Bei der Quantifizierung stellte sich heraus, dass die Ausbeute nach dem Eluieren relativ gering aber für die ISH ausreichend war (121 µg/ml, 148 µg/ml). Zwei Reaktionsgefäße je 1 µg gelöst in 10 µl dienten als Ausgangsmenge für die ISH und vier Reaktionsgefäße je 200 ng gelöst in 5 µl wurden der Sequenziergruppe zur Verifikation der Basenfolge der Sonde übergeben.

Die Protokolle für die RNA In Situ Hybridisierung (ISH) sind hier nicht detailliert aufgeführt, da die Experimente selbst nicht Teil der Arbeit waren, sondern bei metaGen von Nicole Creutzburg im Labor von Edgar Dahl nach deren Standards durchgeführt wurden. Eine exakte Beschreibung der Verfahren mit den entsprechenden Literaturverweisen auf die Originalarbeiten ist Teil der bei metaGen angefertigten Promotion [45]. Ziel der ISH ist die Visualisierung der Verteilung von Transkripten in ihrer zellulären Umgebung. Als Sonde wurden in dieser Arbeit RNA-Moleküle eingesetzt, die man zuvor über In Vitro Transkription (IVT) von genspezifischer cDNA generiert und DIG-markiert hat. Das wie oben beschrieben generierte PCR-Produkt setzt man in zwei separate IVT-Reaktionen ein. Bei der einen synthetisiert die T7-RNA-Polymerase aufgrund der Lage ihres Promoters eine RNA-Sonde, welche revers-komplementär zur Referenzsequenz ist. Im zweiten IVT-Ansatz synthetisiert die SP6-RNA-Polymerase die Sense-Sonde, mit der potenzielle Antisense-Transkripte detektiert werden können. Bei der Reaktion werden Digoxygenin (DIG) -markierte dUTP's in die Sonden eingebaut. Vor der eigentlichen ISH muss man die in Paraffin-eingebetteten Gewebeschnitte mittels der Vorhybridisierung entparaffinieren, rehydrieren und fixieren. Durch die Behandlung mit einer Proteinase schließt man die Zellen auf, so dass sie für die Sonde zugänglich sind. Nun inkubiert man die Schnitte für 12 Stunden bei 65 °C mit der Sonde. In einer weiteren Reaktion lässt man dann eine alkalische Phosphatase, die an einen Anti-DIG-Antikörper gekoppelt ist, an die Sonden binden (12 h bei 4 °C). Gibt man nun nach sorgfältigem Waschen BM-Purple (Roche) für etwa vier Tage auf die Schnitte, so verursacht die alkalische Phosphotase eine blaue Präzipitation, wo die markierte Sonde hybridisiert hat. Im Standardverfahren überprüft man die Spezifität des Hybridisierungssignals durch die zelluläre Lokalisation des Präzipitates und das Signal Sense-Sonde als Negativkontrolle. Wie bereits erwähnt, erwartet man bei der Untersuchung von *DC13* Signale für Sense- und Antisense-Sonde, womit die spezifische Negativkontrolle wegfällt. Am Schluss färbt man die Schnitte noch mit Kernechtrot, was die zellulären Strukturen hervorhebt. Die Bilder wurden mit einer an ein Mikroskop gekoppelten Digitalkamera aufgenommen und durch die Pathologin Dr. Irina Klaman beurteilt. Die Bilder sind im Anhang zu Kapitel 4 beigefügt.

5. Die Wirkung der Exon- und Intronlänge auf die Genexpression in Hefe und Fruchtfliege

Dieses Kapitel untersucht die Frage, ob es spezielle strukturelle Eigenschaften hoch exprimierter Gene gibt. Mit Hilfe der Expressionsdaten sollte man in der Lage sein, generelle Restriktionen der Transkriptionsmaschinerie aufzudecken. Unter anderem wachsen die Kosten für die Expression eines Gens mit dessen Länge. Kürzlich wurde für Mensch und *C. elegans* gezeigt, dass die am höchsten exprimierten Gene im Mittel deutlich kürzere Introns haben [46]. Nachfolgend ist eine sorgfältige Analyse dreier Expressionsdatensets von *S. cereviseae* und eines Datensatzes von *D. melanogaster* dargestellt. Es zeigt sich, dass in Hefe die hochexprimierten Gene im Mittel ebenfalls deutlich verkürzt sind, obwohl Introns relativ selten sind. Die Differenzierung zwischen Genlänge und Transkriptlänge bei der Analyse ergibt, dass bei hochexprimierten Genen die Exons im Mittel etwa um den gleichen Faktor verkürzt sind wie die Introns. Sämtliche Kosten der Expression jedes Gens zu bilanzieren und die limitierenden Ressourcen hoch exprimierter Gene zu bestimmen, ist zurzeit noch nicht möglich. Aber die Analyse zeigt, dass eine evolutionäre Kraft die Länge hoch exprimierter Gene reduziert, indem Exons und Introns gleichermaßen verkürzt werden.

5.1. Einführung

In den letzten Jahren wurden neben dem humanen die Genome für die Forschung wichtiger Modellorganismen weitgehend aufgeklärt. Zu den bekanntesten eukaryontischen Vertretern zählt die Bäckerhefe *Saccharomyces cerevisiae* und die Fruchtfliege *Drosophila melanogaster*. Seit einiger Zeit werden grosse Expressionsstudien dieser Organismen durchgeführt, um die differenzielle Regulation der mRNA-Konzentration bestimmter Zellen in definierten Zuständen in ihrer Gesamtheit zu messen. Die Array-Technologie scheint dafür zurzeit die weitreichendsten Möglichkeiten zu bieten. Züchtet man beispielsweise Hefezellen auf unterschiedlichen Nährmedien, so lassen sich damit die regulatorischen Mechanismen studieren, die für die Umstellung des Metabolismus nötig sind. Dieser Abschnitt beschäftigt sich mit der Frage, ob es strukturelle Eigenschaften von Genen gibt, die ihr transkriptionelles Potential einschränken. Aufgrund dessen, dass man für Hefe beispielsweise bereits Strukturinformation und Expressionsdaten zur Mehrzahl der Gene vorliegen hat, besteht berechtigte Hoffnung, sogar spezifische Merkmale hoch exprimierter Gene zu finden. So steigt der Zeit- und Energiebedarf für die Transkription proportional zur Länge des Gens. Die

Vermutung liegt also nahe, dass Gene die besonders häufig abgeschrieben werden müssen, längenoptimiert sind. Als Genlänge wird die Anzahl Basenpaare von der ersten Position von Exon eins bis zur letzten Position des letzten Exons auf der genomischen Sequenz bezeichnet. Und in der Tat ist in einer kürzlich erschienen Arbeit für *C. elegans* und humane Gehirnzellen gezeigt worden, dass hoch exprimierte Gene eine Tendenz zu kürzeren Introns haben [46]. Interessanterweise stellten die Autoren ebenfalls fest, dass bei der Dichte der Introns (Anzahl pro Transkriptlänge) kein Effekt zu beobachten ist. Und in einer anderen Arbeit konnte für S. cerevisiae sogar gezeigt werden, dass gerade intronenthaltende Gene unter den am häufigsten transkribierten sind [47]. In einer anderen umfangreichen Studie konnte in Hefe bereits ein Zusammenhang zwischen Kodon-Bias und mRNA-Konzentration festgestellt werden [48]. Außerdem fanden die Autoren, dass für Gene mit ähnlichem Kodon-Bias die mRNA-Konzentration und die Proteinlänge negativ korreliert sind.

Welche Kosten verursacht das Spleißen? Was sind die Kosten der Translation, der Faltung und der Modifikation? Welche sind die limitierenden Faktoren für hochexprimierte Gene? Es scheint ein Balance zwischen unterschiedlichen Kosten- und Nutzenfaktoren vorzuliegen, die zum einen die Präsenz von Introns fördert und zum anderen die Länge der Exons wie auch der Introns minimiert. Betrachtet man intronische Sequenzen als mehr oder minder informationslose Platzhalter, die bestenfalls kurze regulatorische Motive enthalten, so würde man erwarten, dass hochexprimierte, längenoptimierte Gene besonders die Introns in Mitleidenschaft ziehen. Um dieser Frage nachzugehen wurden Genstruktur- mit Expressionsdaten von S. cerevisiae und D. melanogaster detailliert untersucht.

5.2. Die Analyse und Resultate

Als erstes wurde ausführlich der Zusammenhang von Expressionsniveau und Struktur der Gene in Saccharomyces cerevisiae untersucht. Als Basisdatensätzen dienten zwei genomweite Expressionsstudien ([49], [50]). Cho und Kollegen setzten für ihre Studie Oligonukletid-Arrays ein und die Arbeitsgruppe von DeRisi verwendete cDNA-Arrays. Wie im Methodenteil erläutert, ist es vorteilhaft, die Analysen auf die charakterisierten Gene einzuschränken. Das führte zu einer Auswahl von 2820 Hefegenen, für die zum Zeitpunkt der Analyse eine Annotation vorlag.

Korreliert man die gemittelten Expressionssignale mit der Länge des Gens, ergibt sich ein deutlicher Trend: Die am höchsten exprimierten Gene sind relativ kurz, und es gibt praktisch

keine langen Gene, die hoch exprimiert sind (Siehe Abbildung 5-1). Der Eindruck, den der Scatterplot vermittelt, kann täuschen. Es sind nämlich nicht nur die meisten hoch exprimierten Gene sondern auch insgesamt die meisten Gene eher kurz. Um zu verfolgen, wie sich die Expressionswerte mit der Genlänge ändern, wurden in einem gleitenden Fenster (500 Werte breit) die 90% und 75% Perzentile errechnet und in die Graphik eingetragen. Obwohl die Tendenz in dem Datensatz von Cho stärker hervortritt, ist sie auch in den cDNA-Array-Daten klare ersichtlich.

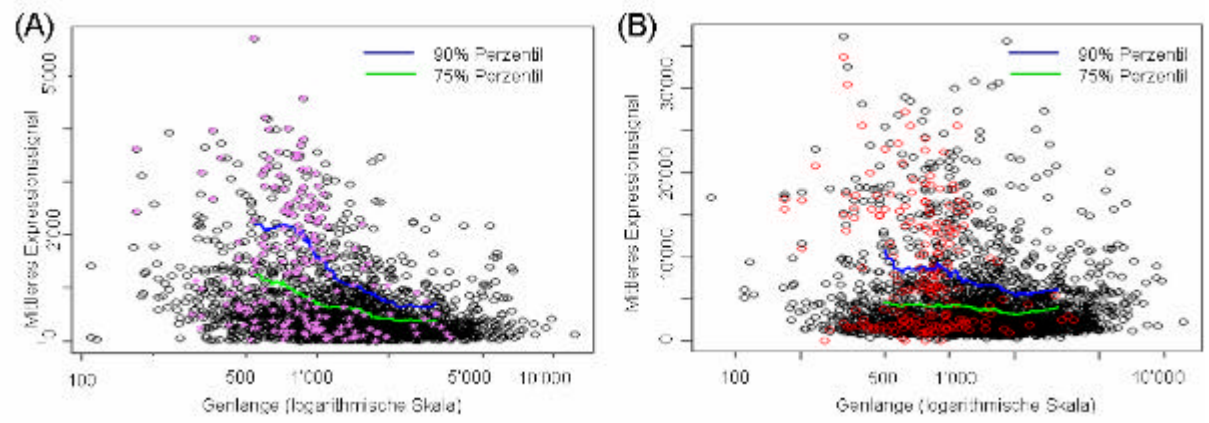


Abbildung 5-1 Abhängigkeit der Genexpression von der Genlänge in Hefe
Für alle gut charakterisierten Gene (2820) sind die Mittelwerte aller Expressionssignale gegen die Genlänge aufgetragen. Die grünen und blauen Linien veranschaulichen den Trend der Verteilung der Expressionswerte in Abhängigkeit von der Genlänge. (A) Zellzyklusdatensatz von Cho auf Oligo-Chips. ***Violette*** *Sterne markieren die Wertepaare der Gene, die Introns enthalten. (B) Metabolismusdatensatz von DeRisi auf cDNA-Arrays. Mit* ***rot*** *sind die Gene gekennzeichnet, die ribosomale Proteine kodieren.*

Als zweite wichtige Frage sollte in dieser Arbeit untersucht werden, ob die beobachtete Verkürzung hochexprimierter Gene hauptsächlich auf Kosten der Intronlängen oder eher auf Kosten der mRNA-Längen. Für die Analyse von Genen mit intronischen Sequenzen spielt die Bäckerhefe eine besondere Rolle. In Hefe konnten nur 197 intronenthaltende Gene identifiziert werden und paradoxerweise sind diese im Mittel sogar kürzer und höher exprimiert als die ohne Introns [47]. Das generelle Muster, dass die langen Gene nicht besonders hoch exprimiert sind, findet man auch für die intronenthaltenden Hefegene wieder. Es gilt aber nicht, dass die Introns im besonderen Maße verkürzt sind.

Analysen dieser Art bergen immer die Gefahr, dass es sich bei der Beobachtung um rein technische Phänomene handelt. Zum Beispiel ist bekannt, dass der mittlere GC-Gehalt zwischen hoch und niedrig exprimierten Genen verschieden ist und dass der GC-Gehalt und die Genlänge korreliert sind. Dazu kommt, dass der GC-Gehalt, der Oligo-Sonden einen

Einfluss auf das Expressionssignal hat. Um die Ergebnisse dahingehend zu überprüfen, kann man die Daten nach GC-Gehalt und Länge faktorisieren: Man teilt dazu den Wertebereich der Genlänge so in 10 Intervalle ein, dass in jeden gleich viele Gene fallen und ebenso verfährt man mit dem Wertebereich des GC-Gehalts. Anschließend lässt sich in jedem der zehnmal zehn Felder der mittlere Expressionswert berechnen und zur Anschauung farbkodieren.

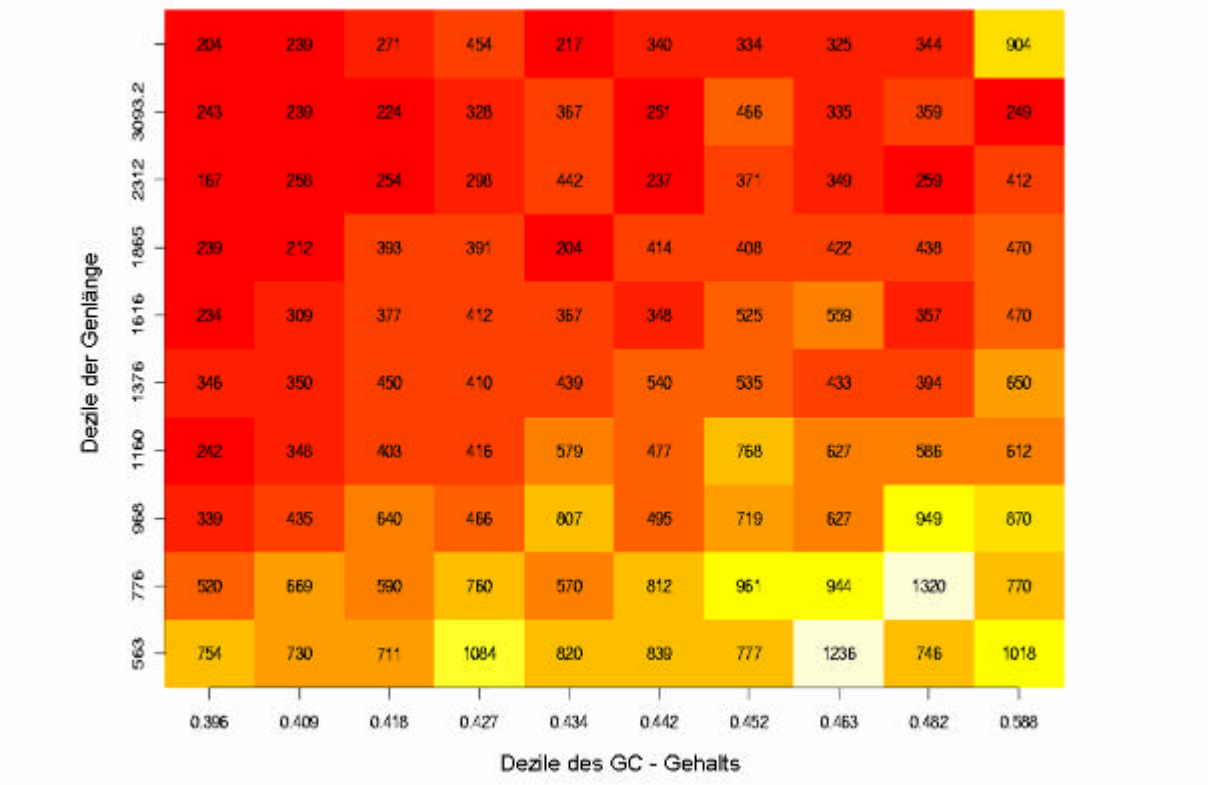


Abbildung 5-2 Der Einfluss des GC-Gehalts und der Genlänge auf das Expressionsniveau

Die Expressionswerte sind zweidimensional nach GC-Gehalt und Genlänge aufgelöst. Hierzu berechnet man die Dezile der Längenverteilung der Gene, also die Perzentile für 10%, 20% und so weiter. Auf ähnliche Weise berechnet man die Dezile für die Verteilung des GC-Gehalts der Sequenzen. Jetzt kann jedes Gen in eines der zehn Felder einsortiert und dann der Durchschnitt ihrer Expressionswerte pro Feld gebildet werden. Die Farben kodieren die mittlere Expression in dem jeweiligen Feld, und die Zahl ist gerade das gerundete Mittel.

In Abbildung 5-2 ist klar der Einfluss des GC-Gehalts zu erkennen, allerdings überwiegt Abhängigkeit des Expressionssignals von der Genlänge. Als zusätzliche Kontrolle konnten noch Daten, die mit einer dritten, der SAGE-Technologie (Siehe Abschnitt 1.6) erzeugt

wurden, hinzugezogen werden, wobei die Summe der gefundenen SAGE-Tags[1] über die Proben als repräsentativer Expressionswert diente [51]. Um dabei vergleichbare und aussagekräftige Resultate zu erhalten, hat es sich als zweckdienlich erwiesen, die Gene nach ihrer Länge zu klassifizieren und anschließend die Verteilungen der Expressionswerte der Klassen zu vergleichen. Aus einem Datensatz extrahiert man dazu die Expressionswerte der 33% kürzesten Gene und bildet die empirische kumulative Verteilung. Dann verfährt man entsprechend mit den Expressionswerten der 33% längsten Gene und erhält damit zwei vergleichbare Verteilungen.

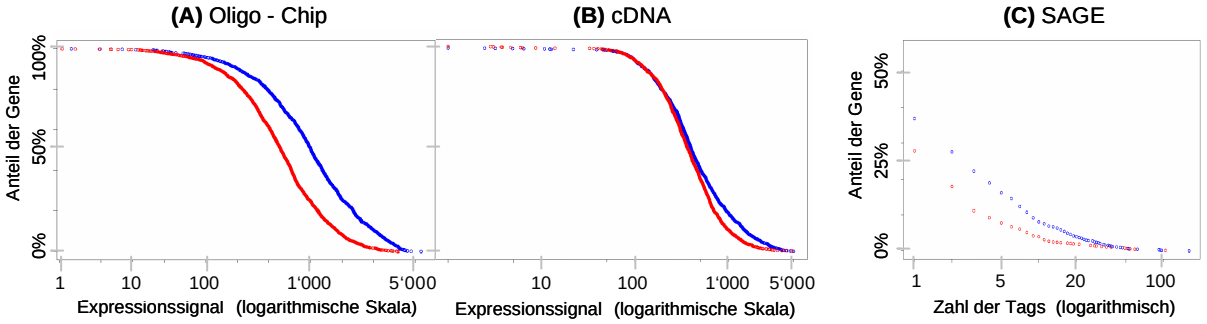


Abbildung 5-3 Unterschiede in der Expression zwischen langen und kurzen Genen in Hefe

*Es wurden in jedem Datensatz (A, B, C) zwei Genpopulationen gewählt: die 33% **längsten (rot)** und die 33% **kürzesten (blau)**. Für jede dieser Gruppen ist die kumulative Verteilung der mittleren Expressionswerte dargestellt (siehe Text). Ein Punkt (x,y) in der Graphik gibt an, wie groß der prozentuale Anteil y der Gene dieser Population ist, die mindestens den Expressionswert x erreichen. (A) Oligo-Chip-Daten (B) cDNA- Daten, (C) Bei den SAGE-Daten ist zu beachten, dass für über 50% der Transkripte überhaupt kein Tag gefunden und damit keine Expression gemessen werden konnte. Die Skala ist entsprechend abgeschnitten und die Werte damit verzerrt.*

In allen drei Datensätzen unterscheiden sich die Verteilungen deutlich, was sich auch durch die Anwendung des U-Tests[2] bestätigen lässt. Aufgrund der logarithmischen Darstellung in der Graphik mag der Unterschied gering erscheinen. Im Datensatz A, der mit Oligo-Arrays erzeugt worden ist, beträgt der Mittelwert der Expressionsniveaus über die kurzen Gene 817

---

[1] SAGE: Serial Analysis of Gene Expression (Siehe Einführung Abschnitt 1.6)

[2] U-Test nach Mann, Whitney und Wilcoxon gilt als nichtparametrische Alternative zum ungepaarten t-Test [8] Sachs, Lothar (1997): Angewandte Statistik, Springer, Berlin Heidelberg..

im Gegensatz zu 402 über die langen. Für die dritten Quartile (75% Perzentile) gilt das Verhältnis 1024 zu 498. Unter den am höchsten exprimierten Genen finden sich besonders häufig solche, die für ribosomale Proteine kodieren. Außerdem enthalten diese Gene oft Intron, sind aber insgesamt relativ kurz. An dieser Stelle sei explizit darauf hingewiesen, dass ausschließlich proteinkodierende Gene untersucht wurden, hohe Signale also nicht von ribosomalen RNA's oder ähnlichen herrühren. Aufgrund der hohen Produktionsrate an Ribosomen und der relativ kurze Halbwertzeit der mRNA's schätzte man, dass etwa 50% der RNA Polymerase II Transkription allein der Biosynthese der Ribosomen dient [52].

Um die Ergebnisse in einen evolutionären Bezug zu stellen, wurden mit Oligo-Arrays erzeugte Expressionsdaten von Drosophila Melanogaster analysiert (Quelle: [53]). Die Zahl der Gene in Drosophila schätzt man auf 13000. Die meisten von ihnen sind beträchtlich länger als in Hefe und enthalten Introns. Auch für diesen Datensatz beschränkt man sich sicherheitshalber auf die annotierten 3889 Gene. Wieder wurden die Expressionswerte pro Gen über die Arrays gemittelt und zur Genlänge in Bezug gesetzt.

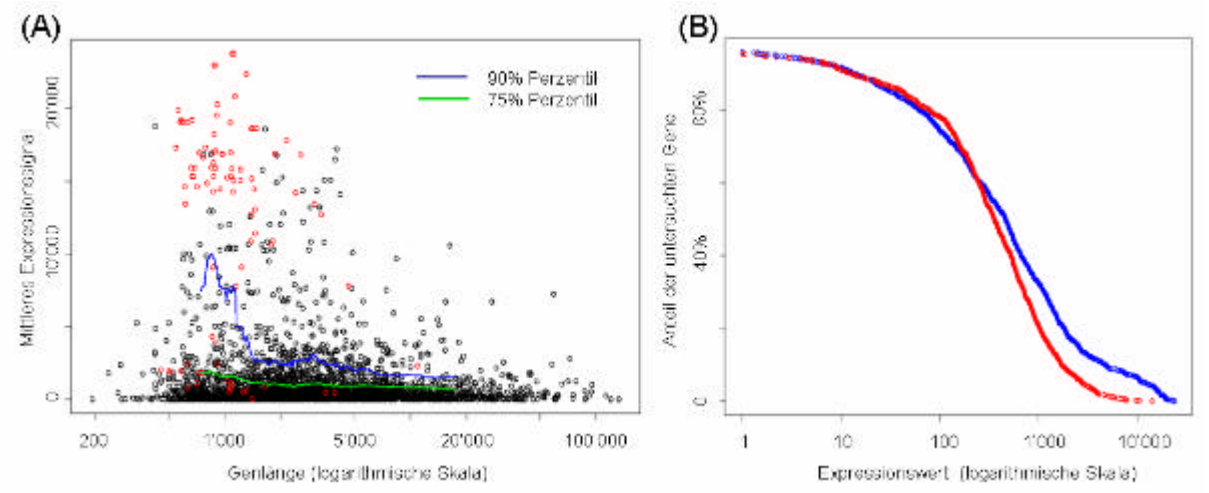


Abbildung 5-4 Die Analyse der Drosophila-Daten

(A) *Darstellung der mittleren Expressionswerte in Abhängigkeit von der Genlänge. Mit rot sind die Wertepaare gekennzeichnet, die zu Genen gehören, die ribosomale Proteine kodieren. Wieder wurden 90%- und 75%-Perzentile in einem gleitenden Fenster (500 Werte breit) berechnet und in die Graphik eingefügt.* ***(B)*** *Vergleichbar der sind wieder die kumulativen Verteilungen der Expressionsniveaus der 33%* ***längsten (rot)*** *und der 33%* ***kürzesten (blau)*** *Gene gegenübergestellt.*

Auch in Drosophila gilt generell, dass die hoch exprimierten Gene im Mittel kürzer sind. Im Gegensatz zur Hefe ist ihr Anteil in Bezug auf alle untersuchten Gene aber geringer. Verfolgt man die mittlere Expression mit einem gleitenden Fenster, so zeigen verglichen mit den 90%-

Perzentilen die 75%-Perzentile nur eine schwache Tendenz. Der Vergleich der Verteilungen der Expressionswerte des längsten und des kürzesten Drittels der Gene ergibt einen signifikanten (p-Wert des U-Tests: 0,00017) aber geringeren Unterschied als in dem Hefedatensatz. Ist die Zahl der hoch exprimierten Gene pro Zelle in Hefe und Drosophila etwa gleich groß, so ließe sich der beobachtete Unterschied erklären: Geht man davon aus, dass die Fruchtfliege 2,5-mal so viele Gene hat wie die Bäckerhefe, so verringert sich der Anteil der hoch exprimierten am Gesamtpool der Gene um diesen Faktor. Um für diesen Datensatz zwischen den Längen intronischer Sequenz und mRNA zu diskriminieren, wurden die Gene nach ihren Expressionsniveaus geordnet und anschließend die 5% beziehungsweise 10% höchsten extrahiert. Der Vergleich der Längenverteilungen für die mRNA's wie für die Summe der Introns ergibt, dass die Introns der hochexprimierten Gene geringfügig stärker verkürzt sind als die mRNA's. In den publizierten Daten für *C. elegans* [46] sind ebenfalls die Längen der mRNA's und der Introns etwa um den gleichen Faktor verkürzt für die hoch exprimierten Gene.

Tabelle 5-1 Korrelation von Sequenzlängen und Expression in Drosophila

Auswahl der Gene nach ihrer Expression	Mittlere Längen		
	Gene	mRNA's	Summe der Introns*
5% höchsten	2642	1305	1533
10% höchsten	3393	1614	1921
Alle	5782	2325	3954

**Für die Berechnung der Werte dieser Spalte wurden nur intronenthaltende Gene benutzt. Die Werte entstanden durch die Addition der Längen aller Introns pro Gen mit anschließender Mittelwertbildung in den jeweiligen Gruppen.*

Für das **humane** Transkriptome ist es schwierig, ein repräsentatives Set von Genen mit vollständig aufgeklärter Struktur zu bekommen. Oft ist zwar der kodierende Bereich eines Transkripts bekannt (*complete CDS*), aber nicht sämtliche Transkriptvarianten mit ihrer Gewebsverteilung und auch nicht das vollständige Gen mit seiner Exons-Intron-Struktur. Auch für klare Signale in einer Expressionsstudien ist nicht immer bekannt, ob sie von Transkripten desselben Typs oder von einer Menge von Transkriptvarianten stammen. Im Vergleich zur Hefe sind die humanen Gene im Mittel etwa zehnmal so lang und häufig alternativ gespleißt beziehungsweise polyadenyliert. In einem Versuch, sich auf gut charakterisierte Gene zu beschränken, konnten 1892 Referenzsequenzen identifiziert werden, die sich vollständig auf eine zusammenhängende genomische Sequenz abbilden ließ. Zur

Abschätzung der Expressionsniveaus wurden die gleichen Lungendaten wie in Kapitel 3 eingesetzt [16]. Als zusätzliches Auswahlkriterium der Gene galt, dass sie auf dem in der Expressionsanalyse verwendeten Array repräsentiert sein müssen. Auch hier lassen sich die Verteilungen der Expressionswerte der längsten und kürzesten Gene gegenüberstellen. Man kann eine leichte jedoch im U-Test nicht signifikante Verschiebung der Verteilungen beobachten. Die mittlere Länge der 138 am höchsten exprimierten Gene beträgt 12640 Basen im Gegensatz zum Gesamtmittel von 21160 Basen. Die Aussage ist eingeschränkt, da ja schätzungsweise nur 5% der humanen Gene in die Analyse eingegangen sind.

5.3. Zusammenfassung und Diskussion der Ergebnisse

In diesem Kapitel ist durch eine differenzierte Betrachtung gezeigt, dass hoch exprimierte Gene im Mittel kürzere Exons und Introns haben. Ausführlich wurden dazu Datensätze von der Bäckerhefe und der Fruchtfliege untersucht. Damit konnte die Arbeit von Castillo-Davis et al auf diese Organismen ausgedehnt werden. In *Drosophila* und *C. elegans* gilt für die am höchsten exprimierten im Vergleich zu allen Genen, dass die Introns und die Exons etwa um den gleichen Faktor kürzer sind (Tabelle 5-1, [46]). In Hefe gibt es nur sehr wenige intronenthaltende Gene, die häufig hoch exprimiert sind. Das Ergebnis der Analyse der Hefedaten zeigt, dass die Genlänge in einem klaren Abhängigkeitsverhältnis zum Expressionsniveau steht und dass die intronischen Sequenzen zumindest nicht in stärkerem Maße verkürzt sind als die Exons. Die Ergebnisse geben deutliche Hinweise darauf, dass es sich positiv auf die Fitness eines Organismus auswirkt, wenn gerade hoch exprimierte Gene Introns enthalten. Allerdings ist die Funktion der Introns in diesem Zusammenhang völlig rätselhaft, da sich der Aufwand während der Transkription ja erhöht.

Aus methodischer Sicht lässt sich zusammenfassend sagen, dass die verbreiteten Hochdurchsatzverfahren zur Messung der Genexpression zumindest bedingt auch eine Aussage über die absolute mRNA-Konzentration erlauben. Vor allem cDNA-Arrays werden meistens mit der Begründung komparativ hybridisiert, dass den Absolutwerten nicht zu trauen ist. Die Analyse der Datensätze und der Vergleich der Technologien zeigen jedoch, dass die hintergrundkorrigierten Signale der einzelnen Farbkanäle (absolute Intensitätssignale) durchaus informativ sind.

5.4. Daten und Methoden

Im Folgenden werden die verwendeten Datensätze beschrieben: Die Arbeitsgruppe um Cho setzte den von Affymetrix (Santa Clara, USA) kommerziell erhältlichen Oligo-Chip Ye6100 ein, um die Expression der 6'218 vorhergesagten Transkripte im Zellzyklus zu studieren. Dazu wurden Hefezellen im mitotischen Zellzyklus synchronisiert, wobei ein Zyklus etwa 160 Minuten dauert. Es wurden Zellen an 17 Messpunkte, also etwa in 10 Minuten Abständen, geerntet und deren RNA auf die Oligo-Chips hybridisiert. Die Daten wurden von den Autoren mit der Analysesoftware des Chipherstellers ausgewertet. Dabei berechnet das Programm für jedes Gen in jedem Experiment ein Expressionswert, in dem die mittlere Differenz der Signale der Oligosonde und der spezifischen Negativkontrolle gebildet wird (AvgDiff-Wert).

Zur Untersuchung der Genexpression während der Änderung der Nahrungsquelle von Glukose auf Ethanol führt eine Arbeitsgruppe um DeRisi Experimente mit cDNA-Arrays durch [50]. Diese Daten der beiden Experimente sind auf der Internetseite http://genome-www5.stanford.edu/MicroArray/SMD bereitgestellt. Normalerweise verwendet man bei Analysen von cDNA-Arrays den Quotient der Signale der beiden komparativ hybridisierten Proben. Da aber in dieser Arbeit eine Aussage über absolute Expressionshöhen der Gene getroffen werden soll, muss man direkt die hintergrundkorrigierten Signalwerte benutzen. In den Dateien im Stanford-Format sind sie als CH1D_MEDIAN und CH2D_MEDIAN bezeichnet.

Ziel der Studie von A.J. Kal und Kollegen in Hefe war die Bestimmung der Änderungen der Genexpression bei der Umstellung der Nahrungsquelle auf Oleat [51]. Dazu fertigten sie eine SAGE-Bibliothek von Hefezellen an, die man vorher 18 Stunden auf Oleat-Medium hatte wachsen lassen. Anschließend konnten dann die Ergebnisse mit einem Datensatz verglichen werden, für den die Zellen auf Glukose-Medium gewachsen waren. Generiert wurden letztendlich etwa 14000 10-Basen-lange Sequenztags zu 1700 verschiedenen Hefegenen. Aus der Anzahl der Tags, die pro Gen gefunden wurden, lässt sich dessen Expressionsniveau schätzen. Die Daten dieser Analyse konnten von der folgenden Internetseite runtergeladen werden:

http://www.molbiolcell.org/cgi/content/full/10/6/1859/DC1

Die Annotation der Hefetranskripte erfolgte durch aktuelle Listen aus Datenbanken SGD [54] und MIPS [55], wobei sich der Nomenklatur des Geneontology-Konsortiums [56] bedient wurde. Transkripte, die allein durch Computerprogramme vorhergesagt worden sind und für

die keine zusätzlichen Evidenzen vorliegen, sind in den Listen als „unknown“ beziehungsweise „not yet annotated“ charakterisiert. Sie sind nicht in die weiteren Analysen eingeflossen, da es deutliche Hinweise für eine hohe Rate an falsch-positiven gerade bei kurzen Genen gibt [57]. Die Expressionswerte der Gene zwischen 300 und 500 Basen Länge gehorchen einer völlig anderen Verteilung als die der restlichen Gene.

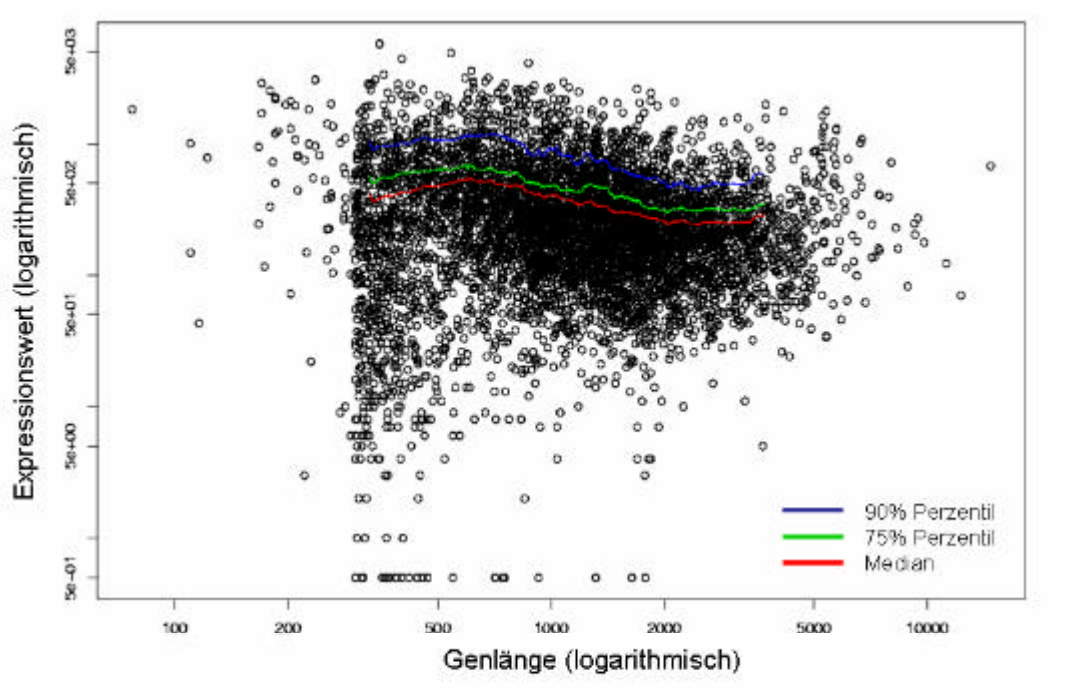


Abbildung 5-5 Hohe Rate falsch Positiver bei kurzen Transkripten
*Dargestellt sind die Expressionswerte **aller** 6218 auf dem ChipYe6000 repräsentierten Sequenzen, dass heißt auch solche, die nicht annotiert sind. Die sich scharf abzeichnende Grenze liegt bei etwa 300 Basenpaaren Länge. Die Perzentile und der Median wurden in einem gleitenden Fenster(500 Werte) berechnet.*

Der Drosophila-Datensatz [58] wurde generiert, um die von dem evolutionär hoch konservierten Transkriptionsfaktor *otd* regulierten Gene zu finden. Die Expression dieser Gene wurde mit der einer Mutante verglichen, die anstelle von *otd* das humane homologe Gen *Otx2* trägt. Das Gen brachte man in einer Form in die Eizellen, dass es in den Embryonen mit Hilfe eines Hitzeschocks von 25 °C auf 37 °C gezielt aktiviert werden konnte. Für die Experimente verwendeten die Autoren Oligo-Arrays von Affymetrix für Drosophila. Insgesamt führte man 16 Chipexperimente durch, deren Daten über die GEO-Seite am NCBI unter der Referenz GPL70 in bereits ausgewerteter und normalisierter Form verfügbar sind. Als Basis für die Annotation der Gene ließen sich die von den Autoren gelieferten Listen und die umfangreichen Datenbanken der öffentlichen Zentren benutzen:

http://www.fruitfly.org/sequence/download.html

Für die Rechnungen und die Erstellung der Graphiken wurde das Statistik-Packet R verwendet ([59], http://www.r-project.org). Für die Berechnungen der gleitenden Perzentile in den Scatterplots erwies sich eine Fensterweite von 500 Werten als angemessen und eine Variation bringt keine neuen Einsichten. Mit Hilfe der R-Funktion *ecdf* (Bibliothek *stepfun*) ließen sich die empirischen kumulativen Verteilungsfunktionen in den Abbildungen berechnen. Der U-Test wurde mit der R-Funktionen *wilcox.test* berechnet.

6. Zusammenfassung und Ausblick

Ein Kernstück der vorliegenden Arbeit ist die Entwicklung eines neuen Verfahrens zur Auswertung von Genexpressionsdaten, die auf der Affymetrix-Plattform generiert worden sind. Entscheidender Unterschied zur Standardmethode von Affymetrix ist die ausschließliche Verwendung der *PM*-Signale zur Schätzung der Genexpression (Kapitel 2). Zusammenfassend lässt sich folgendes feststellen: Die PMQ-Methode liefert vor allem für Rohdaten unterschiedlicher Qualität besser reproduzierbare Resultate. Das gilt sowohl für die absoluten Expressionswerte wie auch für errechnete Expressionsänderungen. Aus der Untersuchung der Spike-Kontrollen ergab sich, dass zwar die Expressionswerte proportional zur eingesetzten Konzentration wachsen aber die tatsächliche Stärke der Expressionsänderung im Allgemeinen aus den Werten nicht genau bestimmt werden kann. Diese Aussage gilt für beide Auswerteverfahren und Gründe dafür könnten in der Sequenzabhängigkeit der Amplifikation der RNA wie auch der Hybridisierung liegen.

Kapitel 3 enthält die Reanalyse und Gegenüberstellung der Expressionsdaten zweier umfangreicher Studien zum Bronchialkarzinom. Besonders interessant ist der Vergleich aus technologischer Sicht, da sich die eine Arbeitsgruppe der Affymetrix-Plattform bediente und die andere cDNA-Arrays einsetzte. Bei der Analyse findet man Gene, die in beiden Datensätzen zwischen Tumor- und Normalgewebe als differenziell exprimiert auffallen. Einige davon sind bereits in der Literatur beschrieben, andere sind völlig unbekannt. Bei der Gegenüberstellung der Ergebnisse der auf unterschiedlichen Array-Plattformen durchgeführten Analysen zeigt sich der deutliche Einfluss der Technologie auf die Expressionssignale. Es empfiehlt sich deshalb, Gene, die in beiden Versuchsreihen konsistent als tumorrelevant auffallen, zuvorderst weiterzubearbeiten. Völlige Sicherheit über die Expression einzelner Gene kann man nur über eine Validierung mit komplementären Methoden wie zum Beispiel Northern-Blots gewinnen. Die umfangreiche Datenbank und das robuste Auswerteverfahren erlauben die Suche nach Gewebs- bzw. tumorspezifische Expressionsprofilen. Zum Beispiel ließen sich die Proben des Plattenepithelkarzinoms von den Normalproben anhand spezifisch exprimierter Gene trennen. Besonders interessant in diesem Zusammenhang wäre eine Subklassifizierung histologisch nicht unterscheidbarer Formen des Bronchialkarzinoms, die unterschiedlich auf Therapien ansprechen oder andere Krankheitsverläufe aufweisen.

Transkripte, die bisher nur am Computer vorhergesagt wurden, können im großen Umfang durch Arrayexperimente bestätigt werden. Besonders interessant ist die Möglichkeit, neue Gene beziehungsweise Transkriptvarianten zu entdecken. Der überraschende Befund, dass für einige Gene Sense- und Antisense-Sonde ein konsistentes Expressionssignal liefern, konnte mit Northern-Blot-Hybridisierungen bestätigt werden (Kapitel 4). Diese Ergebnisse erinnern daran, dass das Transkriptom des Menschen noch nicht vollständig entschlüsselt ist. Sorgfältig konstruierte Arrays können bei der weiteren Analyse eine hilfreiche Stütze sein.

Kapitel 5 beschäftigt sich mit der Abhängigkeit des transkriptionellen Potenzials der Gene von ihren strukturellen Eigenschaften. Für die im Detail untersuchten Organismen, Hefe und Fruchtfliege, konnten Zusammenhänge zwischen dem Expressionsniveau und der Genlänge aufgedeckt beziehungsweise bestätigt werden. Bei hochexprimierten Genen sind die Exons etwa im gleichen Maße verkürzt wie die Introns. An diese Ergebnisse schließen sich interessante Fragen zur Ökonomie der Zelle und zum Zweck von Introns an.

Bibliographie

[1] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. und Lipman, D. J. (1990): Basic local alignment search tool, J Mol Biol (Band 215), Nr. 3, Seite 403-10.

[2] Pruitt, K. D. und Maglott, D. R. (2001): RefSeq and LocusLink: NCBI gene-centered resources, Nucleic Acids Res (Band 29), Nr. 1, Seite 137-40.

[3] Schmitt, A. O.; Specht, T.; Beckmann, G.; Dahl, E.; Pilarsky, C. P.; Hinzmann, B. und Rosenthal, A. (1999): Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues, Nucleic Acids Res (Band 27), Nr. 21, Seite 4251-60.

[4] Schuler (1997): Pieces of the puzzle: expressed sequence tags and the catalog of human genes., J Mol Med (Band 75). URL: http://www.ncbi.nlm.nih.gov/UniGene

[5] Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P. und Trent, J. M. (1999): Expression profiling using cDNA microarrays, Nat Genet (Band 21), Nr. 1 Suppl, Seite 10-4.

[6] Schena (1999): DNA microarrays, Hames, The Practical Approach Series.

[7] Velculescu, V. E.; Zhang, L.; Vogelstein, B. und Kinzler, K. W. (1995): Serial analysis of gene expression, Science (Band 270), Nr. 5235, Seite 484-7. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=7570003

[8] Sachs, Lothar (1997): Angewandte Statistik, Springer, Berlin Heidelberg.

[9] Li, C. und Wong, W. H. (2001): Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, Proc Natl Acad Sci U S A (Band 98), Nr. 1, Seite 31-6. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.pnas.org/cgi/content/full/98/1/31

[10] Bolstad, Irizzary, Astrand, Speed (2002 (akzeptiert)): A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias, Bioinformatics.

[11] Workman, Jensen, Jarmer, Berka, Gautier, Nielsen HB, Saxild, Nielsen C, Brunak, Knudsen (2002): A new non-linear normalization method for reducing variability in DNA microarray experiments, Genome Biology.

[12] Durbin, B. P.; Hardin, J. S.; Hawkins, D. M. und Rocke, D. M. (2002): A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics (Band 18 Suppl 1), Seite S105-10.

[13] Huber, W.; Von Heydebreck, A.; Sultmann, H.; Poustka, A. und Vingron, M. (2002): Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics (Band 18 Suppl 1), Seite S96-S104.

[14] Chu, T. M.; Weir, B. und Wolfinger, R. (2002): A systematic statistical linear modeling approach to oligonucleotide array experiments, Math Biosci (Band 176), Nr. 1, Seite 35-51.

[15] Roeing (1996): Oracle7 Datenbanken erfolgreich realisieren, Härder, Reuter, Datenbanksysteme.

[16] Bhattacharjee, A.; Richards, W. G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.; Gillette, M.; Loda, M.; Weber, G.; Mark, E. J.; Lander, E. S.; Wong, W.; Johnson, B. E.; Golub, T. R.; Sugarbaker, D. J. und Meyerson, M. (2001): Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc Natl Acad Sci U S A (Band 98), Nr. 24, Seite 13790-5. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.pnas.org/cgi/content/abstract/98/24/13790

[17] Garber, M. E.; Troyanskaya, O. G.; Schluens, K.; Petersen, S.; Thaesler, Z.; Pacyna-Gengelbach, M.; van de Rijn, M.; Rosen, G. D.; Perou, C. M.; Whyte, R. I.; Altman, R. B.; Brown, P. O.; Botstein, D. und Petersen, I. (2001): Diversity of gene expression in adenocarcinoma of the lung, Proc Natl Acad Sci U S A (Band 98), Nr. 24, Seite 13784-9. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.pnas.org/cgi/content/abstract/98/24/13784

[18] Minna, J. D.; Roth, J. A. und Gazdar, A. F. (2002): Focus on lung cancer, Cancer Cell (Band 1), Nr. 1, Seite 49-52.

[19] Proctor, Robert N. (2001): Tobacco and the global lung cancer epidemic, Nature Reviews Cancer (Band 1).

[20] Geneser, Finn (1990): Histologie, Deutscher Ärzte- Verlag GmbH, Köln.

[21] Riede, Ursus-Nikolaus und Schaefer, Hans-Eckart (1999): Allgemeine und spezielle Pathologie, Thieme Verlag, Stuttgart, New York.

[22] Mendelsohn; Howley; Israel und Liotta (2001): The Molecular Basis of Cancer, W. B. Saunders Company, Philadelphia.

[23] Seeber, Siegfried und Schütte, Jochen (1998): Therapiekonzepte Onkologie, Springer-Verlag.

[24] Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O. und Weinstein, J. N. (2000): A gene expression database for the molecular pharmacology of cancer, Nat Genet (Band 24), Nr. 3, Seite 236-44.
taf/DynaPage.taf?file=/ng/journal/v3/n3/full/ng0300_236.html
taf/DynaPage.taf?file=/ng/journal/v3/n3/abs/ng0300_236.html. URL:
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.nature.com/cgi-taf/DynaPage.taf%3ffile=/ng/journal/v3/n3/abs/ng0300_236.html

[25] Zochbauer-Muller, S.; Gazdar, A. F. und Minna, J. D. (2002): Molecular pathogenesis of lung cancer, Annu Rev Physiol (Band 64), Seite 681-708. URL:
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://physiol.annualreviews.org/cgi/content/abstract/64/1/681

[26] Zinner, R. G.; Kim, J. und Herbst, R. S. (2002): Non-small cell lung cancer clinical trials with trastuzumab: their foundation and preliminary results, Lung Cancer (Band 37), Nr. 1, Seite 17-27.

[27] Harries, M. und Smith, I. (2002): The development and clinical use of trastuzumab (Herceptin), Endocr Relat Cancer (Band 9), Nr. 2, Seite 75-85. URL:
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://journals.endocrinology.org/erc/009/erc0090075.htm

[28] Kopitar-Jerala, N.; Gubensek, F. und Turk, V. (2000): Recombinant anti-stefin A Fab fragment: sequence analysis of the variable region and expression in Escherichia coli, Biol Chem (Band 381), Nr. 12, Seite 1245-9.

[29] Ebert, E.; Werle, B.; Julke, B.; Kopitar-Jerala, N.; Kos, J.; Lah, T.; Abrahamson, M.; Spiess, E. und Ebert, W. (1997): Expression of cysteine protease inhibitors stefin A, stefin B, and cystatin C in human lung tumor tissue, Adv Exp Med Biol (Band 421), Seite 259-65.

[30] Strojan, P.; Budihna, M.; Smid, L.; Svetic, B.; Vrhovec, I. und Skrk, J. (2001): Cathepsin B and L and stefin A and B levels as serum tumor markers in squamous cell carcinoma of the head and neck, Neoplasma (Band 48), Nr. 1, Seite 66-71.

[31] Park, P. J.; Tian, L. und Kohane, I. S. (2002): Linking gene expression data with patient survival times using partial least squares, Bioinformatics (Band 18 Suppl 1), Seite S120-7.

[32] Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.;

Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; Szustakowki, J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S. und Chen, Y. J. (2001): Initial sequencing and analysis of the human genome, Nature (Band 409), Nr. 6822, Seite 860-921.

[33] Rong, M.; Durbin, R. K. und McAllister, W. T. (1998): Template strand switching by T7 RNA polymerase, J Biol Chem (Band 273), Nr. 17, Seite 10253-60. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jbc.org/cgi/content/full/273/17/10253

[34] Vanhee-Brossollet, C. und Vaquero, C. (1998): Do natural antisense transcripts make sense in eukaryotes?, Gene (Band 211), Nr. 1, Seite 1-9.

[35] Sleutels, F.; Zwart, R. und Barlow, D. P. (2002): The non-coding Air RNA is required for silencing autosomal imprinted genes, Nature (Band 415), Nr. 6873, Seite 810-3. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=11845212

[36] Rougeulle, C. und Heard, E. (2002): Antisense RNA in imprinting: spreading silence through Air, Trends Genet (Band 18), Nr. 9, Seite 434.

[37] Lehner, B.; Williams, G.; Campbell, R. D. und Sanderson, C. M. (2002): Antisense transcripts in the human genome, Trends Genet (Band 18), Nr. 2, Seite 63-5.

[38] Church, Jay Shendure and George M (2002): Computational discovery of sense-antisense transcription in the human and mouse genomes, Genome Biology, Nr. 3(9).

[39] Kasper, Grit (akzeptiert): Differential structural organization of the encephalopsin gene in man and mouse, Gene.

[40] Halford, S.; Freedman, M. S.; Bellingham, J.; Inglis, S. L.; Poopalasundaram, S.; Soni, B. G.; Foster, R. G. und Hunt, D. M. (2001): Characterization of a novel human opsin gene with wide tissue expression and identification of embedded and flanking genes on chromosome 1q43, Genomics (Band 72), Nr. 2, Seite 203-8. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=11401433

[41] Collins, C.; Rommens, J. M.; Kowbel, D.; Godfrey, T.; Tanner, M.; Hwang, S. I.; Polikoff, D.; Nonet, G.; Cochran, J.; Myambo, K.; Jay, K. E.; Froula, J.; Cloutier, T.; Kuo, W. L.; Yaswen, P.; Dairkee, S.; Giovanola, J.; Hutchinson, G. B.; Isola, J.; Kallioniemi, O. P.; Palazzolo, M.; Martin, C.; Ericsson, C.; Pinkel, D.; Gray, J. W. und et al. (1998): Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma, Proc Natl Acad Sci U S A (Band 95), Nr. 15, Seite 8703-8. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=9671742

[42] You, A.; Tong, J. K.; Grozinger, C. M. und Schreiber, S. L. (2001): CoREST is an integral component of the CoREST- human histone deacetylase complex, Proc Natl Acad Sci U S A (Band 98), Nr. 4, Seite 1454-8. URL:

http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.pnas.org/cgi/content/full/98/4/1454

[43] Mandai, K.; Nakanishi, H.; Satoh, A.; Takahashi, K.; Satoh, K.; Nishioka, H.; Mizoguchi, A. und Takai, Y. (1999): Ponsin/SH3P12: an l-afadin- and vinculin-binding protein localized at cell-cell and cell-matrix adherens junctions, J Cell Biol (Band 144), Nr. 5, Seite 1001-17. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jcb.org/cgi/content/full/144/5/1001

[44] Baumann, C. A.; Ribon, V.; Kanzaki, M.; Thurmond, D. C.; Mora, S.; Shigematsu, S.; Bickel, P. E.; Pessin, J. E. und Saltiel, A. R. (2000): CAP defines a second signalling pathway required for insulin-stimulated glucose transport, Nature (Band 407), Nr. 6801, Seite 202-7. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=11001060

[45] Kuner, Ruprecht (2002): Identifizierung differenziell exprimierter Gene bei Brust- und Ovarialkarzinomen in den chromosomalen Regionen

1q32-q41 und 11q12-q23, Biologie, Humboldt, Berlin.

[46] Castillo-Davis, C. I.; Mekhedov, S. L.; Hartl, D. L.; Koonin, E. V. und Kondrashov, F. A. (2002): Selection for short introns in highly expressed genes, Nat Genet (Band 31), Nr. 4, Seite 415-8.

[47] Ares, M., Jr.; Grate, L. und Pauling, M. H. (1999): A handful of intron-containing genes produces the lion's share of yeast mRNA, Rna (Band 5), Nr. 9, Seite 1138-9.

[48] Coghlan, Keneth H Wolfe (2000): Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae, Yest.

[49] Cho, R. J.; Campbell, M. J.; Winzeler, E. A.; Steinmetz, L.; Conway, A.; Wodicka, L.; Wolfsberg, T. G.; Gabrielian, A. E.; Landsman, D.; Lockhart, D. J. und Davis, R. W. (1998): A genome-wide transcriptional analysis of the mitotic cell cycle, Mol Cell (Band 2), Nr. 1, Seite 65-73.

[50] DeRisi, J. L.; Iyer, V. R. und Brown, P. O. (1997): Exploring the metabolic and genetic control of gene expression on a genomic scale, Science (Band 278), Nr. 5338, Seite 680-6. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.sciencemag.org/cgi/content/full/278/5338/680

[51] Kal, A. J.; van Zonneveld, A. J.; Benes, V.; van den Berg, M.; Koerkamp, M. G.; Albermann, K.; Strack, N.; Ruijter, J. M.; Richter, A.; Dujon, B.; Ansorge, W. und Tabak, H. F. (1999): Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources, Mol Biol Cell (Band 10), Nr. 6, Seite 1859-72. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.molbiolcell.org/cgi/content/full/10/6/1859

[52] Warner, Jonathan R (1999): The economics of ribosome biosynthesis in yeast, TIBS.

[53] Ueda, H. R.; Matsumoto, A.; Kawamura, M.; Iino, M.; Tanimura, T. und Hashimoto, S. (2002): Genome-wide transcriptional orchestration of circadian rhythms in Drosophila, J Biol Chem (Band 277), Nr. 16, Seite 14048-52. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jbc.org/cgi/content/full/277/16/14048

[54] Cherry, J. M.; Adler, C.; Ball, C.; Chervitz, S. A.; Dwight, S. S.; Hester, E. T.; Jia, Y.; Juvik, G.; Roe, T.; Schroeder, M.; Weng, S. und Botstein, D. (1998): SGD: Saccharomyces Genome Database, Nucleic Acids Res (Band 26), Nr. 1, Seite 73-9. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.oup.co.uk/nar/Volume_26/Issue_01/gkb024_gml.abs.html

[55] Mewes, H. W.; Frishman, D.; Guldener, U.; Mannhaupt, G.; Mayer, K.; Mokrejs, M.; Morgenstern, B.; Munsterkotter, M.; Rudd, S. und Weil, B. (2002): MIPS: a database for genomes and protein sequences, Nucleic Acids Res (Band 30), Nr. 1, Seite 31-4. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://nar.oupjournals.org/cgi/content/abstract/30/1/31

[56] Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M. und Sherlock, G. (2000): Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet (Band 25), Nr. 1, Seite 25-9. taf/DynaPage.taf?file=/ng/journal/v25/n1/full/ng0500_25.html taf/DynaPage.taf?file=/ng/journal/v25/n1/abs/ng0500_25.html. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.nature.com/cgi-taf/DynaPage.taf%3ffile=/ng/journal/v25/n1/abs/ng0500_25.html

[57] Harrison, P. M.; Kumar, A.; Lang, N.; Snyder, M. und Gerstein, M. (2002): A question of size: the eukaryotic proteome and the problems in defining it, Nucleic Acids Res (Band 30), Nr. 5, Seite 1083-90.

[58] Montalta-He, H.; Leemans, R.; Loop, T.; Strahm, M.; Certa, U.; Primig, M.; Acampora, D.; Simeone, A. und Reichert, H. (2002): Evolutionary conservation of otd/Otx2 transcription factor action: a genome-wide microarray analysis in Drosophila, Genome Biol (Band 3), Nr. 4.

[59] Ross Ihaka, Robert Gentleman (1996): R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics

Anhang

Anhang zu Kapitel 2

Grundlegende Notationen und Funktionen

Die Verteilungsfunktion $F(x)$ weist laut Definition jeder reellen Zahl x die Wahrscheinlichkeit zu, mit der eine Zahl kleiner oder gleich x gezogen wird, für diskrete Verteilungen lässt sich also schreiben:

$$F(x) = \sum_{x_i \leq x} P(X = x_i), \text{ wobei } X \text{ die Zufallsvariable bezeichnet.}$$

Quantile Ein Quantil ist ein Lokalisationsmaß, dass durch $F(x) = p$ definiert ist: $\mathbf{x_p}$ ist derjenige Wert einer stetigen Verteilung $F(x)$, bei dem die Wahrscheinlichkeit für einen kleineren Wert genau p und für einen größeren Wert genau $1-p$ beträgt. Für diskrete Verteilungen $F(x)$ definiert man das Quantil mit Hilfe von Ungleichungen: z_p ist eine Zahl, für die gilt $F(x) \leq p$, für alle $x < z_p$ und $F(x) \geq 1-p$, für alle $x > z_p$. Im diskreten Fall sind zwei Dinge zu beachten: (1) Das Quantil ist nicht immer eindeutig bestimmt. Und (2) die Verteilung kann an der Stelle des Quantils nicht definiert sein.

Beispiel: Sei $G = (1,2,3,4,4,7)$ die Grundgesamtheit aus der in einem Zufallsexperiment eine Stichprobe mit Zurücklegen gezogen wird. Die Verteilungsfunktion nimmt beispielsweise an der Stelle vier den Wert $5/6$ an, da fünf der sechs Zahlen der Grundgesamtheit kleiner oder gleich vier sind. In der für diese Arbeit verwendeten Implementierung ist der Median ($x_{0,5}$) in diesem Beispiel 3,5 und damit kein Element der Grundgesamtheit. Laut Definition ist jede reelle Zahl im Intervall [3,4] ein Median für diese Verteilung.

Einige Quantile sind für die Beschreibung von Verteilungen besonders wichtig, so vor allem der Median bei $p = 1/2$ und das erste und dritte Quartil bei $p = 1/4$ beziehungsweise $p = 3/4$. Ebenfalls verwendet werden Dezile und Perzentile, die es ermöglichen die Stichprobe in zehn beziehungsweise hundert gleichgroße Teile aufzuteilen.

Berechnung: Gegeben seien ein Vektor x mit Zahlen und eine Wahrscheinlichkeit p. Sei weiter ox der gleiche Vektor nur der Größe nach geordnet und n die Länge des Vektors. Dann lässt sich das Quantil mit folgender Funktion berechnen:

$$quantile(x, p) = (1-(r-\lfloor r \rfloor)) \cdot ox[\lfloor r \rfloor] \quad + \quad (r-\lfloor r \rfloor) \cdot ox[\lceil r \rceil],$$

$$\text{wobei } r = 1 + (n-1) \cdot p \text{ und } ox[n+1] = ox[n].$$

In der Arbeit verwendete statistische Tests

Nachfolgend werden der Wilcoxon-Test für Paardifferenzen und der U-Test näher erläutert.

Der Wilcoxon-Test für Paardifferenzen

Für den Vergleich zweier verbundener Stichproben, deren Differenzen nicht normalverteilt sind, ist der Vorzeichen-Rang-Test von Wilcoxon der optimale Test. Er stellt das nicht - parametrische Pendant zum T-Test dar und weist eine Effizienz von 95% auf. Das heißt, benötigt man beispielsweise für das Erreichen einer bestimmten Signifikanzschwelle mit dem Wilcoxon-Test einen Stichprobenumfang von 100, so bräuchte man mit dem T-Test einen von 95 für die gleiche Schwelle.

Es wird geprüft, ob die Differenzen der paarig angeordneten Beobachtungen symmetrisch mit dem Median gleich Null verteilt sind, oder anders ausgedrückt unter der Nullhypothese entstammen die Paardifferenzen d_i einer Grundgesamtheit mit der Dichte $f(d)$, wobei $H_0: f(d) = f(-d)$. Die Ablehnung von H_0 bedeutet, dass entweder der Median der Differenzen ungleich Null ist oder den beiden Stichproben unterschiedliche Verteilungen zugrunde liegen. Um die Testgröße zu erhalten, bildet man zuerst die Differenzen der Wertepaare $d_i = x_{i1} - x_{i2}$ und streicht alle die, deren Differenzen verschwinden ($d_i = 0$). Anschließend werden die Absolutbeträge der verbleibenden *n* Differenzen in ansteigende Rangordnung gebracht und die entsprechenden Rangzahlen bei eins beginnend zugeordnet. Bei gleichgroßen Beträgen ordnet man mittlere Rangzahlen zu. Anschließend werden alle Rangzahlen, deren Differenz ein positives Vorzeichen aufweist, zur positiven Rangsumme und alle, deren Differenz ein negatives Vorzeichen aufweist, zur negativen Rangsumme aufaddiert. Die kleinere der beiden Rangsummen gilt nun für den beidseitigen Test als Testgröße. Für Stichprobenumfänge $n \leq 25$ entnimmt man die kritischen Werte für

bestimmte Signifikanzschwellen aus statistischen Tafeln oder errechnet sie, in dem man alle möglichen Kombinationen von Rängen und Vorzeichen generiert und dann das kritische Perzentil der Verteilung bestimmt. Für größere Stichproben errechnet man die Approximation:

$$z = \frac{\left| R - \frac{n(n+1)}{4} \right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Der z–Wert wird dann anhand der Standardnormalverteilung beurteilt, wobei R die kleinere der beiden Rangsummen bezeichnet.

Der U-Test von Wicoxon, Mann und Whitney

Der U-Test ist ein Rangsummentest für den Vergleich zweier unabhängiger Stichproben bei nicht-normalverteilter Grundgesamtheit. Er bildet das Pendant zum T-Test bei unabhängigen Stichproben und weist ebenfalls eine hohe Effizienz von 95% auf. Als Voraussetzungen gelten, dass die beiden Stichproben aus stetigen Verteilungen mit ähnlicher bis gleicher Form stammen. Der Test prüft die Nullhypothese: Die Wahrscheinlichkeit, dass eine Beobachtung der ersten Grundgesamtheit größer ist als eine beliebige gezogene Beobachtung der zweiten Grundgesamtheit, ist gleich ½ oder als Formel $H_0 : P(X_1 > X_2) = \frac{1}{2}$. Für die einseitige Fragestellung formuliert man die Nullhypothese $H_0 : P(X_1 > X_2) \geq \frac{1}{2}$ oder $H_0 : P(X_1 > X_2) \leq \frac{1}{2}$. Als allgemeine Eigenschaften des Tests gelten, dass er empfindlich gegenüber Medianunterschieden, weniger empfindlich bei unterschiedlichen Schiefen der Verteilungen und gänzlich unempfindlich für Varianzunterschiede ist [8].

Für die Berechnung der Prüfgröße bringt man zuerst die $n = n_1 + n_2$ Beobachtungen beider Stichproben zusammen in Rangordnung und weist den Werten Rangzahlen von 1 bis n zu. Für jede Stichprobe berechnet man dann die Rangsummen R_1, R_2 und anschließend die Größen U_1, U_2 nach der Formel:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Die Prüfgröße ist dann die kleinere der beiden. Es gilt zusätzlich $n_1 n_2 = U_1 + U_2$, was zur Kontrolle der Rechnungen genutzt werden kann. Die kritischen Werte für bestimmte Signifikanzniveaus lassen sich exakt wieder direkt berechnen, falls der Stichprobenumfang klein ist, oder aus statistischen Tafeln ablesen. Für größere Stichproben $(n_1, n_2 \geq 8)$ lässt sich wiederum folgende Approximation nutzen:

$$z = \frac{\left| U - \frac{n_1 n_2}{2} \right|}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Der z–Wert wird dann anhand der Standardnormalverteilung beurteilt, wobei U die kleinere der beiden Prüfgrößen bezeichnet.

Nichtlineare Anpassung von Verteilungen

Ziel ist die Vergleichbarkeit von Chipexperimenten. Aufgrund des Mangels an Wiederholungsexperimenten lässt sich nicht einwandfrei entscheiden, welche Effekte von biologisch interessanten Unterschieden des Ausgangsmaterials stammen und welche allein technischer Natur sind. Bislang konnte auch noch kein überzeugendes mathematisches Modell gefunden werden, was Lage und Streuung der Daten hinreichend genau erklärt. Die Gesamtmenge der in jeden Versuch eingesetzten RNA ist konstant und alle Gene auf dem Chip sollten aufgrund ihrer großen Anzahl eine repräsentative Teilmenge aller Gene des Genoms darstellen. Das legt nahe, auch die den Signalen zugrund liegende Verteilung für jedes Chipexperiment als identisch anzunehmen. Die Verteilungen der Rohintensitäten sind nicht direkt vergleichbar und müssen daher normiert werden.

Voraussetzung: Es gibt annähernd so viele hoch- wie runterregulierte Gene mit vergleichbaren Expressionsniveaus. Zur Anpassung der Werte eines Experiments an die eines anderen wird eine robuste, glatte Regression mit Hilfe der R-Funktion *loess* durchgeführt. Glatt heißt, dass die Kurve, die die Transformation definiert keine Unstetigkeitsstellen aufweist. Die Funktion *loess* fittet eine polynomiale Oberfläche von einem oder mehreren numerischen Prediktoren und nutzt dabei lokale Anpassungen. Laut Venables und Ripley ist der benutzte Algorithmus relativ komplex. Es wird ein Fenster um x gelegt und Datenpunkte, die innerhalb des Fensters liegen, werden so gewichtet, dass diejenigen, die am nächsten an x liegen, die höchsten Gewichte erhalten. Anschließend wird eine robuste, gewichtete

Regression angewendet, um den Wert an der Stelle *x* zu schätzen. Über einen Parameter lassen sich die Fenstergröße und der Anteil der Daten, die in die Analyse eingehen, steuern.

Das Prinzip lässt sich exakter wie folgt formulieren: Man fittet ein Polynom *d*-ten Grades an die Punktwolke um x_0, und nutzt das Ergebnis zur Anpassung in einem einzigen Punkt x_0. Der Beitrag, den jeder Punkt in der Umgebung von x_0 beisteuert, wird mit Hilfe einer so genannten Kernel-Funktion bestimmt, je näher der Punkt an x_0 desto größer sein Gewicht. Die Bewertung der Güte der Anpassung erfolgt mit der Methode der kleinsten Quadrate.

$$\min_{a(x_0),\, b_j(x_0),\, j=1,\mathrm{K},d} \sum_{i=1}^{N} K_l(x_0, x_i) \left[y_i - a(x_0) - \sum_{j=1}^{d} b_j(x_0)\, x_i^j \right]^2$$

Die Lösung ist ein Polynom vom Grad *d*.

$$\hat{f}(x_0) = \hat{a}(x_0) + \sum_{j=1}^{d} \hat{b}(x_0)\, x_0^{\ j}$$

Algorithmen von MAS 5.0

Die Berechnungen der repräsentativen Expressionswerte und der Expressionsänderung zwischen einer Probe und einem Referenzchip sind grundlegend verschieden von denen in Kapitel 2 und sind deshalb genau beschrieben. Die Hintergrundkorrektur ist dahingehend ausgebaut worden, dass die lokalen Hintergrundwerte eine glatte, stufenlose Fläche bilden.

Hintergrundkorrektur

Aufgrund der 2% niedrigsten Werte wird ein lokaler Hintergrund für jede Bildzelle bestimmt, und dieser vom Intensitätswert abgezogen. Das Array wird in *K* (üblicherweise *K=16*) rechteckige Regionen $(R_k), k = 1,...,K$ aufgeteilt. Innerhalb dieser berechnet man dann den Mittelwert HR_k und die Standardabweichung SR_k der 2% niedrigsten Werte. Der Hintergrundwert für eine Bildzelle (*x*,*y*) ergibt sich aus einer gewichteten Summe der Hintergrundwerte der Regionen:

$$hg(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y) \cdot HR_k \qquad (1)$$

Die Gewichtsfunktion ist gegeben durch: $w_k(x,y) = \frac{1}{d_k^{\,2}(x,y) + smooth}$ wobei $d_k(x,y)$ der Euklidische Abstand der Bildzelle vom Zentrum der Region k ist. Und *smooth* ist ein Parameter, der dafür sorgt, dass der Nenner immer größer Null ist (Voreinstellung: *smooth* = 100). Je näher die Bildzelle an einer bestimmten Region, desto größer ist ihr Gewicht. Auf analoge Weise berechnet man einen lokalen Rauschwert für jede Bildzelle *r(x,y)*, in dem man in die Formel (1) anstelle der Mittelwerte HR_k die Standardabweichungen SR_k der Regionen einsetzt. Bezeichne *I(x,y)* die Intensität einer Bildzelle, so berechnet sich der hintergrundkorrigierte Wert durch die Formel:

$$\tilde{I}(x,y) = \max(\ I(x,y) - hg(x,y),\quad NoiseFrac \cdot r(x,y)\ )$$

NoiseFrac ist ein Parameter mit dem man spezifiziert, welcher Bruchteil des Rauschwertes die untere Schranke der Signalwerte bilden soll. (Voreinstellung: *NoiseFrac* = 0,5) Nach dieser Korrektur hat man eine Nivellierung regionaler Schwankungen der Intensitäten innerhalb eines Chips erreicht.

Biweight-Verfahren von Tukey

Dieser Algorithmus ist ein modernes Verfahren zur Berechnung eines Schätzers für das Zentrum der zugrunde liegenden Verteilung, der gegenüber Ausreißern robust ist.

Gegeben seien ein Zahlenvektor $x = x_1, \mathrm{K}, x_n$ der Länge n. M sei der Median der Zahlen und S der Median der absoluten Distanzen der Einzelwerte von M. S stellt ein robustes Streuungsmaß dar.

Das Verfahren wichtet jeden Einzelwert bezüglich seines Abstandes zum Median, je weiter weg desto geringer. Werte die sehr weit vom Median abweichen, werden mit Null gewichtet. Man berechnet zuerst $u_i = \frac{x_i - M}{c \cdot S + e},\quad i = 1, \mathrm{K}, n$ wobei e eine kleine positive Zahl ist um Divisionen durch Null zu vermeiden und c ein fester Parameter mit dem die Stärke der Gewichtung bezüglich der vorliegenden Daten angepasst werden kann. Die Voreinstellung in der MAS 5.0 Software von Affymetrix sind $e = 0{,}0001$ und $c = 5$. Dann lassen sich mit diesem Distanzmaß für jeden Einzelwert des Vektors die Gewichte mit der folgender Funktion bestimmen:

$$w(u_i) = \begin{cases} (1-u_i^{\,2})^2, & |u_i| \le 1 \\ 0, & |u_i| > 1 \end{cases}$$

Tukeys Biweight-Schätzer wird berechnet als $T = \frac{\sum_i w(u_i) \cdot x_i}{\sum_i w(u_i)}$. Die Funktion, die diesen Schätzwert berechnet wird innerhalb der vorliegenden Arbeit mit ***tukey()*** bezeichnet.

Berechnung des repräsentativen Expressionswerts

Ausgangspunkt sind die hintergrundkorrigierten Intensitäten. Für jedes Sondenset auf jedem Chip wird ein so genannter Signalwert berechnet, der die Menge der Transkripte in der Hybridisierungslösung repräsentieren soll. Zuerst werden die Werte jedes Oligopaares (*PM, MM*) miteinander verrechnet. Das Signal des *MM*-Oligos repräsentiert den spezifischen Hintergrund, der sich durch Subtraktion vom *PM*-Signal eliminieren lässt. Das *MM*-Signal wird nur benutzt, wenn es kleiner als das *PM*-Signal ist, um das Auftreten negativer Expressionswerte zu verhindern. Als Zwischenstufe gibt es deshalb einen idealisierten Missmatch (*IM*).

$$V = \max(\ PM - IM, d\ ), \text{ Voreinstellung: } d = 10^{-20}$$

Der idealisierte Missmatch-Wert ergibt sich aus folgender Formel:

$$IM = \begin{cases} MM, \ falls\ MM < PM \\ \frac{PM}{2^{SB}}, \ falls\ MM \ge PM\ und\ SB > contrast \\ \frac{PM}{2^{y}}, \ falls\ MM \ge PM\ und\ SB \le contrast \end{cases}$$

Dabei ist $y = \frac{contrast}{1+\left(\frac{contrast - SB}{scale}\right)}$ eine wenig kleinere Zahl als *contrast*. Die Parameter *contrast* und *scale* dienen der Beschränkung des Wertebereichs. Voreinstellungen: *contrast* = 0,03 und *scale* = 10. *SB* ist der für das Sondenset, zu dem die Intensität gehört, spezifische Kontrastwert zwischen den *PM*- und *MM*-Signalen, der mit Hilfe von Tukeys Biweight-Verfahren berechnet wird:

$$SB = tukey(\log_2(PM_i) - \log_2(MM_i) : \forall\ Oligopaare\ i\ des\ Sondensets)$$

Lässt sich ein Transkript mit dem Sondenset sicher detektieren, so ist die Differenz (*PM-MM*) für die meisten Oligopaare groß und damit auch Tukeys Biweight-Schätzer. Für jedes Sondenset kann nun ein *LogSignal*-Wert berechnet werden:

$$LogSignal = tukey(\log_2(V_i) : \forall\ Oligopaare\ i\ des\ Sondensets)$$

Als Expressionswerte gibt das Programm für jedes Sondenset die rücktransformierten und skalierten Signalwerte aus:

$$reportedValue = nf \cdot sf \cdot 2^{LogSignal}$$

Der Parameter *nf* ist der Normierungsfaktor für paarweisen Chipvergleich und ergibt sich aus dem Quotienten der getrimmten Mittelwerte von Referenzchip und Experiment. Für Einzelchipanalysen gilt *nf* = 1. Der Skalierungsfaktor *sf* transformiert alle Expressionswerte eines Chips so, dass das Zentrum ihrer Verteilung dann etwa bei dem vom Anwender spezifizierbaren Zielwert (*T*, Voreinstellung: *T* = 500) liegt.

$$sf = \frac{T}{trimMean(2^{LogSignal_j} : \forall\ Sondensets\ j; 0{,}02; 0{,}98)}$$

In dieser Formel bezeichnet *trimMean* das getrimmte Mittel: Die höchsten und niedrigsten 2% der Werte gehen nicht in die Mittelwertbildung ein.

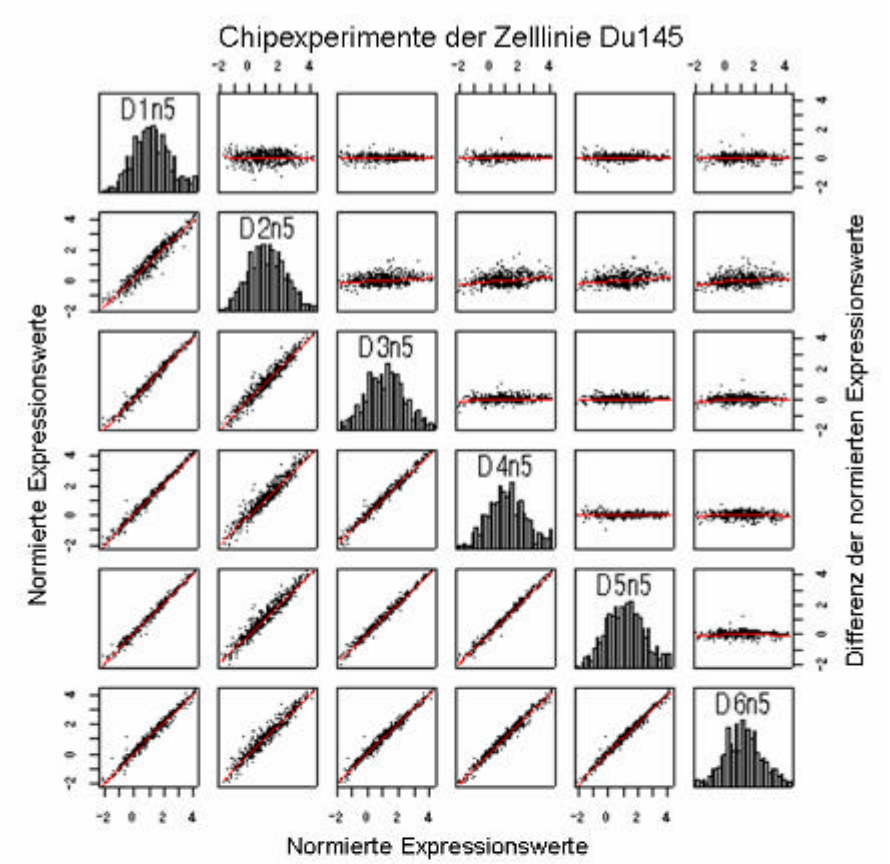
Berechnung relativer Expressionsänderungen

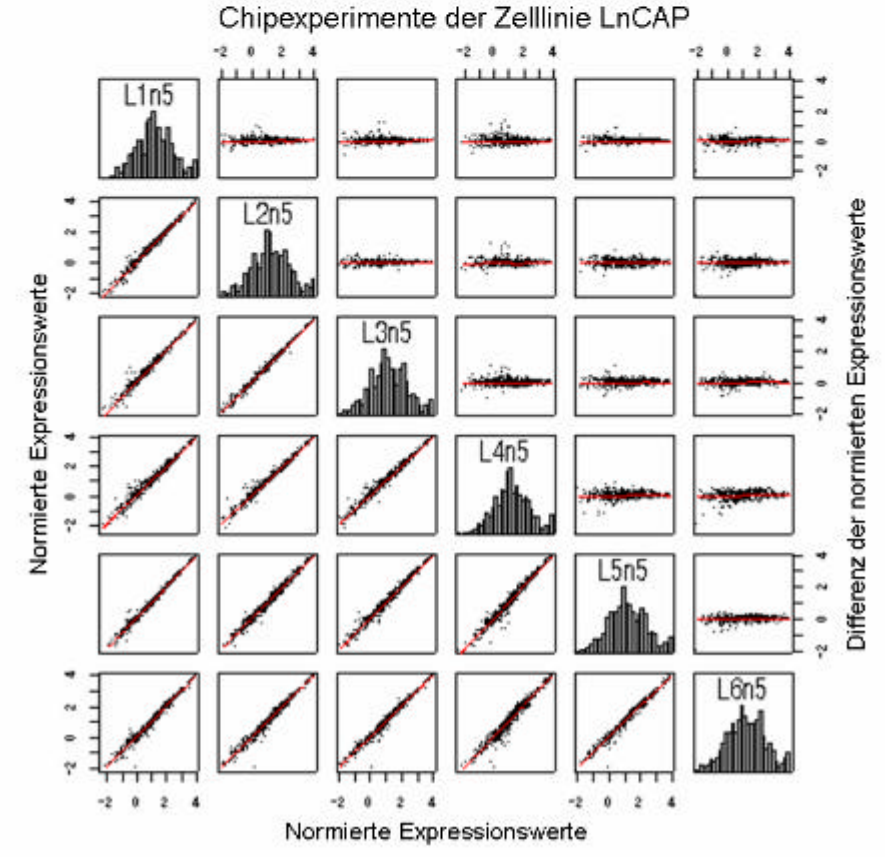
Die Analysesoftware MAS 5.0 bietet die Möglichkeit, Chipexperimente paarweise zu vergleichen. Für jedes Sondenset wird dazu ein Wert für die Expressionsänderung (*LogRatio*) berechnet. Dabei geht man von den verrechneten Signalen der Oligopaare aus: $V = \max(\ PM - IM, d\ )$, siehe oben. Um über die Chips vergleichbare Intensitäten zu erhalten, wendet man auf diese Werte die Skalierung an, die bei der Einzelchipauswertung im letzten Schritt zu den ausgegebenen Expressionswerten führt: $\tilde{V} = nf \cdot sf \cdot V$. Zur Berechnung der Expressionsänderung für ein bestimmtes Sondenset zwischen der Probe (*e - experiment*) und der Referenz (*b - baseline*) bedient man sich abermals des Verfahrens von Tukey:

$$LogRatio = tukey(\log_2(\tilde{V}_i^e) - \log_2(\tilde{V}_i^b) : \forall\ Oligopaare\ i)$$

Parallel wird ein 95%-Vertrauensbereich für die *LogRatio*-Werte bestimmt.

Reproduzierbarkeit der Zelllinienexperimente (Abschnitt 2.5)





Die Tabellen der Datenbank

Es folgt eine Auflistung der Tabellen mit den Attributen in Klammern und eine kurze Beschreibung deren Inhalte, falls die Namen nicht selbsterklärend sind. Die Tabellennamen sind fett gedruckt und die Spaltennamen kursiv. *ID* (identifier) ist, falls vorhanden, immer der Primärschlüssel der Tabelle, dass heißt jede Zeile der Tabelle hat einen Eintrag in dieser Spalte, und es gibt keinen Wert in dieser Spalte, der in zwei Zeilen vorkommt. Sind in einer Spalte Primärschlüssel einer anderen Tabelle aufgeführt, so bezeichnet man diese als Fremdschlüssel. Wenn möglich setzen sich die Namen dieser Spalten immer aus dem Namen der fremden Tabelle und dem Kürzel *ID* zusammen, also zum Beispiel *Patient_ID* für die Patientennummer als Fremdschlüssel in der Tabelle **Case**. Es ist nicht gefordert, dass für Primärschlüssel Nummern vergeben werden müssen. Hat man für bestimmte Entitäten umkehrbar eindeutige Namensbereiche vorliegen, so können diese auch als Primärschlüssel Verwendung finden. Affymetrix vergibt beispielsweise Namen für ihre Chips, die in der Tabelle *Array_Type* in der Spalte *ID* benutzt werden. Mit dem Vorteil, dass man keine zusätzlichen Bezeichner einführen muss, handelt man sich allerdings eventuell Probleme mit der Klein-, Großschreibung und mit Leerzeichen ein. In der Spalte *Operator* speichert man entweder die Person, die das Experiment durchgeführt oder diejenige, die die Daten eingetragen hat. Die Spalten *Description* und *Remark* enthalten Freitext, müssen aber nicht gefüllt sein.

Probendaten

1. **Patient** (*ID, External_ID, Sex, Date_of_Entry, First_Diagnosed, Age_at_Diagnose, Survival, Status, Description, Operator*): In dieser Tabelle sind alle auswertungsrelevanten Personendaten gespeichert, die nicht spezifisch für den Arztbesuch beziehungsweise für die Behandlung sind. Die Patientendaten sind an dieser Stelle schon anonymisiert, Name und Adresse sind also nicht verfügbar. Benötigt man später weitere Informationen zum Krankheitsverlauf, so muss man sich mit der externen Patientennummer an den behandelnden Arzt wenden.

2. **Case** (*ID, Source, Patient_ID, External_ID, T, N, M, R, L, V, TNM_Annex, Summary_Stage, Date_of_Entry, Operator, Smoking, Site_of_Relapse, Description*): Die Falldaten bezeichnen alle Patientendaten, die zum Zeitpunkt einer Behandlung oder bei einem Arztbesuch anfallen und nicht allgemein für den Patienten gelten. Patient_ID verweist auf die Tabelle *Source* verzeichnet den Kooperationspartner oder

die Klinik. Zusammen mit der externen Fallbezeichner (*External_ID*) ist man in der Lage, den Fall zurück zu verfolgen. Die Spalten *T* bis *Summary_Stage* dienen der Dokumentation der Tumorklassifikation[1]. Da das Zigarettenrauchen als Hauptrisikofaktor für Krebserkrankungen anerkannt ist und zwar nicht nur für Lungenkrebs, sind die Rauchgewohnheiten in der Spalte *Smoking* erfassbar. Die Angabe erfolgt in Schachteln pro Jahr. Ist die Erkrankung nicht zum ersten Mal ausgebrochen, sondern nach einer Behandlung zurückgekehrt, so dokumentiert man den Ort des erneuten Auftretens im Organ in der Spalte *Site_of_Relapse*. Krankheitenspezifische Daten sind separate in den Tabellen **Therapy** und **Diagnosis** erfasst.

3. **Therapy** (*Case_ID, ID, Age_at_Therapy, Remark, Operator*): Für die unterschiedlichen Krebserkrankungen gibt es die unterschiedlichsten Therapieformen, die in dieser Tabelle für jeden Fall erfasst werden. Wichtig ist hier, dass die Bezeichnungen (*ID*) einem durch Fachwissenschaftler (in unserem Fall Pathologen) kontrolliertem Vokabular entsprechen, dass sorgfältig gepflegt sein sollte.

4. **Diagnosis** (*Case_ID, ID, Remark, Operator*): Ähnlich der Tabelle Therapy warden hier mit Hilfe eines kontrollierten Vokabulars die für die Auswertung wichtigen diagnostischen Daten für jeden klinischen Fall dokumentiert.

5. **Block** (*ID, Case_ID, Localisation, External_ID, Tissue_Type, Material_Type, Tissue_Evaluation, External_Evaluation, Micro_Dissection, Organ, Metastasis_from, Operator, Date_of_Entry*): Wird ein Tumor operiert und kann davon ein Gewebsstückchen für weitere Untersuchungen entnommen werden, so nennt man dieses einen Block. Von einem Tumorpräparat können mehr als ein Block gewonnen werden, beispielsweise einer von der Invasionsfront am Rand zum gesunden Gewebe und einer vom Resttumor. Zu einem Fall können auch ein Block aus dem Primärtumor und einer aus einer Metastase generiert worden sein. Von den Blöcken fertigt man Gewebsschnitte an, an denen die Befundung stattfindet. In der Spalte *Localisation*

---

[1] Internationale TNM-Klassifikation:

Siehe zum Beispiel http://www.hc-sc.gc.ca/hpb/lcdc/bc/ccocs/lungpoumon/lcg_i_e.html

dokumentiert man die Lage des Blocks im Entnahmeorgan und in *Tissue_Type* eine standardisierte Grobklassifikation für Gewebeproben. In der Spalte *Material_Type* ist aufgeführt in welcher Form die Probe vorliegt als Gefriermaterial in Parafin eingebettet, bereits isolierte RNA oder Zellen aus einer Zellkultur. *Tissue_Evaluation* enthält als Freitext die interne pathologische Beurteilung (Befundung) und *External_Evaluation* die Befundung des behandelnden Arztes. Die Spalte *Micro_Dissection* enthält einen Boolschen Wert (yes/no), der besagt, ob das Gewebe für die Mikrodissektion eignet ist. In *Organ* ist der Ort des Blockes erfasst und handelt es sich dabei um eine Metastase, so trägt man den Ort des Primärtumors in *Metastasis_from* ein.

6. **Pathological_Findings** (*Block_ID, Finding, Percent, Remark, Operator*): Für einen Block können Beurteilungen verschiedener Pathologen vorliegen, die in dieser Tabelle erfasst werden. Außerdem lassen sich in dieser Tabelle Eigenschaften für eine Probenklassifikation speichern. Wurden beispielsweise die Patientinnen einer Brustkrebsstudie auf Mutationen in dem Gen *BRCA1* untersucht, so lässt sich der Status in dieser Tabelle erfassen und später mit den Expressionsdaten in Beziehung setzen. Der Eintrag in *Finding* sollte wieder einem kontrollierten Vokabular unterliegen. In *Percent* wird die Schätzung erfasst, für wie viele Tumorzellen die Beobachtung gilt.

7. **Cell_Line** (*ID, ATCC_ID, Synonym, Provider, Organ, Tissue, Morphology, Growth_Property, Mouse, Supplement, Medium, Remark, Date_of_Provision, Operator*) Zelllinien werden bei metaGen nicht selbst etabliert sondern von komerziellen Anbietern erworben oder von Kooperationspartnern bereitgestellt. ATCC[1] ist ein globaler Anbieter von Zellkulturen, der ein umfangreiches Sortiment an Zelllinien vorhält. Außerdem wurden dort systematisch Bezeichnungen (ATCC_ID) eingeführt. Die Spalte Synonym wird genutzt, falls im Labor ein anderer Name für die Zelllinie gebräuchlich ist. Die Spalten *Organ*, *Tissue* und *Morphology* beschreiben das Gewebe beziehungsweise die Zellen, aus dem die Zelllinie ursprünglich etabliert wurde. In Growth_Property erfasst man, ob die Zellen zum Beispiel nur einschichtig

[1] ATCC: American Type Culture Collection, Manassas, USA

wachsen. In der Spalte *Mouse* wird vermerkt, ob diese Zelllinie bereits für Experimente in Mäusen eingesetzt wurde. In *Supplement* und *Medium* vermerkt man die optimalen Wachstumsbedingungen für Kulturen dieser Zelllinie.

8. **Cell_Line_Property** (*Cell_Line_ID, Property, Operator*): In dieser Tabelle werden weitere spezielle Eigenschaften von bestimmten Zelllinien erfasst.

9. **Cell_Culture** (*ID, Cell_line_ID, Passage, Stadium_on_Harvest, Remark, Operator, Date*): Als Zellkultur bezeichnet man den vorliegenden Zellpool einer bestimmten Zelllinie, also sich teilende und wachsende Zellen vom selben Ursprung. Als Passage bezeichnet man einen Zyklus, in dem die Zellen einer Kultur aufgetaut, im Medium zum Wachsen angeregt und anschließend wieder eingefroren werden. Bringt man die Zellen in ein Gefäß mit Medium, so wachsen die Zellen bis zu einer bestimmten Dichte mit exponentieller Rate und dann gehen sie in den stationären Zustand über. In welcher Wachstumsphase sie geerntet wurden, ist in *Stadium_on_Harvest* gespeichert.

10. **Test_Probe** (*ID, Type, Operator, Date*): In dieser Tabelle sind speziell für Test- und Kontrollzwecke generierte Proben dokumentiert. Beispielsweise wurde bei metaGen mal DNA markiert und auf den Chip gebracht.

11. **Probe_Pool** (*ID, Cell_Pool_ID, Remark, Operator, Date*): Für manche Experimente ist es wichtig, Proben zu poolen. Konnte man beispielsweise bei der Mikrodissektion eines Blockes nicht genug RNA für ein Chipexperiment gewinnen und es wurde bereits vom gleichen Patienten ein zweiter Block bearbeitet, so lassen sich diese Proben eventuell poolen. Bei Experimenten mit cDNA-Arrays erzeugt man oft eine Referenzprobe durch das Zusammenlegen verschiedener Einzelproben. Cell_Pool_ID ist eine Verallgemeinerung und verweist in Abhängigkeit von der Probe auf den Primärschlüssel einer der folgenden Tabellen: **Block, Cell_Culture** oder **Test_Probe.**

12. **RNA** (*ID, Cell_Pool_ID, Cell_Pool, Start_Amount, TaqMan_CT1, Amount_IVT1, Amount_Used, TaqMan_CT2, Amount_IVT2, Amount_IVT3, cRNA, Protocol, Operator, Date*): In dieser Tabelle sind die für die Auswertung wichtigen Parameter der RNA-Aufbereitung aufgeführt. Das Standardprotokoll und die während dessen erhobenen Kontrollparameter sind in Kapitel 1 beschrieben. *Cell_Pool* enthält die Art der eingesetzten Zellen und *Cell_pool_ID* einen Fremdschlüssel auf die Tabellen: **Probe_Pool, Block, Cell_Culture** oder **Test_Probe.** Welches Protokoll man benutzt

hat, vermerkt man in der Spalte *Protocol*. In den Spalten **TaqMan_CT1** und *TaqMan_CT2* vermerkt man die Ergebnisse der quantitativen PCR (Zyklenzahlen) der zwei konstitutiv exprimierten Gene Succinatdehydrogenase und GAP-Dehydrogenase.

Sequenzdaten

1. **Protein** (*ID, DB, Length, Header, Weblink, Sequence, Remark*): Diese Tabelle enthält Informationen über Proteinsequenzen aus öffentlichen Datenbanken. *DB* enthält einen eindeutigen Bezeichner für jede Datenbank und *ID* enthält den Sequenzbezeichner innerhalb dieser. *Header* ist ein Textfeld für eine in der DB verfügbare Beschreibung der Sequenz. In *Remark* können zusätzliche Kommentare eingetragen werden. Falls die Quellen über das Internet erreichbar sind, so wird der Verweis in *Weblink* eingefügt.

2. **Prot_Module** (*Prot_ID, Prot_DB, Module, Module_DB, Module_Length, Description, Method, Total, Number, Start_on_Prot, End_on_Prot, Score, E_Value, Date*): Proteinmodule sind Abschnitte (Position: *Start_on_Prot, End_on_Prot*) in Proteinsequenzen, denen sich über Sequenzähnlichkeit gewisse Eigenschaften zuweisen lassen. *Prot_ID, Prot_DB* verweisen als Fremdschlüssel auf **Protein**. In *Method* erfasst man das Verfahren mit dem die Module in der Sequenz identifiziert wurden und in *Score, E_Value* die Parameter beziehungsweise die Qualität des Treffers. Verschiedentlich findet man auch ein Modul in mehreren Kopien in einer Proteinsequenz, was dann in *Total* (Gesamtzahl), *Number* (Zähler) vermerkt wird.

3. **Prot_TM** (*Prot_ID, Prot_DB, Start_on_Prot, End_on_Prot, Total, Number, Score, Method, Date*): Diese Tabelle enthält die Ergebnisse einer Vorhersage transmembraner Bereiche in Proteinsequenzen. Die Bedeutung der Spaltennamen entspricht der in der Tabelle **Prot_Module**.

4. **Prot_SigPep** (*Prot_ID, Prot_DB, Cleavage_Site, Score, Method, Date*): Es gibt Motive in Proteinsequenzen, die als Transportsignal bei ihrer Synthese innerhalb der Zelle dienen und damit die Lokalisation des Ausgereiften Moleküls bestimmen. Man bezeichnet solche Motive als Signalpeptide und kann sie algorithmisch in Proteinsequenzen vorhersagen. An einer bestimmten Position (*Cleavage_Site*) wird das Peptid später abgespalten. Die Bedeutung der anderen Spaltennamen entspricht der in der Tabelle **Prot_Module**.

5. **cDNA** (*ID, DB, Length, Header, Weblink, Cluster, Cluster_DB, OMIM, Chrom_Location, Remark*): Diese Tabelle enthält Informationen über cDNA-Sequenzen aus öffentlichen Datenbanken. Die Spalte *DB* enthält einen eindeutigen Bezeichner für jede Datenbank, und *ID* enthält den Sequenzbezeichner innerhalb dieser. *Header* ist ein Textfeld für eine öffentlich verfügbare Beschreibung der Sequenz. In *Remark* können zusätzliche Kommentare eingetragen werden. Falls die Quellen über das Internet erreichbar sind, so wird der Verweis in *Weblink* eingefügt. Ist die Sequenz in einer Datenbank von Sequenzclustern wie zum Beispiel UniGene enthalten, so vermerkt man die Datenquelle und den Clusternamen in *Cluster_DB* und *Cluster*. *OMIM*[1] ist eine Datenbank, in der bekannte vererbbare Eigenschaften und die involvierten Gene dokumentiert sind. *Chrom_Location* enthält die chromosomale Lokalisation des Gens.

6. **cDNA_Repeat** (*cDNA_ID, cDNA_DB, Repeat, Repeat_Familie, Start_on_cDNA, End_on_cDNA, Method, Date*): Diese Tabelle dokumentiert bekannte repetitive Sequenzabschnitte, wenn sie in cDNA's gefunden wurden. Die Position auf der cDNA-Sequenz ist in *Start_on_cDNA, End_on_cDNA* festgehalten.

7. **cDNA_Alias** (*ID_a, DB_a, ID_b, DB_b, Method, Date*): Diese Tabelle nutzt man für die Identifizierung von cDNA-Sequenzen, von denen viele in mehreren Datenbanken gespeichert aber unterschiedlich benahmt werden.

8. **cDNA_Protein** (*cDNA_ID, cDNA_DB, Prot_ID, Prot_DB, Method, Date*): Ist zu einer mRNA ein translatiertes Protein bekannt, so wird diese Relation zwischen den Tabellen **cDNA** und **Protein** in dieser Tabelle erfasst.

9. **BLAST** (*Query_ID, Query_DB, Query_Length, Method, Hit_ID, Hit_DB, Hit_Header, Hit_Length, E_Value, Score, HSP_E_Value, HSP_Score, HSP_Length, HSP_Identical, HSP_Percent_Identical, HSP_QStart, HSP_QEnd, HSP_HStart, HSP_HEnd, Frame, Orientation, Date*): Das Computerprogramm BLAST führt einen Sequenzvergleich zwischen der *Query*-Sequenz und der *Hit*-Sequenz durch und erzeugt eine komplexe aber standardisierte Ausgabe. Diese Ausgabedatei lässt sich

---

[1] OMIM: Online Mendelian Inheritance in Man (am NCBI verfügbar)

parsen und die für weitere Analysen wichtigen Parameter können anschließend in die Datenbanktabelle überführt werden. Die einzelnen Einträge sind an dieser Stelle nicht erläutert, sollten aber für jemanden, der mit BLAST vertraut ist, zu verstehen sein.

10. **Oligo** (*Array_Type_ID, Probeset, Reverse_Complement, Probe_Pair, Middle_Position, Sequence, N_A, N_C, N_G, N_T, Remark*): In dieser Tabelle sind die Daten zu allen Oligos der verschiedenen Oligo-Arrays zusammengetragen. *Array_Type_ID* bezeichnet den Chip, um den es sich handelt. Die 25 Basen lange Sequenz ist eingetragen (*Sequence*), sowie ihre Position auf der cDNA (*Middle_Position*). In den Spalten *N_A* bis *N_T* wird die Basenzusammensetzung der Oligosequenz vermerkt und in *Reverse_Complement* die 5'-3'-Orientierung.

11. **Oligo_BLAST** (*Array_Type_ID, Probeset, Probe_Pair, DB, DB_Date, Hit_ID, Hit_Header, Orientation, Hit_Length, Position, Identicals*): Wird mit den einzelnen Oligos eine Suche in einer Sequenzdatenbank durchgeführt, so kann man das Ergebnis in dieser Tabelle ablegen. Die ersten drei Spalten spezifizieren das Oligonukleotid und die folgenden beschreiben die getroffene Sequenz (*Hit*) und die Datenquelle. In der Spalte *Identicals* ist gespeichert, wie viele der 25 Basen tatsächlich exakt die Sequenz treffen. Bei so kurzen Sequenzen ist das meistens ein besseres Maß für die Qualität des BLAST-Treffers als der berechnete Score.

Experimentdaten

1. **Array_Type** (*ID, Type, File_Name, File_Location, Source, Operator, Remark):* In dieser Tabelle findet man alle Arrays, für die Daten in der Datenbank verfügbar sind. Type ist zurzeit Oligo oder cDNA. Meistens ist das Array-Format ursprünglich in einer Datei abgelegt, deren Name, Speicherort und Herkunft in den Spalten *File_Name, File_Location* und *Source* vermerkt sind.

2. **Oligo_Array** (*Array_Type_ID, Probeset, Probe_Pair, Probe_Type, X_Coord, Y_Coord, Property*): *Probeset* ist der Bezeichner des Sondensets auf einem bestimmten Array. Die Spalte *Probe_Pair* enthält die Nummer des Oligopaares innerhalb dieses Sondensets und *Probe_Type* nimmt die Werte „*PM*" (perfect match), „*MM*" (missmatch) oder „C" für Kontrollregionen an. In *X_Coord* und *Y_Coord* sind die physischen Koordinaten der Bildzelle auf dem Chip abgelegt. Und *Property*

schließlich wird für eine Klassifikation der Sondensets genutzt. Beispielsweise haben alle bakteriellen Spike-Kontrollen hier den gleichen Wert.

3. **cDNA_Array** (*Array_Type_ID, Spot, Sequence_ID, X_Coord, Y_Coord, Property*): Spot ist ein eindeutiger Bezeichner einer cDNA-Sonde auf einem bestimmten Array und Sequence_ID ist der Name der cDNA-Sequenz. Diese Trennung ermöglicht die Unterscheidung von mehreren Spots der gleichen cDNA auf ein und demselben Array. In *X_Coord* und *Y_Coord* sind wiederum die physischen Koordinaten des Spots auf dem Array abgelegt. Und *Property* wird für eine Klassifikation der cDNA-Sonden genutzt.

4. **Array_Experiment** (*ID, RNA_ID, File_Name, File_Location, Array_ID, Array_Type_ID, Hyb, Scan, Lot, Operator, Remark, Protocol, Date, QC_Flag*): Genau genommen ist der Name der Tabelle irreführend. Intuitiv würde man eher eine Hybridisierung als Arrayexperiment ansehen. In diese Tabelle wird aber für jedes Bild eines Arrays, das gescannt und durch die Bildanalyse gegangen ist, eine Zeile eingefügt. Beispielsweise könnten auf einem physischen Chip zwei Hybridisierungen vorgenommen werden, wobei pro hybridisiertem Chip zwei gescannte Bilder entstehen. Die Daten liegen in den meisten Fällen vorher als Dateien im Verzeichnisbaum vor, deren Namen und Quellen in den Spalten *File_Name* und *File_Location* gespeichert werden. *RNA_ID* verweist auf die Probe, die auf das Array hybridisiert wurde. *Array_ID* gibt einem die Möglichkeit, die Arrays physisch zu identifizieren und *Lot* enthält die Bezeichnung der Produktionsserie zur Aufdeckung systematischer Fehler. In *Hyb* und *Scan* werden Hybridisierungen und die Scans eines Arrays nummeriert. *Protocol* eröffnet die Möglichkeit, Arrays nach unterschiedlichen Labormethoden zu klassifizieren. Die wichtigste Anwendung ist bisher, Experimente aus weiteren Analysen auszuschließen, die nicht nach dem Standardprotokoll durchgeführt wurden. Mit Hilfe von *QC_Flag* lassen sich Datensätze markieren, für die nachträglich Qualitätsprobleme konstatiert wurden.

5. **cDNA_Array_Raw** (*Experiment_ID, Spot, Channel, Mean_Signal, Stdv_Signal, Mean_BG, Stdv_BG, QC_External, QC_Flag*): In dieser Tabelle werden die unbearbeiteten Intensitätswerte von cDNA-Array-Experimenten gespeichert. *Experiment_ID* ist ein Fremdschlüssel und verweist auf die Tabelle **Array_Experiment**, und damit ist auch spezifiziert um welches Array es sich handelt.

Spot ist für einen bestimmten Arraytyp ein eindeutiger Bezeichner. Bei komparativer Hybridisierung kann durch *Channel* der Farbkanal identifiziert werden. Das Signal und der Hintergrund sind von der Bildanalysesoftware jeweils in Form eines Lage- und eines Streuungsmaßes bereits geschätzt worden und in den Spalten *Mean_Signal, Stdv_Signal, Mean_BG* und *Stdv_BG* abgelegt. In den *QC*-Spalten sind wiederum Qualitätsparameter gespeichert, die ein Filtern der Daten für Analysezwecke unterstützt.

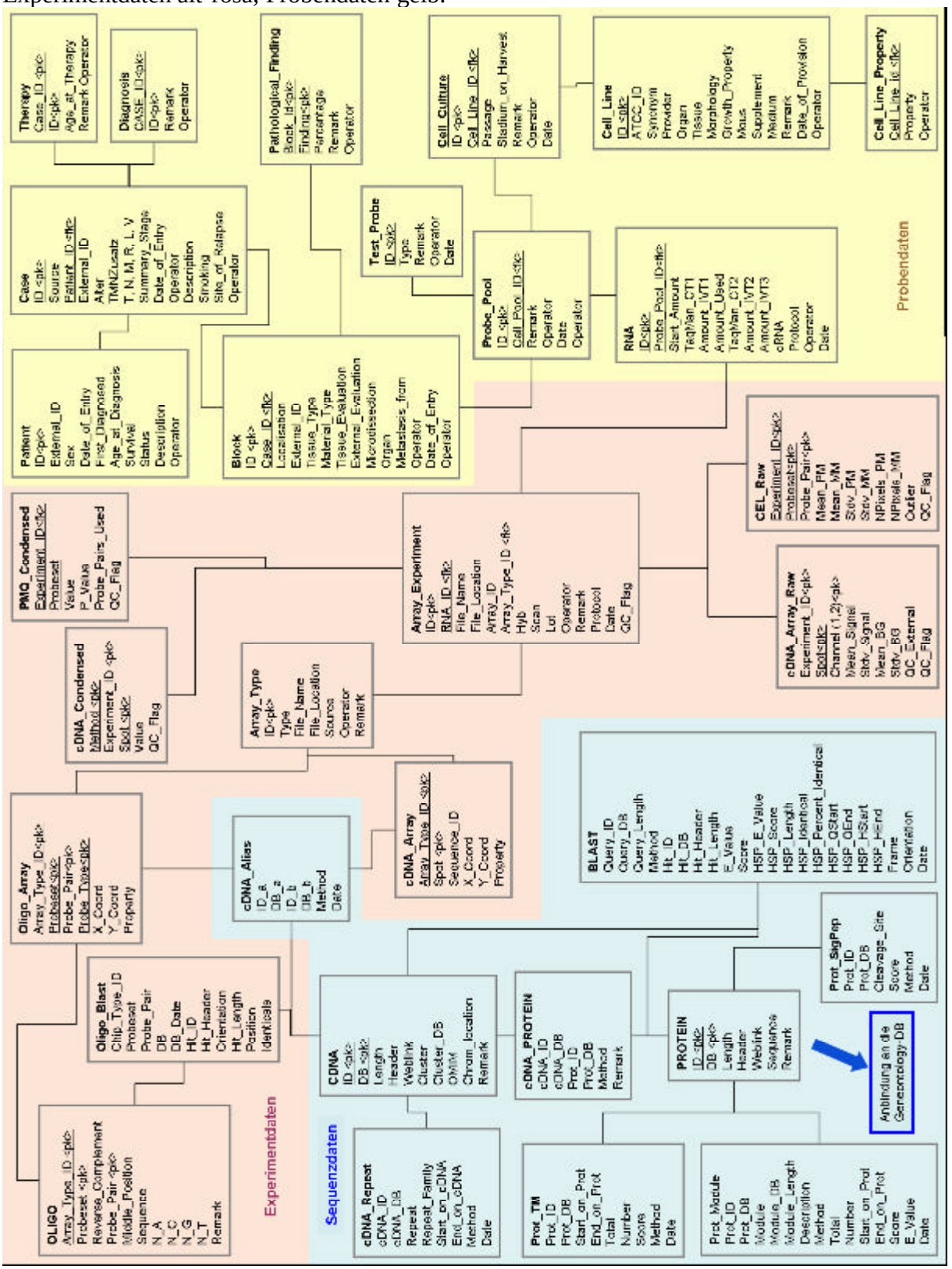
6. ***CEL*_Raw** (*Experiment_ID, Probeset, Probe_Pair, Mean_PM, Mean_MM, Stdv_PM, Stdv_MM, NPixels_PM, NPixels_MM, Outlier, QC_Flag*): In dieser Tabelle werden die unbearbeiteten Intensitätswerte von Oligo-Chip-Experimenten gespeichert. *Experiment_ID* ist ein Fremdschlüssel und verweist auf die Tabelle **Array_Experiment**, und damit ist auch spezifiziert um welchen Chip es sich handelt, beziehungsweise um welches gescannte Bild. Die Ergebnisse der Bildanalyse sind pro Bild in einer Datei mit der Extension „*.CEL*“ gespeichert, was den Namen der Tabelle erklärt. *Probeset* und *Probe_Pair* spezifizieren das *PM-MM*-Oligopaar, also zwei Bildzellen. Für jede Bildzelle sind ein Lage- und ein Streuungsmaß und die Anzahl der verwendeten Pixel in der Datenbank abgelegt. *Outlier* ist ein binärer Qualitätsparameter aus der Bildanalyse und in *QC_Flag*- kann später ein weiterer Qualitätsparameter gespeichert werden.

7. **cDNA_Condensed** (*Method, Experiment_ID, Spot, Value, QC_Flag*): Diese Tabelle dient zur Speicherung vorverarbeiteter Daten. Wurde beispielsweise für einen größeren Datensatz eine komparative Hybridisierung einer Probe gegen eine Referenz vorgenommen, so können die Daten gegen die Referenz verrechnet und normalisiert in diese Tabelle zurück geschrieben werden. Weitere Analysen und das einfache Abfragen der Daten fallen dann erheblich leichter. In der Spalte *Method* lassen sich die verwendeten Verfahren vermerken. Das ist wichtig, da man damit die Möglichkeit hat dieselben Rohdaten unterschiedlich zu analysieren.

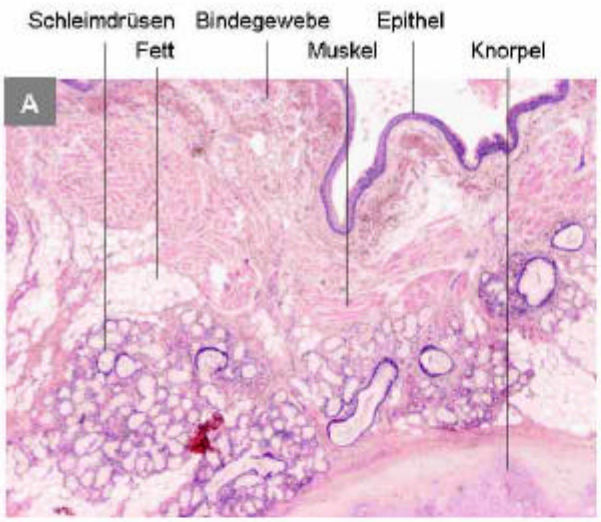
8. ***PM*Q_Condensed** (*Experiment_ID, Probeset, Value, P_Value, Probe_Pairs_Used, QC_Flag*): In dieser Tabelle sind die nach dem Standardverfahren ausgewerteten Datensätze gespeichert. Die angelegte Struktur verbietet das Speichern von Daten, die mit anderen Methoden verdichtet wurden. Für ein bestimmtes Sondenset (*Probeset*) auf einem Chipbild (*Experiment_ID*) gibt es genau einen repräsentativen

Expressionswert *Value* (*PM*-Quartil) den Detektionsscore *P_Value* (p-Wert des Wilcoxon-Tests) und die Qualitätsparameter *Probe_Pairs_Used, QC_Flag*.

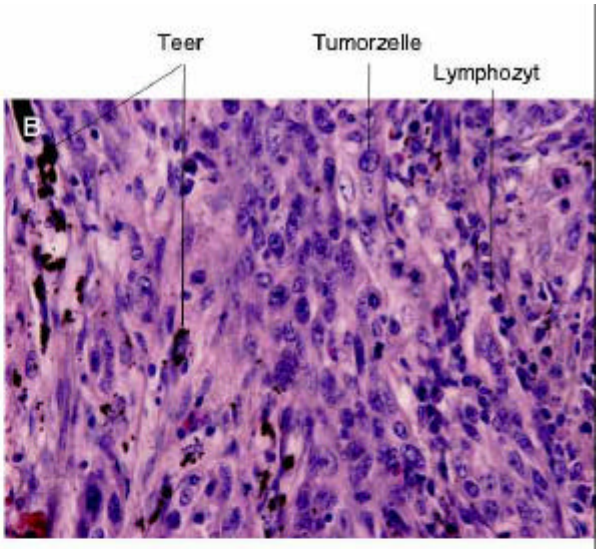
Das Schema der Datenbank: Die folgende Abbildung stellt das Datenbankschema dar. Jede Tabelle ist als Kasten zusammengefasst. Die Schlüsselelemente sind unterstrichen und mit <pk> für *primary key* gekennzeichnet. Zwischen zwei Tabellen wurde eine Linie eingefügt, falls eine der Tabellen ein Schlüsselelement der anderen als Attribut enthält. Die drei inhaltlichen Datenbereiche sind farblich voneinander abgesetzt: Sequenzdaten blau-grau, Experimentdaten alt-rosa, Probendaten gelb.



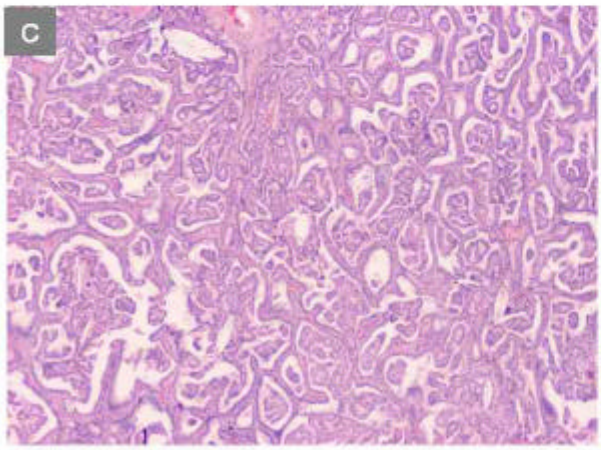
Anhang zu Kapitel 3

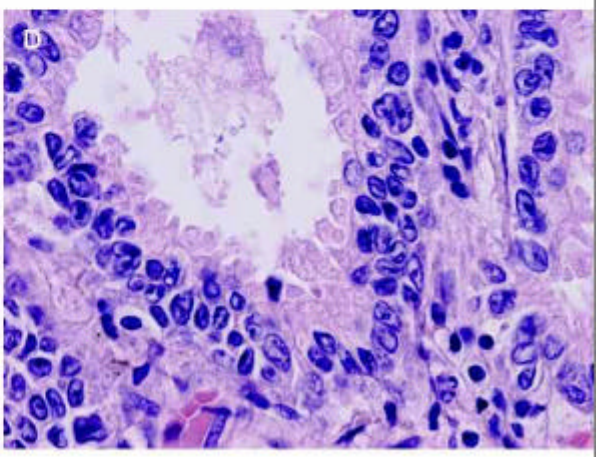


Bronchialwand der proximalen Atemwege
Die Hauptbronchien sind wie die Luftröhre mit einem mehrreihigen hochprismatischen Epithel ausgekleidet. Neben den seromukösen Drüsen findet man glatte Muskelatur in der Bronchialwand, die zusammen mit den Knorpelelementen einen Ring um die Atemwege bilden. (HE-Färbung, 2X)



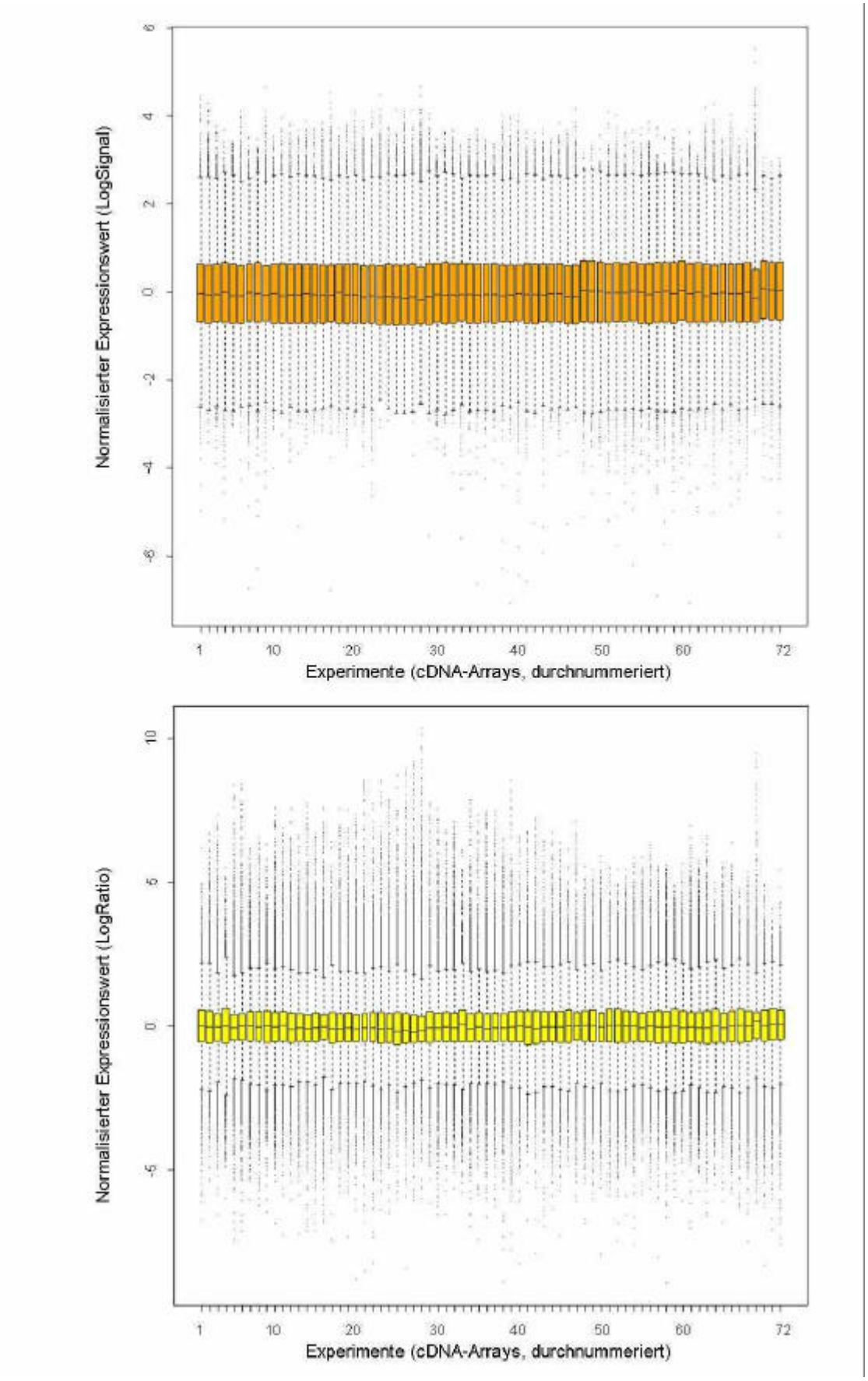
Plattenepithelkarzinom
Der Tumor ist bereits fortgeschritten (G3). Epitheliale Strukturen sind nicht mehr erkennbar. Die Kerne der Tumorzellen sind vergrößert und zeigen eine unregelmäßige Färbung. Die kleinen dunklen Kerne gehören zu den zahlreichen Lymphozyten. Schwarz zu erkennen sind Teer- und Kohlenstaubeinlagerungen. (HE-Färbung, 20X)



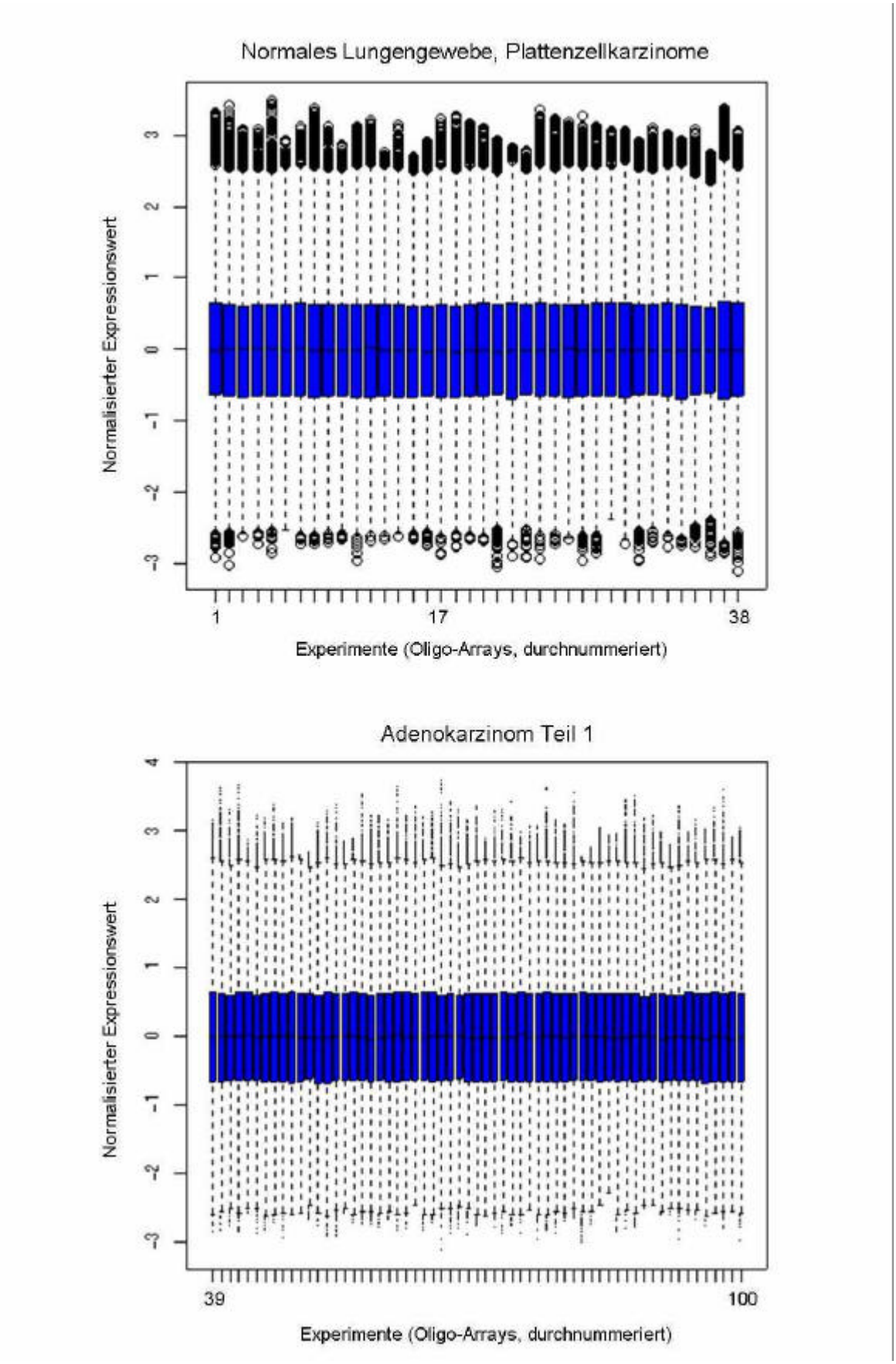


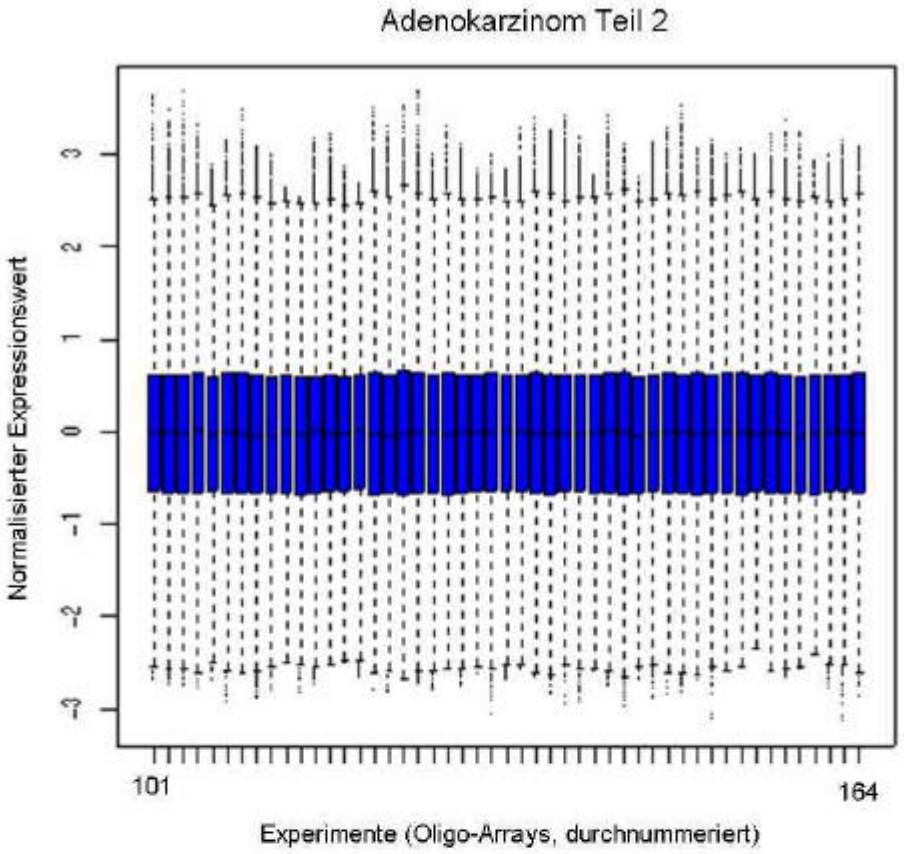
Adenokarzinom
Der Tumor befindet sich in einem Frühstadium (G1-G2). Die Zellen sind noch zu einem gewissen differenziert und zeigen Reste epithelialer Struktur. Abbildung **C:** HE-Färbung (2X), Abbildung **D:** Ausschnitt aus C, HE-Färbung (40X)

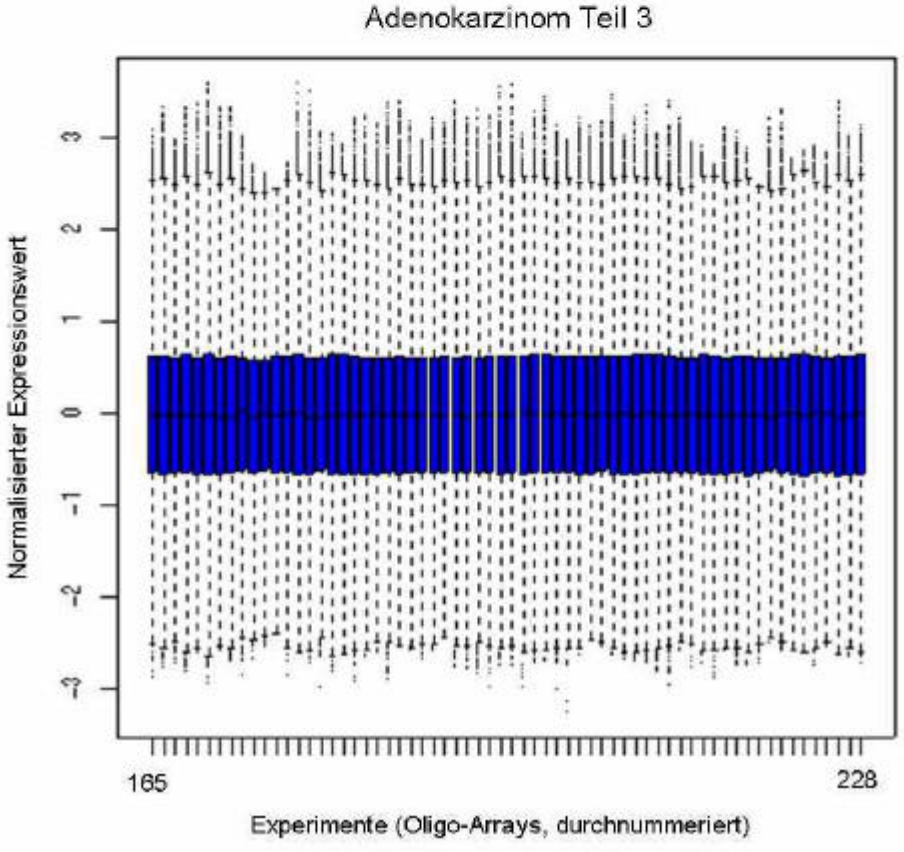
Box-Plots für die cDNA-Daten zu Abschnitt 3.6



Box-Plots für die Oligo-Array-Daten zu Abschnitt 3.6




Adenokarzinom Teil 2
Normalisierter Expressionswert
3
2
1
0
-1
-2
-3
101
164
Experimente (Oligo-Arrays, durchnummeriert)


Adenokarzinom Teil 3
Normalisierter Expressionswert
3
2
1
0
-1
-2
-3
165
228
Experimente (Oligo-Arrays, durchnummeriert)

Beispiele für differenzielle Gene im Plattenepithelkarzinom

HGNC	Sondenset	Spot	G_P	B_P	G_mÄ	B_mÄ	mÄ	Chr. Lokal.	Beschreibung
SPRR1B	37160_at	757	0	0	2,9	0,89	1,895	1q21-q22	small proline-rich protein 1B (cornifin)
GPX2	35194_at	15186	0,014	0	2,21	1,03	1,62	14q24.1	glutathione peroxidase 2 (gastrointestinal)
SPRR2C	36242_at	16641	0	0	2,18	0,12	1,15	1q21-q22	small proline-rich protein 2C
KRT17	34301_r_at	18725	0,002	0	2,05	0,68	1,365	17q12-q21	keratin 17
PLAU	37310_at	10409	0	0	1,73	0,61	1,17	10q24	plasminogen activator, urokinase
KRT13	36883_at	19888	0,001	0	1,63	0,82	1,225	17q21-q23	keratin 13
DIO2	31902_at	3878	0,003	0	1,54	0,67	1,105	14q24.2-q24.3	deiodinase, iodothyronine, type II
GJB5	38903_at	11055	0,002	0	1,53	0,41	0,97	1p35.1	gap junction protein, beta 5 (connexin 31.1)
STAT1	33338_at	1343	0,006	0,009	1,48	0,12	0,8	2q32.2	signal transducer and activator of transcription 1
HMGIY	39704_s_at	9894	0,032	0	1,47	1,29	1,38	6p21	high-mobility group protein isoforms I and Y
CDC45L	37458_at	17171	0	0	1,47	0,23	0,85	22q11.21	CDC45 cell division cycle 45-like (S. cerevisiae)
EYA2	35226_at	9850	0	0,002	1,26	0,13	0,695	20q13.1	eyes absent homolog 2 (Drosophila)
FAT2	38202_at	3345	0	0,005	1,39	0,18	0,785	5q32-q33	FAT tumor suppressor homolog 2 (Drosophila)
ABCC5	41428_at	20	0	0	1,34	1,3	1,32	3q27	ATP-binding cassette, sub-family C (CFTR/MRP), 5
AGER	35868_at	4067	0	0	-2,02	-2,41	-2,215	6p21.3	advanced glycosylation end product-specific receptor
FABP4	38430_at	14944	0	0	-2,88	-2,05	-2,465	8q21	fatty acid binding protein 4, adipocyte
TITF1	33754_at	24160	0	0	-2,11	-1,91	-2,01	14q13	Human thyroid transcription factor-1
FHL1	32542_at	811	0	0	-1,32	-1,82	-1,57	xq26	four and a half LIM domains 1
DLC1	37951_at	15386	0,007	0	-0,66	-1,65	-1,155	8p22-p21.3	mRNA for KIAA1723 protein
ABCA3	35183_at	16163	0	0	-2,1	-1,63	-1,865	16p13.3	ATP-binding cassette, sub-family A (ABC1), 3
WIF1	35178_at	18464	0	0	-2,25	-1,6	-1,925	12q13.13	WNT inhibitory factor 1
SFTPB	37004_at	12065	0,004	0	-1,68	-1,6	-1,64	2p12-p11.2	surfactant, pulmonary-associated protein B
MRC1	36908_at	18198	0	0	-2,23	-1,57	-1,9	10p13	mannose receptor, C type 1
RNASE1	37402_at	10621	0,002	0	-0,91	-1,51	-1,21	14q11.1	ribonuclease, RNase A family, 1 (pancreatic)
AQP1	36156_at	23721	0	0	-1,39	-1,38	-1,385	7p14	aquaporin 1 (channel-forming integral protein, 28kD)
GRO2	37187_at	3055	0,019	0	-1,03	-1,38	-1,205	4q21	GRO2 oncogene
CAV1	36119_at	84	0	0	-1,54	-1,32	-1,43	7q31.1	caveolin 1, caveolae protein, 22kD
KRT7	41294_at	11390	0,008	0	-0,86	-1,31	-1,085	12q12-q21	keratin 7
LPL	41209_at	7061	0	0	-1,58	-1,29	-1,435	8p22	lipoprotein lipase
OLR1	37233_at	10092	0	0	-1,92	-1,25	-1,585	12p13.2-p12.3	oxidised low density lipoprotein (lectin-like) receptor 1
MARCO	40331_at	18437	0	0	-1,75	-1,25	-1,5	2q12-q13	macrophage receptor with collagenous structure
HPGD	32570_at	22468	0	0	-1,31	-1,25	-1,28	4q34-q35	hydroxyprostaglandin dehydrogenase 15-(NAD)
EMP2	39631_at	1784	0,001	0	-0,86	-1,24	-1,05	16p13.2	epithelial membrane protein 2
IGSF4	35829_at	5256	0	0	-1,24	-1,15	-1,195	11q23.2	nectin-like protein 2
FBP1	36495_at	19940	0	0	-1,47	-1,09	-1,28	9q22.3	fructose-1,6-bisphosphatase 1

HGNC – Genname in HUGO-Nomenklatur, G – cDNA-Array-Datensatz (Garber et al), B – Oligo-Array-Datensatz (Bhattacharjee et al), P – P-Wert des T-Tests, mÄ – mittlere Änderung der logarithmierten Expressionswerte von Normal- zu Tumorgewebe (mÄ>0 – im Tumor überexprimiert), Chr. Lokal. – chromosomale Lokalisation

Beispiele für differenzielle Gene im Adenokarzinom

HGNC	Sondenset	Spot	G_P	B_P	G_mÄ	B_mÄ	mÄ	Chr. Lokal.	Beschreibung
COL1A2	32305_at	18222	0,001	0	0,35	0,96	0,655	7q22.1	collagen, type I
PAICS	39056_at	1013	0	0	0,54	0,82	0,68	4pter-q21	similar to SAICAR synthetase
COL3A1	32488_at	9142	0	0,006	1,03	0,73	0,88	2q31	collagen, type III
CLDN3	33904_at	23701	0,047	0	0,34	0,69	0,515	7q11.23	claudin 3
HMGIY	39704_s_at	9894	0,037	0	1,14	0,66	0,9	6p21	high-mobility group
LCN2	32821_at	23438	0,005	0	0,46	0,56	0,51	9q34	lipocalin 2
KRT19	40899_at	8895	0,005	0	0,62	0,55	0,585	17q21	keratin 19
ZWINT	35995_at	17169	0	0	0,73	0,52	0,625	10q21-q22	ZW10 interactor
H1F0	33386_at	10745	0,003	0	0,33	0,51	0,42	22q13.1	H1 histone family, member 0
PFKP	39175_at	16679	0	0	0,95	0,51	0,73	10p15.3-p15.2	phosphofructokinase, platelet
HDGF	36446_s_at	18164	0,012	0	0,93	0,5	0,715	xq25	hepatoma-derived growth factor
PFN2	38839_at	542	0,002	0	0,54	0,47	0,505	3q25.1-q25.2	profilin 2
SSR4	38635_at	5362	0,01	0	0,41	0,45	0,43	xq28	cDNA FLJ32555 fis
TPI1	34003_at	15365	0,02	0	0,6	0,45	0,525	12p13	triosephosphate isomerase 1
STAT1	33339_g_at	1343	0,001	0,002	1,77	0,41	1,09	2q32.2	signal transducer, activator of transcription 1
PABPC1	31950_at	9120	0,017	0	0,91	0,44	0,675	8q22.2-q23	poly(A) binding protein, cytoplasmic 1
WFDC2	33933_at	9138	0,034	0	0,45	0,43	0,44	20q12-q13.2	WAP four-disulfide core domain 2
PLAU	37310_at	10409	0	0	2,21	0,42	1,315	10q24	plasminogen activator, urokinase
AGER	35868_at	4067	0	0	-1,38	-2,15	-1,765	6p21.3	adv. glycosylation end product-spec. rec.
FABP4	38430_at	14944	0	0	-2,63	-2,11	-2,37	8q21	fatty acid binding protein 4, adipocyte
FHL1	32542_at	811	0	0	-1,72	-1,85	-1,785	xq26	four and a half LIM domains 1
CAV1	36119_at	13136	0	0	-1,05	-1,58	-1,315	7q31.1	caveolin 1, caveolae protein, 22kD
WIF1	35178_at	18464	0	0	-2,06	-1,55	-1,805	12q13.13	WNT inhibitory factor 1
EMP2	39631_at	1784	0	0	-0,95	-1,25	-1,1	16p13.2	epithelial membrane protein 2
MRC1	36908_at	18198	0	0	-1,96	-1,21	-1,585	10p13	mannose receptor, C type 1
SPARCL1	36627_at	9899	0,001	0	-0,8	-1,21	-1,005	4q21.3	SPARC-like 1 (mast9, hevin)
S100A8	41096_at	21415	0,001	0	-0,82	-1,17	-0,995	1q21	S100 calcium binding p. A8, calgranulin A
DLC1	37951_at	15386	0	0	-3,14	-1,16	-2,15	8p22-p21.3	mRNA for KIAA1723 protein
AQP1	36156_at	23721	0	0	-0,81	-1,15	-0,98	7p14	aquaporin 1, channel-forming integral protein
PALM2	35985_at	21588	0	0	-1,3	-1,06	-1,18	9q31-q33	mRNA for AKAP-2 protein
OLR1	37233_at	10092	0	0	-1,41	-1,05	-1,23	12p13.2-p12.3	oxidised low density lipoprotein receptor 1
LPL	41209_at	7061	0	0	-1,1	-1,05	-1,075	8p22	lipoprotein lipase
ITM2A	40775_at	20202	0	0	-0,82	-1,03	-0,925	xq13.3-xq21.2	integral membrane protein 2A
IGSF4	35829_at	5256	0	0	-0,81	-0,89	-0,85	11q23.2	nectin-like protein 2
SFTPD	31775_at	9797	0	0	-1,56	-0,88	-1,22	10q22.2-q23.1	surfactant, pulmonary-associated protein D

HGNC – Genname in HUGO-Nomenklatur, G – cDNA-Array-Datensatz (Garber et al), B – Oligo-Array-Datensatz (Bhattacharjee et al), P – P-Wert des T-Tests, mÄ – mittlere Änderung der logarithmierten Expressionswerte von Normal- zu Tumorgewebe (mÄ>0 – im Tumor überexprimiert), Chr. Lokal. – chromosomale Lokalisation

Vollständige Liste der analysierten Gene aus [25] A

Gen	Typ	Mechanismus	Genexpression	Proteinexpression	SQ Garber	SQ Bhatt.	AD Garber	AD Bhatt.
GRP/BN	Wachstumsfaktor	Hochregulation	20-60% (SCLC), selten (NSCLC)	20-60% (SCLC), selten (NSCLC)	neutral	NA	neutral	NA
GRPR	WF-Rezeptor	Hochregulation	20-60% (SCLC), selten (NSCLC)	20-60% (SCLC), selten (NSCLC)	NA	*?rauf*	NA	*?rauf*
NMB	Wachstumsfaktor	Hochregulation	100%	100%	NA	NA	NA	NA
neuromedin B Rez. (NMBR)	WF-Rezeptor	Hochregulation	häufig in NSCLC	häufig in NSCLC	NA	neutral	NA	neutral
bombesin rec. Subtyp 3 (BRS-3)	WF-Rezeptor	Hochregulation	20-60% (SCLC), selten (NSCLC)	20-60% (SCLC), selten (NSCLC)	NA	neutral	NA	neutral
IGF I and II	Wachstumsfaktor			erhöht im Blut	rauf, NA	runter, neutral	**rauf, NA**	**runter**, rauf
IGF-binding protein-3				IGF/IGF-binding protein-3 erhöht	rauf	**rauf**	**rauf**	**rauf**
c-erbB-1 (EGFR)	Onkogen	Genamplifikation, Punktmutation	überexpremiert, 13% (NSCLC)	korreliert mit schlechter Prognose	rauf	**rauf**	**rauf**	rauf
c-erbB-2 (Her2/neu)	Onkogen	Genamplifikation	überexpremiert	korreliert mit schlechter Prognose	neutral	runter	rauf	rauf
hepatocyte GF (HGF)	Wachstumsfaktor			viele NSCLC	runter	NA	runter	NA
MET (HGFR)	Onkogen, WF-Rezeptor	Triploidie mit Punktmutation, Translokation			neutral	neutral	neutral	rauf
estrogen receptor (ESR1)	Hormonrezeptor				**rauf**	neutral	**RAUF**	neutral
progesterone receptor (PGRMC1)	Hormonrezeptor				rauf	rauf	rauf	neutral
p29 (ESR1 related)				korreliert mit schlechter Prognose (NSCLC)	NA	NA	NA	NA
RAS (KRAS 90%)	Onkogen	Punktmutation		konstitutiv, dereguliert in 20-30%	NA	rauf	NA	rauf
HRAS	Onkogen	Punktmutation			runter	rauf	**runter**	neutral
MYC (MYCN, v-MYC)	Onkogen	Genamplifikation, -deregulation, Promotertranslokation	überexpremiert	überexpremiert, häufiger in SCLC	rauf	runter, rauf	**rauf**	rauf, neutral
BCL-2	Onkogen	Promotertranslokation		überexpremiert	**rauf**	NA	**rauf**	NA
Notch-3	Onkogen	Translokation(19p,15q)	überexpremiert		neutral	**rauf**	neutral	neutral
Cyclin D1	Onkogen	Genamplifikation, Promotertranslokation			**runter**	runter	runter	**runter**
CDK4, CDK6	Onkogen	Genamplifikation, Punktmutation			rauf, rauf	NA, neutral	rauf, neutral	NA, neutral

Vollständige Liste der analysierten Gene aus [25] B

Gen	Typ	Mechanismus	Genexpression	Proteinexpression	SQ Garber	SQ Bhatt.	AD Garber	AD Bhatt.
p53	Tumorsupressor	Punktmutation + LOH, erhöhte Halbwertzeit		überexpremiert in 40-70%(SCLC), 40-60%(NSCLC)	*runter*	NA	neutral	NA
AIS (p53 paralog)		Genamplifikation		überexpremiert im Plattenepithelk.	NA	NA	NA	NA
RB (p107, RB2/p130)	Tumorsupressor	Deletion, Mutation, fehlerhaftes Spleißen		90%(SCLC), 15-30%(NSCLC)	NA (*runter*)	NA (runter)	NA(neutral)	NA (runter)
CDKN2A (p16), p14	Tumorsupressor	fehlerhafte Methylierung, Punktmutation + LOH, homozygote Deletion	runterreguliert		NA	NA	NA	NA
FHIT	Tumorsupressor	fehlerhafte Methylierung	runterreguliert		neutral	NA	neutral	NA
RASSF1A	Tumorsupressor	fehlerhafte Methylierung	runterreguliert		NA	neutral	NA	neutral
APC	Tumorsupressor	fehlerhafte Methylierung, Punktmutation + LOH oder 2. Mutation	runterreguliert		*runter*	neutral	**runter**	neutral
CDH13		fehlerhafte Methylierung	runterreguliert		***rauf***	NA	neutral	NA
RAR-beta		fehlerhafte Methylierung	runterreguliert, 50% NSCLC		rauf	*?rauf*	**rauf**	neutral
TIMP-3		fehlerhafte Methylierung	runterreguliert		**rauf**	NA	**rauf**	NA
MGMT		fehlerhafte Methylierung	runterreguliert		NA	NA	NA	NA
DAPK1 (2,3)		fehlerhafte Methylierung	runterreguliert		NA	runter, runter, neutral	NA	runter, runter, neutral
LRP-DIT (LRP1B)		homozygote Deletion, verkürzte Transkripte		17%, 30% (NSCLC)	*leicht runter*	NA	neutral	NA
TSLC1		fehlerhafte Methylierung	runterreguliert, häufig in NSCLC		NA	NA	NA	NA
PPP2R1B		Mutation			neutral	neutral	neutral	*runter*
VEGF, VEGFB, PDendothGF (FLT)	Wachstumsfaktor				NA, NA, rauf	neutral, runter, rauf	NA, neutral	rauf
VEGFR	WF-Rezeptor				NA	neutral	NA	*rauf?*
factor VIII (Gerinnung)	Angiogenese-marker				**runter**	neutral, runter	**runter**	runter
HYAL2					**runter**	runter	**runter**	runter
CD-44					*runter*	neutral	**runter**	**runter**
CD-40					NA	NA	NA	NA
Cyclin E1, CCNE2				überexpremiert im Plattenepithelk.	rauf, *rauf*	rauf, rauf	**rauf**, neutral	rauf, *rauf*
GSTP1				hochreguliert	rauf	*rauf*	**runter**	rauf

Anhang zu Kapitel 4

Auflistung der 102 Gene aus *RefSeq*, für die Sense- und Antisense-Sonde auf metaGen-Chip I vorliegen, Teil I

RefSeq	Wie oft detektiert				5'-EST's		3'-EST's		ETS's	PMQ-Mittel		Beschreibung der Sequenz
Bezeichner	S,A	S,A	S,A	S,A	A	S	S	A	gesamt	S	A	
NM_018011	10	2	24	274	4	31	118	17	184	8,9632	2,642	hypothetical protein FLJ10154 (FLJ10154)
NM_016127	3	3	34	270	56	1	2	117	229	10,7746	1,0367	HSPC035 protein (LOC51669)
NM_018509	9	3	32	266	0	35	93	11	155	10,5713	1,2585	hypothetical protein PRO1855 (PRO1855)
NM_006526	26	19	37	228	3	6	19	1	39	1,938	0,9468	zinc finger protein 217 (ZNF217)
NM_018975	4	0	97	209	5	28	113	5	177	6,0006	1,0738	TRF2-interacting telomeric RAP1 protein (RAP1)
NM_016629	19	2	98	191	14	1	0	37	61	8,7378	1,1901	hypothetical protein (LOC51323)
NM_024026	35	65	32	178	16	7	35	4	64	2,333	0,8784	mitochondrial ribosomal protein 63 (MRP63)
NM_016617	8	2	125	175	2	6	1	67	78	3,9674	0,3897	hypothetical protein (BM-002)
NM_020188	7	2	139	162	53	2	5	83	176	3,9187	2,9169	DC13 protein (DC13)
NM_021238	5	1	146	158	1	16	110	8	143	7,893	1,0875	TERA protein (TERA)
NM_030912	7	1	147	155	12	2	2	34	57	7,953	1,6552	tripartite motif protein TRIM8 (TRIM8)
NM_022349	38	41	81	150	5	1	0	17	25	4,2719	1,4491	CD20-like precusor (LOC64166)
NM_015070	39	93	31	147	0	3	6	5	47	1,0976	1,7779	KIAA0853 protein (KIAA0853)
NM_006283	14	2	156	138	0	29	106	0	153	12,0625	2,2953	transforming, acidic coiled-coil containing protein 1
NM_015385	101	34	43	132	1	11	30	0	53	2,4284	0,7084	SH3-domain protein 5 (ponsin) (SH3D5)
NM_014050	67	55	64	124	0	95	30	2	153	2,9389	0,6677	mitochondrial ribosomal protein L42 (MRPL42)
NM_017689	62	121	16	111	0	9	45	2	75	9,2354	1,9437	hypothetical protein FLJ20151 (FLJ20151)
NM_023037	86	11	122	91	0	9	14	1	29	2,0853	0,5399	putative gene product (13CDNA73)
NM_022781	25	6	201	78	0	7	54	3	70	2,2709	0,6197	hypothetical protein FLJ21343 (FLJ21343)
NM_007106	36	2	194	78	1	26	35	2	72	2,738	0,6532	ubiquitin-like 3 (UBL3)
NM_016052	1	232	0	77	0	18	2	1	22	1,411	3,0785	CGI-115 protein (LOC51018)
NM_031453	18	2	220	70	14	0	0	56	137	3,8456	1,161	hypothetical protein MGC11034 (MGC11034)
NM_025226	68	3	188	51	2	81	55	3	145	5,0197	2,3364	MSTP032 protein (MSTP032)
NM_019000	45	0	235	30	1	4	41	4	50	4,6037	1,2788	hypothetical protein (FLJ20152)
NM_018330	13	0	271	26	10	1	2	37	63	3,9468	0,6575	KIAA1598 protein (KIAA1598)
NM_031455	82	10	196	22	0	15	54	2	75	4,4275	2,953	hypothetical protein DKFZp761F241 (DKFZP761F241)
NM_014787	172	13	107	18	2	7	27	0	38	0,6271	0,4076	KIAA0473 gene product (KIAA0473)
NM_005033	92	187	13	18	0	50	8	3	68	1,7136	2,2937	polymyositis/scleroderma autoantigen 1 (75kD)
NM_024040	99	1	195	15	64	0	1	45	128	4,3082	3,737	hypothetical protein MGC2491 (MGC2491)
NM_017687	110	12	173	15	0	11	0	1	12	0,9791	0,7774	hypothetical protein FLJ20147 (FLJ20147)
NM_022497	136	159	0	15	41	0	3	39	94	5,6718	2,2382	mitochondrial ribosomal protein S25 (MRPS25)
NM_015678	182	2	112	14	0	7	8	0	17	1,0153	1,4875	neurobeachin (NBEA)
NM_024343	27	1	269	13	15	25	37	8	113	1,7305	7,4905	hypothetical protein MGC10764 (MGC10764)
NM_000391	2	0	296	12	12	34	39	9	158	5,5026	1,2185	ceroid-lipofuscinosis, neuronal 2 (CLN2)
NM_024038	5	0	293	12	114	1	5	56	231	14,0019	2,1392	hypothetical protein MGC2803 (MGC2803)
NM_018204	44	2	252	12	0	7	6	0	32	2,514	1,1396	cytoskeleton associated protein 2 (CKAP2)
NM_017657	47	3	250	10	2	22	70	0	132	4,2266	1,0263	hypothetical protein FLJ20080 (FLJ20080)
NM_024626	257	0	45	8	1	5	24	0	33	3,719	1,2022	hypothetical protein FLJ22418 (FLJ22418)
NM_032847	82	6	215	7	0	20	4	0	30	2,091	2,4248	hypothetical protein FLJ14825 (FLJ14825)
NM_006023	52	1	251	6	101	0	3	107	232	2,13	1,3247	D123 gene product (D123)
NM_014398	201	2	101	6	5	1	0	13	21	2,2578	1,2632	similar to lysosome-associated membrane glycoprotein
NM_019058	11	0	294	5	4	36	174	6	283	4,9438	1,9522	hypothetical protein (FLJ20500)
NM_022473	47	3	255	5	0	18	1	1	29	1,0788	1,1241	zinc finger protein 106 (ZFP106)
NM_024056	177	2	126	5	126	0	5	26	169	2,5996	1,0471	hypothetical protein MGC5576 (MGC5576)
NM_022484	118	4	185	3	0	9	2	0	13	1,3868	1,9186	hypothetical protein FLJ13576 (FLJ13576)
NM_024695	291	0	16	3	0	3	11	0	18	3,5321	11,3478	hypothetical protein FLJ13993 (FLJ13993)
NM_032712	55	0	253	2	20	1	2	59	97	5,8172	2,9459	hypothetical protein MGC13170 (MGC13170)
NM_024747	98	1	209	2	4	2	0	25	32	3,7978	2,2956	hypothetical protein FLJ22501 (FLJ22501)
NM_031307	191	5	112	2	3	1	0	9	16	3,3514	1,7795	hypothetical protein FKSG32 (FKSG32)
NM_024561	198	4	106	2	5	8	7	1	21	1,2467	0,8953	hypothetical protein FLJ22054 (FLJ22054)
NM_024736	204	1	103	2	0	12	40	3	67	7,3535	2,5509	hypothetical protein FLJ12150 (FLJ12150)

Die Struktur der Tabelle entspricht der von Tabelle 4-2.

Auflistung der 102 Gene aus *RefSeq*, für die Sense- und Antisense-Sonde auf metaGen-Chip I vorliegen, Teil II

RefSeq	Wie oft detektiert				5'-EST's		3'-EST's		ETS's	PMQ-Mittel		Beschreibung der Sequenz
Bezeichner	S,A	S,A	S,A	S,A	A	S	S	A	gesamt	S	A	
NM_019895	4	0	305	1	128	2	6	10	163	11,2852	1,0613	chromosome 3 open reading frame 4 (C3ORF4)
NM_000435	152	0	157	1	23	2	8	84	125	4,5606	1,3003	Notch (Drosophila) homolog 3 (NOTCH3)
NM_021020	225	4	80	1	0	8	0	0	14	8,228	2,7532	F37/Esophageal cancer-related gene-coding leucine-zipper motif
NM_032907	242	1	66	1	0	48	84	2	157	3,9407	3,1716	hypothetical protein MGC14421 (MGC14421)
NM_024099	11	0	299	0	0	39	59	2	132	5,795	5,3376	hypothetical protein MGC2477 (MGC2477)
NM_021830	17	0	293	0	0	5	26	3	39	3,1521	3,5706	hypothetical protein FLJ21832 (FLJ21832)
NM_014320	24	0	286	0	0	63	16	0	94	3,3292	1,9987	putative heme-binding protein (SOUL)
NM_001247	28	0	282	0	33	0	0	102	156	14,069	1,9919	ectonucleoside triphosphate diphosphohydrolase 6
NM_017606	31	0	279	0	17	7	0	33	64	3,7714	1,444	hypothetical protein DKFZp434K1210 (DKFZp434K1210)
NM_012193	65	1	244	0	0	11	48	1	83	10,0296	2,3867	frizzled (Drosophila) homolog 4 (FZD4)
NM_022083	85	0	225	0	9	3	7	19	41	11,0768	1,937	niban protein (NIBAN)
NM_021639	84	1	225	0	0	21	0	0	26	2,2555	1,9676	hypothetical protein SP192 (SP192)
NM_024101	91	4	215	0	0	16	1	1	32	3,6554	2,4038	hypothetical protein MGC2771 (MGC2771)
NM_006321	122	1	187	0	25	1	0	42	74	4,4816	1,4846	ariadne (Drosophila) homolog 2 (ARIH2)
NM_018075	126	0	184	0	9	0	0	37	55	9,0071	1,748	hypothetical protein FLJ10375 (FLJ10375)
NM_018246	124	5	181	0	25	0	3	37	69	3,9894	0,9978	hypothetical protein FLJ10853 (FLJ10853)
NM_031484	130	0	180	0	1	13	0	0	17	8,2038	2,0381	hypothetical protein MGC4415 (MGC4415)
NM_007286	135	1	174	0	20	1	1	23	67	16,936	1,8456	synaptopodin (KIAA1029)
NM_024329	147	0	163	0	22	1	0	70	104	4,7134	1,6701	hypothetical protein MGC4342 (MGC4342)
NM_032391	174	0	136	0	0	3	5	2	14	4,9403	3,0608	small nuclear protein PRAC (PRAC)
NM_001168	242	0	68	0	3	132	18	8	170	2,8954	1,2686	baculoviral IAP repeat-containing 5 (survivin) (BIRC5)
NM_012405	254	0	56	0	0	13	109	8	134	8,2261	2,893	isoprenylcysteine carboxyl methyltransferase (ICMT)
NM_018487	254	0	56	0	17	1	10	42	97	6,9994	1,3238	hepatocellular carcinoma-associated antigen 112 (HCA112)
NM_003579	262	0	48	0	14	0	0	19	34	1,6763	1,5143	RAD54 (S.cerevisiae)-like (RAD54L)
NM_013258	271	0	39	0	0	17	25	0	48	3,9597	2,5642	apoptosis-associated speck-like protein containing a CARD
NM_017815	281	0	29	0	0	84	5	2	99	1,4454	1,2212	hypothetical protein FLJ20424 (FLJ20424)
NM_023073	284	0	26	0	1	0	0	12	15	3,428	1,7046	hypothetical protein FLJ13231 (FLJ13231)
NM_018119	286	0	24	0	1	21	0	0	25	3,4434	2,6198	hypothetical protein FLJ10509 (FLJ10509)
NM_032842	286	3	21	0	0	6	8	0	24	1,0276	1,289	hypothetical protein FLJ14803 (FLJ14803)
NM_016647	289	1	20	0	1	20	27	0	55	4,8185	2,6382	mesenchymal stem cell protein DSCD75 (LOC51337)
NM_025149	292	0	18	0	0	14	0	0	15	3,4588	5,2377	hypothetical protein FLJ20920 (FLJ20920)
NM_032323	282	13	15	0	0	0	13	0	26	4,9796	2,7313	hypothetical protein MGC13102 (MGC13102)
NM_014186	297	0	13	0	0	64	3	1	73	3,16	1,3197	HSPC166 protein (HSPC166)
NM_018836	303	0	7	0	0	0	2	2	4	2,875	3,3773	hypothetical protein (MOT8)
NM_016046	304	0	6	0	0	1	1	0	3	0,4684	1,1724	homolog of yeast exosomal core protein CSL4 (CSL4)
NM_014795	304	1	5	0	0	5	0	4	11	5,1324	1,5704	zinc finger homeobox 1B (ZFHX1B)
NM_015310	282	23	5	0	17	0	0	0	20	2,1892	0,6137	KIAA0942 protein (KIAA0942)
NM_016625	306	0	4	0	27	0	0	4	56	1,4726	1,5122	hypothetical protein (LOC51319)
NM_032659	219	89	2	0	16	0	0	33	54	4,0759	2,5416	hypothetical protein MGC11138 (MGC11138)
NM_030927	308	0	2	0	0	30	7	0	59	1,9497	1,4056	hypothetical protein MGC11352 (MGC11352)
NM_014654	252	58	0	0	38	0	1	1	48	3,8192	4,1748	KIAA0468 gene product (KIAA0468)
NM_024722	308	2	0	0	3	0	1	22	31	6,6633	2,838	hypothetical protein FLJ13322 (FLJ13322)
NM_024709	296	14	0	0	10	1	0	2	29	2,9322	4,2147	hypothetical protein FLJ14146 (FLJ14146)
NM_024832	310	0	0	0	2	6	6	12	26	2,3519	10,6219	hypothetical protein FLJ22439 (FLJ22439)
NM_000500	310	0	0	0	11	3	25	3	46	3,2692	1,5044	cytochrome P450, subfamily XXIA, polypeptide 2
NM_031431	300	10	0	0	11	2	0	2	18	0,9911	2,6519	tethering factor SEC34 (SEC34)
NM_014943	183	127	0	0	11	0	0	1	15	2,3954	2,6361	KIAA0854 protein (KIAA0854)
NM_018660	310	0	0	0	19	1	4	1	27	4,5602	3,4936	papillomavirus regulatory factor PRF-1 (LOC55893)
NM_015490	307	3	0	0	8	0	2	0	20	2,0341	2,1664	DKFZP434M183 protein (SEC31B-1)
NM_016458	310	0	0	0	10	1	0	0	20	2,817	3,248	hypothetical protein (LOC51236)
NM_031429	306	4	0	0	0	4	20	0	25	3,8688	2,2005	retbindin (RTBDN)

Die Struktur der Tabelle entspricht der von Tabelle 4-2.

Oligonukleotide für die Northern Blot - Hybridisierungen

Die verwendeten Primer und Oligonukleotide wurden ausnahmslos von metabion (Martinsried, BRD) bezogen.

Oligonukleotide für *KMO-OPN3*

OPN3-Sonde	KMO-Sonde
5'-CTTCTGAGGGTCTGAAATTGAATAA-3'	5'-TTATTCAATTTCAGACCCTCAGAAG-3'
5'-TCATCATCTTCTAATGTGTTGGAGA -3'	5'-TCTCCAACACATTAGAAGATGATGA-3'
5'-CTCAAAGCTCTTTTTCTTTGTTTTG-3'	5'-CAAAACAAAGAAAAAGAGCTTTGAG-3'
5'-TCCCTACAAACTGAACATGGATTAT -3'	5'-ATAATCCATGTTCAGTTTGTAGGGA-3'
5'-TGTTGGACTCTATTCAGTGTCATGT -3'	5'-ACATGACACTGAATAGAGTCCAACA-3'
5'-AAAATTTACTGTTCTTTGTCGATGC-3'	5'-TCAAAGTTGTCTCTGAAACTCCTCT-3'
5'-TAATTCAACGGGTGCTTTACATAAT-3'	5'-ATATATGTGGGAAATACAGGGGAAT-3'
5'-CTGTTAAAGAAGGATGAACCAAAGA-3'	5'- GCATCGACAAAGAACAGTAAATTTT-3'

Oligonukleotide für *ZNF217*

Antisense-Sonde (detektiert ZNF217)	Sense-Sonde (detektiert Antisense-Transkript)
5'-GTTAATCTTCAAAAATAGGCTATAA-3'	5'-ACATATAATAGAGGTACAATTCGTT-3'
5'-TTGAGTCAATCATCTTGCAAATGTG-3'	5'-TAACCACATTTCTGAATGTATAAAA-3'
5'-CTCTGTAACACTCATTGGATTAGGC-3'	5'-TGTATTATGCTGGAATTTTTTGGGC-3'
5'-TTATGGTTCTAGTCACAGCAAGCTC-3'	5'-ATGTGATAATAGAGGGCTGGAATTT-3'
5'-CTTGTTTTCAAATAGATTTGGTGAT-3'	5'-AGCTTGCTGTGACTAGAACCATAAA-3'
5'-AAATTCCAGCATAATACAAATCGAC-3'	5'-TCTTTGCCTAATCCAATGAGTGTTA-3'
5'-TGGTTAAAAAGAGAAATCTGAAAGC-3'	5'-TATAGCCTATTTTTGAAGATTAACA-3'
5'-GGACACAAAAACATATTTTGAAGTA-3'	5'-TGTTCACAAGCTGAGCCATATGTAC-3'
5'-ACCAGTAGTATACCAATAGTTAATA-3'	
5'-ATAACTTTAAGCTGAAATATATCAT-3'	

Oligonukleotide für *Ponsin*

Antisense-Sonde (detektiert Ponsin)	Sense-Sonde (detektiert Antisense-Transkript)
5'-TACATGGGCTATACAAAGTTAAATA-3'	5'-ACTTTATTCAGAATTGCGTGGCAAA-3'
5'-TGTAGGTGAGTTATTTGGAATTCCT-3'	5'-GCAGTGGATTGATTTGATAAATAGA-3'
5'-GTAACTGAGATCAGAGTTACCTTTC-3'	5'-CACACTGAGAGCATGTCCTATGCAG-3'
5'-TCTAGCAATAGGACTTAATACGACT-3'	5'-CAGTCCTATTTATCTGCAGTCGTAT-3'
5'-GGACATGCTCTCAGTGTGTAATTTA-3'	5'-AAATTTACGATAAGTATTCTATTGG-3'
5'-ATCAAATCAATCCACTGCAATGAAG-3'	5'-TAGCACAAAAATAGCCATTGTAAAG-3'
5'-ATGTACAGCAAATGTAGTAATTCAA-3'	5'-ATAGCCCATGTACCTACCTTGTATA-3'
5'-TTGCCACGCAATTCTGAATAAAGTT-3'	5'-ATAATGTGCTTAAGAAGTGGGACTG-3'

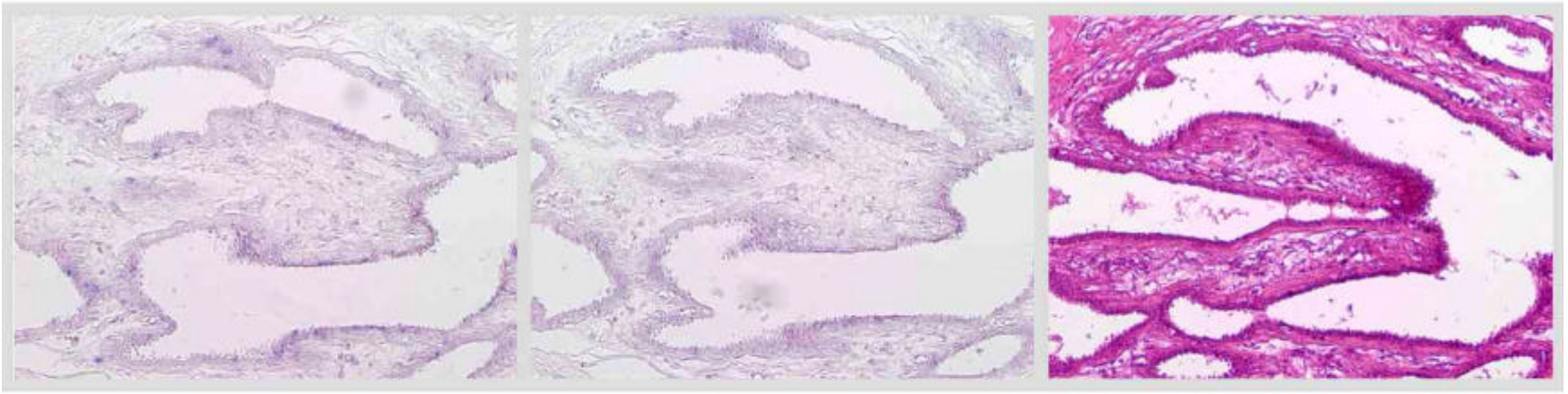
Oligonukleotide für *CGI-115*

Antisense-Sonde (detektiert CGI-115)	Sense-Sonde (detektiert Antisense-Transkript)
5'- CGACGCTACGGCTCCAGTTACCATT -3'	5'- ATGGCTAAAGTCCTCAACAAGAAAA -3'
5'- ACTATCAGAAGTGTCTGTGGCCTCG -3'	5'- GGAGAATCAAGTGTTGGGACTAATA -3'
5'- CTTTTCCGATCCACAGCTTCCTTCA -3'	5'- GATGCTGAGCCCTGTGACAAAGAAA -3'
5'- TGCGTCATCATCAGAATAAAAGTGG -3'	5'- GATGACGCAATAGAAGCTGACAGTG -3'
5'- ATCACCCTCACTGTCAGCTTCTATT -3'	5'- AGCTGTGGATCGGAAAAGGACCACT -3'
5'- TTTCTTGTTGAGGACTTTAGCCATA -3'	5'- GAAGACGAGGCCACAGACACTTCTG -3'
5'- GACCAGAATAGTAGGTTTACTTTCA -3'	5'- ATGGTAACTGGAGCCGTAGCGTCGG -3'

Oligonukleotide für *DC13*

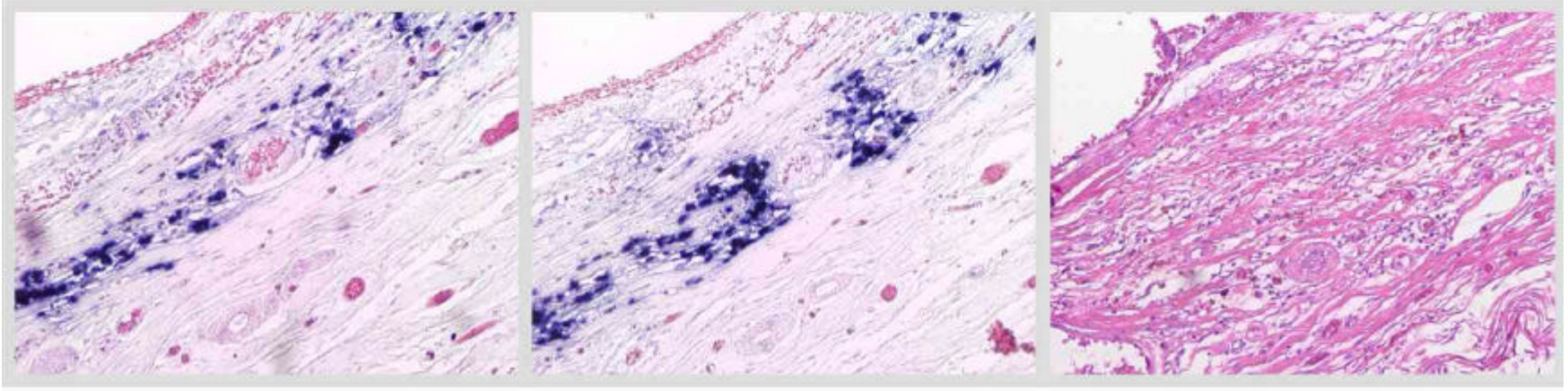
Antisense-Sonde (detektieren DC13)	Sense-Sonde (detektieren Antisense-Transkript)
5'-GTGACCTGTGAGAGATTGAACCATG-3'	5'- TAGGGAGCAGACAGCTGAACCGCTT -3'
5'-TGATACCTGAAAGAATCCTGTCTTA-3'	5'- ACAGCAGTGCAACCCGACTCGTGGC -3'
5'-TGGCTGAGAGAAGACCTAAAGACTC-3'	5'- GGATGCATCTTTAGGAGATGAGGAT -3'
5'-GAGCAGGGAGCATGGCATTGCAATG-3'	5'- TCTTCAGTGTGCAAGTGTGGAGATA -3'
5'-TGTTGATCGGGAGTTGAGAAAATGC -3'	5'- AGCAAGTTAATCAAGACGTTGCATT -3'
5'-TGAAGAATGCAACGTCTTGATTAAC-3'	5'- CTCAACTCCCGATCAACATCATTAC -3'
5'-TCCTCATCTCCTAAAGATGCATCCT-3'	5'- ATGCCATGCTCCCTGCTCTTGGTCC -3'
5'-CCACGAGTCGGGTTGCACTGCTGTG-3'	5'- AGGCATCGAGTGAAAATACAATTTA -3'
5'-CGGCGTCTGGCAAGCGGTTCAGCTG-3'	5'- AAGACAGGATTCTTTCAGGTATCAA -3'
	5'-ATGGTTCAATCTCTCACAGGTCACT -3'
	5'-TCAGATATTAACTGGTTGTAGGCAA-3'

Brustgewebe, normales Drüsenepithel



DC13 Antisense Sonde (Sense wird detektiert.) DC13 Sense Sonde (Antisense wird detektiert.) HE-Färbung

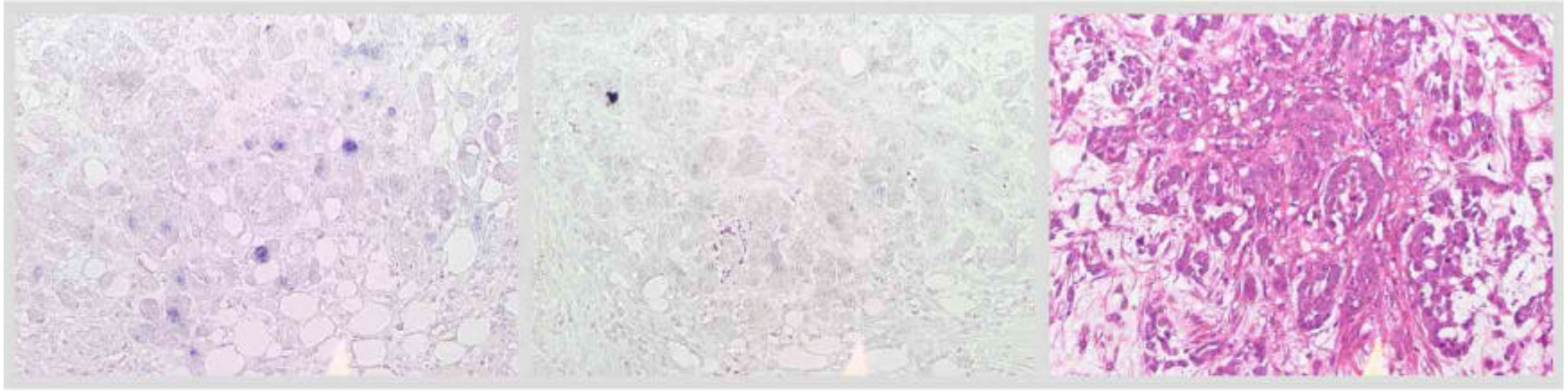
Brustkrebsgewebe, muzinöses Karzinom



DC13 Antisense Sonde (Sense wird detektiert.) DC13 Sense Sonde (Antisense wird detektiert.) HE-Färbung

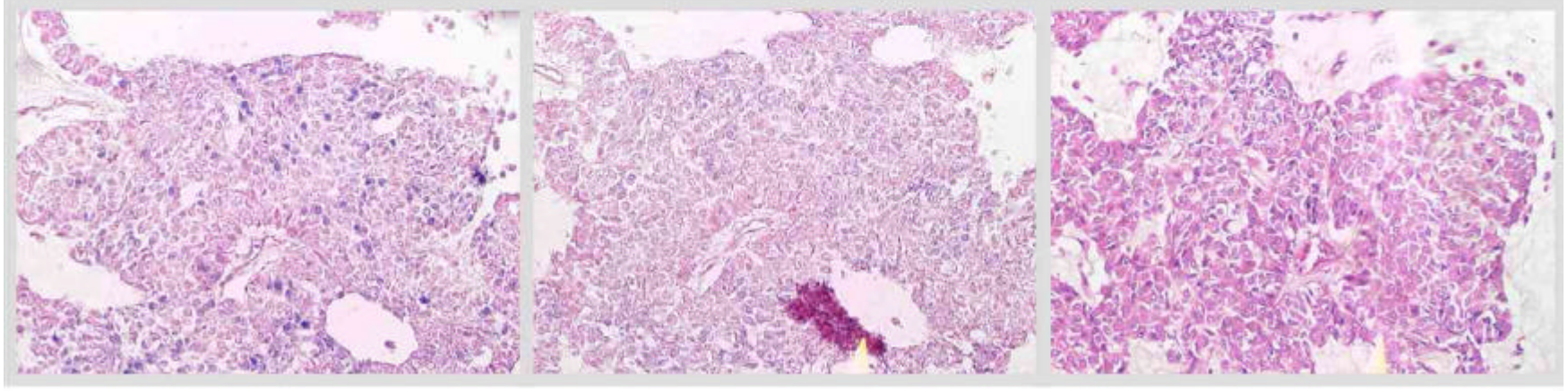
RNA In Situ Hybridisierungen für dc13 A

Brustkrebsgewebe, invasives, duktales Karzinom



DC13 Antisense Sonde (Sense wird detektiert.) DC13 Sense Sonde (Antisense wird detektiert.) HE-Färbung

Brustkrebsgewebe, muzinöses Karzinom



DC13 Antisense Sonde (Sense wird detektiert.) DC13 Sense Sonde (Antisense wird detektiert.) HE-Färbung

RNA In Situ Hybridisierungen für dc13 B