Aus dem Universitätsklinikum Münster
Institut für Medizinische Informatik und Biomathematik
– Direktor: Univ.-Prof. Dr. W. Köpcke –

Intensitätsbasierte Qualitätskontrolle und Skalierung von Genexpressionsdaten

INAUGURAL-DISSERTATION
zur
Erlangung des doctor rerum medicinalium

der Medizinischen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von Martin Eisenacher
aus Dorsten, Nordrhein-Westfalen

2005

Gedruckt mit Genehmigung der Medizinischen Fakultät der Westfälischen Wilhelms-Universität Münster

Dekan: Univ.-Prof. Dr. H. Jürgens

1. Berichterstatter: Univ.-Prof. Dr. W. Köpcke

2. Berichterstatter: Univ.-Prof. Dr. H. Funke

Tag der mündlichen Prüfung: 11. Mai 2005

Aus dem Universitätsklinikum Münster
Institut für Medizinische Informatik und Biomathematik
– Direktor: Univ.-Prof. Dr. W. Köpcke –

Referent: Univ.-Prof. Dr. W. Köpcke
Koreferent: Univ.-Prof. Dr. H. Funke

ZUSAMMENFASSUNG

Intensitätsbasierte Qualitätskontrolle und Skalierung
von Genexpressionsdaten

Martin Eisenacher

In den vergangenen Jahren wurde die Sequenzierung der Genome verschiedener Spezies abgeschlossen. Mit diesen Informationen ermöglicht die Affymetrix GeneChip-Technologie heute die gleichzeitige Expressionsmessung von bis zu 47.000 Transkripten, deren *Signal*-Werte aus den Intensitäten von Teilmessungen zusammengesetzt werden. In dieser Arbeit wird zunächst eine SPLUS-Funktionsbibliothek zur Handhabung von GeneChip-Daten entwickelt. Sie wird ergänzt um Algorithmen zum automatischen Gruppieren, Benennen und Kombinieren von Experimenten in Abhängigkeit von ihren experimentellen Merkmalen, was bei der Visualisierung experimenteller Merkmale verwendet wird.

Die Messergebnisse für ein Gen unterliegen einerseits einer zu messenden biologischen Varianz, die zur Beantwortung klinischer Fragestellungen (z. B. Klassifikation von Tumortypen, Identifikation von Signalmechanismen) dienen soll, und andererseits einer technischen Varianz durch Ungenauigkeiten in den Aufarbeitungsschritten. Um das Ausmaß der technischen Varianz der GeneChip-Methode zu quantifizieren, wurde hier die Relevanz der empfohlenen Qualitätskriterien in eigenen und öffentlich zugänglichen Datensätzen untersucht. Darüber hinaus wurde das neue Qualitätskriterium „Gesamtintensität“ entwickelt, welches bisherigen Qualitätskriterien in wichtigen Aspekten überlegen ist, da es den globalen Chip-Zustand auf Intensitätsebene berücksichtigt, also näher an der Quelle technischer Varianz liegt.

Selbst gut reproduzierte Experimente müssen für weiter gehende Auswertungen vergleichbar gemacht werden („Skalierung“). Sechs hier formulierte Bewertungsaspekte ermöglichen die objektive Beurteilung der Standardskalierung im Vergleich zu drei eigenen Entwicklungen. Hinweise auf eine Überlegenheit einer dieser eigenen Entwicklungen („sum.of.intens“-Skalierung) gegenüber der Standardskalierung können in weiteren Experimenten unter Verwendung spezieller *probe sets* (*PolyA Controls*) überprüft werden.

Tag der mündlichen Prüfung: 11. Mai 2005

Inhaltsverzeichnis

Verzeichnis der Abbildungen

Verzeichnis der Tabellen

1 Einleitung

Im Laufe der letzten Jahre wurden große Anstrengungen unternommen, die Sequenz des menschlichen Genoms und die anderer Organismen zu ermitteln (Lander et al.[58]; Venter et al.[92]; Waterston et al.[95]). Die Sequenzinformationen und weitere Merkmale wurden in großen Datenbanken systematisch gespeichert und so der Forschergemeinde zugänglich gemacht (*National Center for Biotechnology Information (NCBI) – Entrez Nucleotide*[69]; RefSeq[76]; UniGene[90]; *NCBI – Entrez Genome*[68]). Mit der Sequenzinformation ist nun die strukturelle Zusammensetzung der Genome bekannt.

Dies liefert jedoch noch kein Gesamtbild möglicher physiologischer Vorgänge, da zur Zeit nur ein kleiner Teil des menschlichen Genoms hinsichtlich seiner funktionell aktiven Einheiten – dies sind im Wesentlichen die Gene und regulative Sequenzen – hinreichend charakterisiert ist. Darüber hinaus werden in einer Zelle nicht immer alle Gene abgelesen. Die Transkription von DNA in mRNA erfolgt zeitlich reguliert und abhängig vom Gesamtstoffwechsel.

Die Microarray-Technologie (Fodor et al. (1991)[36]; Fodor et al. (1993)[35]; Lorkowski und Cullen[62]; Lockhart et al.[61]) ermöglicht die simultane vergleichende Messung der Konzentrationen von bis zu 47.000 mRNAs in einer Probe und liefert somit Aussagen über den Expressionszustand von Genen in Zellen oder Geweben. Damit hat diese Technik das Potenzial, bisher unbekannte Zusammenhänge zwischen Genexpressionsmuster und Krankheit, bei der Regulation von Genen und in der Wirkung genetischer und extragenetischer Faktoren auf Stoffwechselzusammenhänge zu erschließen und damit auch für die Entwicklung neuer diagnostischer Verfahren. Bei der von mir verwendeten GeneChip-Technologie der Firma Affymetrix handelt es sich um eine besonders hoch integrierte Microarray-Technologie, bei der bis zu eine Million verschiedene als *capture probes* fungierende Oligonukleotide auf einer Fläche von wenigen Quadratzentimetern untergebracht sind. Diese Oligonukleotide repräsentieren etwa 50.000 Gene. Eine Analyse liefert damit in einer einzigen Messung Informationen über den Expressionszustand nahezu aller Gene der untersuchten Zellen oder Gewebe.

Die durch eine Messung erhaltenen Daten sind sehr umfangreich und erfordern effiziente Methoden bei der Handhabung und Auswertung. Die mit der Technologieplattform gelieferte Software bietet hierfür einen Satz grundlegender Funktionen an, die aber den Bedürfnissen der Anwender häufig nur unzureichend gerecht werden, da die Einbeziehung anderer und neuer Methoden nicht vorgesehen ist. Außerdem

sind im Wesentlichen Manipulationen von Einzelexperimenten vorgesehen, nicht jedoch die Berücksichtigung von erfassten Merkmalen zur Zusammenfassung von Experimenten zu Gruppen und zur automatischen Untersuchung von Zusammenhängen zwischen diesen Experimentgruppen. Die ersten Ziele dieser Arbeit sind daher

- die Modellierung und Implementierung einer Schnittstelle von GeneChip-Daten zu den Programmiersprachen SPLUS und C++ und
- die Entwicklung einer Programmbibliothek zum Gruppieren von Experimenten nach ihren Merkmalen.

Die Reproduzierbarkeit der experimentellen Bedingungen und der Messwerte ist bei einer absoluten Messmethode wie der GeneChip-Technologie, die auf eine Zweikanalmessung verzichtet, eine wichtige Voraussetzung, um Experimente miteinander vergleichen zu können. Trotz einer exakten Standardisierung bei der Probengewinnung und -aufarbeitung unterliegen die gemessenen Daten einer erheblichen Varianz. Diese ist einerseits technisch bedingt und resultiert aus unterschiedlichsten Quellen und ergibt sich andererseits aus der biologischen Varianz. Dabei ist zu beachten, dass die Erfassung der biologischen Varianz, sei es als natürliche oder induzierte Varianz, häufig Ziel des Experimentes ist, dass jedoch gleichzeitig unerwünschte Kovarianzen auftreten können, die den Analyseprozess erheblich beeinflussen. Um möglichst verlässliche Ergebnisse und genaue Aussagen über die biologische Varianz zu erhalten, sollten eventuelle Ursachen technischer Varianz und unerwünschter biologischer Kovarianz ermittelt und idealerweise ausgeschaltet oder ihr Einfluss vermindert oder zumindest erfasst werden. Ein weiteres Ziel dieser Arbeit ist es deshalb,

- durch Meta-Analysen der vorhandenen Datensätze Methoden zur Quantifizierung der interexperimentellen Varianz zu entwickeln, die beobachtete Varianz zu beschreiben und Qualitätskriterien für durchgeführte Messungen zu identifizieren und anzuwenden.

Sollte es nicht gelingen, die technische Varianz zu verhindern, so können ihre Auswirkungen möglicherweise mehr oder weniger vollständig vor der Ermittlung und Bewertung der biologischen Varianz korrigiert werden, was als Skalierung bezeichnet wird. Abschließendes Ziel dieser Arbeit ist es,

- die existierenden Skalierungsverfahren zu analysieren, um eventuell neue, verbesserte Verfahren zu entwickeln und deren Vorteile darzustellen.

2 Grundlagen –Technologie und verwendete Datensätze

In diesem Kapitel werden in Unterkapitel 2.1 zunächst die den Experimenten zugrunde liegenden biologischen Methoden skizziert, elementare Begrifflichkeiten der GeneChip-Technologie aufgeführt und der Weg eines Experimentes von der Probenentnahme bis zum Scannen eines Chips beschrieben (Protokollskizze). Unterkapitel 2.2 enthält die Beschreibung der für das weitere Verständnis notwendigen Komponenten der Software- und Datenumgebung der Technologieplattform. Ferner werden die verwendeten Datensätze vorgestellt (Unterkapitel 2.3) sowie die Eigenschaften von Graphen des Typs „Box Plot“ (Unterkapitel 2.4).

2.1 Biologische Methoden

Eine ausführliche Sammlung verschiedener Methoden zur Messung der Genexpression findet sich bei Lorkowski und Cullen[62]. Alizadeh et al.[17] (Klassifikation von Leukämietypen) und Golub et al.[40] (Klassifikation von Lymphomtypen) haben erste Arbeiten vorgelegt, die den Nachweis der Eignung von Hochdurchsatzmethoden der Expressionsanalyse für die Beantwortung klinischer Fragestellungen erbringen.

2.1.1 DNA-Microarrays

Charakteristikum der GeneChip-Technologie ist die massenhafte simultane Durchführung von Hybridisationsreaktionen an der Oberfläche einer Festphase (*array*). Bei diesen Hybridisationsreaktionen bilden einzelsträngige, an der Oberfläche des Arrays immobilisierte DNA-Moleküle Doppelstränge mit DNA- oder RNA-Molekülen, die in der Flüssigkeit vorhanden sind, mit der ein solcher Array inkubiert wird (*target*). Die Bildung der Doppelstränge setzt voraus, dass die Bindungspartner, die eine solche Hybridbildung eingehen, zueinander komplementär sind, also eine Watson-Crick-Basenpaarung miteinander eingehen können. Nach Entfernen der flüssigen Phase bleiben die geformten Doppelstränge an der Festphase zurück und können dort nach Art und Menge vermessen werden. Wird beispielsweise ein Chip, der an seiner Oberfläche DNA-Sonden (*probes*) enthält, die 10.000 verschiedene Gene spezifisch repräsentieren, unter stringenten Bedingungen mit einer Lösung inkubiert, die die mRNA aus einem bestimmten Tumorgewebe enthält, so kann anhand der gebildeten DNA/RNA-Hybride bestimmt

werden, welche der 10.000 auf dem Chip vorhandenen Gene in dem Tumor in welcher Menge exprimiert werden.

Die einzelnen Technologien unterscheiden sich unter anderem in der verwendeten festen Oberfläche (z. B. Glas, Plastik oder Nylonmembran) und dem davon abhängigen Verfahren der Immobilisierung, in der Länge der *probes*, in der *probe*-Dichte – also in der Anzahl der pro Array gleichzeitig messbaren *probes* – und in der Detektionsmethode (z. B. Ein- oder Zwei-Farben-Fluoreszenz). Eine Übersicht über Färbe-, Label- und Detektions-Strategien von Nukleinsäuren findet sich bei Kricka[57]. Nicht zuletzt unterscheiden sich Microarrays in den zugrunde liegenden Sequenzinformationen (Art der Gendatenbank, Version der Gendatenbank, Selektionsverfahren der *probes* usw.).

Die wichtigsten Anwendungen von DNA-Microarrays sind die Reanalyse einer bekannten Gensequenz, die Erkennung von Mutationen / SNPs (*single nucleotide polymorphisms*) und die Bestimmung der mRNA-Konzentration spezifischer Gene.

Schulze et al.[80] bezeichnen den *spotted cDNA microarray* als eine der am häufigsten verwendeten Microarray-Technologien. Dabei werden zunächst mit einem speziellen Algorithmus geeignete Sequenzabschnitte (*probes*) der selektierten Gene bestimmt. Die durch PCR amplifizierten Oligonukleotide oder alternativ klonierte cDNA-Fragmente werden mit Robotern als *spots* auf speziell beschichtete Glasträger aufgebracht, wo anhand chemischer Linker-Substanzen eine kovalente Bindung hergestellt wird.

Bei der Probenaufarbeitung für die Detektion mit der Zwei-Farben-Fluoreszenz-Methode wird die mRNA zweier Zellen oder Gewebeproben in cDNA umgeschrieben, mit einem jeweils anderen Fluoreszenzfarbstoff gefärbt und dann in einem Hybridisierungspuffer auf die *spots* gegeben. Dort findet eine kompetitive Reaktion der unterschiedlich gefärbten cDNAs statt, und zwar nur mit den zur jeweiligen cDNA komplementären *spots*. Durch Detektion der mit den *spots* assoziierten Farbstoffintensitäten mit zwei unterschiedlichen Wellenlängen kann bestimmt werden, in welchem Verhältnis die ursprünglichen mRNA-Mengen eines Gens in beiden Proben zueinander stehen.

Eine Vielzahl von Veröffentlichungen befasst sich mit der Qualitätskontrolle (z. B. Tran et al.[87] und Wang et al.[94]) und der Normalisierung (Yang et al.[98] und Park et al.[71])

von Microarray-Experimenten. Die jeweiligen Ergebnisse sind aufgrund technologiespezifischer Problemfelder – wie z. B. der Ermittlung der *spot accuracy* oder der Normalisierung der beiden Farbkanäle aufeinander – nicht ohne weiteres auf die in dieser Arbeit verwendeten GeneChips zu übertragen.

2.1.2 GeneChip-Technologie

Die GeneChip-Technologie (Lockhart et al.[61]; Lipshutz et al.[60]) – in der Literatur oft zitiert als „*high-density oligonucleotide microarrays*" – basiert ebenfalls auf der Hybridisierung von Nukleinsäurefragmenten an immobilisierten *probes*. Die *probes* sind dabei 25mere Oligonukleotide, die an einem Silizium-Wafer immobilisiert sind. Durch photolithographische Methoden (siehe Abbildung 1) kann die pro Flächeneinheit definierte Sequenz sehr exakt positioniert werden; auf diese Weise wird eine sehr hohe *spot*-Dichte erreicht (über 400.000 *probe*-Einheiten auf 1,28 cm^2 bei HG-U95-Arrays). Die *spots* sind hierbei quadratisch, liegen räumlich direkt nebeneinander und werden *probe cells* genannt.

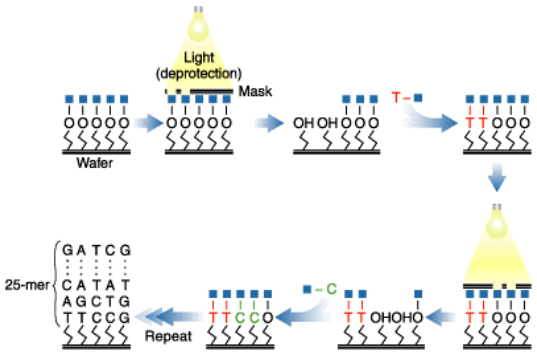


Abbildung 1: Photolithografische Synthese der Oligonukleotide einer *probe cell* (Quelle: Affymetrix, Inc.[2])

Bei einer *probe*-Länge von 25 Oligonukleotiden und einer nicht hundertprozentigen Ausbeute der einzelnen Syntheseschritte werden mehrere *probes* benötigt, um ein Gen spezifisch detektieren zu können. Außerdem wird neben der *probe cell* mit der Zielsequenz (*Perfect Match-probe cell*, PM) eine *probe cell* (*MisMatch-probe cell*, MM) mit einer Sequenz benötigt, an deren mittlerer Position 13 ein homologer Basenaustausch stattgefunden hat, um Kreuzhybridisierungen durch unspezifische Nukleinsäurefragmente erkennen zu können (siehe Abbildung 2).

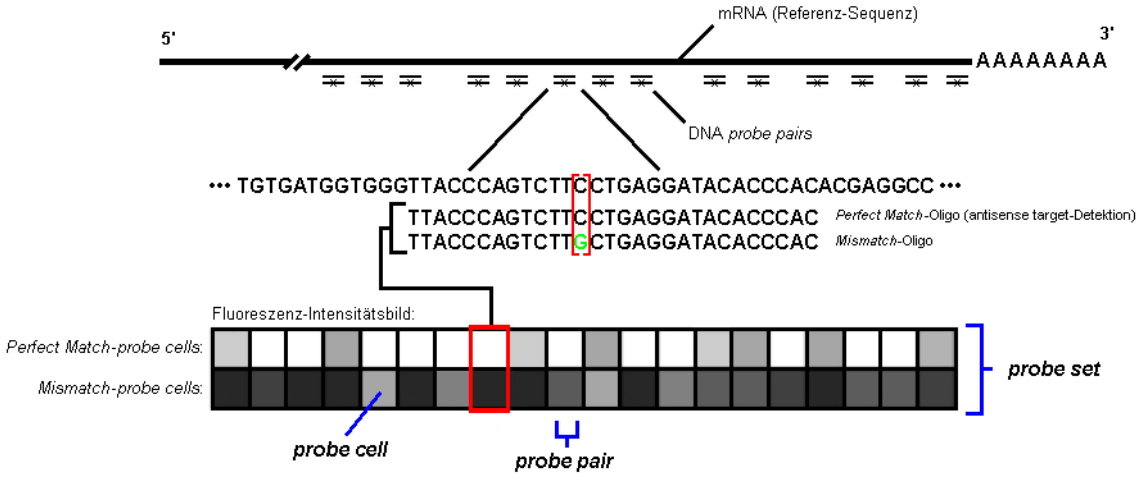


Abbildung 2: Nomenklatur: *probe cell, Perfect Match, Mismatch, probe pair, probe set*
(Abb. modifiziert nach: Affymetrix, Inc.[6])

Die *probe pairs* eines *probe sets* sind bei neueren Array-Typen über den Chip verteilt, damit größere räumlich begrenzte Störungsquellen (Flusen, Ausfällungen) nach Möglichkeit nicht mehrere *probe pairs* eines *probe sets* beeinträchtigen. Die Auswahl geeigneter Oligonukleotide stellt eine große Herausforderung dar. Für den Array-Typ HG-U133 wird dieser Vorgang in einem *Technical Report* (Affymetrix, Inc.[10]) näher erläutert. Arrays verschiedener Spezies unterscheiden sich grundsätzlich durch ihre Sequenzinformationen. Bisher verfügbar sind Expressions-Arrays für Mensch, Maus, Ratte und verschiedene Prokaryoten. Chismar et al.[24] untersuchen die Tauglichkeit von humanen Arrays bei ihrer Verwendung für eine verwandte Spezies (*rhesus macaque*).

Ist das zu untersuchende Sample aufgearbeitet, gefärbt und auf einen Chip hybridisiert (Beschreibung des Protokolls siehe nächster Abschnitt 2.1.3), wird mit einem GeneChip-Scanner zunächst mithilfe eines Lasers und eines Detektors eine Datei mit den **Bilddaten** erzeugt (DAT-Datei). Die Auflösung des Scanners ist größer als die Ausdehnung der einzelnen *probe cells*. Auf jedem Array befindet sich ein Oligo B2-Rand, der durch Zugabe eines nicht mit dem zu testenden Genom homologen Oligonukleotids und Hybridisation an sein randständig auf den Chip aufgebrachtes Komplementär entsteht. Mithilfe dieses Oligo B2-Rands wird ein virtuelles Gitter auf die Bilddaten gelegt, für jede *probe cell* eine Durchschnittsintensität der Bildpixel berechnet und diese in einer Datei mit den **Intensitätsdaten** abgelegt (CEL-Datei). Mit einem Kondensierungsalgorithmus (siehe Abschnitt 2.2.1) wird aus den Einzelintensitäten der *probe cells* eine Datei mit den

Maßzahlen der **Primäranalyse** für das entsprechende *probe set* berechnet (CHP-Datei). Qualitätskriterien und andere Eigenschaften eines Experimentes werden in einer **Berichtsdatei** (RPT-Datei) zusammengefasst. Bei lokalem Arbeiten wird zusätzlich eine Datei mit Informationen zum Chip-Experiment angelegt (EXP-Datei), deren Inhalt beim Arbeiten mit einem Server in Datenbanken gespeichert wird.

Über eine Workflow-Verwaltung können mehrere Experimente angelegt und die Fortschritte der einzelnen Arbeitsschritte verfolgt werden.

2.1.3 Beschreibung der Aufarbeitung eukaryotischer Proben (Protokollskizze)

Im Folgenden werden die wichtigsten Schritte des Standardprotokolls für die Aufarbeitung eukaryotischer Proben aufgeführt. Ein detailliertes Protokoll findet sich im *Affymetrix Expression Manual*[15]. Abwandlungen des Protokolls für geringe Mengen an Ausgangsmaterial finden sich bei Mahadevappa et al.[63] und Eberwine et al.[32]; Baugh et al.[20] schlagen weitere Modifikationen vor dem Hintergrund konkreter Untersuchungen von GeneChip-Experimenten vor. Affymetrix stellt mittlerweile ebenfalls ein Protokoll für geringe Mengen an Ausgangsmaterial zur Verfügung (Affymetrix, Inc.[12]).

RNA-Isolierung

Das konkrete Verfahren zur RNA-Isolierung variiert je nach verwendetem Material. Ziel ist es, ausreichend poly(A)$^+$-mRNA (0,2-5 µg) für die Reverse Transkription zu erhalten. Da RNA relativ instabil ist und schnell von RNAsen angegriffen wird, muss an einem vollständig RNAse-freien Arbeitsplatz gearbeitet werden. Weil auch kleinste Flusen und Partikel beim Scannen Störungen erzeugen, ist es erforderlich, dass die verwendeten Gefäße und Substanzen hohen Anforderungen an die Reinheit genügen. Nach der Isolation muss eine Quantifizierung der RNA-Ausbeute zur Erfolgskontrolle stattfinden. Außerdem wird unter Umständen eine Fällung mit NaO-Acetat und Ethanol zur Konzentrierung der RNA notwendig.

cDNA-Synthese (Reverse Transkription)

Ziel dieses Schritts ist es, aus den poly(A)$^+$-mRNA-Molekülen über eine Reverse Transkription (RT) und die Synthese eines zweiten Strangs eine doppelsträngige cDNA zu

gewinnen, die im nächsten Schritt als Vorlage für eine *in-vitro*-Transkription dient. Abbildung 3 zeigt eine schematische Darstellung der einzelnen Vorgänge.

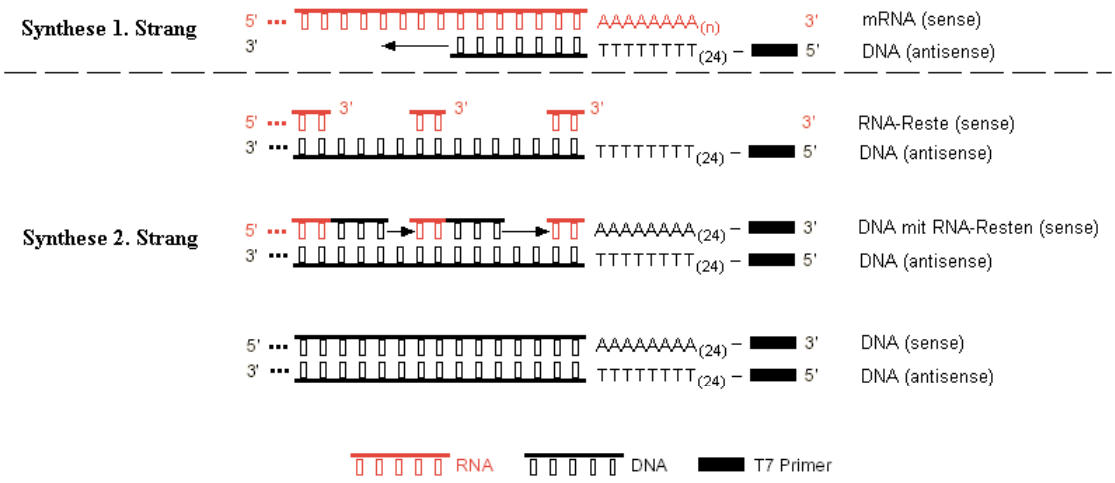


Abbildung 3: cDNA-Synthese (Reverse Transkription)
(Abb. modifiziert nach: Affymetrix[12])

Zur Synthese des ersten Strangs dient ein PolyT-Primer, an den eine T7-Promotorsequenz angehängt ist. Nach Abschluss der RT liegt ein mRNA/cDNA-Hybrid vor. Zur Synthese des zweiten DNA-Strangs werden mit RNAse H große Teile des RNA-Strangs entfernt, es verbleiben jedoch RNA-Reste am bisher synthetisierten cDNA-Molekül. Mit DNA-Polymerase I und T4-DNA-Polymerase wird ausgehend von den RNA-Resten als Primer der zweite DNA-Strang synthetisiert. Zum Abschluss dieses Schrittes werden mit einer DNA-Ligase die neu synthetisierten DNA-Abschnitte des zweiten Strangs miteinander verbunden, sodass nun ein doppelsträngiges cDNA-Molekül vorliegt.

Synthese Biotin-gelabelter cRNA (*in-vitro*-Transkription)

Mithilfe einer T7-RNA-Polymerase und dem cDNA-Doppelstrang als Template werden unter teilweiser Verwendung biotinylierter Ribonukleotide Biotin-gelabelte cRNA-Moleküle synthetisiert (siehe Abbildung 4).

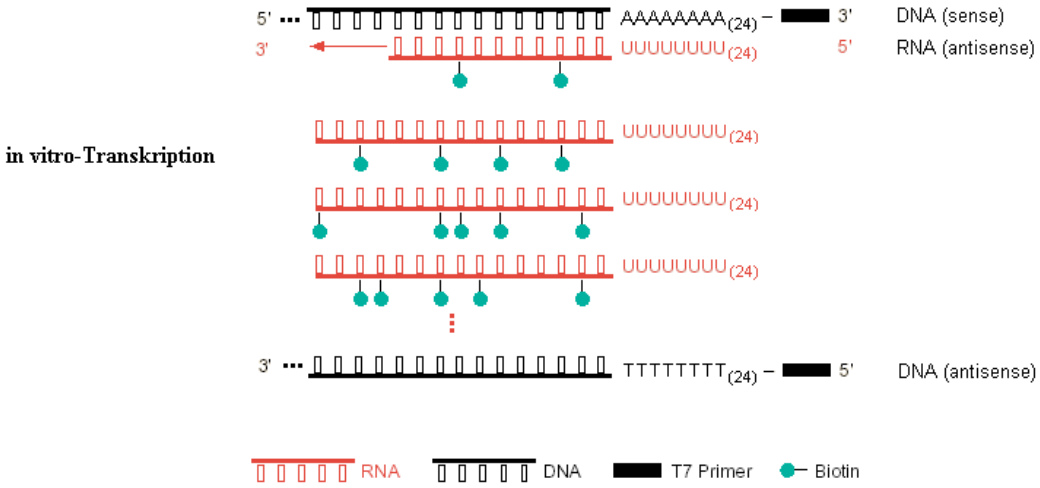


Abbildung 4: Synthese Biotin-gelabelter cRNA (*in-vitro*-Transkription) (Abb. modifiziert nach: Affymetrix[12])

Danach werden in einem Reinigungsschritt die Reaktionsteilnehmer herausgewaschen. Die gewonnene cRNA wird mittels einer Ethanol-Fällung und Resuspension in einem kleineren Volumen aufkonzentriert. Abschließend wird die cRNA-Konzentration bestimmt.

Fragmentierung der cRNA

Mithilfe von Metallsalzen und Hitze wird eine Fragmentierung der cRNA-Moleküle durchgeführt. Im Anschluss daran erfolgt eine Längenkontrolle (Agarosegel oder Agilent 2100 Bioanalyzer).

Hybridisierung

Hierbei wird zunächst der Hybridisierungscocktail aus der fragmentierten cRNA, dem Control-Oligo B2 (für die Hybridisierung des Chip-Rands zum Legen des Gitters), den *Hybridization Controls* (siehe Abschnitt 4.2.2) und weiteren Substanzen in wässriger Lösung hergestellt. Dieser wird dann auf den Chip aufgebracht. In einem Hybridisierungsofen (45°C, 16h) findet die Hybridisierung der Biotin-gelabelten fragmentierten cRNA an die Oligonukleotide der *probe cells* statt.

Waschen, Färben, Scannen

In der Fluidics-Station, einem automatischen Chip-Waschmodul, wird zunächst die überschüssige Hybridisationslösung entfernt und der Chip mehrmals gewaschen. Anschließend wird R-Phycoerythrin-Streptavidin an die Biotin-Reste der cRNA gebunden und danach das durch Phycoerythrin erzeugte Fluoreszenzsignal durch die sequenzielle Zugabe von biotinylierten Anti-Streptavidin-Antikörpern und R-Phycoerythrin-Streptavidin verstärkt. Beim Scannen des Chips wird die durch einen Argon-Ionen-Laser induzierte Fluoreszenz des R-Phycoerythrin ortsabhängig und somit *probe cell*-spezifisch vermessen.

2.2 Software- und Datenumgebung

Neben der oben beschriebenen Hardware ist die Software- und Datenumgebung essenzieller Bestandteil der GeneChip-Plattform. Abbildung 5 fasst diese schematisch zusammen.

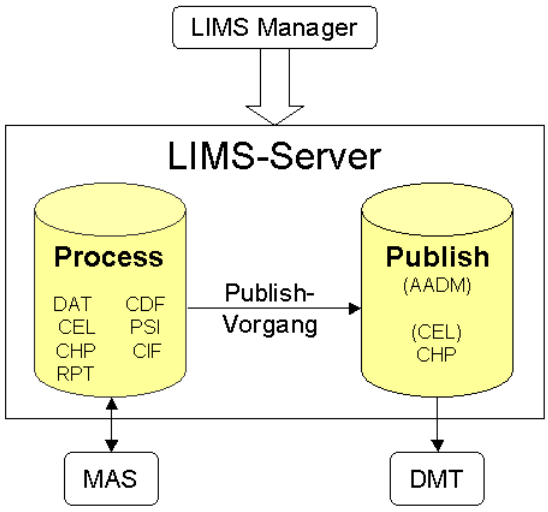


Abbildung 5: Übersicht über das LIMS-System

In der so genannten *process*-Datenbank auf dem LIMS-Server werden die Experimente gespeichert. Zu den Informationen eines Experimentes gehören der Projektname, der Sample-Name und der Sample-Typ. Außerdem werden hier Template-Informationen gespeichert, anhand derer weitere Eigenschaften eines Experimentes spezifiziert werden, sowie User Sets, die benutzerangepasste Parametersätze für Primäranalysen enthalten.

Die Chip-Daten, die zu einem Experiment gehören, also DAT-, CEL- und CHP-Dateien, werden im Server-Dateisystem in einer Dateifreigabe `GCLims` gespeichert und mit ihrem entsprechenden Experimentnamen über Datenbankeinträge verknüpft. Auch die RPT-Datei wird hier abgelegt. Die Dateien mit den Layout- und *probe set*-Informationen zu einzelnen Arrays (CDF, CIF, PSI) werden ebenfalls einmalig im Dateisystem gespeichert und ihr Vorhandensein in der Datenbank eingetragen.

Neben der *process*-Datenbank kann es eine oder mehrere *publish*-Datenbanken geben. Primäranalysen, die als Basis für weiter gehende statistische Auswertungen dienen sollen, können mit einem *publish*-Vorgang in eine *publish*-Datenbank auf dem LIMS-Server übertragen werden. Hier können zusätzlich CEL-Informationen und – mit speziellen Anwendungen – sogar Daten gespeichert werden, die mit der Spotted Array-Technologie gewonnen wurden. Die Spezifikation einer *publish*-Datenbank wird als AADM (*Affymetrix Analysis Data Model*) bezeichnet. Sie ist öffentlich zugänglich und somit für eigenständige Entwicklungen nutzbar.

Die Client-Anwendungen Microarray Suite (MAS), Data Mining Tool (DMT) und LIMS Manager, mit denen die Benutzer des Systems vorwiegend arbeiten, werden in Abschnitt 2.2.1 näher beschrieben. Dort wird auch der Kondensierungsalgorithmus zur Primäranalyse eingehend dargestellt, der von der MAS ausgeführt wird. Der technische Hintergrund, soweit er zum Verständnis dieser Arbeit wichtig ist, wird in Abschnitt 2.2.2 erläutert. Hierzu zählen beispielsweise die Datenbank- und Dateiformate sowie die Schnittstelle zwischen SPLUS und C++.

2.2.1 Front-End-Software und Kondensierungsalgorithmus

Die Microarray Suite (MAS, verwendete Version: 5.1) stellt die zentrale Client-Anwendung der GeneChip-Technologie dar. Sie ermöglicht die Ansteuerung des Scanners, das Abspeichern der Experimente auf dem Server, die Durchführung der Sichtkontrolle der gescannten Chips und der manuellen oder automatischen Primäranalyse der Intensitätsdaten zu statistischen Maßzahlen (**Kondensierungsalgorithmus**), sowie das Anstoßen des *publish*-Vorgangs. Außerdem wird über die MAS die Report-Datei eines Experimentes erzeugt und angezeigt. Details zu diesen Vorgängen finden sich im Benutzerhandbuch[9].

Das Benutzerhandbuch der MAS bietet des Weiteren eine Übersicht über den Kondensierungsalgorithmus. Die folgenden fünf grundsätzlichen Schritte werden dabei durchgeführt:

1. Ermittlung von *background* und *noise* für räumliche Unterteilungen des Chips (*zones*). Preprocessing der *probe cell*-Intensitäten: *background subtraction* und *noise correction*.
2. Berechnung eines *Ideal Mismatch*-Wertes für jede *probe cell*; Subtraktion desselben von der *Perfect Match*-Intensität (Justierung).
3. Log-Transformation der justierten *Perfect Match*-Werte zur Stabilisierung der Varianz.
4. Errechnung eines robusten Durchschnitts der justierten *Perfect Match*-Werte über die *probe pairs* eines *probe sets* durch einen *Tukey's Biweight Estimator*-Algorithmus. Der unskalierte *Signal*-Wert ergibt sich als der Delogarithmus dieses Wertes.
5. Ermittlung des skalierten *Signal*-Wertes durch Anwenden eines Skalierungsfaktors auf den robusten Durchschnitt und anschließendes Delogarithmieren. Zur Berechnung des Skalierungsfaktors wird ein randbereinigter Durchschnitt (*trimmed mean*) verwendet.

Eine ausführliche Beschreibung zusammen mit den verwendeten mathematischen Formeln enthält das *Statistical Algorithms Description Document*[14]. Eine Reihe von Veröffentlichungen befasst sich mit einer Verbesserung der Kondensierung: Li und Wong[59] stellen einen modellbasierten Ansatz vor, der mehrere Chips in die Kondensierung einbezieht, ebenso wie Strand et al.[84]; Naef et al. entwickeln eine Methode speziell für gepaarte Chips[66] (z. B. vorher / nachher-Replikate) und zur Ergebnisverbesserung bei hohen Konzentrationen[67]; Sasik et al.[77] berücksichtigen Sample-Serien. Irizarry et al.[50,51] zeigen anhand extensiver „Spike-in"-Experimente die Überlegenheit eines eigenen Kondensierungsalgorithmus gegenüber der MAS 5.0. Vergleiche verschiedener Kondensierungsalgorithmen finden sich bei Rajagopalan[74] und Zhou und Abagyan[99]; Giles und Kipling[39] weisen darüber hninaus nach, dass die kondensierten Expressionswerte der Methoden von Li und Wong, MAS 4 und MAS 5

größtenteils normalverteilt sind, was eine wichtige Voraussetzung für parametrische Tests in weiter gehenden Auswertungen ist.

Die Nachfolgeversion der MAS heißt GCOS (*GeneChip Operating Software*). Zu den wichtigsten Änderungen gehören hierbei die Möglichkeit zur Ansteuerung eines neuen, höher auflösenden Scanners und die Bereitstellung eines Templates, mit dessen Hilfe Experimentinformationen nach dem MIAME-Standard[23] (*Minimum Information About A Microarray Experiment*) eingegeben werden können. Ferner wird der Export von Expressionsdaten im standardisierten XML-basierten MAGE-ML-Format[83] (*Microarray Gene Expression Markup Language*) ermöglicht.

Das Data Mining Tool (DMT, verwendete Version: 3.0) ist die Client-Anwendung, mit der grundlegende statistische Größen wie Durchschnitt, Median, Standardabweichung und *Fold Change* berechnet, sowie weiter gehende Auswertungen wie t-Test, Mann-Whitney-Test, SOM-Clustering und *Correlation Coefficient*-Clustering durchgeführt werden können. Darüber hinaus wird die Anfertigung verschiedener Arten von Graphen mit den Maßzahlen aller oder ausgewählter *probe sets* ermöglicht. Nicht zuletzt können Kandidatengenlisten zu *probe lists* zusammengefasst und exportiert werden. Das Benutzerhandbuch[7] erklärt die Funktionalität von DMT näher.

Der LIMS Manager (verwendete Version: 3.0) dient zur Verwaltung des Servers. Mit ihm können Experimente in *process*- oder *publish*-Datenbank nach verschiedenen Kriterien aufgelistet, gelöscht, archiviert oder importiert und Templates und User Sets verwaltet werden. Eine weitere wichtige Funktion ist die Benutzerverwaltung mit der Zuweisung von Benutzerrechten. Der LIMS Manager wird im Gegensatz zu den anderen Anwendungen weniger von den Benutzern als vielmehr von den Administratoren eingesetzt.

2.2.2 Back-End-Software und zugrunde liegende Konzepte

Der LIMS-Server und die Server-Software (verwendete Version: 3.0) liegen aus Sicht der Benutzer im Hintergrund. Details zur Installation und Wartung des Servers finden sich im *LIMS Server Installation and Administration Guide*[8]. Neben dem Server-

Betriebssystem sind ein Datenbanksystem (hier verwendet: Microsoft SQL Server) und die LIMS-Software installiert. Letztere umfasst hauptsächlich Dienste, Datenbanktabellen und ODBC-Einträge (*Open DataBase Connectivity*).

Im Folgenden werden die wichtigsten technischen Konzepte vorgestellt, die zum Verständnis der in Kapitel 3 vorgestellten SPLUS-Funktionen unerlässlich sind. Detailliertere Informationen finden sich in den jeweils angegebenen weiterführenden Dokumenten.

Datenbankstruktur

Die Struktur der *process*-Datenbank ist nicht dokumentiert. Das AADM-Schema, welches die Struktur einer *publish*-Datenbank beschreibt, findet sich zusammen mit einer ausführlichen Dokumentation auf der Affymetrix-Website[1]. Über ODBC-Verbindungen können Datenbankabfragen von *publish*-Datenbanken vorgenommen werden. Abbildung 6 enthält einen Ausschnitt des kompletten AADM-Schemas.

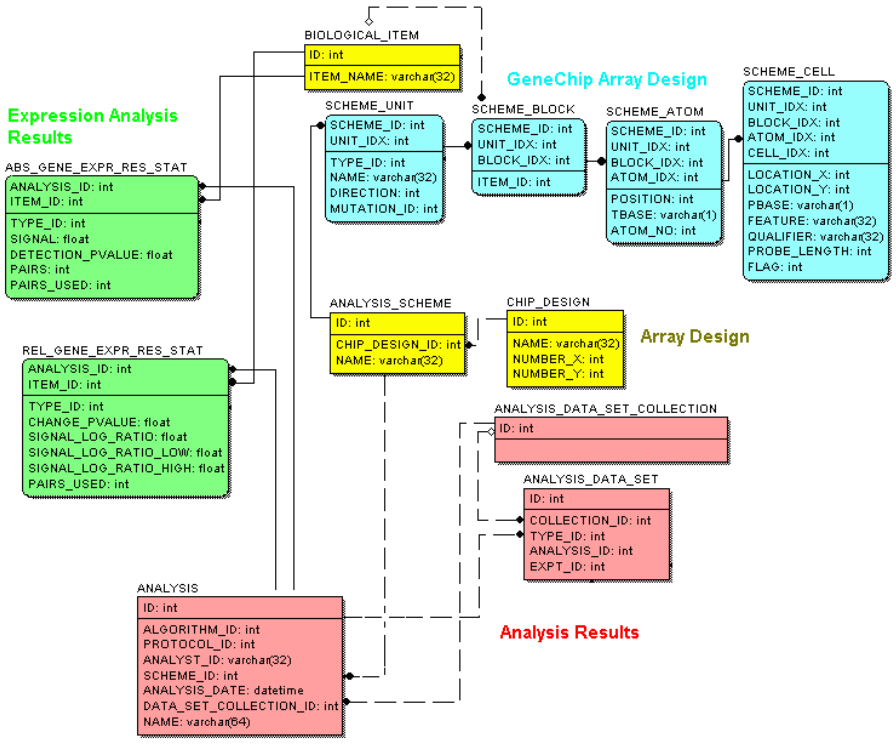


Abbildung 6: AADM-Schema (Ausschnitt)
(Abb. modifiziert nach: Affymetrix[1])

Eine zentrale Stellung nimmt die ANALYSIS-Tabelle ein. Sie enthält unter anderem die Namen aller in der *publish*-Datenbank gespeicherten Primäranalysen (CHP-Dateien). Außerdem kann über eine SCHEME_ID-Spalte mithilfe der Tabelle ANALYSIS_SCHEME der Array-Typ eines Experimentes ermittelt werden. Dabei enthält die Spalte NAME den sprechenden Namen (wie z. B. „HG-U95A“). Zudem ist die interne Chip-ID hier abgelegt. Über die Spalte DATA_SET_COLLECTION_ID der ANALYSIS-Tabelle und über die Tabellen ANALYSIS_DATA_SET und EXPERIMENT werden die Namen der zugehörigen Experimente referenziert. Die Maßzahlen einer Primäranalyse sind in den Tabellen ABS_GENE_EXPR_RES_STAT und REL_GENE_EXPR_RES_STAT (für eine *comparison analysis*) enthalten.

Das Layout eines Arrays, also die *probe cell*-Informationen einer (X,Y)-Position, kann mit den Tabellen SCHEME_UNIT, SCHEME_BLOCK, SCHEME_ATOM und SCHEME_CELL nachvollzogen werden. Dabei ermöglicht die Tabelle BIOLOGICAL_ITEM ein Mapping eines *probe set*-Namens (z. B. „100_g_at“) auf seinen entsprechenden ITEM_ID-Bezeichner. Äquivalente Informationen sind in der CDF-Datei eines Arrays enthalten.

Durch Re-Engineering konnte die Tabelle BIOLOGICAL.DESCRIPTION der GENEINFO-Datenbank der *process*-Datenbank als Speicherort für die *probe set descriptions* identifiziert werden.

Dateiformate

DAT- und CHP-Dateien sind binär codiert, ihre Informationen können nur über das *Affymetrix File Software Development Kit*[4] (File SDK) extrahiert werden, welches hier nicht zur Verfügung stand. Die Informationen einer CHP-Datei können trotzdem über den Umweg einer *publish*-Datenbank zugänglich gemacht werden.

Die Intensitätsdaten eines Experimentes sind in der CEL-Datei als tabulatorgetrennter Text gespeichert. Der nachfolgend erklärte Aufbau einer CEL-Datei ist von Affymetrix nicht dokumentiert und kann sich daher in zukünftigen Versionen des LIMS-Systems ändern.

Es liegen sechs Abschnitte vor, die jeweils mit den Zeilen [CEL], [HEADER], [INTENSITY], [MASKS], [OUTLIERS] und [MODIFIED] eingeleitet werden. [CEL]-

und [HEADER]-Abschnitt enthalten Meta-Informationen wie Versionsnummer des CEL-Formates und Anzahl der *probe cell*-Zeilen und -Spalten. Der [INTENSITY]-Abschnitt enthält neben der Anzahl der *probe cells* eine Zeile pro *probe cell*. Die Anzahl der Zeilen kann also sehr groß sein, beim HG-U95A-Array handelt es sich beispielsweise um 409600 Zeilen. Für jede *probe cell* existieren fünf Spalten X, Y, MEAN, STDV und NPIXELS. Während X und Y die Position der *probe cell* auf dem Array spezifizieren, enthält MEAN mit der durchschnittlichen Intensität der verwendeten Pixel aus der DAT-Datei die wesentliche Information der CEL-Datei (siehe Abbildung 7). Man beachte, dass hier keine Zuordnung einer (X,Y)-Position zu einem *probe set* stattfindet. Für diesen Zweck werden die Layout-Informationen (CDF-Datei) benötigt, deren Struktur im nächsten Abschnitt beschrieben wird.

Nach dem [INTENSITY]-Abschnitt folgt im [MASKS]-Abschnitt eine Auflistung der (X,Y)-Positionen aller maskierten *probe cells* und im [OUTLIERS]-Abschnitt eine Auflistung aller outlier-*probe cells*. Der [MODIFIED]-Abschnitt wird hier nicht verwendet.

```
[INTENSITY]
NumberCells=285156
CellHeader=X    Y       MEAN        STDV        NPIXELS
  0             0        1844.5      302.8       36
  1             0       46224.0     2765.3       36
  2             0        1894.3      229.2       36
  3             0       46224.0     3127.1       36
  4             0         706.8      144.6       36
  5             0        1477.8      253.8       36
  6             0       45422.5     4925.1       36
  7             0        1468.3      197.7       36
  8             0       45402.0     5774.7       36
  9             0        1604.0      468.0       36
 10             0       45162.5    12993.8       36
 11             0        1573.0      295.7       36
 12             0       45680.5    12583.2       36
 13             0        1518.3      228.1       36
 14             0       44643.5    12009.2       36
 15             0        1367.3      232.9       36
 16             0       45255.5    10672.0       36
 17             0        1496.0      224.3       36
 18             0       45897.5     9758.8       36
 19             0        1348.5      222.6       36
```

Abbildung 7: [INTENSITY]-Abschnitt einer CEL-Datei (Ausschnitt)

Array-Layout

Auch der Aufbau der Layout-beschreibenden Dateien ist von Affymetrix nicht dokumentiert. In den Funktionen aus Kapitel 3 wird jedoch auf Layout-Informationen zugegriffen, daher ist der Aufbau der CDF- und PSI-Dateien in der aktuellen Version von Bedeutung.

Neben *probe cells*, die zu *probe sets* gehören, existieren auf jedem Array auch *probe cells*, die zu so genannten *Quality Features* gehören. Hierzu zählen beispielsweise der Oligo B2-Rand zur automatischen Ausrichtung des Gitters bei der Umwandlung der Bilddaten (DAT-Datei) in Intensitätsdaten (CEL-Datei) sowie die *probe cells*, welche den Array-Typ als Text aus *probe cells* nachbilden. Außerdem ist eine geringe Anzahl von *probe cells* vorhanden, die weder zu *probe sets* noch zu *Quality Features* gehören, also leer sind. Sowohl *probe sets* als auch *Quality Features* werden in der CDF-Datei als Units bezeichnet und erhalten einen Unit-Bezeichner.

Die CDF-Datei eines Arrays enthält zunächst Abschnitte, die von den Zeilen [CDF] und [Chip] angeführt werden, in denen Meta-Informationen gespeichert sind, wie beispielsweise die Anzahl der *probe cell*-Zeilen und -Spalten, die Anzahl der Units und die Anzahl der *Quality Features* des Arrays. Da die Units nicht fortlaufend nummeriert sein müssen, ist außerdem ein Eintrag `MaxUnit` enthalten.

Nach den beiden einleitenden Abschnitten folgen zunächst Abschnitte, die die *Quality Features* beschreiben und dann Abschnitte, die die eigentlichen Units, also die *probe sets*, beschreiben (siehe Abbildung 8). Jeder Abschnitt enthält eine Zeile für jede *probe cell* einer Unit. Der wichtigste Eintrag ist jeweils die (X,Y)-Position auf dem Array. Außerdem kann ermittelt werden, welche *probe cells* ein *probe pair* bilden. Auf diese Weise können die Koordinaten aller *probe cells* eines *probe sets* bestimmt werden sowie die *Perfect Match* / *Mismatch*-Paare. Darüber hinaus ist ablesbar, welche Base den *Mismatch* bildet.

```
[Unit10_Block1]
Name=AFFX-BioB-5_at
BlockNumber=1
NumAtoms=20
NumCells=40
StartPosition=1
StopPosition=20
CellHeader=X        Y       PROBE     FEAT      QUAL      EXPOS     POS       CBASE     PBASE     TBASE
Cell1=1   12        N       control   AFFX-BioB-5_at      0         13        T         A         T
Cell2=1   13        N       control   AFFX-BioB-5_at      0         13        T         T         T
Cell3=2   12        N       control   AFFX-BioB-5_at      1         13        G         C         G
Cell4=2   13        N       control   AFFX-BioB-5_at      1         13        G         G         G
Cell5=3   13        N       control   AFFX-BioB-5_at      2         13        C         C         C
Cell6=3   12        N       control   AFFX-BioB-5_at      2         13        C         G         C
Cell7=4   13        N       control   AFFX-BioB-5_at      3         13        A         A         A
Cell8=4   12        N       control   AFFX-BioB-5_at      3         13        A         T         A
Cell9=5   12        N       control   AFFX-BioB-5_at      4         13        G         C         G
Cell10=5  13        N       control   AFFX-BioB-5_at      4         13        G         G         G
Cell11=6  12        N       control   AFFX-BioB-5_at      5         13        G         C         G
Cell12=6  13        N       control   AFFX-BioB-5_at      5         13        G         G         G
Cell13=7  13        N       control   AFFX-BioB-5_at      6         13        A         A         A
Cell14=7  12        N       control   AFFX-BioB-5_at      6         13        A         T         A
Cell15=8  13        N       control   AFFX-BioB-5_at      7         13        A         A         A
Cell16=8  12        N       control   AFFX-BioB-5_at      7         13        A         T         A
Cell17=9  13        N       control   AFFX-BioB-5_at      8         13        C         C         C
Cell18=9  12        N       control   AFFX-BioB-5_at      8         13        C         G         C
Cell19=10 13        N       control   AFFX-BioB-5_at      9         13        C         C         C
Cell20=10 12        N       control   AFFX-BioB-5_at      9         13        C         G         C
Cell21=11 13        N       control   AFFX-BioB-5_at      10        13        C         C         C
Cell22=11 12        N       control   AFFX-BioB-5_at      10        13        C         G         C
Cell23=12 13        N       control   AFFX-BioB-5_at      11        13        C         C         C
```

Abbildung 8: Unit-beschreibender Abschnitt einer CDF-Datei (Ausschnitt)

LIMS Software Development Kit (LIMS SDK)

Das LIMS SDK stellt eine Schnittstelle zur Software- und Datenumgebung des LIMS-Servers dar und bietet eigenen Programmen die Möglichkeit, die LIMS-Funktionalität bezüglich Experimentdatenbank, Workflow, Benutzerverwaltung, *publish*-Vorgang, Import, und Primäranalysen zu nutzen. Dem in dieser Arbeit verwendeten LIMS SDK lag eine ausführliche Dokumentation bei, die auch im Internet eingesehen werden konnte. Zwischenzeitlich hat der Hersteller eine Versions- und Namensänderung vorgenommen.

Das SDK ist als C++- oder Java-Version erhältlich. Im Rahmen dieser Arbeit wird die C++-Version eingesetzt, wobei die Objekte des SDK über den Mechanismus der *Type Libraries* bekannt gemacht werden. Die Objekte wiederum verwenden den DCOM-Mechanismus (*Distributed Component Object Model*) zur Kommunikation mit dem Server. Dabei können Objekte lokal instanziiert werden, die eigentlich *remote* angelegt werden.

Es existieren zwei Arten von Objekten: Die *Action Objects* und die *Data Container Objects*. Erstere sind Objekte, mit deren Hilfe Aktionen ausgeführt und Veränderungen des bestehenden Zustands vorgenommen werden können (siehe Tabelle 1). Letztere bilden den bestehenden Zustand in Objekten ab oder ermöglichen die Speicherung von Daten. Hierzu gehören beispielsweise Chip-, ChipType-, Experiment-, FileType- und Sample-Objekt. Außerdem existiert ein spezielles `CGcdoClient`-Objekt, mit dem eine vereinfachte Nutzung des DCOM-Mechanismus und der Verbindung zum LIMS-Server ermöglicht wird.

Objektbezeichner	**Funktion**
Admin	Benutzer- und Rollenverwaltung, Task-Verwaltung (*publishing* mehrerer Analysen)
Connection	Verbindung zur *process*-Datenbank des LIMS-Servers
Manager	Zusammenstellen von Data Container-Objekten nach verschiedenen Kriterien
PublishData	*publishing* einer Analyse
Workflow	Verwalten der Workflow-Informationen (kein physikalisches Erzeugen von Dateien)
Import	Importieren eines Experimentes (nur Einträge in der *process*-Datenbank, kein Kopieren oder Erzeugen von Dateien)
Analysis	Anstoßen einer Primäranalyse

Tabelle 1: LIMS SDK - Action Objects

Zur Bekanntmachung der LIMS SDK-Objekte über die *Type Libraries* müssen diese in einem Unterverzeichnis `imports` abgelegt sein und in einer globalen Header-Datei (`StdAfx.h`) die entsprechenden `import`-Befehle aufgenommen werden, also beispielsweise:

```
#import "imports\GcdoAdmin.tlb" no_namespace, named_guids
```

Ein Beispiel für die Instanziierung eines Manager-Objektes mithilfe des `CGcdoClient`-Hilfsobjektes `m_Gcdo` in C++-Syntax ist:

```
pManag=(IManagerPtr) m_Gcdo.GetObj(CLSID_Manager,&IID_IManager);
```

2.3 Verwendete Datensätze

In diesem Unterkapitel werden die hier verwendeten Datensätze vorgestellt. Dabei handelt es sich sowohl um Datensätze aus Münster, die alle in demselben Labor erzeugt wurden, als auch um Datensätze, die öffentlich über das Internet zugänglich waren.

Drei der Datensätze bestehen aus Experimenten mit HG-U95A / HG-U95Av2-Arrays, welche mittlerweile durch neuere Array-Typen abgelöst wurden (HG-U133), deren prinzipielle Eigenschaften sich aber nur wenig geändert haben. Aufgrund der Verfügbarkeit einer großen Anzahl technischer Replikate in einem Datensatz mit mausgenomspezifischen Arrays (Mu11KsubA- und Mu11KsubB) wurden diese Daten in die Untersuchung einbezogen.

Alle Experimente fanden vor der Scanner-Neujustierung auf einen anderen dynamischen Bereich vom 18.1.2002 statt und werden daher mit der Affymetrix-Skalierung auf einen Ziel-*Signal*-Wert von 1000 skaliert. Die Unterschiede zwischen HG-U95A und HG-U95Av2 sind für die Betrachtungen in dieser Arbeit irrelevant, daher werden sie wie derselbe Array-Typ behandelt.

2.3.1 Datensätze aus Münster

MS1

Der MS1-Datensatz umfasst 23 Chip-Experimente mit HG-U95A-Arrays, bei denen aus dem Vollblut von zehn Probanden (fünf männlich, fünf weiblich) RNA extrahiert und aufgearbeitet wurde. Diese Experimente waren als *proof of principle*-Untersuchung angelegt. Mit ihnen sollte gezeigt werden, dass zwischen dem Genexpressionsprofil von

Männern und Frauen grundsätzliche Unterschiede bestehen, die nicht durch geschlechtschromosomal kodierte Gene hervorgerufen wurden. Tabelle 2 enthält eine Aufstellung der Experimente.

N	EXPERIMENT.NAME	ARRAY.TYPE	PROBAND	VERSUCH	GESCHLECHT
093	BF-280201	HG-U95A	B	Mann/Frau	w
095	EL-280201	HG-U95A	EL	Mann/Frau	w
097	RV-280201	HG-U95A	R	Mann/Frau	m
100	26-102-150301	HG-U95A	H	Basis	m
106	26-103-150301	HG-U95A	R	Basis	m
137	S29-113-280601	HG-U95A	S	Basis	w
139	S29-115-280601	HG-U95A	R	Basis	m
141	S29-116-280601	HG-U95A	EB	Basis	m
143	S29-117-280601	HG-U95A	B	Basis	w
145	S29-118-280601	HG-U95A	EL	Basis	w
147	S30-120-050701	HG-U95A	B	Basis	w
149	S30-121-050701	HG-U95A	EL	Basis	w
151	S30-122-050701	HG-U95A	R	Basis	m
153	S30-123-050701	HG-U95A	EB	Basis	m
155	S31-125-250701	HG-U95Av2	H	Basis	m
159	S31-127-250701	HG-U95Av2	EB	Basis	m
161	S32-128-250701	HG-U95Av2	H	Basis	m
163	S32-129-250701	HG-U95Av2	S	Basis	w
165	S33-130-300801	HG-U95Av2	S	Basis	w
183	S-41.154-131101	HG-U95Av2	P	Basis	m
185	S-41.155-131101	HG-U95Av2	G	Basis	m
187	S-41.156-131101	HG-U95Av2	K	Basis	w
189	S-41.157-131101	HG-U95Av2	A	Basis	w

Tabelle 2: Experimente des MS1-Datensatzes

Gruppiert nach GESCHLECHT und PROBAND, ergeben sich zehn Gruppen (siehe Tabelle 3).

Gruppen-nummer	Gruppen-bezeichner	GESCHLECHT	PROBAND	Größe
1	"m|EB"	m	EB	3
2	"m|G"	m	G	1
3	"m|H"	m	H	3
4	"m|P"	m	P	1
5	"m|R"	m	R	4
6	"w|A"	w	A	1
7	"w|B"	w	B	3
8	"w|EL"	w	EL	3
9	"w|K"	w	K	1
10	"w|S"	w	S	3

Tabelle 3: Gruppierung des MS1-Datensatzes

MS_MuA und MS_MuB

Diese Datensätze bestehen aus 34 Mu11KsubA- bzw. 34 Mu11KsubB-Experimenten, bei denen die RNA-Expression in Maus-Knochenmarkszellen (Granulozyten) gemessen wurde. Dabei wurde zum einen der Wildtyp (Plus) und zum anderen eine MRP14(S100A9)-Knockout-Maus (Minus) untersucht (Beschreibung der Knockout-Maus siehe Hobbs et al.[44] und Manitz et al.[64]).

Mit den Experimenten sollten – basierend auf den gefundenen Ergebnissen – Rückschlüsse auf die Funktion des MRP14-Proteins gezogen werden. Bekannt ist, dass die S100-Proteine MRP8 und MRP14 gewebespezifisch nur in Granulozyten und Monozyten exprimiert werden bzw. in entzündlich aktivierten Keratinozyten (Nacken et al.[65]). Während der Differenzierung von Monozyten zu Makrophagen werden beide Proteine sowohl auf RNA- als auch auf Protein-Ebene herunterreguliert. Der Nachweis der Proteine in der Frühphase von entzündlichen Reaktionen lässt die Annahme zu, dass diese proinflammatorische Aktivität aufweisen, also vermutlich an der Immunabwehr beteiligt sind. So konnten *in-vitro* beispielsweise bakterienhemmende Funktionen und eine Inhibition der Casein-KinaseII gezeigt werden.

Für die beiden Proteine werden sowohl intra- als auch extrazelluläre Funktionen beschrieben (Donato[30]). Die oben erwähnte antibakterielle Wirkung beruht darauf, dass MRP8/MRP14 Zink chelatiert, welches daraufhin nicht mehr für das Wachstum der Bakterien zur Verfügung steht (Clohessy and Golden[28]; Sohnle et al.[82]). Zur dafür notwendigen Sezernierung der Proteine wird ein neuer Tubulin-abhängiger Transportweg benutzt (Rammes et al.[75]).

Die intrazelluläre Funktion ist noch unklar. Zum einen fungieren die Proteine als Fettsäuretransporter im Arachidonsäure-Stoffwechsel (Kerkhoff et al.[52]). Zum anderen lässt sich nachweisen, dass der Komplex sowie dessen phosphorylierte Isoform an Migrationsvorgängen in Monozyten beteiligt sind. Der Mechanismus verläuft vermutlich über die Beeinflussung des Umbaus des Zytoskeletts während der Migration. In diesem Zusammenhang spielen auch Änderungen im Oligomerisierungsgrad des Komplexes nach Calciumbindung eine wichtige Rolle (Strupat et al.[85]; Vogl et al.[93]).

Es wurden Knochenmarkszellen (ohne Erythrozyten) isoliert und in Teflonbags kultiviert. Zwei Kontrollversuche (Kontrolle0h bzw. Kontrolle4h, mehrere Wiederholungen) fanden statt, bei denen diese Zellen direkt nach der Entnahme bzw. nach

vier Stunden Ruhe lysiert und daraus die Gesamt-RNA aufgearbeitet wurden (Kontrolle4h: Diese Zeitspanne wurde gewählt, damit die Zellen über einen Zeitraum von vier Stunden durch verschiedene Stimuli aktiviert werden konnten. Diese Stimulationszeit gilt allgemein als ausreichend zur Erfassung der meisten Änderungen in der RNA-Expression. Sehr frühe bzw. späte Ereignisse in der Genregulation können dadurch jedoch nicht komplett erfasst werden). Zusätzlich zu den beiden Kontrollversuchen wurden die Knochenmarkszellen mit verschiedenen Stimulanzien aktiviert, wie z. B. LTB-4, LPS (Zellwandbestandteil des gram-negativen Bakteriums), A23187 (Calcium-Ionophor) und anderen.

Mit der RNA einer Aufarbeitung wurde sowohl ein Mu11KsubA-, als auch ein Mu11KsubB-Chip befüllt, indem mithilfe des Affymetrix-Standardprotokolls die doppelte Menge an RNA aufgearbeitet und dieses Sample erst beim Befüllen der Chips A und B getrennt wurde. Die auf diese Weise erzeugten technischen Replikate ermöglichen Folgerungen bezüglich technischer oder biologischer Varianz.

Eine Aufstellung der Mu11KsubA-Experimente (MS_MuA) findet sich in Tabelle 4.

N	EXPERIMENT.NAME	ARRAY.TYPE	MAUS.GENOTYP	STIMULATION
66	V13MLT4-280601A	Mu11KsubA	Minus	LTB-4
70	V13PLT4-280601A	Mu11KsubA	Plus	LTB-4
74	V14MK4-270601A	Mu11KsubA	Minus	Kontrolle4h
78	V14PK4-270601A	Mu11KsubA	Plus	Kontrolle4h
82	V15MK4-270601A	Mu11KsubA	Minus	Kontrolle4h
86	V15PK4-270601A	Mu11KsubA	Plus	Kontrolle4h
90	V16MA4A-260601A	Mu11KsubA	Minus	A23.4h
94	V16MA4B-260601A	Mu11KsubA	Minus	A23.4h
98	V16PA4A-260601A	Mu11KsubA	Plus	A23.4h
102	V16PA4B-260601A	Mu11KsubA	Plus	A23.4h
110	V17ML4-130701A	Mu11KsubA	Minus	LPS4h
118	V17PL4-130701A	Mu11KsubA	Plus	LPS4h
126	V18ML4-130701A	Mu11KsubA	Minus	LPS4h
134	V18PL4-130701A	Mu11KsubA	Plus	LPS4h
138	V19MA4-200701A	Mu11KsubA	Minus	A23.4h
142	V19PA4-200701A	Mu11KsubA	Plus	A23.4h
146	V1MK1-260301A	Mu11KsubA	Minus	Kontrolle0h
150	V1PK1-260301A	Mu11KsubA	Plus	Kontrolle0h
154	V20MK1-200701A	Mu11KsubA	Minus	Kontrolle0h
158	V20PK1-200701A	Mu11KsubA	Plus	Kontrolle0h
162	V3MK4-260301A	Mu11KsubA	Minus	Kontrolle4h
166	V3PK4-260301A	Mu11KsubA	Plus	Kontrolle4h
170	V4MK4-260301A	Mu11KsubA	Minus	Kontrolle4h
174	V4PK4-260301A	Mu11KsubA	Plus	Kontrolle4h
178	V6MP4-270301A	Mu11KsubA	Minus	PMA4h
182	V6PP4-270301A	Mu11KsubA	Plus	PMA4h
187	V7MK4-200401A	Mu11KsubA	Minus	Kontrolle4h
195	V7ML4-200401A	Mu11KsubA	Minus	LPS4h
203	V7PK4-270301A	Mu11KsubA	Plus	Kontrolle4h
207	V7PL4-200401A	Mu11KsubA	Plus	LPS4h
215	V8MA4-040401A	Mu11KsubA	Minus	A23.4h
219	V8PA4-040401A	Mu11KsubA	Plus	A23.4h
223	V9MM1-040401A	Mu11KsubA	Minus	Milz
227	V9PM1-040401A	Mu11KsubA	Plus	Milz

Tabelle 4: Experimente des MS_MuA-Datensatzes

Gruppiert nach MAUS.GENOTYP und STIMULATION ergeben sich daraus 14 Gruppen (siehe Tabelle 5).

Gruppen-nummer	Gruppen-bezeichner	MAUS.GENOTYP	STIMULATION	Größe
1	"Minus|A23.4h"	Minus	A23.4h	4
2	"Minus|Kontrolle0h"	Minus	Kontrolle0h	2
3	"Minus|Kontrolle4h"	Minus	Kontrolle4h	5
4	"Minus|LPS4h"	Minus	LPS4h	3
5	"Minus|LTB-4"	Minus	LTB-4	1
6	"Minus|Milz"	Minus	Milz	1
7	"Minus|PMA4h"	Minus	PMA4h	1
8	"Plus|A23.4h"	Plus	A23.4h	4
9	"Plus|Kontrolle0h"	Plus	Kontrolle0h	2
10	"Plus|Kontrolle4h"	Plus	Kontrolle4h	5
11	"Plus|LPS4h"	Plus	LPS4h	3
12	"Plus|LTB-4"	Plus	LTB-4	1
13	"Plus|Milz"	Plus	Milz	1
14	"Plus|PMA4h"	Plus	PMA4h	1

Tabelle 5: Gruppierung des MS_MuA-Datensatzes

Eine Aufstellung der Mu11KsubB-Experimente (MS_MuB) findet sich in Tabelle 6.

N	EXPERIMENT.NAME	ARRAY.TYPE	MAUS.GENOTYP	STIMULATION
68	V13MLT4-280601B	Mu11KsubB	Minus	LTB-4
72	V13PLT4-280601B	Mu11KsubB	Plus	LTB-4
76	V14MK4-270601B	Mu11KsubB	Minus	Kontrolle4h
80	V14PK4-270601B	Mu11KsubB	Plus	Kontrolle4h
84	V15MK4-270601B	Mu11KsubB	Minus	Kontrolle4h
88	V15PK4-270601B	Mu11KsubB	Plus	Kontrolle4h
92	V16MA4A-260601B	Mu11KsubB	Minus	A23.4h
96	V16MA4B-260601B	Mu11KsubB	Minus	A23.4h
100	V16PA4A-260601B	Mu11KsubB	Plus	A23.4h
104	V16PA4B-260601B	Mu11KsubB	Plus	A23.4h
112	V17ML4-130701B	Mu11KsubB	Minus	LPS4h
120	V17PL4-130701B	Mu11KsubB	Plus	LPS4h
128	V18ML4-130701B	Mu11KsubB	Minus	LPS4h
136	V18PL4-130701B	Mu11KsubB	Plus	LPS4h
140	V19MA4-200701B	Mu11KsubB	Minus	A23.4h
144	V19PA4-200701B	Mu11KsubB	Plus	A23.4h
148	V1MK1-260301B	Mu11KsubB	Minus	Kontrolle0h
152	V1PK1-260301B	Mu11KsubB	Plus	Kontrolle0h
156	V20MK1-200701B	Mu11KsubB	Minus	Kontrolle0h
160	V20PK1-200701B	Mu11KsubB	Plus	Kontrolle0h
164	V3MK4-260301B	Mu11KsubB	Minus	Kontrolle4h
168	V3PK4-260301B	Mu11KsubB	Plus	Kontrolle4h
172	V4MK4-260301B	Mu11KsubB	Minus	Kontrolle4h
176	V4PK4-260301B	Mu11KsubB	Plus	Kontrolle4h
180	V6MP4-270301B	Mu11KsubB	Minus	PMA4h
184	V6PP4-270301B	Mu11KsubB	Plus	PMA4h
189	V7MK4-200401B	Mu11KsubB	Minus	Kontrolle4h
197	V7ML4-200401B	Mu11KsubB	Minus	LPS4h
205	V7PK4-270301B	Mu11KsubB	Plus	Kontrolle4h
209	V7PL4-200401B	Mu11KsubB	Plus	LPS4h
217	V8MA4-040401B	Mu11KsubB	Minus	A23.4h
221	V8PA4-040401B	Mu11KsubB	Plus	A23.4h
225	V9MM1-040401B	Mu11KsubB	Minus	Milz
229	V9PM1-040401B	Mu11KsubB	Plus	Milz

Tabelle 6: Experimente des MS_MuB-Datensatzes

Gruppiert nach MAUS.GENOTYP und STIMULATION ergeben sich daraus 14 Gruppen analog zu denen in Tabelle 5.

MS2

Dieser Datensatz entstand bei der Etablierung und Charakterisierung eines Zellkulturmodells für M-Zellen. M-Zellen sind spezielle Zellen im Epithel des Dünndarms, die eine wesentliche Rolle bei der Initiierung einer primären, schnellen Immunantwort des Organismus innehaben (Kraehenbuhl et al.[56]). Auf einem Costar-Transwell-Filter werden zunächst Zellen einer Zelllinie aus einem Colon-Adenokarzinom (Caco-2) ausgesät. Sobald diese ausdifferenziert sind, werden auf der Gegenseite des Caco-Zellrasens Lymphozyten aus humanem venösen Blut zugefügt. Poren im Filter ermöglichen es den Lymphozyten, in Kontakt mit Caco-Zellen zu kommen. Als Folge dieses Kontaktes werden einige Zellen des Epithelrasens zu so genannten „M-Zell-ähnlichen Zellen“ umdifferenziert, die im Wesentlichen die Morphologie und Funktion von M-Zellen im Dünndarm aufweisen (Kerneis et al.[53]; El Bahi et al.[33]).

Zur Untersuchung des Unterschiedes zwischen Caco-2-Zellen in Anwesenheit und Abwesenheit von humanen Blutlymphozyten wurden sieben Experimente durchgeführt (Aufstellung siehe Tabelle 7). In fünf Experimenten wurden Caco-Zellen auf einem Filter angezüchtet. In drei von diesen Experimenten wurden Lymphozyten zur Induktion hinzugefügt. Von diesen dreien sind zwei Experimente technische Replikate, sie wurden aus einem Sample hybridisiert. Bei einem der Experimente handelt es sich um ein biologisches Replikat. Es wurde aus einer anderen Kultur unter denselben Bedingungen gewonnen. In weiteren zwei Experimenten wurden einerseits Zellen der Caco-2-Zelllinie in einer Flasche angezüchtet, um eine Gegenprobe zur Zellzucht auf den Filtern zu erhalten und andererseits das Expressionsmuster von Lymphozyten bestimmt.

N	EXPERIMENT.NAME	ARRAY.TYPE	KULTUR	INDUZIERT	REPLIKAT
2	Coculture-Caco-Co-16-10	HG_U95A	FILTER	NEIN	BIOLOGISCH
4	Coculture-Caco1_08_01	HG_U95A	FILTER	JA	TECHNISCH
6	Coculture-Caco2_08_01	HG_U95A	FILTER	JA	TECHNISCH
7	Coculture-Caco_08_01_Kontr	HG_U95A	FILTER	NEIN	BIOLOGISCH
11	Coculture-Cacos-Lym-16-10	HG_U95A	FILTER	JA	BIOLOGISCH
15	Coculture-Cacos-pur-16-10	HG_U95A	FLASCHE	NEIN	
17	Coculture-Lymphos-16-10	HG_U95A	LYMPHOZYTEN	NEIN	

Tabelle 7: Experimente des MS2-Datensatzes

Gruppiert nach KULTUR und INDUZIERT ergeben sich daraus vier Gruppen (siehe Tabelle 8).

Gruppen-nummer	Gruppenbezeichner	KULTUR	INDUZIERT	Größe
1	"FILTER|JA"	FILTER	JA	3
2	"FILTER|NEIN"	FILTER	NEIN	2
3	"FLASCHE|NEIN"	FLASCHE	NEIN	1
4	"LYMPHOZYTEN|NEIN"	LYMPHOZYTEN	NEIN	1

Tabelle 8: Gruppierung des MS2-Datensatzes

2.3.2 Öffentlich zugängliche Datensätze

Affymetrix

Der *Latin square*-Datensatz von Affymetrix wurde eingesetzt, um den statistischen Kondensierungsalgorithmus der MAS 5.0 zu entwerfen und zu testen und um ihn mit dem empirischen Algorithmus der MAS 4.0 zu vergleichen. Gleichzeitig erlaubt dieser Datensatz die Prüfung des Vorliegens einer linearen Beziehung zwischen Stoffmenge in der Hybridisationslösung und dem *Signal*-Wert (siehe *Affymetrix Technical Note*[11]). Die Experimente dieses Datensatzes (HG-U95A-Arrays) wurden erstellt, indem 14 bekannte Transkripte in festen Konzentrationen (0 bis 1024 pM) zu einem komplexen menschlichen RNA-Sample hinzugefügt wurden. Die Konzentrationen der Transkripte variierte zwischen den Experimenten, sodass keine Konzentrationskombination doppelt vorkam. Sie folgt damit der mathematischen Definition eines *Latin square* als quadratische Matrix, in deren Zeilen und Spalten alle Elemente unterschiedlich sind. Nach Ausschluss von zwei Transkripten und einem Experiment schlechter Qualität verblieben 59 Experimente im Affymetrix-Datensatz, der auf der Affymetrix-Website[5] zugänglich ist.

Der Affymetrix-Datensatz dient in den Kapiteln 4 und 5 nur für Betrachtungen, die prinzipiell mit den anderen Datensätzen nicht möglich sind; die Einzelheiten dieser Betrachtungen werden dort genannt.

Klein

In Klein et al.[55] werden bestimmte Zellfraktionen von Patienten mit chronischer lymphatischer Leukämie vom B-Zelltyp mit denen von Patienten mit anderen Erkrankungen und mit verschiedenen Fraktionen normaler B-Zellen auf der Ebene ihrer

Expressionsprofile verglichen. Die CEL-Dateien dieser Experimente sind öffentlich zugänglich (siehe entsprechende Web-Seite[54]).

Von den normalen B-Zellen wurden in jeweils fünf Experimenten vier Fraktionen untersucht: naive B-Zellen vor Wanderung durch das Keimzentrum eines Lymphknotens (NAIV), Zentroblasten (CB) und Zentrozyten (CC) des Keimzentrums und memory-B-Zellen nach Wanderung durch das Keimzentrum (MEM). Diese 20 Experimente werden für die Betrachtungen dieser Arbeit verwendet (siehe Tabelle 9).

N	EXPERIMENT.NAME	ZELLTYP
1	CB 2-23	CB
2	CB 3-10	CB
3	CB 3-30	CB
4	CB 3-7	CB
5	CB 6-8	CB
6	CC 3-28	CC
7	CC 4-14	CC
8	CC 4-6	CC
9	CC 7-25	CC
10	CC 7-7	CC
11	M 4-12	MEM
12	M 4-14	MEM
13	M 4-26	MEM
14	M 5-2	MEM
15	M 6-8	MEM
16	N 4-13	NAIV
17	N 4-14	NAIV
18	N 4-7	NAIV
19	N1 4-21	NAIV
20	N2 4-21	NAIV

Tabelle 9: Experimente des Klein-Datensatzes

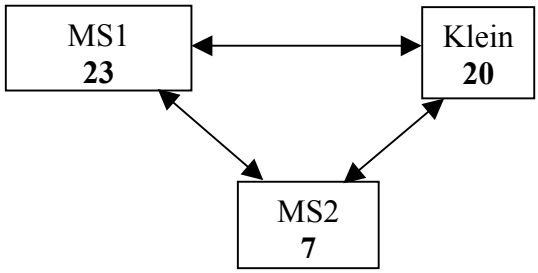
Gruppiert nach ZELLTYP ergeben sich daraus vier Gruppen (siehe Tabelle 10).

Gruppen-nummer	Gruppen-bezeichner	ZELLTYP	Größe
1	"CB"	CB	5
2	"CC"	CC	5
3	"MEM"	MEM	5
4	"NAIV"	NAIV	5

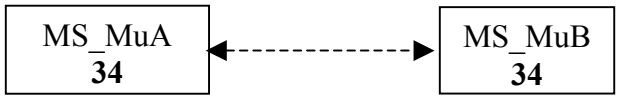
Tabelle 10: Gruppierung des Klein-Datensatzes

2.3.3 Vergleichsgruppen

Die Datensätze MS1, Klein und MS2 sind aufgrund ihres übereinstimmenden Array-Typs grundsätzlich miteinander vergleichbar; in ihren Experimenten wurden dieselben *probe sets* gemessen. Man beachte, dass die in den genannten Datensätzen untersuchten Proben sich sowohl innerhalb eines Datensatzes, als auch gerade zwischen den Datensätzen im experimentellen Setting mehr oder weniger stark unterscheiden. In Kapitel 4 und 5 werden Vergleiche zwischen diesen Datensätzen angestellt; sie gehören zur Vergleichsgruppe der U95A-Datensätze.



Die MS_Mu-Datensätze fallen streng genommen nicht in eine Vergleichsgruppe, da die verwendeten Arrays aus unterschiedlichen *probe sets* bestehen. Jedoch stellt jeweils ein Experiment aus MS_MuA ein quasi-technisches Replikat eines Experiments aus MS_MuB dar, weil sie mit einem Sample derselben Aufarbeitung hybridisiert und gescannt wurden. Es besteht die Hoffnung, dass mit diesem Wissen aus vergleichenden Betrachtungen Folgerungen gezogen werden können:



2.4 Eigenschaften von Box Plots

Mit einem Box Plot können in SPLUS Verteilungseigenschaften kontinuierlicher Variablen einer Stichprobe visualisiert werden. Ähnlich wie bei einem Histogramm ist es möglich, Häufungspunkt, Verteilungsbreite, Rechts- / Linksschiefe und Ausreißer zu identifizieren. Ein Box Plot kann als „Aufsicht“ auf ein Histogramm aufgefasst werden. Während sich mehrere Histogramme in einem Graphen gegenseitig verdecken können, liegen Box Plots prinzipiell separat voneinander (siehe Abbildung 9).

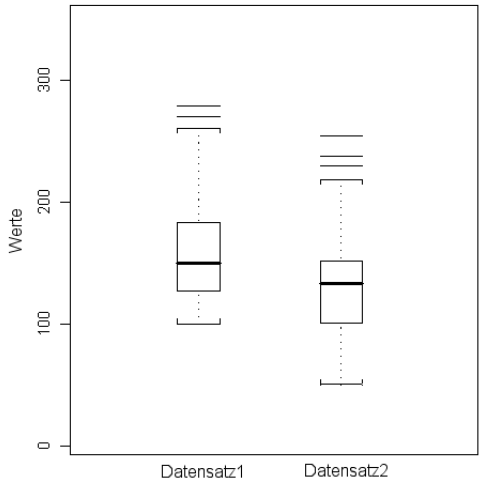


Abbildung 9: Beispiele für Box Plots

Die waagerechte Linie (**fett**) bezeichnet den Median der Daten; die „Box“ um ihn herum umfasst die „mittlere Hälfte“ der Daten vom ersten bis zum dritten Quartil einschließlich (Interquartilsabstand); die gepunkteten senkrechten Linien nach oben und unten (mit abschließender Klammer) werden „whiskers“ („Schnurrhaare“) genannt. Die Klammern sind dabei auf dem äußersten Datenpunkt platziert, der nicht jenseits eines Standardspanns vom Rand der Box entfernt liegt. Ein Standardspann ist dabei das 1,5-fache des Interquartilsabstandes. Datenpunkte, die jenseits davon liegen (so genannte Ausreißer oder *Outlier*), werden als separate waagerechte Linien eingezeichnet.

3 Material und Methoden – Handhabung der Daten

In diesem Kapitel werden die entwickelten Funktionen und Anwendungen sowie deren Schnittstellen skizziert. Eine ausführlichere Beschreibung der meisten Funktionen einschließlich ihrer wichtigsten Parameter findet sich im Anhang.

Unterkapitel 3.1 diskutiert die Vor- und Nachteile von SPLUS als Entwicklungsumgebung. In Unterkapitel 3.2 wird eine Übersicht über die implementierten Datenobjekte und deren Abhängigkeiten gegeben. Als Basis für den Großteil der SPLUS-Funktionalität nimmt der Algorithmus zum Gruppieren, Benennen und Kombinieren von Experimenten in Unterkapitel 3.3 eine zentrale Position ein. Eine Übersicht über weitere SPLUS-Funktionen findet sich in Unterkapitel 3.4. Das Unterkapitel 3.5 enthält die Beschreibung der C++-Komponenten der für diese Arbeit entwickelten Programmbibliothek. Dazu zählt auch die Client-Server-Anwendung „Evaluation Server“ zur effizienten Abfrage der LIMS-Datenbank, welche detailliert in Unterkapitel 3.6 beschrieben wird.

3.1 SPLUS als Entwicklungsumgebung

Eine Entwicklungsumgebung für Funktionen zur Handhabung von GeneChip-Daten sollte idealerweise bereits einen umfangreichen Satz an statistischen Funktionen mitbringen. Für die Ergänzung eventuell fehlender Funktionalität und die Anbindung an das GeneChip-System sollte aber auch eine mächtige Programmiersprache mit Bibliotheken für mathematische Operationen und für Betriebssysteminteraktionen zur Verfügung stehen. Um neue Eigenschaften von GeneChip-Experimenten aufzudecken, müsste auch eine möglichst einfache Visualisierung beliebiger Daten mithilfe verschiedener Graphen existieren. Die Betriebssystemumgebung Windows 2000 war durch das GeneChip-System und das vorgefundene Umfeld vorgegeben. Diese Faktoren zusammengenommen führten zur Verwendung des Statistikpakets SPLUS zunächst in der Version 2000 und später in der Version 6.1.

SPLUS bietet durch die Basis-Datenobjekte `numeric`, `double`, `character` und `logical` bereits eine maschinenferne Grundausstattung. Durch bereits implementierte komplexe Datenobjekte wie Vektor, Matrix und data.frame und effizient implementierte Operationen auf diesen Datenobjekten wird die einfache Umsetzung mathematisch formulierter Algorithmen in Programme ermöglicht. Typinferenz statt strenger

Typisierung vereinfacht den Übergang von nominalen in kontinuierliche Variablen. SPLUS enthält eine Vielzahl implementierter Statistik- und Graphen-Funktionen. Essenziell wichtig für die Entwicklung der Schnittstelle zum GeneChip-System sind die vorhandenen Möglichkeiten zur Anbindung von C++-Code. Dadurch stehen alle Betriebssystem- und LIMS SDK-Bibliotheken zur Verfügung. Das hierarchisch gegliederte datenbankähnliche Konzept für die permanente Speicherung einmal angelegter Datenobjekte kommt den unterschiedlichen Dateitypen des LIMS-Systems entgegen.

Während der Entwicklung mit SPLUS wurden neben den aufgeführten positiven Eigenschaften auch einige Nachteile deutlich: Im funktionalen Konzept der SPLUS-Programmiersprache stehen keine Pointer zur Verfügung, d. h. alle Zuweisungen führen zu einer internen Kopie auch großer Datenobjekte. Zwar existiert eine *garbage collection*, doch ist bei der Größe der in GeneChip-Experimenten anfallenden Datenmengen bei sehr langen Schleifen oder tiefen Rekursionen mit einem Voll- oder Überlaufen des Speichers zu rechnen. Dies trat beispielsweise bei der Anwendung komplexer Operationen auf Kombinationsobjekten (siehe weiter unten) auf. Da es sich bei SPLUS um eine Interpretersprache handelt, treten einige Fehler erst zur Laufzeit auf, und die Ausführungszeit ist in der Regel länger als bei einer Compiler-Sprache. Außerdem steht für SPLUS-Programme nicht ohne weiteres ein Front-End zur Interaktion mit dem Benutzer zur Verfügung. Alle Vorgaben und Spezifikationen werden über die Funktionsparameter gemacht. Ein Benutzer muss eine gewisse Programmiererfahrung mitbringen, um die entwickelten Funktionen sinnvoll zu nutzen.

3.2 Objektmodell – Datenstrukturen

Im Folgenden werden die verwendeten Datenstrukturen beschrieben. Sie bilden einerseits Objekte ab, die von der GeneChip-Technologie vorgegeben sind, wie beispielsweise Chip-Layout, Format der *publish*-Datenbanken, Intensitätsdaten und Primäranalysedaten, andererseits Objekte, die für die implementierte Funktionalität notwendig sind. Dazu zählen der beschreibende data.frame (desc.df) und Objekte zur Speicherung von Gruppierungen und Kombinationen.

Eine Übersicht über das Objektmodell bietet Abbildung 10.

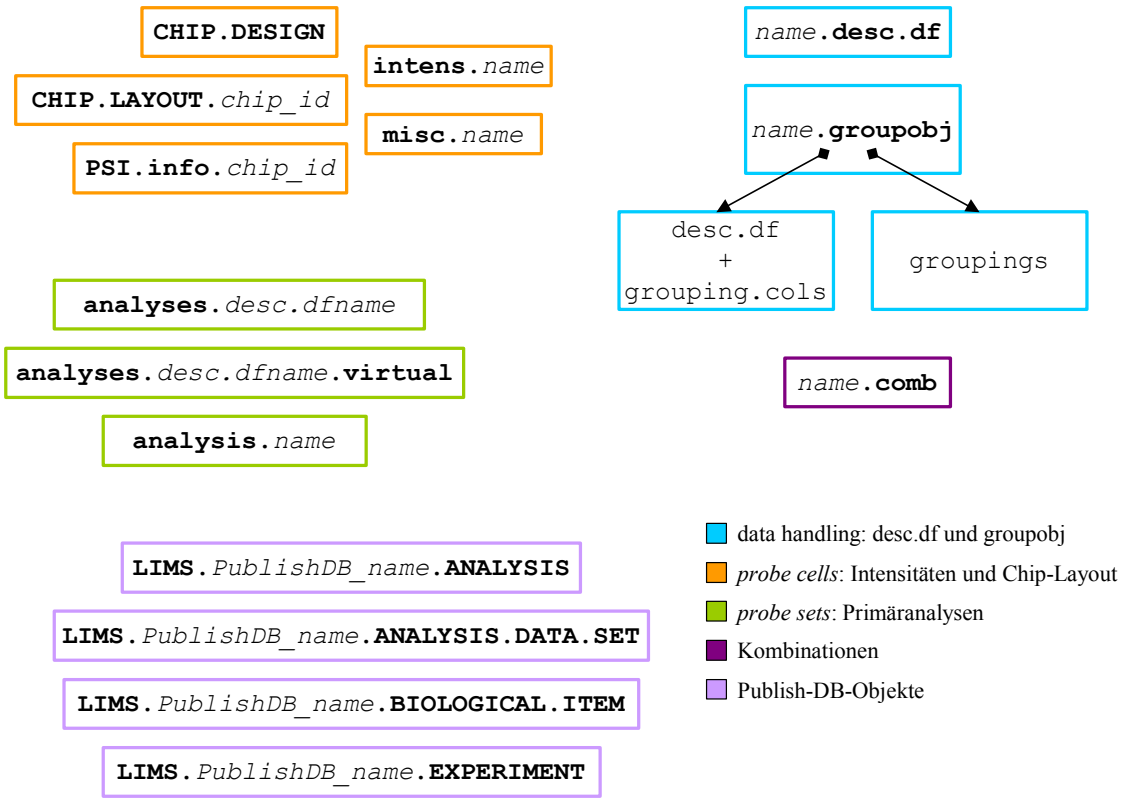


Abbildung 10: Übersicht über das Objektmodell

3.2.1 Chip-Layout und Intensitätsdaten

In einem data.frame `CHIP.DESIGN` wird der Bezeichner des Array-Typs (`CHIP.ID`), die Anzahl der Zeilen und Spalten und die Anzahl der *probe sets* gespeichert. Der data.frame wird durch einmaligen Aufruf der Funktion `create.CHIP.design.table` initialisiert. Dabei kann spezifiziert werden, welches Verzeichnis die Layout-beschreibenden Dateien enthält.

CHIP.DESIGN
CHIP.ID
NAME
NUMBER.X
NUMBER.Y
NUMBER.PROBE.SETS

Für jeden Array-Typ gibt es eine Liste `CHIP.LAYOUT.`*chip_id*.

`CHIP.LAYOUT.`*chip_id*

`CEL.Layout`
`PPInd`
`PMflag`
`NumProbeSets`
`NumQC`

Momentan gültige Werte für *chip.id* finden sich in Tabelle 11.

chip.id	**Array-Typ**
1	Hu6800
5	HC_G110
6	HG_U95A
10	HG_U95Av2
16	HG_U95B
17	HG_U95C
18	HG_U95D
19	HG_U95E
26	HG-U133A
25	Test1
9	Test2
24	Test3
14	Mu11KsubA
15	Mu11KsubB
11	MG_U74A
22	MG_U74Av2
12	RG_U34A
20	RG_U34B
21	RG_U34C
13	RN_U34

Tabelle 11: Mögliche Werte für *chip.id*

Die Komponente `CEL.Layout` ist ein Vektor, der für jede *probe cell* verzeichnet, zu welcher Unit sie gehört. `PPInd` listet für jede *probe cell* auf, welchem *probe pair* sie innerhalb des *probe sets* angehört. `PMflag` ist ein boolscher Vektor, der für jede *probe cell* bezeichnet, ob es sich um eine *Perfect Match-probe cell* (1) oder eine *Mismatch-probe cell* (0) handelt. Diese drei Vektoren haben die Länge `NUMBER.X * NUMBER.Y`. Die Komponente `NumProbeSets` enthält die Anzahl der *probe sets* auf dem Chip. `NumQC` enthält die Anzahl der *Quality Features* des Chips.

In der MAS beginnt die Nummerierung der Koordinaten einer *probe cell* mit 0. Dieser Festlegung folgend findet sich die Information einer *probe cell* mit den Koordinaten

(X, Y) an Position $Y \cdot \texttt{Number.X} + X + 1$ des Vektors (da dessen Nummerierung mit eins beginnt).

Der Vektor `CEL.Layout` enthält eine Null, wenn die *probe cell* weder einem *probe set* noch einem *Quality Feature* zugeordnet ist. Er enthält `maxint` – QCNum, wenn sie einem *Quality Feature* zugeordnet ist und eine *unit_number* mit $0 < \textit{unit\_number} < (\texttt{maxint} - \texttt{NumQC})$, wenn sie einem *probe set* zugeordnet ist. Über das nachfolgend beschriebene Objekt `PSI.info.`*chip_id* kann eine Zuordnung der *unit_number* zum *probe set*-Bezeichner (wie z. B. 35509_at) erfolgen.

Für jeden Array-Typ gibt es eine Liste `PSI.info.`*chip_id*.

PSI.info.*chip_id*
UNIT.number
probeset.name
number.cells

Die Komponente `UNIT.number` ist ein Vektor und enthält alle UNIT-Bezeichner des Chips. Die entsprechende Position des Vektors `probeset.name` enthält den *probe set*-Bezeichner im Klartext (z. B. „100_g_at“). Im Vektor `number.cells` ist vermerkt, aus wie vielen *probe pairs* das *probe set* besteht.

Die Objekte `CHIP.LAYOUT.`*chip_id* und `PSI.info.`*chip_id* werden durch einmaligen Aufruf von `import.chip.layout` initialisiert. Zum Befüllen der Objekte werden die CDF- und PSI-Dateien des jeweiligen Array-Typs benutzt.

Die Daten der eingelesenen CEL-Datei eines Experimentes mit dem Namen *name* werden in intens- und misc-Objekten gespeichert. Die fünf Spalten `X, Y, MEAN, STDV, NPIXELS` aus dem [INTENSITY]-Abschnitt mit den eigentlichen Intensitätsdaten in der `MEAN`-Spalte werden in `intens.`*name* abgelegt.

intens.*name*
X
Y
MEAN
STDV
NPIXELS

Die restlichen Abschnitte der CEL-Datei, also [HEADER]-, [MASKED]-, [OUTLIER]- und [MODIFIED] werden im Objekt `misc.`*name* gespeichert. Sie werden unter anderem benötigt, um eine in SPLUS eingelesene (und eventuell modifizierte) CEL-Datei wieder im Dateisystem abzuspeichern (siehe Funktion `export.CEL.file`).

misc.*name*
HEADER MASKED.OUTLIER.MODIFIED

3.2.2 Primäranalysen

Zu den Analysen der Experimente eines desc.df oder groupobj-Objekts werden Informationen in eigenen Datenobjekten gespeichert. Ein data.frame `analyses.`*hybridname* enthält die Informationen zu allen Primäranalysen, die von Experimenten durchgeführt wurden, die in einem desc.df oder groupobj-Objekt mit dem Namen *hybridname* enthalten sind.

analyses.*hybridname*
ANALYSIS.name EXPERIMENT.name emp.flag BF NF TGT SF ...

analysis.*name*
Probe.Set.Name Signal Detection.pvalue Abs.Call Pairs Pairs.Used

Angelegt wird ein solcher data.frame von der Funktion `import.analyses.descriptions`. Die Spalte `ANALYSIS.name` enthält danach die Namen der Analysen, wie sie in der *process*-Datenbank gespeichert sind. `EXPERIMENT.name` enthält den Namen des Experimentes, aus dem die Analyse entstanden ist. Von einem Experiment können mehrere Primäranalysen existieren, die mit

unterschiedlichen Parametern durchgeführt wurden. Die wichtigsten Analyseparameter werden in den weiteren Spalten `BF`, `NF`, `TGT` und `SF` aufgeführt. Sie enthalten die Informationen „Baseline File“ und „Normalization Factor“ und / oder „Target Intensity“ und „Scale Factor“. Mithilfe dieser Spalten können über die Funktion `get.analyses.of.experiment` die Namen von Analysen mit gleichen Analyseparametern extrahiert werden. Die Spalte `emp.flag` bezeichnet, ob die gespeicherte Analyse mit dem empirischen Algorithmus der früheren MAS-Versionen durchgeführt wurde.

Werden eigene Skalierungsmethoden auf Intensitätsebene angewandt und daraus resultierende Analysen in der *process*-Datenbank gespeichert (siehe Kapitel 5), so werden Informationen zu Skalierungsfaktor und Methode automatisch in dem data.frame `analyses.`*hybridname* gespeichert. Ihre Merkmale können naturgemäß nicht aus der *process*-Datenbank ermittelt werden. Daher kann die Funktion `import.analyses.descriptions` nach der Anwendung einer eigenen Skalierungsmethode nicht mehr aufgerufen werden, ohne diese Informationen zu überschreiben. Ein Weiterarbeiten mit ihnen in weiter gehenden Auswertungen oder grafischen Funktionen ist jedoch wie mit den Informationen aus MAS-Analysen möglich. Darüber hinaus können analog zu der Möglichkeit, einem desc.df weitere Spalten wie `sum.of.intens` hinzuzufügen, auch den gespeicherten Analysen zusammenfassende Merkmale wie z. B. `mean.of.Signal` hinzugefügt werden. Außerdem können Spalten von `analyses.`*hybridname* auf den desc.df übertragen werden (z. B. `map.SF.from.analyses.desc.df.to.hybrid` zur Übertragung der Skalierungsfaktoren einer bestimmten Skalierung).

Primäranalysen lassen sich in der vorliegenden Umgebung nur über den Umweg einer *publish*-Datenbank importieren, da CHP-Dateien binär kodiert sind und nur mit dem File SDK direkt eingelesen werden können. Mit der Funktion `import.LIMS.analyses` ist es beispielsweise möglich, Primäranalysen in die analysis-Datenbank von SPLUS zu importieren. Eine Analyse *name* wird in einem data.frame `analysis.`*name* gespeichert. In einer Zeile werden der Name und die berechneten Maßzahlen eines *probe sets* gespeichert.

3.2.3 Informationen über *publish*-Datenbanken und *probe set descriptions*

Da die Maßzahlen von Primäranalysen nur aus *publish*-Datenbanken importiert werden können, müssen einige Informationen zu den *publish*-Datenbanken in eigenen data.frames vorgehalten werden:

LIMS.*PublishDB name***.ANALYSIS**

LIMS.*PublishDB name***.ANALYSIS.DATA.SET**

LIMS.*PublishDB name***.BIOLOGICAL.ITEM**

LIMS.*PublishDB name***.EXPERIMENT**

Die Funktion zum Importieren von Primäranalysen nutzt diese Informationen zum Erstellen der ODBC-Datenbankabfrage, die die Maßzahlen aus den Tabellen `ABS_GENE_EXPR_RES_STAT` und `REL_GENE_EXPR_RES_STAT` extrahiert. Die vier Objekte müssen durch Aufruf von `initialize.LIMS.import` initialisiert werden, sobald sich der Inhalt der jeweiligen *publish*-Datenbank geändert hat.

Analog zur `ANALYSIS`-Tabelle des AADM-Schemas (siehe Abbildung 6) nimmt `LIMS.`*PublishDB_name*`.ANALYSIS` eine zentrale Position ein. Hierin sind alle enthaltenen Primäranalysen und die Experimente, aus denen sie entstanden sind, aufgelistet. Eine Unterscheidung zwischen Primäranalysen und Experimenten sowie die Ermittlung aller enthaltenen Primäranalysen eines Experimentes kann unter Zuhilfenahme von `LIMS.`*PublishDB_name*`.EXPERIMENT` erfolgen. Das Objekt `LIMS.`*PublishDB_name*`.BIOLOGICAL.ITEM` ermöglicht die Umsetzung einer *probe set*-ID in einen *probe set*-Namen. Die Funktionen zum Import / Export von Primäranalysen, welche die *publish*-DB-Objekte nutzen, sind in Anhang A.4 beschrieben.

LIMS.GENEINFO.BIOLOGICAL.DESCRIPTION

LIMS.GENEINFO.EXTERNAL.DATABASE.LINK

Zwei LIMS-globale Tabellen enthalten zusätzliche Informationen über *probe sets*: `LIMS.GENEINFO.BIOLOGICAL.DESCRIPTION` speichert *probe set*-Bezeichner und

die zugehörigen *probe set description* für alle auf dem LIMS-System installierten Array-Typen. `LIMS.GENEINFO.EXTERNAL.DATABASE.LINK` enthält Verweise auf externe Datenbanken für jedes *probe set*.

3.2.4 „Beschreibender data.frame (desc.df)“ und Gruppierungsobjekte

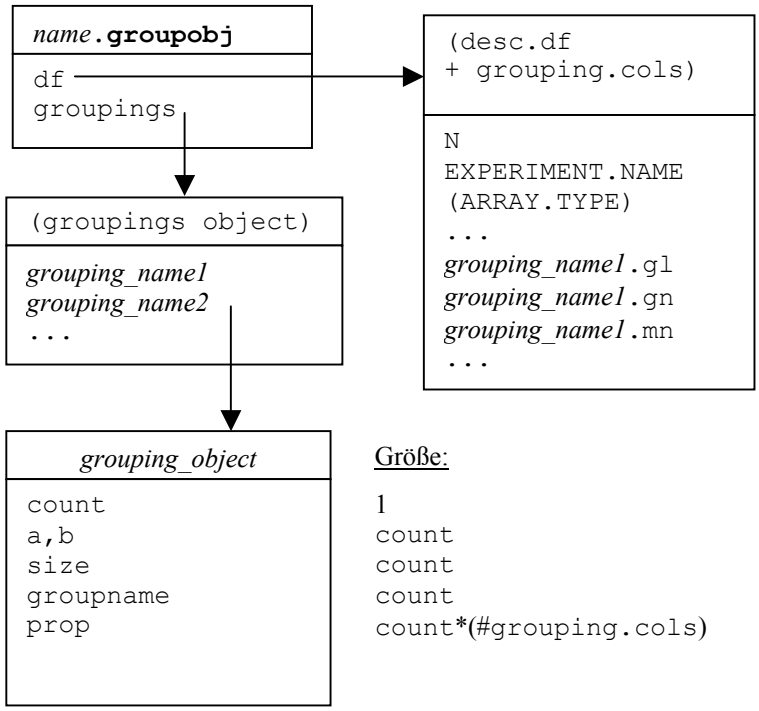
Alle Merkmale, die ein GeneChip-Experiment charakterisieren, werden in einem so genannten beschreibenden data.frame (*descriptional data.frame*, kurz: desc.df) erfasst.

name`.desc.df`

`N`
`EXPERIMENT.NAME`
`(ARRAY.TYPE)`
`...`

Ein desc.df enthält in der Regel alle Experimente einer Fragestellung. Er besteht mindestens aus den Spalten `N` und `EXPERIMENT.NAME`, in welchen keine Einträge doppelt vorkommen dürfen. Mit der Funktion `import.Experiment.descriptions` kann ein desc.df aus einer Excel-Datei oder einer Access-Datenbank importiert werden. Ein desc.df enthält in der Regel weitere Spalten mit Merkmalen des Experimentes (Beispiele für einen desc.df siehe nächstes Unterkapitel).

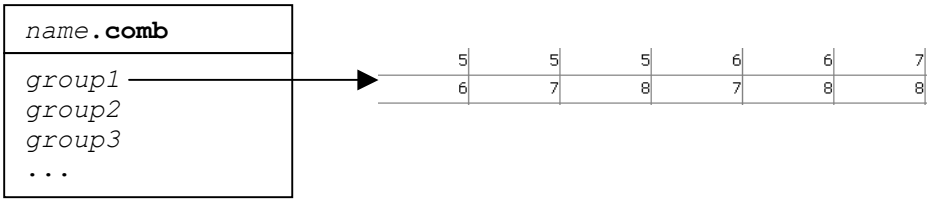
Der Gruppierungsalgorithmus (siehe nächstes Unterkapitel) fügt dem ursprünglichen desc.df drei Spalten pro Gruppierung hinzu und erstellt ein `grouping`-Objekt, sodass andere Funktionen die Gruppierungsinformationen nutzen können. Beide Komponenten werden dann in einer `groupobj`-Datenstruktur zur späteren Verwendung zusammengefasst.



Das `groupings`-Objekt enthält pro Gruppierung eine *grouping_object*-Komponente, welche die jeweilige Gruppierung beschreibt. Sie enthält ein Element `count`, das die Anzahl der Gruppen speichert. Des Weiteren speichert es die Komponenten `a`, `b`, `size` und `groupname`, die für jede Gruppe den Anfangs- und Endindex im sortierten desc.df sowie die Größe und den Gruppenbezeichner speichern. Im `grouping`-Objekt gibt es überdies eine Matrix `prop`, die so viele Spalten wie gruppierende Merkmale und so viele Zeilen wie Gruppen besitzt und für jede Gruppe die Ausprägung der gruppierenden Merkmale enthält. Zusammen mit den zusätzlichen Spalten im desc.df können die Gruppierungsinformationen in den Funktionen, die mit Gruppierungen arbeiten (z. B. `get.experiments.of.groups`), effizient genutzt werden.

3.2.5 Kombinationsobjekte

Im Rahmen weiter gehender statistischer Auswertungen entstand die Anforderung, Experimente paarweise miteinander zu kombinieren. Selbst wenn die Kombinationsfunktion (z. B. Scatter Plot) noch nicht bekannt ist, können die Kombinationsmöglichkeiten abgespeichert werden.



Der bislang implementierte Ansatz ermöglicht die Kombination von Experimenten innerhalb einer Gruppe einer Gruppierung (Funktion `create.combinations`). Die resultierende comb-Datenstruktur besteht für jede Gruppe der zugrunde liegenden Gruppierung aus einer Matrix, in der alle kombinierten Paare als zweireihige Spalte aufgeführt sind. Ausführlich erläutert wird dieses Konzept im folgenden Unterkapitel.

3.3 Algorithmen zum Gruppieren, Benennen und Kombinieren von Experimenten

3.3.1 Motivation

Bei der Bearbeitung einer Fragestellung werden einige Merkmale der dafür durchgeführten Experimente so gut wie möglich reproduziert – so in der Regel die Parameter des Laborprotokolls. Je nach Fragestellung werden andere Merkmale bewusst variiert, wie z. B. der Proband, der Zeitpunkt der Gewebe- oder Blutentnahme oder die Art des Gewebes (gesund/krank). Indem auch diese Merkmale reproduziert werden, entstehen Replikat-Experimente. Wieder andere Merkmale lassen sich nicht gezielt (oder nur bedingt) variieren, sondern höchstens im Nachhinein durch Messung bestimmen, wie beispielsweise die Konzentration der total-RNA nach der RNA-Gewinnung oder die Leukozytenzahl nach Blutentnahme.

Durch die Ausprägung all dieser Merkmale ergeben sich die logischen Gruppen, die in weiter gehenden statistischen Auswertungen benutzt werden. So könnten in einem t-Test

für unverbundene Stichproben beispielsweise alle Experimente eines Probanden mit denselben Merkmalen zu einer Gruppe zusammengefasst werden, und alle Experimente eines anderen Probanden mit denselben Merkmalen zu einer anderen Gruppe. Wünschenswert ist es darüber hinaus, logische Gruppen bzw. einzelne Experimente mit Bezeichnern zu versehen, aus denen hervorgeht, wie sie entstanden sind bzw. welcher Gruppe sie angehören. Auch beim Anfertigen von Graphen ist bei der Beschriftung der Achsen oder bei der Erstellung der Legende die übersichtliche Kennzeichnung der logischen Gruppen und einzelner Experimente essenziell. Wird auf demselben Datensatz eine andere Fragestellung angewandt, so ändert sich im Allgemeinen auch die logische Gruppierung. Damit sollte eine Änderung der Bezeichnung von logischen Gruppen und Experimenten einhergehen, da der Name eines Experimentes, der beim Scannen mit der MAS fest vergeben wird, im Zusammenhang der neuen Fragestellung unter Umständen wenig aussagekräftig ist.

Liegt nur ein kleiner Datensatz vor, kann das Gruppieren vom Auswerter manuell oder „im Kopf“ vorgenommen werden. Bei größeren Datensätzen und solchen, die über eine abgegrenzte Fragestellung hinausgehen, oder beim Arbeiten auf großen Expressionsdatenbanken mit vielen variierten Merkmalen ist dies nicht mehr möglich. Außerdem kann es bei größeren Datensätzen notwendig werden, nur Experimente mit bestimmten Merkmalen in die Betrachtung einzubeziehen (Fixieren von Merkmalen) oder umgekehrt einige Experimente gar nicht einzubeziehen (Filtern von Merkmalen). Für diese Aufgabenstellungen ist ein automatischer Ansatz hilfreich.

Im Folgenden wird die Implementierung eines solchen Ansatzes beschrieben. Den Funktionen und Auswertungen der Kapitel 4 und 5 liegt dieses Konzept zugrunde.

Beim Umgang mit den beschriebenen Funktionen ergab sich eine Erweiterung des Konzeptes: auch die Experimente sollen zu einer logischen Gruppe zusammenzufassen sein, die sich in genau einem Merkmal unterscheiden. Dieses „primär variierte Merkmal (pvm)“ kann sowohl bei der Benennung der Experimente berücksichtigt werden als auch bei der Kombination von Experimenten (siehe weiter unten).

3.3.2 Implementierung des Gruppierungsalgorithmus

Ausgangspunkt des Algorithmus (Funktion `add.grouping`) ist ein beschreibender data.frame (desc.df), in welchem für jedes zu verwendende Experiment mindestens der (eindeutige) MAS-Name (ohne Endung) und eine eindeutige Nummer vermerkt ist (Spalten `EXPERIMENT.NAME` und `N`). Jedes weitere Merkmal wird in einer separaten Spalte mit einem eindeutigen Namen erfasst.

Weitere Eingabeparameter für den Algorithmus sind:

- `groupingname`: der Name der zu erstellenden Gruppierung,
- `fixcolnames`: die Namen der Filter- / Fixierungs-Spalten,
- `include`: ein TRUE / FALSE-Vektor, der angibt, ob fixiert / gefiltert werden soll,
- `fixvalues`: eine Liste aus Vektoren mit den Ausprägungen, die dem Filtern / Fixieren zugrunde liegen,
- `groupcolnames`: die Namen der Spalten, nach denen gruppiert werden soll,
- `sortcolnames` (optional): die Namen der Spalten, nach denen zusätzlich sortiert werden soll,
- `leastgroupsize`: Mindest-Gruppengröße (Standardwert: 1),
- `pvm` (optional): Name der Spalte des primär variierten Merkmals,
- `only.diff.pvm` (optional): TRUE: keine Gruppen erzeugen, in denen es nur Experimente mit gleichen Ausprägungen des primär variierten Merkmals gibt.

Die Rückgabe der Funktion ist eine groupobj-Datenstruktur, die den um drei Spalten erweiterten desc.df und die grouping-Struktur zusammenfasst. Da ein bestehender desc.df auf mehr als eine Weise gruppiert werden kann, kann ein groupobj-Objekt seinerseits Eingabe für den Algorithmus sein.

Die prinzipiellen Schritte, die der Algorithmus durchführt, sind die folgenden:

1. Fixieren / Filtern von Experimenten.
2. Hinzufügen von Experimenten zu Gruppen.
3. Löschen von Gruppen mit zu kleiner Größe.
4. Erstellen der grouping-Struktur (Gruppengrößen, Gruppenbezeichner, gruppierende Merkmalsausprägungen).
5. Erstellen der Experimentbezeichner.

Diese prinzipiellen Schritte lassen sich wie folgt weiter unterteilen:

1. Fixieren / Filtern von Experimenten:

   a. Für fixierte Spalten Hilfsspalten hinzufügen.
   b. Für Gruppierungsspalten Hilfsspalten hinzufügen.

2. Hinzufügen von Experimenten zu Gruppen:

   a. Sortieren nach Fixierungshilfsspalten, Gruppierungsshilfsspalten, Sortierungsspalten, pvm.
   b. Für alle Experimente: Wenn nicht ausgefiltert: Zeilen/Experimente mit gleichen Merkmalsspalten in gemeinsame Gruppe => Gruppennummer in Spalte *grouping_name*`.gn`, Mitgliedsnummer in Spalte *grouping_name*`.mn`.

3. Löschen von Gruppen mit zu kleiner Größe:

   a. Ausfiltern von Gruppen, die kleiner als die Mindestgröße sind (Überschreiben der Werte in den Spalten *grouping_name*`.gn` und *grouping_name*`.mn` mit NA).
   b. Sortieren nach *grouping_name*`.gn` und *grouping_name*`.mn` => eindeutige Sortierung nach Gruppennummern und Mitgliedsnummer.

4. **Erstellen der grouping-Struktur:**
    a. Erneut Zeile für Zeile durchgehen und aktuelle grouping-Struktur erstellen:
        i. Anfangs- und End-Zeile jeder Gruppe bestimmen (in Komponenten `a` und `b` eintragen).
        ii. Größe jeder Gruppe bestimmen (in Komponente `size` eintragen).
        iii. Für jede Gruppe Namen der Gruppierungsspalten und Werte der Merkmalsausprägungen (*per definition* identisch in der Gruppe) abspeichern (in Komponente `prop` eintragen).
        iv. Aus den Merkmalsausprägungen Gruppenbezeichner erstellen (Trennzeichen: „|“) (in Komponente `groupname` eintragen).
        v. Anzahl der Gruppen bestimmen (in Komponente `count` eintragen).
    b. Aktuelle grouping-Struktur zu eventuell bereits vorhandener hinzufügen.
5. **Erstellen der Experimentbezeichner:**
    a. Erstellen der Experimentbezeichner und Abspeichern derselben in der Spalte *grouping_name*`.gn`. Wenn der Parameter `pvm` nicht benutzt wird, werden Experimente mit „N*N*.g*gn*“ bezeichnet, wobei *N* die eindeutige Nummer aus der Spalte `N` ist und *gn* die Nummer der Gruppe, also z. B. „N1.g2“. Wird der Parameter `pvm` verwendet, wird diesem Bezeichner noch die Ausprägung des primär variierten Merkmals und eine laufende Nummer vorangestellt. Gibt es beispielsweise in einer Gruppe mit der Nummer 2 drei Experimente 3,4,5, von denen zwei für das primär variierte Merkmal `REPLIKAT` die Ausprägung `TECHNISCH` und eines die Ausprägung `BIOLOGISCH` besitzen, so lauten die Experimentbezeichner beispielsweise „TECHNISCH1.N3.g2“, „TECHNISCH2.N4.g2“ und „BIOLOGISCH.N5.g2“.
    b. Wiederherstellen der ursprünglichen Sortierung des desc.df.

Das folgende Beispiel verdeutlicht die Benutzung der Funktion `add.grouping`. Der desc.df für die Experimente des MS2-Datensatzes ist:

N	EXPERIMENT.NAME	ARRAY.TYPE	STAINING	DATUM	KULTUR	INDUZIERT	REPLIKAT
1	Coculture-Caco-Co-16-10-pre	HG_U95A	Simple	16-10	FILTER	NEIN	BIOLOGISCH
2	Coculture-Caco-Co-16-10	HG_U95A	AK	16-10	FILTER	NEIN	BIOLOGISCH
3	Coculture-Caco1_08_01-pre	HG_U95A	Simple	8-10	FILTER	JA	TECHNISCH
4	Coculture-Caco1_08_01	HG_U95A	AK	8-10	FILTER	JA	TECHNISCH
5	Coculture-Caco2_08_01-pre	HG_U95A	Simple	8-10	FILTER	JA	TECHNISCH
6	Coculture-Caco2_08_01	HG_U95A	AK	8-10	FILTER	JA	TECHNISCH
7	Coculture-Caco_08_01_Kontr	HG_U95A	AK	8-10	FILTER	NEIN	BIOLOGISCH
8	Coculture-Caco_08_01_Kontr_pre	HG_U95A	Simple	8-10	FILTER	NEIN	BIOLOGISCH
9	Coculture-Cacos-Co-Bed-15-10	Test3	AK	16-10			
10	Coculture-Cacos-Lym-16-10-pre	HG_U95A	Simple	16-10	FILTER	JA	BIOLOGISCH
11	Coculture-Cacos-Lym-16-10	HG_U95A	AK	16-10	FILTER	JA	BIOLOGISCH
12	Coculture-Cacos-Lymph-15-10	Test3	AK	16-10			
13	Coculture-Cacos-normal-pur-15-10	Test3	AK	16-10			
14	Coculture-Cacos-pur-16-10-pre	HG_U95A	Simple	16-10	FLASCHE	NEIN	NEIN
15	Coculture-Cacos-pur-16-10	HG_U95A	AK	16-10	FLASCHE	NEIN	NEIN
16	Coculture-Lymphos-16-10-pre	HG_U95A	Simple	16-10	LYMPHOZYTEN	NEIN	NEIN
17	Coculture-Lymphos-16-10	HG_U95A	AK	16-10	LYMPHOZYTEN	NEIN	NEIN
18	Coculture-Lymphos-pur-15-10	Test3	AK	16-10			

Im Folgenden wird auf diesem `MS2.desc.df` eine neue Gruppierung `PVMTEST` erstellt. Dabei sollen nur Experimente verwendet werden, die für das Merkmal `STAINING` die Ausprägung `AK` besitzen. Gruppiert werden soll nach den Merkmalen `KULTUR` und `INDUZIERT`. Vor dem Erstellen der Gruppen soll zusätzlich nach `N` sortiert werden. Die Mindestgruppengröße sei eins und das primär variierte Merkmal `REPLIKAT`, wobei auch Gruppen erlaubt sein sollen, die nur Experimente mit gleicher Ausprägung für dieses Merkmal haben.

```
MS2.groupobj<-add.grouping (    hybrid.name="MS2.desc.df",
                                groupingname="PVMTEST",
                                fixcolnames=c("STAINING","ARRAY.TYPE"),
                                include=c(TRUE,FALSE),
                                fixvalues=list(c("AK"),c("Test3")),
                                groupcolnames=c("KULTUR","INDUZIERT"),
                                sortcolnames=c("N"),
                                leastgroupsize=1,
                                pvm=c("REPLIKAT"),
                                only.diff.pvm=FALSE)
```

Mit diesem Aufruf ergeben sich vier Gruppen. Die Funktion `summary.of.groups` gibt eine Zusammenfassung einer Gruppierung aus:

```
[1] Gruppen: 4
     Gruppenbezeichner  KULTUR         INDUZIERT #EXP
[1,] "FILTER|JA"        "FILTER"       "JA"      "3"
[2,] "FILTER|NEIN"      "FILTER"       "NEIN"    "2"
[3,] "FLASCHE|NEIN"     "FLASCHE"      "NEIN"    "1"
[4,] "LYMPHOZYTEN|NEIN" "LYMPHOZYTEN" "NEIN"    "1"

[1] Experimente:
 Experimentbezeichner   EXPERIMENT.NAME
  TECHNISCH1.N4.g1       Coculture-Caco1_08_01
  TECHNISCH2.N6.g1       Coculture-Caco2_08_01
  BIOLOGISCH1.N11.g1     Coculture-Cacos-Lym-16-10
  BIOLOGISCH1.N2.g2      Coculture-Caco-Co-16-10
  BIOLOGISCH2.N7.g2      Coculture-Caco_08_01_Kontr
  NEIN1.N15.g3           Coculture-Cacos-pur-16-10
  NEIN1.N17.g4           Coculture-Lymphos-16-10
```

Die Komponente `df` des Objekts `MS2.groupobj` stellt sich wie folgt dar:

N	EXPERIMENT.NAME	PVMTEST.gI	PVMTEST.gn	PVMTEST.mn	ARRAY.TYPE	STAINING	DATUM	KULTUR	INDUZIERT	REPLIKAT
1	Coculture-Caco-Co-16-10-pre		NA	NA	HG_U95A	Simple	16-10	FILTER	NEIN	BIOLOGISCH
2	Coculture-Caco-Co-16-10	BIOLOGISCH1.N2.g2	2	1	HG_U95A	AK	16-10	FILTER	NEIN	BIOLOGISCH
3	Coculture-Caco1_08_01-pre		NA	NA	HG_U95A	Simple	8-10	FILTER	JA	TECHNISCH
4	Coculture-Caco1_08_01	TECHNISCH1.N4.g1	1	1	HG_U95A	AK	8-10	FILTER	JA	TECHNISCH
5	Coculture-Caco2_08_01-pre		NA	NA	HG_U95A	Simple	8-10	FILTER	JA	TECHNISCH
6	Coculture-Caco2_08_01	TECHNISCH2.N6.g1	1	2	HG_U95A	AK	8-10	FILTER	JA	TECHNISCH
7	Coculture-Caco_08_01_Kontr	BIOLOGISCH2.N7.g2	2	2	HG_U95A	AK	8-10	FILTER	NEIN	BIOLOGISCH
8	Coculture-Caco_08_01_Kontr_pre		NA	NA	HG_U95A	Simple	8-10	FILTER	NEIN	BIOLOGISCH
9	Coculture-Cacos-Co-Bed-15-10		NA	NA	Test3	AK	16-10			
10	Coculture-Cacos-Lym-16-10-pre		NA	NA	HG_U95A	Simple	16-10	FILTER	JA	BIOLOGISCH
11	Coculture-Cacos-Lym-16-10	BIOLOGISCH1.N11.g1	1	3	HG_U95A	AK	16-10	FILTER	JA	BIOLOGISCH
12	Coculture-Cacos-Lymph-15-10		NA	NA	Test3	AK	16-10			
13	Coculture-Cacos-normal-pur-15-10		NA	NA	Test3	AK	16-10			
14	Coculture-Cacos-pur-16-10-pre		NA	NA	HG_U95A	Simple	16-10	FLASCHE	NEIN	
15	Coculture-Cacos-pur-16-10	1.N15.g3	3	1	HG_U95A	AK	16-10	FLASCHE	NEIN	
16	Coculture-Lymphos-16-10-pre		NA	NA	HG_U95A	Simple	16-10	LYMPHOZYTEN	NEIN	
17	Coculture-Lymphos-16-10	1.N17.g4	4	1	HG_U95A	AK	16-10	LYMPHOZYTEN	NEIN	
18	Coculture-Lymphos-pur-15-10		NA	NA	Test3	AK	16-10			

Die Komponente `groupings$PVMTEST` des Objekts `MS2.groupobj` hat die nachstehende Struktur (ohne Details):

Object	Pos	Data Class	Dimensions
count	1	numeric	1
a	2	integer	4
b	3	integer	4
size	4	numeric	4
groupname	5	character	4
pvm	6	character	1
prop	7	matrix	4x2

Mit der Funktion `remove.grouping` kann eine Gruppierung *grouping_name* aus einem groupobj-Objekt gelöscht werden. Die entsprechenden *grouping_name*`.gl`-, *grouping_name*`.gn`- und *grouping_name*`.mn`-Spalten werden ebenso wie die `groupings`-Komponente *grouping_name* entfernt.

3.3.3 Berechnen von Kombinationen von Experimenten und Primäranalysen

Weiter gehende statistische Auswertungen erfordern in der Regel die Herstellung eines Zusammenhangs zwischen zwei oder mehreren Experimenten bzw. Primäranalysen. Dazu gehören z. B. das Anwenden einer Distanzfunktion, das Berechnen eines Vektors aus paarweisen Quotienten oder Differenzen und nicht zuletzt das Darstellen von Maßzahlen oder experimentellen Maßzahlen durch Graphen. Ab einer gewissen Anzahl von Experimenten oder bei komplizierteren Kombinationsarten und gerade auch bei der Berücksichtigung von Gruppierungen muss die Berechnung der Kombinationen automatisch erfolgen. Selbst wenn die letztendlich anzuwendende Funktion noch nicht bekannt ist, können trotzdem schon Kombinationsmöglichkeiten berechnet und in comb-Objekten gespeichert werden (siehe Unterkapitel 3.2).

Die möglichen Kombinationen werden von der Funktion `create.combinations` berechnet. Im Prinzip könnte die Funktion jede – auch implizite – Kombinationsfunktion $comb : N^m \rightarrow \{0,1\}$ durch explizites Abspeichern der *m*-dimensionalen Kombinationen mit $comb^{-1}(1)$ umsetzen. Bisher implementiert ist das Berechnen der paarweisen Kombinationen (*m*=2) innerhalb aller Gruppen, ohne dabei gleiche Experimente in anderer Reihenfolge zu kombinieren („Dreiecksmatrix ohne Diagonale“).

Befinden sich in einer Gruppe beispielsweise die Experimente 5, 6, 7 und 8, so ergeben sich die rot markierten paarweisen Kombinationen:

	5	6	7	8
5	5,5	6,5	7,5	8,5
6	5,6	6,6	7,6	8,6
7	5,7	6,7	7,7	8,7
8	5,8	6,8	7,8	8,8

In Listenform notiert sind dies die Kombinationen (5,6),(5,7),(5,8),(6,7),(6,8),(7,8). Abgespeichert werden diese Kombinationen im comb-Objekt in der Komponente *gruppennummer*, welche eine zweizeilige Matrix ist, deren Spalten die Kombinationen enthalten (siehe auch Abschnitt 3.2.5). Dieses Konzept lässt sich auf *m*-dimensionale Kombinationen erweitern, da diese in *m*-zeiligen Matrizen abspeicherbar sind.

Ebenfalls implementiert ist die Berücksichtigung des primär variierten Merkmals bei der Bildung paarweiser Kombinationen. So kann über den Parameter `samepvm=FALSE` gesteuert werden, dass ein Paar (*a*,*b*) nicht in die Liste der Kombinationen aufgenommen wird, wenn *a* und *b* dieselbe Ausprägung des primär variierten Merkmals zeigen.

3.3.4 Anwenden von Funktionen auf gespeicherten Kombinationen

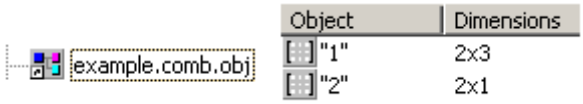
Mithilfe der Funktion `apply.to.combinations` kann auf die gegebenenfalls zuvor gespeicherten Kombinationen eine konkrete Funktion *fun* angewendet werden. Mit dem Parameter `comb.struct` erwartet diese ein comb-Objekt, im Parameter `groupnos` wird eine Liste der Gruppennummern übergeben, auf deren Kombinationen *fun* angewendet werden soll. Mit dem Parameter `FUN` wird *fun* übergeben. Dann folgen weitere beliebig viele Parameter, die an *fun* weitergereicht werden. Einzige Forderung an die Funktion *fun* ist, dass sie als ersten Parameter ein Objekt behandeln können muss, das aus *m* Komponenten besteht.

Liefert die Funktion *fun* einen Wert und eventuell einen Bezeichner zurück, dann ist das Ergebnis von `apply.to.combinations` ein Objekt, das dem comb-Objekt

ähnelt: Für jede Gruppe gibt es eine Komponente und für jede Kombination innerhalb dieser Gruppe gibt es ein Ergebnis, welches eventuell mit einem Bezeichner versehen ist.

Beispiel:

Seien für eine Gruppe 1 mit drei Experimenten und eine Gruppe 2 mit zwei Experimenten die paarweisen Kombinationen errechnet und in einem Objekt `example.comb.obj` gespeichert worden (drei Kombinationen für Gruppe 1 und eine Kombination für Gruppe 2):

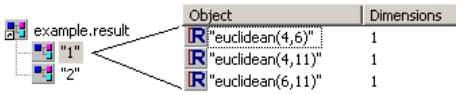


Sei *fun* nun eine Funktion `calc.intens.distances`, die die euklidische Distanz zwischen den Intensitätsvektoren zweier über ihre Experimentnummern *N1* und *N2* gegebene Experimente errechnet und zurückgibt zusammen mit dem Bezeichner „`euclidean(`*N1*`,`*N2*`)`“.

Der Aufruf

```
example.result<-apply.to.combinations ( comb.struct=example.comb.obj,
                                        groupnos="@ALL",
                                        FUN=calc.intens.distances,
                                        metric="euclidean",
                                        hybrid.name="example.groupobj")
```

liefert dann das folgende Ergebnis:



Nach Gruppennummern getrennt wurde also für jede Kombination der euklidische Abstand im Objekt `example.result` gespeichert.

Die Funktion *fun* kann auch so angelegt sein, dass sie keinen expliziten Rückgabewert liefert, sondern nur einen Seiteneffekt hat, wie beispielsweise das Zeichnen eines Graphen.

Beispiel (Fortsetzung):

Wird die Funktion `scatter.plot.intensities` auf die oben beschriebenen gespeicherten Kombinationen angewandt, so ergeben sich mit dem Aufruf

```
oldpar<-par(mfrow=c(3,4))
example.result<-apply.to.combinations ( comb.struct=example.comb.obj,
                                        groupnos="@ALL",
                                        FUN=scatter.plot.intensities,
                                        maintitle="@NUM",
                                        hybrid.name="MS2.groupobj",
                                        fit.line=T)
par(oldpar)
```

vier Scatter Plots als Seiteneffekt (siehe Abbildung 11).

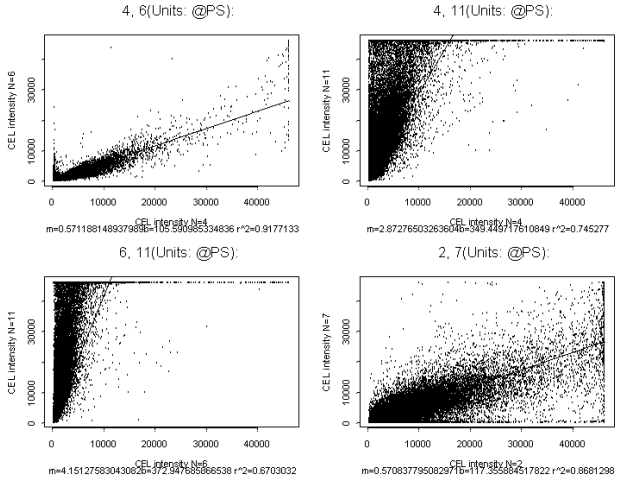


Abbildung 11: `apply.to.combinations` mit Seiteneffekt: Scatter Plots der Intensitäten jeweils zweier Experimente

3.4 Weitere SPLUS-Funktionen

Dieses Unterkapitel enthält in Abschnitt 3.4.2 einen Überblick über die wichtigsten implementierten SPLUS-Funktionsgruppen. Detailliertere Informationen zu jeder Funktion finden sich im Anhang. Einige allgemeine Konzepte bezüglich Datenstrukturen und Parameterübergabe gelten für eine Vielzahl von Funktionen. Sie werden im direkt folgenden Abschnitt beschrieben.

3.4.1 Allgemeine Konzepte

Die meisten Funktionen verwenden mindestens eines der in Unterkapitel 3.2 vorgestellten Objekte. Zur Vermeidung einer unübersichtlich großen working-Datenbank liegt es nahe, die Möglichkeit von SPLUS zur Definition einer Hierarchie von Datenbanken zu nutzen. Inhaltlich zusammengehörende Objekte werden in der gleichen Datenbankebene zusammengefasst. Eine natürliche Einteilung ergibt sich wie folgt:

1. working-Datenbank (Hierarchieebene 1): desc.df-, groupobj- und comb-Objekte
2. intensity-Datenbank: intens- und misc-Objekte, CHIP.DESIGN-, CHIP.LAYOUT- und PSI.info-Objekte
3. analysis-Datenbank: analyses.desc.df-Objekte, analysis-Objekte, Publish.DB-Objekte

Alle Funktionen, die Objekte aus der zweiten und dritten Kategorie nutzen, verfügen über einen Parameter `intens.DB` oder `analysis.DB`, über welchen die konkrete Ebene der tatsächlich verwendeten Datenbank übergeben wird. Standardwerte für diese Parameter sind die globalen Variablen `default.intens.DB` und `default.analysis.DB`, die sinnvollerweise in der `.First`-Funktion gesetzt werden, zusammen mit dem `attach`-Befehl zum Einfügen der intensity- und analysis-Datenbanken in die Datenbankhierarchie von SPLUS.

Zur Übergabe von Experimenten an Funktionen können entweder die Experimentnamen verwendet werden oder ein desc.df und die eindeutigen Nummern aus der Spalte `N`. Der verwendete Variablenbezeichner für beide Möglichkeiten ist `WHAT`. Eine Fehlermeldung wird zurückgegeben, wenn `WHAT` numerische Werte enthält und kein desc.df übergeben wurde, weil dann die Umsetzung der Nummern in Experimentnamen nicht möglich ist.

Funktionen, die mit einem desc.df arbeiten, können *per definition* auch auf einem groupobj-Objekt arbeiten, da dieses einen erweiterten desc.df enthält. Als Variablenbezeichner wird dann `hybrid.name` verwendet. Funktionen, die mit Gruppierungsinformationen arbeiten, benötigen explizit ein groupobj-Objekt. Der Variablenbezeichner ist dann `groupobj.name`. In der Regel existieren dabei auch

Parameter `groupingname` und `groupnos` zur Übergabe des Gruppierungsnamens und der Gruppennummern.

Funktionen, die auf *probe cell*-Ebene arbeiten, besitzen in der Regel einen Parameter `UNITS`, mit dem die Untermenge der zu verwendenden *probe cells* spezifiziert werden kann. Mögliche Werte dieses Parameters sind in Tabelle 12 aufgelistet.

Parameter `UNITS`	**spezifizierte *probe cell*-Fraktion**
@ALL	*probe cells* aller *probe sets*, aller *Quality Features* und aller nicht näher beschriebenen (offenen) *probe cells*
@PS	nur *probe cells*, die zu *probe sets* gehören
@QC	nur *probe cells*, die zu *Quality Features* gehören
@OPEN	nur offene (leere) *probe cells*
@PM / @MM	nur *probe cells*, die *Perfect Match* / *Mismatch*-Zellen sind
unit_numbers (numerischer Vektor)	*probe cells*, die zu den *unit_numbers* entsprechenden *probe sets* gehören
logical_vector	Explizite Angabe, ob *probe cell* eingeschlossen (TRUE) oder ausgeschlossen (FALSE) sein soll (Länge muss Anzahl der *probe cells* entsprechen)

Tabelle 12: Werte des Parameters `UNITS`

Die meisten Graphenfunktionen besitzen Parameter mit angepassten Standardwerten (beispielsweise x- und y-Achsenabschnitt oder Granulierung der Histogramm-Balken). Auch existiert in der Regel der generische Parameter „...“, mit dem weitere Parameter an die SPLUS-Graphenroutinen durchgereicht werden können. Einige Graphenfunktionen kennen den Parameter `sep.groups`. Ist er FALSE; werden die spezifizierten Gruppen in einem Graphen erfasst – in der Regel durch verschiedene Farben unterscheidbar. Hat der Parameter jedoch den Wert TRUE, wird ein separater Graph für jede spezifizierte Gruppe gezeichnet.

3.4.2 SPLUS-Funktionsgruppen

Detailliertere Informationen zu den meisten Funktion finden sich im Anhang.

Funktionen zur Handhabung von Merkmalen von GeneChip-Experimenten

Zu dieser Funktionsgruppe zählen Funktionen zum Import des beschreibenden data.frame eines Datensatzes nach SPLUS, zum Extrahieren von Experimentnamen oder

-nummern von Experimenten mit bestimmten Eigenschaften, zur Ermittlung des Experiment-Labels unter einer bestimmten Gruppierung, zur Bestimmung der `chip.id` bestimmter Experimente und zur Ausgabe statistischer Zusammenfassungen eines Merkmals.

Funktionen im Zusammenhang mit Gruppierungen und Kombinationen

In Unterkapitel 3.3 wird das Gruppierungskonzept beschrieben. Die Funktionen in dieser Funktionsgruppe dienen zum Erstellen von Gruppierungen, Entfernen von Gruppierungen, zur Rückgabe der Gruppenbezeichner einer bestimmten Gruppierung, zum Extrahieren des beschreibenden data.frame aus einem groupobj-Objekt, zur Zusammenfassung von Gruppierungsinformationen, zum Erstellen von Kombinationen und zur Anwendung von Funktionen auf diese Kombinationen.

Funktionen zum Umgang mit Intensitäten

Diese Funktionsgruppe beinhaltet Funktionen zum Importieren von Intensitätsdaten, zum Arbeiten mit Intensitätsobjekten, zum Berechnen von Gesamtintensitäten und anderen Merkmalen, die die Intensitätsdaten eines Experimentes zusammenfassen, zum Modifizieren und zum Export von CEL-Dateien in das Dateisystem und zum Importieren dieser CEL-Dateien in das LIMS-System. Für einige dieser Funktionen werden Layout-Informationen des Chips benötigt, sodass auch Funktionen zur Handhabung von Layout-Informationen zu dieser Gruppe gezählt werden.

Funktionen zum Import / Export von Primäranalysen

Um in SPLUS mit den Maßzahlen und Daten der Primäranalysen von Affymetrix arbeiten zu können, wurden Funktionen zum Importieren aus dem LIMS-System und zum Exportieren in das Dateisystem implementiert. Der Import erfolgt über den Umweg einer *publish*-Datenbank, da CHP-Dateien nicht direkt, sondern nur mit dem File SDK eingelesen werden können. Der *publish*-Vorgang muss vom Anwender manuell durchgeführt werden. Er muss mit der MAS dafür sorgen, dass die zu betrachtenden Analysen aus der *process*-Datenbank in die entsprechende *publish*-Datenbank transferiert werden.

Funktionen zum Umgang mit Primäranalysen

Abgeleitet vom desc.df, der die Beschreibungen aller Experimente eines Datensatzes enthält, kann mit der Funktion `import.analyses.descriptions` ein analyses.desc.df angelegt werden, der Informationen zu den Primäranalysen dieser Experimente enthält. Zusätzlich zählen zu dieser Gruppe Funktionen, mit deren Hilfe die Namen von Primäranalysen unter verschiedenen Bedingungen aus dem analyses.desc.df ermittelt werden können. Dazu gehören überdies Funktionen, die zusammenfassende Größen (z. B. `mean.of.Signal`) von Primäranalysen berechnen und im analyses.desc.df speichern und eine Funktion, die die `SF`-Spalte vom analyses.desc.df in den desc.df verschiebt.

Graphenfunktionen

Eine Vielzahl von Funktionen dient zum Erstellen verschiedener Graphentypen (Balkendiagramm, Histogramm, Box Plot, Scatter Plot), mit deren Hilfe verschiedene GeneChip-Daten (Experimentmerkmale, Intensitäten, Merkmale oder Maßzahlen von Primäranalysen) visualisiert werden können. Die meisten Funktionen können Gruppierungsinformationen bei der Erstellung des Graphen und der Legende berücksichtigen.

Sonstige Funktionen

Zu den sonstigen Funktionen gehören Konvertierungsfunktionen (unter anderem zur Umwandlung einer `chip.id` in ARRAY.TYPE und von CDF-Dateinamen oder von LIMS-Bezeichnern in SPLUS-Bezeichner) und vor allem die Funktion `.First`. Diese Sonderfunktion von SPLUS wird nach ihrer Definition bei jedem Start der Entwicklungsumgebung aufgerufen, sodass hier globale Aktionen wie das Einfügen von Datenbanken in die Datenbankhierarchie von SPLUS stattfinden können (`attach`-Befehl für die intensity- und analysis-Datenbank).

3.5 C++-Komponenten

In diesem Unterkapitel werden die Funktionen beschrieben, die in C++ mit dem LIMS SDK programmiert wurden. Auf sie wird mit der `.C`-Funktion von SPLUS

zugegriffen. Sie dienen einerseits zur Ermittlung der `chip.id` eines Array-Typs, zum Importieren von Layout-Informationen, zur Ermittlung der Namen und der Merkmale der Primäranalysen eines Experimentes und zum Import der Komponenten einer CEL-Datei. Andererseits sind dies Funktionen, mit denen (modifizierte) CEL-Dateien von SPLUS ins Dateisystem exportiert und dann aus dem Dateisystem ins LIMS-System importiert werden können.

Tabelle 13 enthält die implementierten C++-Funktionen jeweils zusammen mit der aufrufenden SPLUS-Funktion, welche im Anhang detaillierter beschrieben werden.

C++-Funktionen	aufrufende SPLUS-Funktion
`get_chip_idC`	`get.chip.id.of.experiments`
`get_cell_tagsC, read_PSI_fileC`	`import.chip.layout`
`get_analyses_names_of_experimentC, get_BF_NF_TGT_SF_C`	`import.analyses.descriptions`
`get_header_of_CELC`	`import.header.by.filename`
`get_number_of_masked_outlier_modified_of_CELC, get_masked_outlier_modified_of_CELC`	`import.masked.outlier.modified. by.filename`
`export_CEL_file`	`export.CEL.file`
`import_CEL_file_C`	`perform.LIMS.CEL.file.import`

Tabelle 13: C++- und aufrufende SPLUS-Funktionen

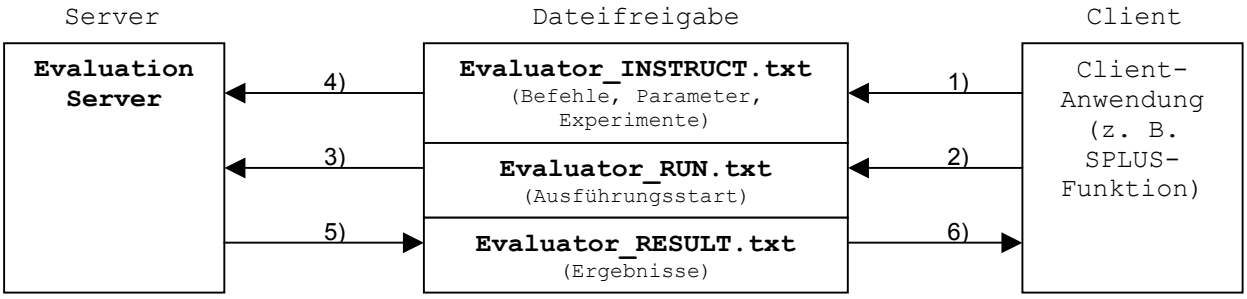
3.6 Evaluation Server – Client-Server-Anwendung zur Abfrage der *process*-Datenbank

Parallel zur Entwicklung der oben beschriebenen SPLUS-Funktionen wurde ein alternativer Ansatz verfolgt. Motiviert wurde er zunächst durch Überlegungen zur Laufzeit- und Speicherplatz-Effizienz. Zur Berechnung der Gesamtintensität nach dem bisher vorgestellten Konzept ist eine zeitintensive Übertragung der großen CEL-Dateien über das Netzwerk notwendig, dann wird die CEL-Datei lokal in der intensity-Datenbank von SPLUS abgespeichert und schließlich clientseitig aus den Intensitätsdaten beispielsweise eine Gesamtintensität ausgerechnet. Effizienter wäre es, die Gesamtintensität serverseitig zu berechnen und nur die errechnete Zahl über das Netzwerk zu übertragen. Hierbei entsteht fast kein Netzwerkverkehr und überdies gibt es kein lokales Duplikat großer Datenmengen. Der beschriebene Mechanismus kann durch eine sehr einfache Form der Interprozesskommunikation realisiert werden, nämlich auf der Ebene von Textdateien.

Ein wichtiger Nebeneffekt dieser Art der Interprozesskommunikation ist die Plattformunabhängigkeit. Clients unter verschiedenen Betriebssystemen können so Anfragen an den Server stellen.

Ergebnis des alternativen Ansatzes ist die so genannte Evaluation Server-Anwendung. Es handelt sich hierbei um eine in C++ programmierte Anwendung, die einem im Hintergrund laufenden Dienst ähnelt, jedoch über eine einfache Windows-Benutzerschnittstelle zur Steuerung verfügt. Die bisher implementierte Funktionalität kann beliebig erweitert werden. Damit ist Evaluation Server gerade in Umgebungen sinnvoll, in denen nicht direkt mit dem LIMS SDK entwickelt werden kann (z. B. Solaris). Im Folgenden werden Architektur und Funktionalität der Anwendung sowie einer der implementierten Beispiel-Clients ausführlicher beschrieben.

3.6.1 Architektur der Evaluation Server-Anwendung



Die Client-Anwendung schreibt den auszuführenden Befehl zusammen mit eventuellen Parametern zeilenweise in eine Textdatei `Evaluator_INSTRUCT.txt` in eine vorher festgelegte Dateifreigabe (Schritt 1). Eine Beschreibung der implementierten Befehle findet sich im nächsten Abschnitt. Nachdem die Textdatei geschlossen wurde, wird eine Datei `Evaluator_RUN.txt` angelegt (Schritt 2). Nach dem Starten der Evaluation Server-Anwendung und der Abfrageschleife (*Evaluator Loop*) über die Benutzerschnittstelle, wird das Vorhandensein der Datei `Evaluator_RUN.txt` detektiert (Schritt 3) und daraufhin die Befehlsdatei gelesen (Schritt 4). Abhängig vom zu bearbeitenden Befehl wird die passende Subroutine aufgerufen, welche sukzessive die übergebenen Parameter (beispielsweise Experimentnamen) aus der Befehlsdatei ausliest, die entsprechende Berechnung durchführt und das Ergebnis (eine Spalte) bzw. die

Ergebnisse (mehrere Spalten) in die Datei `Evaluator_RESULT.txt` schreibt (Schritt 5). Ist der Befehl abgearbeitet, wird die Ergebnisdatei geschlossen und die Datei `Evaluator_RUN.txt` gelöscht. Dies wird von der Client-Anwendung festgestellt, welche daraufhin mit dem Auslesen der Ergebnisse beginnen kann (Schritt 6).

3.6.2 Funktionalität der Evaluation Server-Anwendung

Benutzerschnittstelle

Auf dem Server wird die ausführbare EXE-Datei zusammen mit eventuell benötigten DLL-Dateien gespeichert und gestartet. Daraufhin erscheint ein einfaches Hauptfenster zur Steuerung der Aktivität (siehe Abbildung 12).

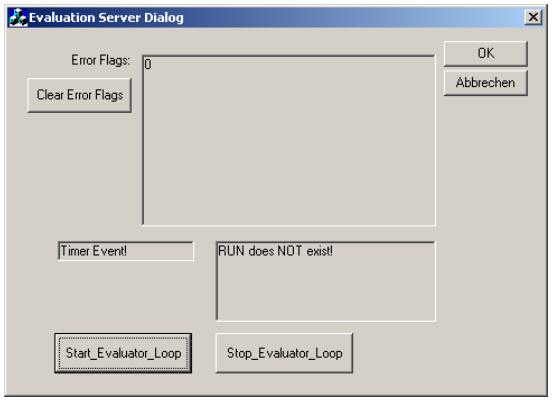


Abbildung 12: Evaluation Server: Hauptfenster zur Aktivitätssteuerung

Die wichtigsten Steuerelemente sind die beiden Buttons am unteren Fensterrand. **Start_Evaluator_Loop** startet die Abfrageschleife, welche wiederholt testet, ob die Datei `Evaluator_RUN.txt` existiert. In dem rechten der beiden unteren Meldungsfenster werden Statusmeldungen angezeigt, so beispielsweise „RUN does NOT exist!“ oder Meldungen des aktuell abgearbeiteten Befehls. **Stop_Evaluator_Loop** beendet die Abfrageschleife. Die aktuelle Befehlsdatei wird dabei allerdings zunächst vollständig abgearbeitet. Das obere Meldungsfenster und der obere Button **Clear Error Flags** hatten hauptsächlich während des Debuggings der Anwendung eine relevante Funktion. Ist die Abfrageschleife gestoppt, kann die Anwendung mit **OK** beendet werden.

Implementierte Befehle

Die folgenden Befehle arbeiten nur mit CEL- und RPT-Dateien, da zum Zeitpunkt ihrer Entwicklung das LIMS SDK noch nicht zur Verfügung stand. Die Befehle `sum_of_intens` und `minmax_of_intens` arbeiten der Einfachheit halber nur auf allen *probe cells* und können nicht auf eine Untermenge der *probe cells* eingeschränkt werden.

Übergebene Parameter sind in der Regel die Namen der Experimente, für welche die übergebene Funktion ausgeführt werden soll.

- `sum_of_intens`: Berechnen der Gesamtintensität (eine Ergebnisspalte)
- `minmax_of_intens`: Ermitteln des Minimums und des Maximums der *probe cell*-Intensitäten (zwei Ergebnisspalten).
- `number_of_masked_and_outliers`: Ermitteln der Anzahl der in der CEL-Datei gespeicherten maskierten und Outlier-*probe cells* (zwei Ergebnisspalten).
- `get_3_5_quotient`: Ermitteln eines 3'/5'-Quotienten. Über einen zusätzlichen Parameter wird die Position des Quotienten innerhalb der Liste der 3'/5'-Quotienten in der RPT-Datei angegeben. So können unterschiedliche 3'/5'-Quotienten und unterschiedliche Array-Typen adressiert werden (eine Ergebnisspalte).
- `get_number_of_total_A_M_P`: Ermitteln der Anzahl aller *probe sets* auf dem Array, und der Anzahl der *probe sets*, deren *Detection Call Absent*, *Marginal* bzw. *Present* lautet (vier Ergebnisspalten).

Im weiteren Verlauf der Entwicklung stand das LIMS SDK zur Verfügung. Dieses ermöglicht beispielsweise das Erstellen von Listen von Experimenten mit bestimmten Eigenschaften in der *process*-Datenbank. Bisher implementiert sind die Befehle:

- `get_all_projects_in_LIMS`: Ermitteln der Namen aller Experimente in der *process*-Datenbank (eine Ergebnisspalte).
- `get_experiments_of_project`: Ermitteln aller Experimente eines als Parameter übergebenen Projektes (eine Ergebnisspalte).

3.6.3 Beispiel-Client: SPLUS-Funktion

Die SPLUS-Funktion `add.property.col.with.Evaluator` stellt eine Implementierung eines Clients für Evaluation Server dar. Mit ihrer Hilfe können einem desc.df weitere Eigenschaftsspalten wie Gesamtintensität, Minimum und Maximum der Intensitäten hinzugefügt werden (siehe vorangehenden Abschnitt für mögliche Befehle).

Als Parameter werden der Funktion der Name des desc.df, die zu betrachtenden Experimente, der auszuführende Befehl und die Namen der Ergebnisspalten übergeben. Daraufhin wird die Datei Evaluator_INSTRUCT.txt in die vorher vereinbarte Freigabe geschrieben und die Datei Evaluator_RUN.txt erzeugt. Daraufhin wird in einer Schleife darauf gewartet, dass die Server-Anwendung diese Datei wieder löscht, um anzuzeigen, dass die Ergebnisse vorliegen. Anschließend wird die Datei Evaluator_RESULT.txt eingelesen. Die Anzahl der Ergebnisspalten muss mit der Anzahl der übergebenen Namen der Ergebnisspalten übereinstimmen. Die Ergebnisspalten werden dem desc.df hinzugefügt, sofern sie noch nicht vorhanden sind, und dann werden die Ergebnisse der Berechnung für das jeweilige Experiment eingetragen.

Beispiel:

Sei ein desc.df wie in Abbildung 13 gegeben.

N	EXPERIMENT.NAME	STAINING	DATUM
1.00	Coculture-Caco-Co-16-10-pre	Simple	16-10
2.00	Coculture-Caco-Co-16-10	AK	16-10
3.00	Coculture-Caco1_08_01-pre	Simple	8-10
4.00	Coculture-Caco1_08_01	AK	8-10
5.00	Coculture-Caco2_08_01-pre	Simple	8-10
6.00	Coculture-Caco2_08_01	AK	8-10
7.00	Coculture-Caco_08_01_Kontr	AK	8-10
8.00	Coculture-Caco_08_01_Kontr_pre	Simple	8-10
9.00	Coculture-Cacos-Co-Bed-15-10	AK	16-10
10.00	Coculture-Cacos-Lym-16-10-pre	Simple	16-10
11.00	Coculture-Cacos-Lym-16-10	AK	16-10
12.00	Coculture-Cacos-Lymph-15-10	AK	16-10
13.00	Coculture-Cacos-normal-pur-15-10	AK	16-10
14.00	Coculture-Cacos-pur-16-10-pre	Simple	16-10
15.00	Coculture-Cacos-pur-16-10	AK	16-10
16.00	Coculture-Lymphos-16-10-pre	Simple	16-10
17.00	Coculture-Lymphos-16-10	AK	16-10
18.00	Coculture-Lymphos-pur-15-10	AK	16-10

Abbildung 13: Beispiel SPLUS-Client: desc.df vor Funktionsaufruf

Durch folgenden Aufruf werden zunächst die Experimentnummern ermittelt, die zu HG-U95A- und Test3-Arrays gehören und die mit der Färbung AK durchgeführt wurden:

```
subsetN<-get.experiments( df=Test.Data,
                          properties=c("ARRAY.TYPE","STAINING"),
                          include=c(T,T),
                          values=list(c("HG_U95A","Test3"),c("AK")),
                          return.property="N")
```

Dann wird für diese Experimente der Evaluation Server mit dem Befehl `number_of_masked_and_outliers` verwendet, um die Anzahl der maskierten und der Outlier-*probe cells* zu ermitteln, welche in die Ergebnisspalten mit den Namen „`number.masked`“ und „`number.outlier`“ geschrieben werden (siehe Abbildung 14):

```
Test.Data<-add.property.col.with.Evaluator ( subsetN,
                          "Test.Data",
                          instruction.string="number_of_masked_and_outliers",
                          result.colnames=c("number.masked","number.outliers"),
                          timeout=6000,
                          read.RESULT.file.only=F)
```

N	EXPERIMENT.NAME	number.masked	number.outliers	STAINING	DATUM
1.00	Coculture-Caco-Co-16-10-pre	NA	NA	Simple	16-10
2.00	Coculture-Caco-Co-16-10	729	1691	AK	16-10
3.00	Coculture-Caco1_08_01-pre	NA	NA	Simple	8-10
4.00	Coculture-Caco1_08_01	0	628	AK	8-10
5.00	Coculture-Caco2_08_01-pre	NA	NA	Simple	8-10
6.00	Coculture-Caco2_08_01	0	59	AK	8-10
7.00	Coculture-Caco_08_01_Kontr	0	181	AK	8-10
8.00	Coculture-Caco_08_01_Kontr_pre	NA	NA	Simple	8-10
9.00	Coculture-Cacos-Co-Bed-15-10	NA	NA	AK	16-10
10.00	Coculture-Cacos-Lym-16-10-pre	NA	NA	Simple	16-10
11.00	Coculture-Cacos-Lym-16-10	764	2626	AK	16-10
12.00	Coculture-Cacos-Lymph-15-10	NA	NA	AK	16-10
13.00	Coculture-Cacos-normal-pur-15-10	NA	NA	AK	16-10
14.00	Coculture-Cacos-pur-16-10-pre	NA	NA	Simple	16-10
15.00	Coculture-Cacos-pur-16-10	0	1839	AK	16-10
16.00	Coculture-Lymphos-16-10-pre	NA	NA	Simple	16-10
17.00	Coculture-Lymphos-16-10	0	2651	AK	16-10
18.00	Coculture-Lymphos-pur-15-10	NA	NA	AK	16-10

Abbildung 14: Beispiel SPLUS-Client: Hinzugefügte Ergebnisspalten `number.masked` und `number.outliers`

Da unter Umständen die Berechnung der Ergebnisse längere Zeit in Anspruch nehmen kann und der Client während der Abfrageschleife in der vorliegenden einfachen Implementierung mit Warten beschäftigt ist, wurde ein Parameter `timeout` eingeführt, mit dem die maximale Wartezeit festgelegt werden kann. Wird diese überschritten, kommt es zum Abbruch der Abfrageschleife und der gesamten Funktion. Um danach oder nach einem Abbrechen der Funktion durch den Benutzer mit der ESC-Taste trotzdem die

berechneten Ergebnisse in den desc.df einzufügen, kann über einen Parameter `read.RESULT.file.only` das alleinige Lesen der Ergebnisdatei erzwungen werden.

Neben diesem Client wurde ein Perl-Webclient unter Solaris implementiert, mit dem eine plattformübergreifende Abfrage der *process*-Datenbank möglich ist.

4 Meta-Analysen zur Untersuchung und Entwicklung von Qualitätskriterien

GeneChip-Experimente sollen *per definition* quantitative Messwerte für die Einheit *probe set* liefern. Die Analyse dieser Messgrößen über mehrere Experimente wird als Meta-Analyse bezeichnet. Bei einem GeneChip-Experiment fallen pro Chip zusätzlich sowohl zusammenfassende Größen (z. B. 3'/5'-Quotient) an, als auch Messungen für kleinere Einheiten (z. B. *probe cells*). Auch die Analysen der zusammenfassenden Messgrößen unter Berücksichtigung mehrerer Experimente und die Analysen der Messwerte der kleineren Einheiten über mehrere oder einzelne Experimente werden hier als Meta-Analysen verstanden. Das Ziel von Meta-Analysen ist es, das Verhalten der Messgrößen zu beschreiben, eine eventuell beobachtete Varianz zu quantifizieren, und gegebenenfalls durch Definition eines Sollbereiches Qualitätskriterien zu entwickeln, anhand derer Experimente fragwürdiger Qualität identifiziert werden können. Außerdem besteht die Hoffnung, die hierdurch gewonnenen Erkenntnisse für alternative Skalierungsmethoden (siehe Kapitel 5) nutzen zu können.

Unterkapitel 4.1 diskutiert zunächst Quellen biologischer und technischer Varianz. Unterkapitel 4.2 stellt daraufhin vorhandene Qualitätskriterien sowie deren Verhalten vor. In Unterkapitel 4.3 wird schließlich die zusammenfassende Größe „Gesamtintensität" untersucht und ihre Verwendung als zusätzliches Qualitätskriterium diskutiert.

4.1 Biologische und technische Varianz

Unter Varianz wird hier in Abweichung zur ursprünglichen mathematischen Definition die Veränderung einer gemessenen Größe innerhalb der Zeit oder über mehrere Experimente verstanden. Dabei geht es beispielsweise um die Größe „Konzentration einer mRNA", wenn der *Signal*-Wert eines *probe sets* betrachtet wird und um die Größe „Konzentration eines Oligonukleotidfragments" bei Betrachtung der Intensität einer *probe cell*. Die Veränderung der hier gemessenen Größen hängt von einer Vielzahl von Faktoren ab: Zunächst ist sie natürlich von dem biologischen Zustand der Zelle abhängig, da dieser die Konzentration einer mRNA determiniert. Die Veränderungen, die die Konzentration einer mRNA in der Zelle betreffen, werden unter dem Begriff **biologische Varianz** zusammengefasst. Ziel von Genexpressionsanalysen ist es, gerade diese Veränderungen in Abhängigkeit von unterschiedlichen Einflussfaktoren zu quantifizieren.

Sei nun die Konzentration einer mRNA zu einem bestimmten Zeitpunkt fest. Um diese mithilfe der GeneChip-Technologie messen zu können, müssen diverse Zwischenschritte (die Aufarbeitung) durchgeführt werden. Unwägbarkeiten in diesen Schritten führen zu einem variablen Signal. Diese Veränderungen des Signals trotz konstanter Konzentration einer mRNA werden unter dem Begriff **technische Varianz** zusammengefasst.

In einer formalisierten Darstellung lesen sich diese Zusammenhänge wie folgt: Für die Maßzahl *Signal* eines *probe sets i* gilt:

$$Signal(probe\ set_i) = (f_i \circ \mathrm{var}_i^{\mathrm{tech}})\, c_{Zelle}(\mathrm{mRNA}_i),$$

wobei f_i und $\mathrm{var}_i^{\mathrm{tech}}$ Funktionen sind. Im Idealfall würde es sich bei der Messfunktion f_i um eine bekannte lineare Funktion handeln, deren Umkehrfunktion bekannt ist und die eine Abbildung einer Konzentration auf einen *Signal*-Wert ist, und bei der Funktion der technischen Varianz $\mathrm{var}_i^{\mathrm{tech}}$ um die Identität (oder zumindest eine Funktion mit bekannter Umkehrfunktion), damit die biologische Varianz vollständig charakterisiert werden kann. Im Realfall werden jedoch große nicht-lineare Anteile vorhanden sein und die Umkehrfunktionen gerade aufgrund von Nicht-Linearitäten an den Rändern von Definitions- bzw. Wertebereichen (Sättigungseffekte) nicht bekannt sein. Angemerkt sei, dass $\mathrm{var}_i^{\mathrm{tech}}$ durch Einflüsse wie unspezifische Hybridisierung und Kreuzhybridisierung vom gesamten Sample und von den Bedingungen während der Aufarbeitung abhängt und sich daher von Experiment zu Experiment unterscheidet.

Analog zum *Signal* gilt für die Intensität einer *probe cell j*:

$$Intensity(probe\ cell_j) = (g_j \circ \mathrm{var}'^{\mathrm{tech}}_j)\, c_{Sample}(\mathrm{oligo}_j).$$

Mit den Intensitäten der *probe cells* wird mit dem Kondensierungsalgorithmus folgende Funktion *cond* realisiert:

$$Signal(probe\ set_i) = cond(\mathrm{CEL}, Intensity(probe\ cell_k), Intensity(probe\ cell_l) \mid \forall\ (\mathrm{PM},\mathrm{MM})\text{-}probe\ pairs\ (k,l))$$

Dabei fließen zusätzlich zu den Intensitäten aller (PM,MM)-*probe pairs* eines *probe sets* noch globale Chip-Eigenschaften wie *noise* und *background* ein, was hier mit einem zusätzlichen Parameter CEL formuliert wurde. Die Nicht-Linearität des Kondensierungsalgorithmus hat zur Folge, dass $\mathrm{var}_i^{\mathrm{tech}}$ zwar prinzipiell durch Funktionen $\mathrm{var}'^{\mathrm{tech}}_j$ ausgedrückt werden kann, jedoch die Formulierung nicht-trivial ist.

Verschiedene Veröffentlichungen behandeln die letztlich resultierende technische Varianz, ohne die einzelnen Varianzquellen näher untersucht haben zu müssen: Huang und Pan[47] versuchen eine Abschätzung der auftretenden Varianz, um die Voraussetzungen für einen t-Test zu prüfen; Huber et al.[48] stellen unter anderem eine Maßnahme zur Stabilisierung der Varianz vor; Welle et al.[96] reduzieren die Varianz durch Nutzung der *comparison analysis* zweier Chips von Affymetrix.

Im folgenden Unterkapitel werden unterschiedliche Arten von Varianz definiert, während das Unterkapitel 4.1.2 eine detaillierte Diskussion möglicher Varianzquellen enthält.

4.1.1 Varianzarten

Im Folgenden werden unterschiedliche Arten von Varianz beschrieben, die sich aber nicht gegenseitig ausschließen.

Bias

Hierunter werden in der Regel Varianzen verstanden, die systematischer Natur sind und bei Wiederholungsmessungen reproduziert werden (Churchill und Oliver[27]). Naturgemäß handelt es sich dabei eher um Anteile der technischen Varianz. Bias bezeichnet hier zusätzlich eine Varianzart, die alle *probe cells* eines Chips in gleicher Quantität betrifft, also Auswirkungen auf die Messgrößen der gesamten Messeinheit hat (**globaler Bias**).

Ein Beispiel für Bias ist die konstante Messung eines Null-*Signal*-Wertes für eine bestimmte mRNA, obwohl sie eigentlich in einer messbaren Konzentration vorliegt. In diesem Fall existiert auch keine Umkehrfunktion. Ein weiteres Beispiel ist die Verfälschung des *Signal*-Wertes um einen bestimmten Faktor. Dabei mag die Umkehrfunktion nicht unbedingt bekannt sein; trotzdem hätte dieser Bias keine Auswirkungen auf den *Fold Change* eines Gens (Ausmaß der Regulation zwischen zwei Experimenten).

Neben dem Auftreten eines konstanten Faktors kann sich Varianz auch durch Addition eines konstanten Summanden äußern (**multiplikative / additive Varianz**).

Bei der Entdeckung eines globalen Bias besteht die Hoffnung, durch Bestimmung der korrekten Umkehrfunktion eine Skalierung zu erhalten, die diese Art technischer Varianz aus dem Experiment weitestgehend herausrechnet.

Noise

Hierunter fallen Varianzen, die bei Wiederholungsmessungen nicht reproduziert werden. Sie treten zufällig auf, folgen dabei aber eventuell einer Verteilung. Die Ursachen lassen sich nicht ohne weiteres aufklären, sie werden unter dem Begriff „Messungenauigkeiten“ zusammengefasst. Churchill et al.[27] zählen auch die biologische Varianz dazu (Bezeichnung dort: „*variance*“). Beim Kondensierungsalgorithmus wird der Noise der *background-probe cells* eines Chips im Schritt *noise correction* verwendet.

Noise kann nicht durch eine Skalierung herausgerechnet werden, sondern muss durch andere Maßnahmen verhindert (Erhöhung der Reproduktionsgenauigkeit) oder durch Mehrfachexperimente gehandhabt (t-Test) oder quantifiziert (Konfidenzintervall) werden. Im Allgemeinen tritt Bias nicht ohne Noise auf.

Lokale Varianz

Probleme beim Hybridisieren oder Scannen wie helle oder dunkle Flecken (*light spots* bzw. *dark spots*) sowie dunkle Chip-Ränder oder Artefakte im Intensitätsbild, die durch Kratzer auf der Chip-Oberfläche bzw. durch Präzipitate oder Flusen im Sample entstehen, werden als lokale Varianz bezeichnet. Charakteristisch ist, dass physikalisch benachbarte *probe cells* betroffen sind. Sie werden in der Regel automatisch erkannt und maskiert. Damit nicht automatisch mehrere *probe cells* eines *probe sets* betroffen sind, hat Affymetrix bei den neueren Array-Typen die *probe pairs* aller *probe sets* über den Chip verteilt.

Schadt et al.[78] stellen Korrekturen für *background*- und Intensitätsgradienten vor, welche als Zwischenform von globaler und lokaler Varianz gelten können.

4.1.2 Quellen biologischer und technischer Varianz

Zu den Quellen biologischer Varianz gehören neben vielen anderen beispielsweise die folgenden Einflussgrößen:

- Stoffwechsellage (Nährstoffsituation, Tagesabhängigkeit, Energiebedarf),
- Alter der Zelle / des Gewebes,
- Position im Zellzyklus,
- genetische Vorgaben (z. B. Genotyp, Polymorphismen, Gendefekte, Chromosomenaberrationen usw.),
- äußere Reize wie Temperatur, biologische oder chemische Substanzen,

oder auf Individuen / Probanden bezogen etwa:

- intraindividuelle Varianz (Unterschiede zwischen verschiedenen Zuständen eines Individuums wie Tageszeiten, Ernährungslage, Leistungsstatus),
- interindividuelle Varianz (Unterschiede zwischen unterschiedlichen Individuen),
- Varianz zwischen unterschiedlichen Geschlechtern.

Die Effekte dieser Einflussgrößen überlagern sich in der Regel, sodass die exakten Auswirkungen einer Einflussgröße nur unter Beibehaltung gleicher Werte für die anderen Einflussgrößen bestimmt werden könnten. Für einige Einflussgrößen ist dies jedoch prinzipiell nicht möglich, da sie das Vorhandensein anderer Einflussgrößen bedingen. Beispielsweise kann die Varianz zwischen Geschlechtern nur bei gleichzeitig vorhandener interindividueller Varianz betrachtet werden.

Ebenfalls ist zu beachten, dass bei Aufarbeitung eines Zellverbandes oder eines Tumors ein Durchschnittsbild vieler Einflussgrößen entsteht. So befinden sich die Zellen eines Gewebes in der Regel an unterschiedlichen Position im Zellzyklus (außer sie werden synchronisiert wie beispielsweise in Cho et al.[25]). Um die Durchschnittsbildung zu vermeiden, besteht die Möglichkeit, durch Mikrodissektion einzelne Zellen zu gewinnen (Ohyama et al.[70]) und mithilfe eines modifizierten Protokolls (Eberwine et al.[32]) trotzdem genügend Material für ein GeneChip-Experiment zu gewinnen. Dabei besteht allerdings die Gefahr, dass Bias und Noise verstärkt werden.

Die Quellen biologischer Varianz sind vielfältig. Um sie möglichst genau quantifizieren zu können, muss – abgesehen von dem Ansatz, viele Replikatexperimente desselben biologischen Zustands durchzuführen – die technische Varianz so gering wie möglich gehalten werden. Um die Ergebnisse gut reproduzieren zu können, ist zunächst eine möglichst genaue Kenntnis der Quellen technischer Varianz notwendig. Piper et al.[72] untersuchen hierzu sogar die Varianz zwischen verschiedenen Laboratorien. Werden diese Varianzquellen ausgeschaltet und in einem einzigen Labor gearbeitet, verbleiben weitere Varianzquellen in der eigentlichen Aufarbeitung.

Abbildung 15 enthält ein detailliertes Bild der diskutierten Ursachen. Es kann leicht durch zusätzliche Aspekte erweitert werden.

Auf der linken Seite finden sich in Fettschrift die einzelnen Schritte eines Experimentes von der Entnahme der Zellen aus ihrer natürlichen bzw. aus ihrer Kultur-Umgebung bis hin zur Kondensierung der Intensitäten zu Maßzahlen. Die umrandeten Textbereiche enthalten die jeweils zu betrachtenden Messeinheiten. Dabei handelt es sich zunächst um Größen, die für eine bestimmte mRNA – also in der Regel für eine Spleißvariante eines Gens – erfasst werden können. Nach der Fragmentierung teilen sie sich in Größen auf, die für jedes mögliche Fragment gelten. Nach der Hybridisierung findet eine Reduzierung auf Messgrößen statt, die eine *probe cell* charakterisieren. Der Kondensierungsalgorithmus kondensiert die Teilgrößen zu Maßzahlen, die je *probe set* berechnet werden. Insgesamt sind also intermediäre Messgrößen aufgeführt, die bei der Bestimmung des Expressionslevels betrachtet werden können, welches letztlich zur ursprünglichen Konzentration der mRNA proportional sein soll. Auf der rechten Seite sind in rot die Einflussgrößen aufgelistet, die zu der technischen Varianz des jeweiligen Schrittes der Aufarbeitung beitragen. Oben sind einige Stichworte zur biologischen Varianz wiederholt, die die Ausgangskonzentration einer mRNA in der Zelle determinieren.

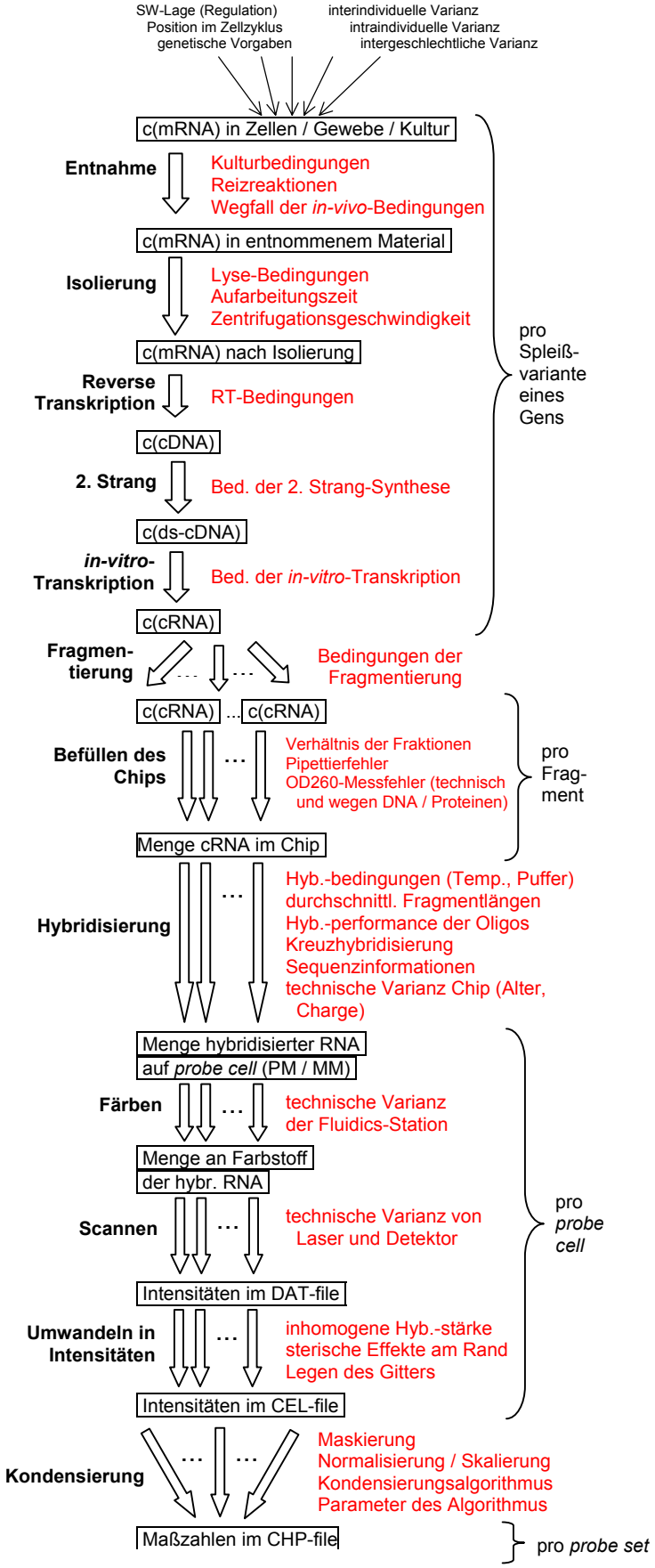


Abbildung 15: Quellen technischer Varianz

Im Folgenden sollen einige der Varianzquellen näher erläutert und die wichtigsten dieser Quellen herausgestellt werden.

Die **Entnahme** von Zellen führt zum Wegfall der *in-vivo*-Bedingungen. Je nachdem, wie die Kulturbedingungen waren und wie schnell die Zellen weiterverarbeitet werden können, fallen die Effekte der Reizreaktionen der Zellen mehr oder weniger stark aus.

Bei der **Isolierung** der RNA müssen zunächst die Zellwände aufgeschlossen werden. Die Substanzen dieser Lyse und die Bedingungen, unter der sie stattfindet, haben Auswirkungen auf die Menge, Zusammensetzung und Qualität der gewonnenen mRNA. Menge und Qualität werden entscheidend durch die Freisetzung endogener RNAsen und die Induktion von Gentranskription durch Zellmanipulationen beeinflusst, was ein möglichst rasches Vorgehen erfordert. Die Zentrifugationsgeschwindigkeit und vor allem die Aufarbeitungszeit, aber auch ein zwischen den Experimenten veränderter zeitlicher Ablauf sind besonders kritische Faktoren.

Bei der **Reversen Transkription (RT)** der mRNAs mag das Bindungsverhalten des PolyT-Primers an die PolyA-Bereiche relativ wenig zur technischen Varianz beitragen, jedoch kann die Polymerisation des cDNA-Strangs an nicht vorhersagbaren unterschiedlichen Positionen abbrechen. Die *probes* eines Gens sind aus diesem Grund möglichst nahe am 3'-nichtkodierenden Bereich lokalisiert. Auf die Abwesenheit von RNAsen ist besonders zu achten. Dies gilt im selben Maße für die **Synthese des 2. Strangs**.

Für die ***in-vitro*-Transkription (IVT)** gilt wie für die Reverse Transkription, dass die Synthese der gelabelten RNA an nicht vorhersagbaren unterschiedlichen Positionen abbrechen kann. Besonders kritisch sind hier die relativen Verhältnisse zwischen Template, Nukleotiden und Enzymaktivitäten. Zu lange Reaktionszeiten führen zu einer überproportionalen Amplifikation kurzer RNAs.

Die **Fragmentierung** der synthetisierten gelabelten RNA durch Salze und Hitze wird durch die vorherrschenden Bedingungen wie Temperatur, Konzentration und Zeitdauer

des Vorgangs beeinflusst. Hierdurch wird die durchschnittliche Länge der im Sample vorhandenen Oligonukleotide variiert.

Die Herstellung des Hybridisierungscocktails und das **Befüllen des Chips** mit einer bestimmten Menge mRNA-Material erfordert eine Konzentrationsbestimmung, welche technischen Messungenauigkeiten unterliegt (der entscheidende Fehler resultiert aus dem Pipettieren sehr kleiner Volumina). Die Konzentrationsmessung basiert beim Standard-protokoll auf dem Absorptionsmaximum von Nukleinsäuren bei einer Wellenlänge von 260nm. Dabei können zwei grundlegende Fehler unterlaufen. Zum einen wird durch die Messung die Konzentration aller RNAs im Sample bestimmt. Selbst wenn die Konzentration in zwei unterschiedlichen Experimenten als gleich bestimmt wird, kann die Zusammensetzung der beiden Samples aus unterschiedlichen RNA-Fraktionen variieren. Dies kann gerade durch fraktionsabhängige Effizienzen in den bis hierhin durchgeführten Aufarbeitungsschritten vorkommen. Wichtig ist in diesem Zusammenhang die Tatsache, dass rRNA, die in den Zellen vorhanden war, durch das verwendete Protokoll nicht aus dem Sample entfernt wird und hierfür ein mittlerer prozentualer Schätzwert (basierend auf der Ausgangsmenge an total-RNA) abgezogen wird. Zum anderen kann die optische Dichte durch DNA-Reste im Sample beeinflusst werden. Zu den größten Unsicherheiten beim Befüllen eines Chips zählt das Pipettieren, welches in der Regel von Hand ausgeführt wird.

Bei der **Hybridisierung** der Oligonukleotide im Sample an den Oligonukleotidrasen auf dem Siliziumträger im Hybridisierungsofen gibt es zahlreiche Varianzquellen. Zunächst sind die Hybridisierungsbedingungen wie Temperatur, Puffereigenschaften und Hybridisierungsdauer zu nennen, des Weiteren die durchschnittlichen Fragmentlängen der Oligonukleotide im Sample und die daraus resultierenden Unterschiede in der Hybridisierungsperformance zwischen zwei Experimenten.

Das Hybridisierungsverhalten wird ganz wesentlich beeinflusst durch Gene, die mit dem eigentlichen Zielgen für ein bestimmtes an die Festphase gebundenes Oligonukleotid im Bereich der vorgesehenen Hybridbildung zumindest abschnittsweise eine hohe Sequenzähnlichkeit aufweisen. Hierdurch kann eine Kreuzhybridisierung entstehen, die durch Kompetition mit dem eigentlichen Bindungspartner oder durch Vortäuschen einer spezifischen Bindung die Messwerte verfälscht. Ähnliche Auswirkungen haben

Hybridbildungen zwischen den *probes* in der Flüssigphase bei Vorliegen invers komplementärer Sequenzen zweier verschiedener Gene. Findet sich eine solche invers komplementäre Region innerhalb eines Gens, können sich stabile intragenische Strukturen ausbilden, die eine Bindung an das auf der Festphase immobilisierte Oligonukleotid erschweren oder unmöglich machen. Um die letztgenannten Phänomene zu reduzieren, werden die erzeugten cRNAs im Fragmentierungsschritt auf eine mittlere Länge von 200 Basen reduziert. Eventuellen Fehlhybridisationen wird beim Array-Entwurf vorgebeugt durch die geeignete Selektion von immobilisierten Oligonukleotidsequenzen und durch die Repräsentation eines einzelnen Gens mit in der Regel 16 verschiedenen *probe pairs* (HG-U95-Arrays). Die *Mismatch-probe cells* sollen die Quantifizierung der Fehlhybridisation und ihre Berücksichtigung bei der Berechnung des *Signal*-Wertes ermöglichen.

Eine weitere Quelle technischer Varianz im Hybridisierungsschritt ist die spezifische Hybridisierung von Fragmenten an die *Mismatch*-Zellen. Dies soll zwar durch die Auswahl der *probe sequences* aus den Sequenzinformationen der Spezies verhindert werden, kann jedoch durch unvollständige oder fehlerhafte Sequenzdatenbanken oder durch Mutationen des Genoms vorkommen. Mutationen können zudem die Ursache dafür sein, dass spezifische Hybridisierung an einer *Perfect Match*-Zelle eines anderen Gens stattfindet. Letztere Effekte sollen durch den Kondensierungsalgorithmus, insbesondere durch die gewichtete Durchschnittsbildung (*One-Step Tukey's Biweight Estimator*) justierter PM-MM-Differenzen an Bedeutung verlieren.

Ein zusätzliches relevantes Problem stellt die technische Varianz zwischen zwei Chips dar, zu der das Alter, die Charge und die Aufbewahrungsbedingungen beitragen.

Das **Färben** der hybridisierten Biotin-gelabelten Oligonukleotide unterliegt den möglichen Auswirkungen technischer Varianzen der Fluidics-Station.

Beim **Scannen** des Chips – also beim Erstellen der DAT-Datei – ist die technische Varianz von Laser und Detektor die wichtigste Einflussgröße. Für reproduzierbare Ergebnisse sind eine genügend hohe Auflösung des Lasers (Pixelgröße) und ein ausreichend starker Kontrast notwendig.

Das **Umwandeln** des Intensitätsbildes in *probe cell*-Intensitäten (DAT- −> CEL-Datei) geschieht vollständig determiniert durch einen Algorithmus. Jedoch haben Effekte wie eine inhomogene Hybridisierungsstärke über die räumliche Ausdehnung der *probe cell* und Intensitätsübergänge am Rande des Oligonukleotidrasens aufgrund sterischer Effekte einen Einfluss auf die Berechnung einer repräsentierenden Intensität jeder *probe cell*. Auch das automatische Legen eines Gitters (*grid*) zur Abgrenzung der *probe cell*s untereinander kann bei mehr oder weniger schief liegenden Chips eine Variation in den Intensitäten hervorrufen.

Der Schritt, in dem die **Kondensierung** von Intensitäten in Maßzahlen erfolgt, ist zum einen von der Zusammensetzung und Anzahl der maskierten *probe cells* abhängig und zum anderen vom verwendeten Kondensierungsalgorithmus und seinen Parametern. Einen wichtigen Bestandteil des Kondensierungsalgorithmus von Affymetrix stellt die Normalisierung bzw. Skalierung dar, die eine Art von globaler Varianz herausrechnen soll, nämlich die Verschiebung des gesamten Bereiches der *Signal*-Maßzahlen. Dieses Thema wird in Kapitel 5 behandelt.

4.2 Verhalten der bisherigen Affymetrix-Qualitätskriterien

In diesem Unterkapitel werden Qualitätskriterien beschrieben, die zur Beurteilung von GeneChip-Experimenten herangezogen werden. Die folgenden Abschnitte enthalten Beobachtungen, wie sich die verschiedenen Qualitätskriterien bei den betrachteten Datensätzen verhalten. Außerdem wird diskutiert, welche Rückschlüsse daraus auf die im vorigen Unterkapitel aufgeführten Varianzquellen möglich sind.

4.2.1 *background*, *noise* und *Noise(Q)*

Die Begriffe *background* und *noise* tauchen im Benutzerhandbuch zur Microarray Suite[9] und in der Dokumentation des Kondensierungsalgorithmus [14] in unterschiedlichen Zusammenhängen auf[†]. Einerseits werden *background* und *noise* für die Schritte *background subtraction* und *noise correction* des Kondensierungsalgorithmus verwendet, andererseits werden *background* und *noise* im Report zu einer CHP-Datei angegeben. Die

[†] Eine weitere Mehrdeutigkeit existiert zum Noise-Begriff aus dem vorangegangenen Unterkapitel: *noise* (und im weiteren Verlauf *Noise(Q)*) stellen Qualitätsmerkmale für einen Chip dar, während der vorab definierte Noise-Begriff eine bestimmte Art von Chip-übergreifender Varianz bezeichnet.

so genannten *background-probe cells* sind die intensitätsschwächsten zwei Prozent aller nach Intensitäten sortierten *probe cells*, die zu *probe sets* gehören; der *background* ist der Durchschnitt der Intensitäten der *background-probe cells*. Der Wert *noise* entspricht der Standardabweichung der Intensitäten der *background-probe cells*. Für *background subtraction* und *noise correction* werden diese Kenngrößen jeweils für räumliche Unterteilungen des Chips, die so genannten *zones*, berechnet, für den Report wird offenbar der gesamte Chip für die Berechnung zugrunde gelegt.

Mit den Kenngrößen *background* und *noise* aus dem Report besteht die Hoffnung, zu einer ersten Einschätzung einiger Varianzquellen zu gelangen, die teils als Bias, teils als Noise zur technischen Varianz beitragen. Anzumerken ist, dass die Art und Weise der Berechnung des *background* lediglich aus der Tatsache erwächst, dass es auf dem Chip inmitten der *probe cells* keine systematisch vorgesehenen Bereiche ohne Oligonukleotide gibt, an denen tatsächlich unspezifische Hybridisierung oder Autofluoreszenz quantifiziert werden könnten. Es lässt sich also nicht ausschließen, dass mit der verwendeten Definition des *background* die Folgen spezifischer, wenn auch sehr schwacher Hybridisierung, gemessen werden. Die *a priori*-Festlegung auf die zwei Prozent intensitätsschwächsten *probe cells* ist lediglich ein Erfahrungswert.

Die Abbildungen 16 bzw. 17 stellen *background* bzw. *noise* für die betrachteten Datensätze dar.

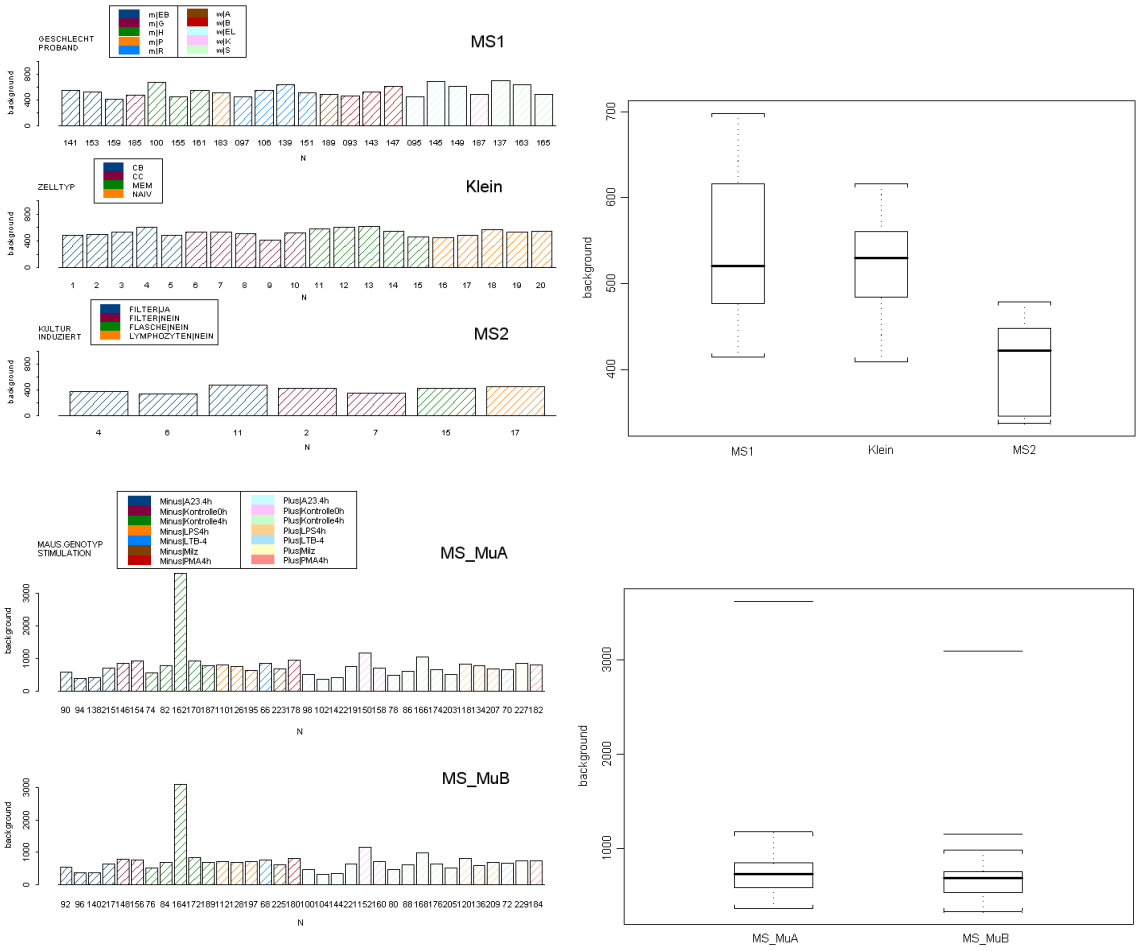


Abbildung 16: Qualitätskriterium *background*
(links: experimentweise, rechts: Zusammenfassung pro Datensatz)

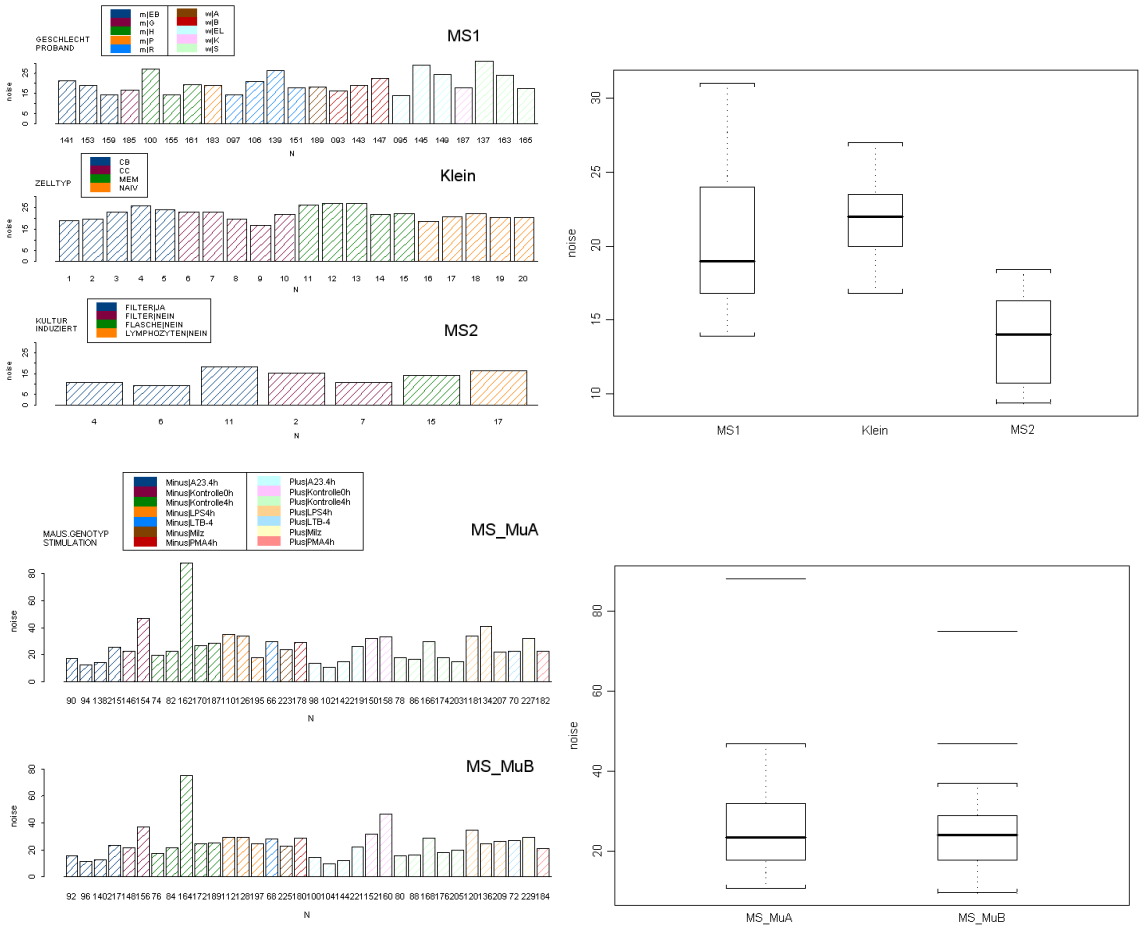


Abbildung 17: Qualitätskriterium *noise*
(links: experimentweise, rechts: Zusammenfassung pro Datensatz)

Durchschnitt und Standardabweichung von *noise* und *background* ergeben sich damit wie in Tabelle 14.

	background		***noise***	
	Durch-schnitt	**Std.-Abw.**	**Durch-schnitt**	**Std.-Abw.**
MS1	541,78	84,80	20,23	4,87
Klein	525,17	54,74	22,05	2,82
MS2	403,36	53,87	13,53	3,34
MS_MuA	795,12	532,94	26,33	13,81
MS_MuB	730,36	452,28	25,01	11,84
MS_MuA (ohne Outlier)	709,62	191,22	24,46	8,60
MS_MuB (ohne Outlier)	658,72	176,01	23,50	8,01

Tabelle 14: Durchschnitt und Standardabweichung von *background* und *noise*

Der Durchschnitt des *background* unterscheidet sich nur wenig zwischen Klein- und MS1-Datensatz. Mit einem ungepaarten t-Test, der nach Welch modifiziert keine identischen Standardabweichungen erfordert, kann kein signifikanter Unterschied belegt

werden. Die Gleichheit der Durchschnitte des *background* von MS1 und MS2 kann allerdings mit diesem t-Test mit p=0.0001 und die von Klein und MS2 mit p=0.0004 verworfen werden. Die Standardabweichungen des *background* von Klein und MS2 unterscheiden sich untereinander wenig, sind jedoch verschieden von MS1. Ursache hierfür könnte das zugrunde liegende biologische Material sein: Bei Klein wurden mit vier Arten von B-Zellen homogenere Samples verwendet als bei MS1, dessen Samples aus mehreren Fraktionen von Blutzellen zusammengesetzt sind. Der MS2-Datensatz enthält reine und mit Lymphozyten aktivierte und gemischte Caco-2-Zellen und kann damit ebenfalls immer noch als homogener gelten als die MS1-Samples.

MS_MuA und MS_MuB zeigen statistische Kenngrößen von *background* und *noise*, die über denen der U95A-Datensätze liegen. Durch deutliche *outlier*-Experimente (162 bzw. 164) sind Durchschnitt und Standardabweichung nach oben verfälscht, Tabelle 14 enthält daher auch die Werte bei Auslassung des *outliers*. Interessant ist, dass sowohl *background* als auch *noise* in den Replikatpaaren in MS_MuA und MS_MuB gut korreliert sind (r=0.99 bzw. r=0.93, siehe Abbildung 18). Hieran lässt sich bereits die wichtige Beobachtung festmachen, dass die größte Quelle der hier dargestellten Varianz nicht aus den Schritten „Befüllen des Chips“, „Hybridisierung“, „Färben“ und „Scannen“ resultieren kann.

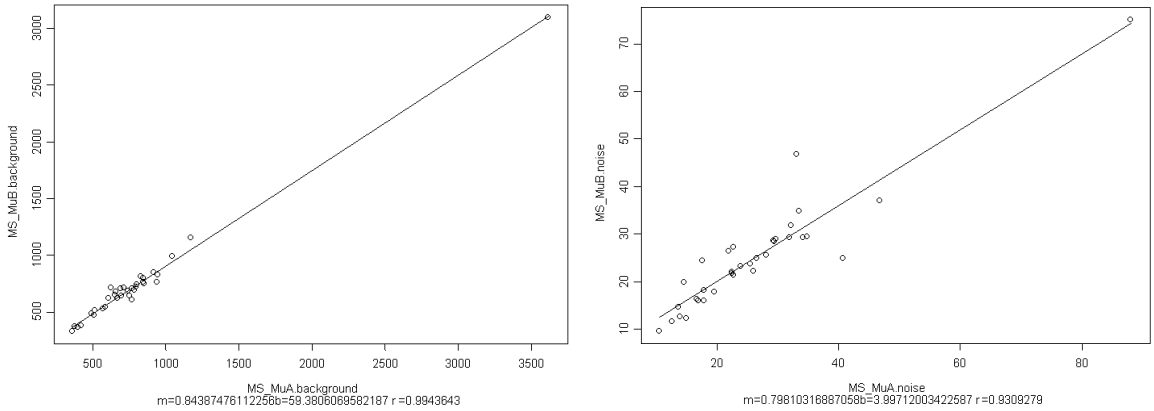


Abbildung 18: Korrelationen von *noise* und *background* zwischen Replikatpaaren in MS_MuA und MS_MuB

Für den *noise* gilt jeweils Analoges wie das oben gesagte für den *background*. Bei genauerer Betrachtung des Zusammenhangs zwischen *background* und *noise* zeigt sich eine gute bis sehr gute Korrelation ($0{,}81 \leq r \leq 0{,}99$) zwischen beiden (siehe

Abbildung 19). Zusammen mit der Tatsache, dass der *noise per definition* die Standardabweichung der Intensität der *background-probe cells* eines Chips ist, ergibt sich damit ein Hinweis darauf, dass die beobachtete Varianz sich im Wesentlichen multiplikativ verhält. Bei einer additiven Varianz würde die Intensität der *background-probe cells* variieren, der *noise* jedoch auf einem vergleichbaren Niveau bleiben.

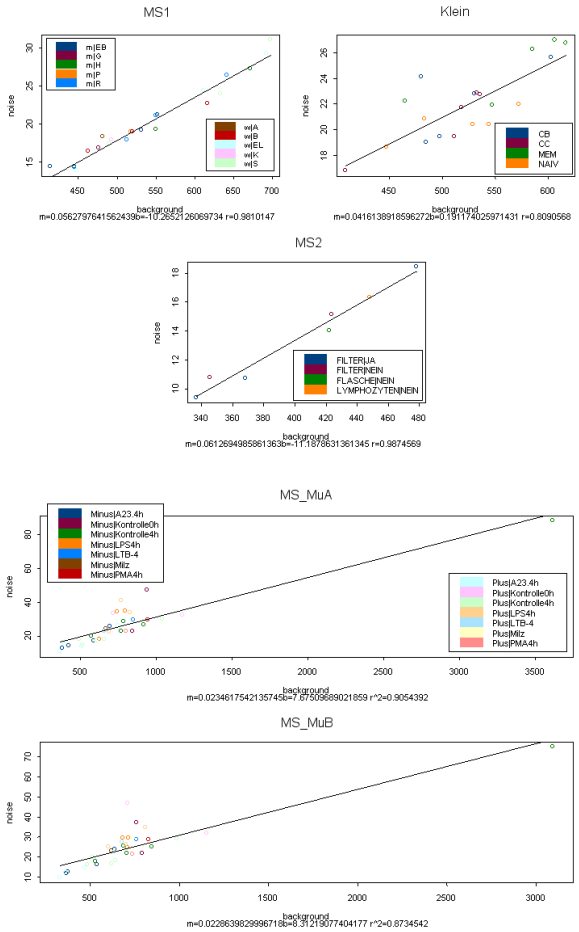


Abbildung 19: Korrelationen zwischen *background* und *noise* in allen Datensätzen

Neben *noise* und *background* wird im Report zu einer CHP-Datei ein Wert *Noise(Q)* angegeben, dessen Berechnung in der Dokumentation des Kondensierungsalgorithmus beschrieben ist (*Statistical Algorithms Description Document*[14], Anhang III, Seite 27). Er wird im Kondensierungsalgorithmus selbst nicht verwendet und kann laut Affymetrix als ein Qualitätskriterium für das Legen des Gitters zur Berechnung der CEL-Datei angesehen werden, da er auf den Pixel-Informationen der DAT-Datei basiert. Definiert ist er als $Q = \frac{1}{N}\left(\sum_{i}^{N} \frac{stdev_i}{\sqrt{pixel_i}}\right) * SF * NF$, wobei also Skalierungs- und Normalisierungsfaktor als

Faktoren einfließen. Jedoch wird im Report der Wert `Raw(Q)` angegeben, welcher offenbar unter Auslassung dieser beiden Faktoren berechnet wird. Aufsummiert wird über alle *N background-probe cells*. Der Quotient nach dem Summenzeichen ist im Wesentlichen eine Quantifizierung der Standardabweichung der Intensitäten aller Pixel einer *probe cell* normiert auf die Anzahl dieser Pixel.

In den betrachteten Datensätzen verhält sich *Noise(Q)* aus den Report-Dateien wie in Abbildung 20 dargestellt.

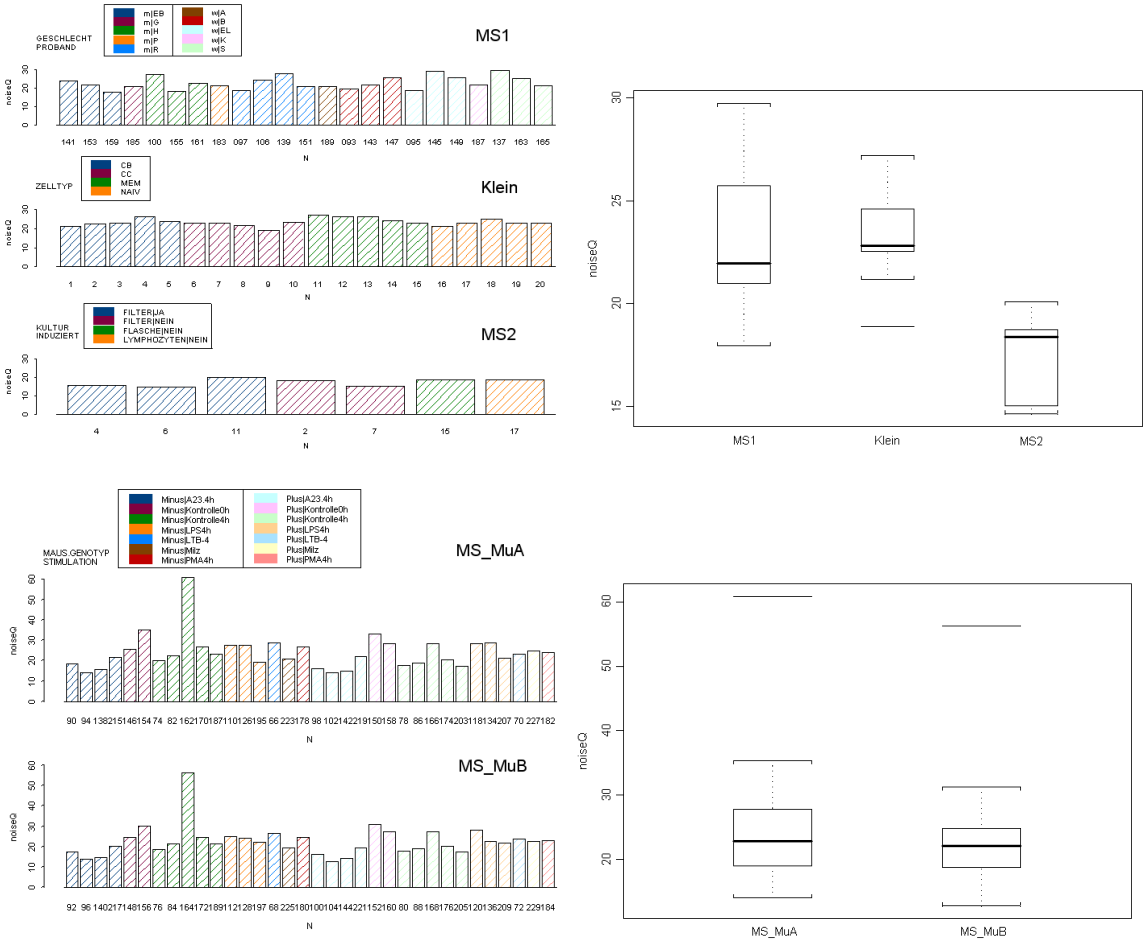


Abbildung 20: Qualitätskriterium *Noise(Q)*
(links: experimentweise, rechts: Zusammenfassung pro Datensatz)

Durchschnitt und Standardabweichung von *Noise(Q)* ergeben sich damit wie in Tabelle 15.

	Noise(Q)	
	Durch-schnitt	**Std.-Abw.**
MS1	22,96	3,48
Klein	23,38	2,06
MS2	17,27	2,16
MS_MuA	24,15	8,43
MS_MuB	22,76	7,50
MS_MuA (ohne Outlier)	23,03	5,46
MS_MuB (ohne Outlier)	21,75	4,69

Tabelle 15: Durchschnitt und Standardabweichung von *Noise(Q)*

Durchschnitt und Standardabweichung von *Noise(Q)* unterscheiden sich zwischen MS1 und Klein nur wenig. Zwischen MS2- und MS1-Datensatz allerdings sind die Unterschiede im Durchschnitt des *Noise(Q)* signifikant (p=0.0001), ebenso wie zwischen MS2- und Klein-Datensatz (p=0.0001). Die MS_Mu-Datensätze unterscheiden sich bezüglich *Noise(Q)* nur wenig, ihre Kenngrößen werden leicht geringer, wenn der *outlier* jeweils herausgenommen wird.

Der Vergleich von *noise* und *Noise(Q)* zeigt, dass der Durchschnitt des *Noise(Q)* für die Mu11K-Datensätze geringer, für die U95A-Datensätze größer ist als der Durchschnitt des *noise*. Eine Erklärung hierfür kann die kleinere *probe cell*-Größe von 20 µm des U95A-Arrays im Gegensatz zu 24 µm bei den Mu11K-Arrays sein. Bei gleicher Scanner-Auflösung ergeben sich so beim U95A weniger Pixel pro *probe cell*. Da bei der Berechnung des *Noise(Q)* die Anzahl der Pixel einfließt, ergibt sich für den U95A bei vorausgesetzter gleicher Standardabweichung der Pixelintensitäten ein größerer *Noise(Q)*-Wert. Eine andere Erklärung, die auch den angegebenen Nutzen des *Noise(Q)* von Affymetrix stärker berücksichtigt, wäre, dass das Legen des Gitters bei den U95A-Arrays größere Probleme verursacht als bei den anderen Array-Typen. Durch die geringere Anzahl von Pixeln bei gleicher Größe der Randbereiche zwischen den *probe cells* haben Ungenauigkeiten eines angepassten Gitters beim U95A eine größere Durchschlagskraft als bei den anderen Arrays.

Trotz der eben beschriebenen Beobachtungen sind *noise* und *Noise(Q)* gut bis sehr gut korreliert ($0{,}88 \leq r \leq 0{,}99$, siehe Abbildung 21). Das lässt darauf schließen, dass mit diesen Qualitätskriterien größtenteils die gleichen Varianzquellen erfasst werden.

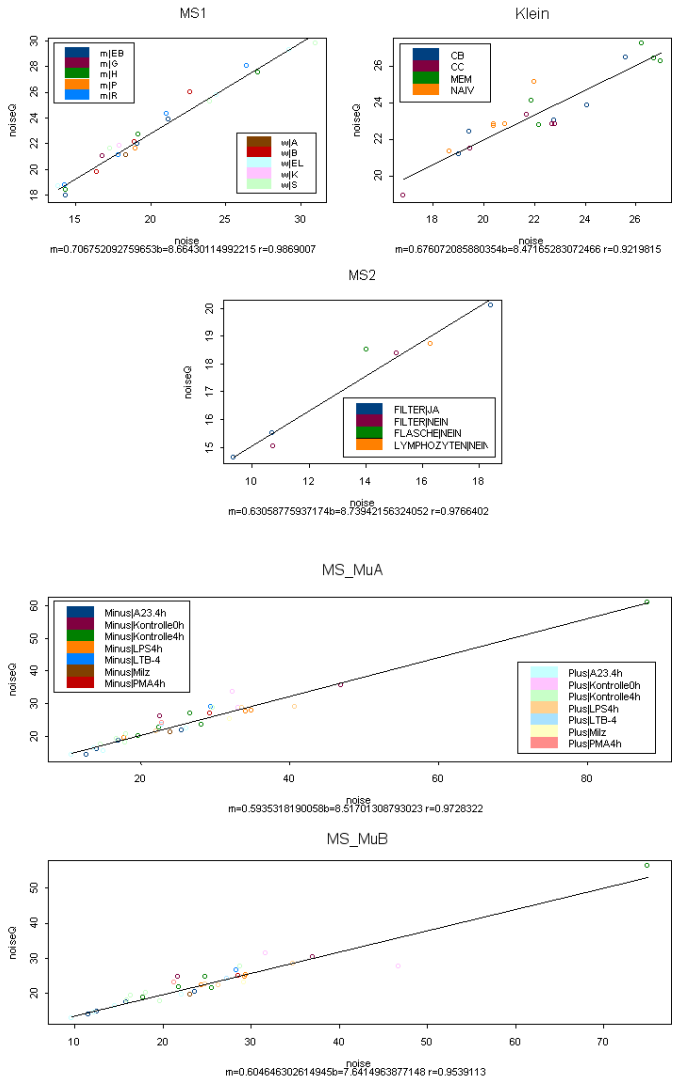


Abbildung 21: Korrelationen zwischen *noise* und *Noise(Q)*

4.2.2 *Probe sets* zur Qualitätskontrolle

Als Ergänzung zu den im vorherigen Abschnitt beschriebenen allgemeinen Qualitätskriterien sind auf jedem Array *probe sets* vorhanden, die eine Einschätzung spezieller RNA-Produkte liefern. Im Folgenden findet sich eine Beschreibung und Bewertung dieser *probe sets* gefolgt von Betrachtungen zu ihrem Verhalten in den betrachteten Datensätzen.

Dabei werden sowohl *Signal*-Werte, als auch 3‘/5‘-Quotienten berücksichtigt. Letztere sind unabhängig von der Affymetrix-Skalierung. Der *Signal*-Wert wird hier bei einem Skalierungsfaktor von eins betrachtet. Auf einigen der *probe sets* hätte eine Skalierung ohnehin keine sinnvolle Funktion (*Hybridization Controls*). Die Auswirkungen einer Skalierung auf andere *probe sets* (*Housekeeping Controls*) wird in Kapitel 5 untersucht.

Hybridization Controls

Die *probe sets* für diese vier RNAs (BioB, BioC, BioDN, Cre) dienen zur Kontrolle der Hybridisierung. Ihre vorgefertigten cRNAs (*Eukaryotic Hybridization Control Kit*) werden gegen Ende der Aufarbeitung zusammen mit der fragmentierten cRNA, dem Oligo B2 (für die Hybridisierung des Gitters) und anderen Komponenten zum Hybridisierungscocktail zusammengestellt.

Für alle *Hybridization Controls* finden sich mehrere *probe sets*, die aus unterschiedlichen Bereichen der RNA stammen. In der Regel liegt ein *probe set* im Bereich des 5’-Endes und eines im Bereich des 3’-Endes. Bei BioB liegt zusätzlich ein *probe set* in der Mitte der RNA (dem so genannten M’-Bereich). Bei den weiter unten beschriebenen *Housekeeping Controls* können die *probe sets* zur Kontrolle der Reversen Transkription und der *in-vitro*-Transkription dienen, hier jedoch ist es nicht möglich, diesbezüglich Folgerungen zu ziehen. Die *Hybridization Controls* dienen vielmehr dazu, Probleme bei der Hybridisierung zu identifizieren. Diese lassen sich beispielsweise in der Report-Datei eines Experimentes daran ablesen, dass der *Detection Call* eines dieser *probe sets Absent* lautet und der *Signal*-Wert relativ gering ist. Da BioB in der geringsten Konzentration vorliegt, kann ein *Absent* von BioB gegebenenfalls ignoriert werden[15].

Über diese vorgeschlagene Nutzung hinaus können semiquantitative Betrachtungen angestellt werden: Die cRNAs werden in unterschiedlichen Konzentrationen hinzugefügt, was sich in den *Signal*-Werten niederschlagen sollte. Wegen unterschiedlicher Hybridisierungsperformance der einzelnen *probe cells* und *probe sets* ist jedoch nicht zu erwarten, das Konzentrationsverhältnis der cRNAs (BioB:BioC:BioDN:Cre = 1,5:5:25:100) in gleichen Quantitäten im *Signal*-Verhältnis wiederzufinden. Man beachte außerdem, dass das Hinzufügen der cRNAs unabhängig von der bisherigen Aufarbeitung stattfindet. So mögen etwaige (auch große) Varianzquellen zwar einen Einfluss auf das

gesamte Sample der fragmentierten cRNAs haben, sie wirken sich jedoch nicht direkt auf die *Hybridization Controls* aus, sondern unterliegen eigenen Varianzquellen wie zum Beispiel Pipettierfehlern.

Es handelt sich bei den *Hybridization Controls* für eukaryotische Arrays um RNAs prokaryotischer Gene (siehe Tabelle 16).

Abkürzung	***probe set*-Bezeichner**	**Beschreibung**
BioB	AFFX-BioB-5_at, AFFX-BioB-M_at AFFX-BioB-3_at	E coli bioB gene biotin synthetase (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
BioC	AFFX-BioC-5_at AFFX-BioC-3_at	E coli bioC protein (-5 and -3 represent transcript regions 5 prime and 3 prime respectively)
BioDN	AFFX-BioDn-5_at AFFX-BioDn-3_at	E coli bioD gene dethiobiotin synthetase (-5 and -3 represent transcript regions 5 prime and 3 prime respectively)
Cre	AFFX-CreX-5_at AFFX-CreX-3_at	Bacteriophage P1 cre recombinase protein (-5 and -3 represent transcript regions 5 prime and 3 prime respectively)

Tabelle 16: Bezeichner und Beschreibungen der *Hybridization Controls*

Einige dieser *probe sets* sind auf den HG-U95A-Arrays zusätzlich statt mit der Endung „`_at`“ mit der Endung „`_st`“ vorhanden. Bei diesen wurden für die *probe cells* Oligonukleotide des *sense*-Strangs verwendet, daher werden sie auch *Negative Controls* genannt. An ihnen darf keine Hybridisierung stattfinden, sie müssen also alle *Absent* sein und ihre *Signal*-Werte müssen nahezu Null sein. Ist dem nicht so, ist es zu unspezifischer Hybridisierung oder anderen Effekten gekommen und das Experiment sollte für eine verlässliche Auswertung nicht verwendet werden.

Im Folgenden wird zunächst das Verhalten der *Hybridization Controls* der U95A-Datensätze (MS1, Klein, MS2) betrachtet. Für die MS_Mu-Datensätze gelten ähnliche Beobachtungen.

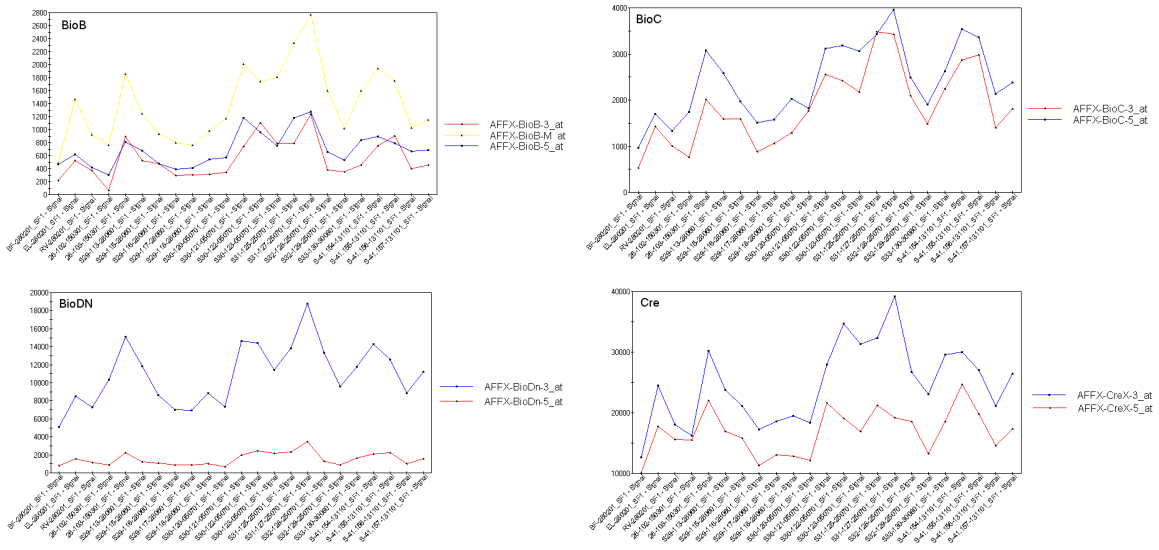


Abbildung 22: *Hybridization Controls*: 3'- (, M'-) und 5'-*Signal*-Profile (MS1)

Die *Signal*-Werte der einzelnen *probe sets* (siehe Abbildung 22) unterliegen starken Schwankungen, was aufgrund der gleichen Konzentration der jeweiligen cRNA im Hybridisierungscocktail nicht in diesem Maße erwartet wurde. Außerdem unterscheiden sich die Abstände der 3'- und 5'-*Signal*-Werte einer cRNA oftmals stark voneinander. Bei BioB überkreuzen sich die Profile der 3'- und 5'-Werte sogar. Dies könnte an der geringen Konzentration von BioB liegen, welche in der Nähe der Detektionsgrenze liegt[15]. Nichtsdestotrotz wird durch die Überkreuzungen klar, dass die *Hybridization Controls* nicht nur Varianzen durch Pipettierfehler unterliegen, sondern auch Qualitätsprobleme bei ihrer Herstellung oder Einflüsse des übrigen Hybridisierungscocktails eine wichtige Varianzquelle darstellen.

Das Verhalten der *Hybridization Controls* beim Klein-Datensatz ähnelt prinzipiell dem eben beschriebenen Verhalten des MS1-Datensatzes. Allerdings gab es drei Ausreißer nach oben beim BioB-*Signal*-Wert, und das Cre-Profil zeigt einige Überkreuzungen der 3'- und 5'-*Signal*-Werte und einen Ausreißer nach unten (siehe Abbildung 23).

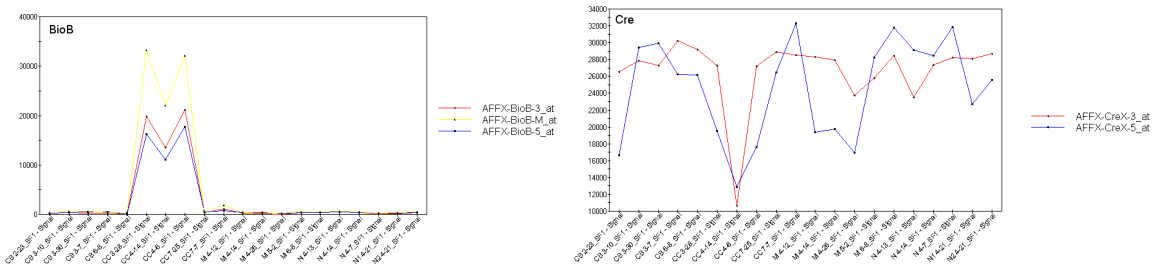


Abbildung 23: *Hybridization Controls*: Auffälligkeiten des Klein-Datensatzes

Da die Profile von BioC und BioDN keine derartigen Auffälligkeiten zeigen, kann die Ursache nicht in einem globalen Handhabungsfehler des *Hybridization Control*-Kits oder in Pipettierfehlern begründet liegen. Wiederum liegt die Vermutung nahe, dass Qualitätsprobleme bei der Herstellung der *Hybridization Controls* oder Einflüsse des Hybridisierungscocktails die Gründe für die Beobachtungen sind. Da es sich hierbei um einen externen Datensatz handelt, wurden bisher keine weiteren Nachforschungen nach einer konkreten Ursache angestellt.

Das Verhalten der anderen *Hybridization Controls* des Klein-Datensatzes und der des MS2-Datensatzes zeigt keine prinzipiellen Abweichungen zu den Betrachtungen beim MS1-Datensatz und wird daher hier nicht gezeigt.

In Abbildung 24 werden die *Signal*-Werte der einzelnen *Hybridization Controls* für die drei U95A-Datensätze jeweils nebeneinander dargestellt.

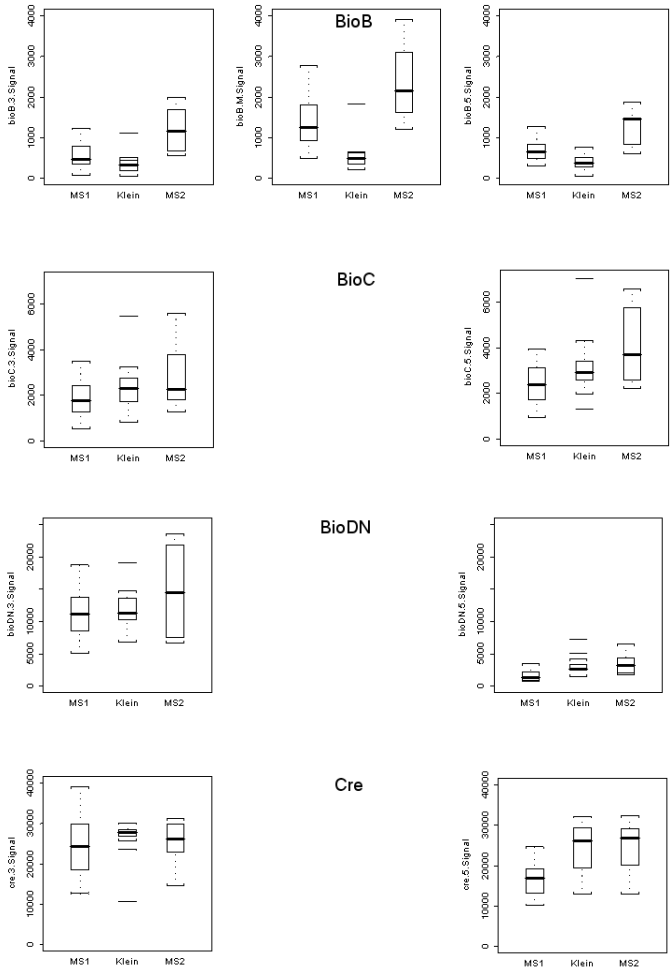


Abbildung 24: *Signal*-Bereiche der *Hybridization Controls* (U95A-Datensätze) (BioB ohne Outlier des Klein-Datensatzes)

Die Bereiche, in denen sich die *Signal*-Werte für die *Hybridization Controls* bewegen, können sich also – trotz gleicher Konzentrationen – zwischen den Datensätzen erheblich unterscheiden (siehe beispielsweise BioB, M' zwischen Klein und MS2). Außerdem können sich die 3'- und 5'-*Signal*-Bereiche auch innerhalb eines Datensatzes stark voneinander unterscheiden (z. B. MS1, BioDN).

Erwartungsgemäß findet sich die Konzentrationsreihe der *Hybridization Controls* beim MS1-Datensatz tendenziell in den *Signal*-Verläufen wieder. Die 3'-*Signal*-Werte bewegen sich für BioB ungefähr im Bereich von 50 bis 1100, für BioC im Bereich von 500 bis 3500, für BioDN im Bereich von 5000 bis 19000 und für Cre im Bereich von 12500 bis 40000 (siehe Abbildung 25).

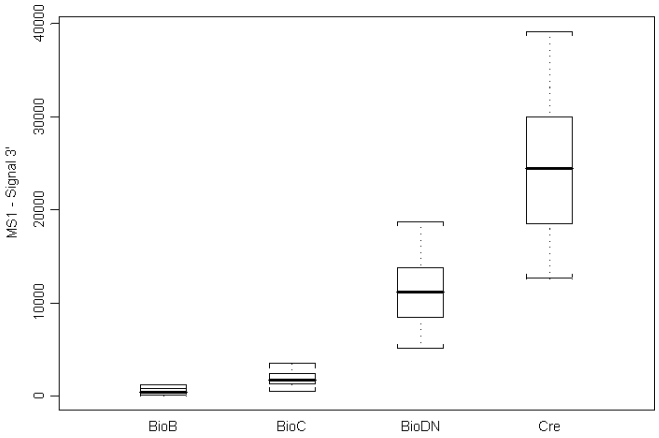


Abbildung 25: Konzentrationsreihe der *Hybridization Controls* (3'-*probe set*)

Die cRNA-Konzentration im Sample von BioB:BioC:BioDN:Cre = 1,5 : 5 : 25 : 100 wird also erwartungsgemäß nicht quantitativ in den *Signal*-Werten gefunden. Die Mediane der 3'-*Signal*-Werte von BioB:BioC:BioD:Cre verhalten sich wie 6,7 : 12,5 : 48,1 : 100.

Die Mediane der zu einem Durchschnittswert zusammengefassten 3'/M'/5'-*Signal*-Werte (Bezeichner in Report-Datei: „`Sig(all)`“) verhalten sich wie 8 : 13,8 : 33,3 : 100. Beim Klein-Datensatz verhalten sie sich wie 5,5 : 13,7 : 28,8 : 100 und beim MS2-Datensatz wie 9,8 : 15,4 : 38,5 : 100. Zwischen den U95A-Datensätzen werden also auch im Mittel über mehrere Experimente die Konzentrationsverhältnisse der ursprünglichen RNAs nur tendenziell und nicht quantitativ reproduziert.

Die *Hybridization Controls* der MS_Mu-Datensätze verhalten sich ebenfalls prinzipiell wie in den bisher beschriebenen Datensätzen und werden daher nicht im Einzelnen betrachtet. Die MS_Mu-Datensätze bieten aber zusätzlich die Möglichkeit, die Varianz zwischen technischen Replikaten zu quantifizieren. Das Sample eines Replikatpaares wurde in seiner Gesamtheit bis einschließlich zur Fragmentierung und Zugabe der *Hybridization Controls* in doppelter Menge hergestellt und erst vor dem Befüllen der entsprechenden Mu11KsubA- und Mu11KsubB-Chips getrennt. Die zwischen diesen Experimenten beobachtete Varianz in den *Signal*-Werten der *Hybridization Controls* kann nur in den Protokollschritten „Befüllen des Chips“, „Hybridisierung“, „Färben“ und „Scannen“ aufgetreten sein. Abbildung 26 visualisiert diese beobachtete Varianz mit einem Scatter Plot der zueinander gehörenden Replikatexperimente aus MS_MuA und MS_MuB zusammen mit der Regressionsgeraden und dem Korrelationskoeffizienten *r*.

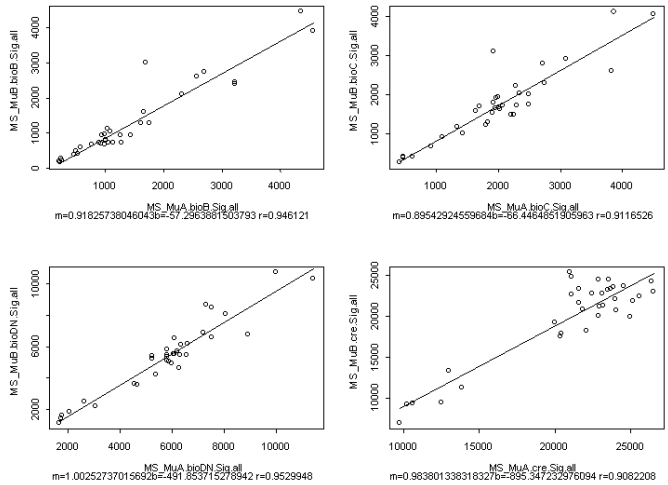


Abbildung 26: *Hybridization Controls*: Scatter Plots der Durchschnitts-*Signal*-Werte von Replikatpaaren (MS_Mu)

Die Durchschnitts-*Signal*-Werte der *Hybridization Controls* der Replikatpaare sind mit einem Korrelationskoeffizienten $r > 0.91$ gut korreliert, die beobachtete Varianz zwischen Replikatpaaren also relativ klein. Der Großteil der Varianz in den *Hybridization Controls* liegt also *nicht* in der eigentlichen Befüllung des Chips oder in den Schritten „Hybridisierung“, „Färben“ und „Scannen“ begründet. Als Ursache verbleibt ein Einfluss des Samples, die Herstellungsqualität der *Hybridization Controls* und das Pipettieren der *Hybridization Controls* in das Sample. Letzteres kann seine Ursache in Pipettierungenauigkeiten, aber auch in Messungenauigkeiten bei der Bestimmung der Konzentration haben.

Zusammenfassend enthält Abbildung 27 Box Plots der Durchschnitts-*Signal*-Werte aller Datensätze für alle vier *Hybridization Controls*.

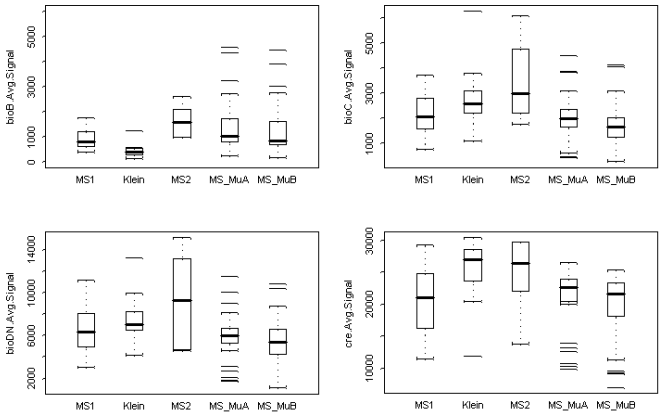


Abbildung 27: *Hybridization Controls*: Bereiche der Durchschnitts-*Signal*-Werte (alle Datensätze)

Zusammenfassend lässt sich beobachten, dass die *Hybridization Controls* einer wechselnden, unter Umständen großen Streuung unterliegen und sich die Mediane zwischen den Datensätzen stark voneinander unterscheiden können. Hier sind sogar einige t-Tests auf Unterschiede in den Durchschnitten signifikant bis hoch signifikant (hier nicht aufgeführt), selbst wenn ein nach Welch modifizierter t-Test angewendet wird, bei dem die Annahme nach gleicher Varianz wegfällt.

Ein Vergleich mit dem *Latin square*-Datensatz von Affymetrix war leider nicht möglich, da dort offensichtlich die Zugabe der *Hybridization Controls* nicht nach dem vorgestellten Protokoll stattfand. Zwar finden sich dort dieselben Beobachtungen bezüglich der Lage der 3'-, M'- und 5'-Profile zueinander, jedoch ist in Richtung BioB→BioC→BioDN→Cre keine aufsteigende Konzentrationsreihe vorhanden (siehe Abbildung 28). Diese Beobachtung bedarf weiterer Klärung.

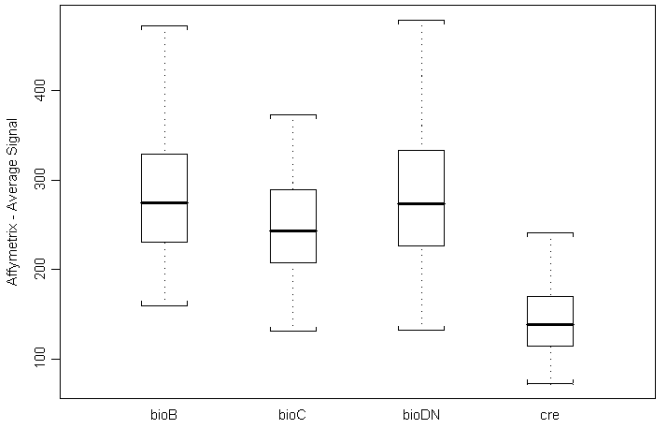


Abbildung 28: *Hybridization Controls* des Affymetrix-Datensatzes

PolyA Controls

Die *probe sets* für diese fünf RNAs (Übersicht siehe Tabelle 17) dienen laut *Affymetrix Expression Manual*[15] zur Überwachung der Aufarbeitung, der Hybridisierung und des Färbens. Es handelt sich um Bacillus subtilis-Gene, die auf Plasmiden in Bakterien lokalisiert sind, welche zunächst gezüchtet werden müssen. Mithilfe einer *in-vitro*-Transkription kann sense-RNA gewonnen werden, die mit einem Poly(A)-Ende versehen ist und die dann nach der RNA-Isolierung in festgelegten Konzentrationen zum Sample hinzugegeben wird. Eine Untersuchung von Chudin et al.[26] (Test2-Arrays) findet einen linearen Zusammenhang von Ausgangskonzentrationen und *Signal*-Werten für mittlere bis hohe Konzentrationen.

Da die RNAs aller fünf *PolyA Controls* die gesamte Aufarbeitung durchlaufen, besteht die Hoffnung, mit ihnen Varianzquellen auf die Spur zu kommen, die nach dem Isolierungsschritt liegen. Außerdem könnten sie bei einer Skalierung hilfreich sein (siehe Kapitel 5).

Abkürzung	***probe set*-Bezeichner**	**Beschreibung**
dap	AFFX-DapX-5_at AFFX-DapX-M_at AFFX-DapX-3_at	B subtilis dapB, jojF, jojG genes corresponding to nucleotides 1358-3197 of L38424 (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
thr	AFFX-ThrX-5_at AFFX-ThrX-M_at AFFX-ThrX-3_at	B subtilis thrC, thrB genes corresponding to nucleotides 248-2229 of X04603 (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
trpn	AFFX-TrpnX-5_at AFFX-TrpnX-M_at AFFX-TrpnX-3_at	B subtilis TrpE protein, TrpD protein, TrpC protein corresponding to nucleotides 1883-4400 of K01391 (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
phe	AFFX-PheX-5_at AFFX-PheX-M_at AFFX-PheX-3_at	B subtilis pheB, pheA genes corresponding to nucleotides 2017-3334 of M24537 (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
lys	AFFX-LysX-5_at AFFX-LysX-M_at AFFX-LysX-3_at	B subtilis lys gene for diaminopimelate decarboxylase corresponding to nucleotides 350-1345 of X17013 (-5, -M, -3 represent transcript regions 5 prime, Middle, and 3 prime respectively)

Tabelle 17: Bezeichner und Beschreibungen der *PolyA Controls*

Bei der Messung der Datensätze MS1, MS2 und MS_Mu wurden die *PolyA Controls* nicht genutzt, da ohne Erfahrungswerte Bedenken gegenüber dem Gewinnungsverfahren, der Konzentrationsbestimmung und der Pipettiergenauigkeit bestanden. Affymetrix hat die *PolyA Controls* in den Experimenten des *Latin square*-Datensatzes zur Bestimmung der Detektionsgenauigkeit verwendet wie im *Expression Manual*[15] beschrieben, d.h. in gleichen Endkonzentrationen. Dies ergab eine Nachfrage beim Affymetrix-Support[18].

Im Folgenden wird in Ermangelung direkt vergleichbarer Daten die Variabilität der *PolyA Controls* im Affymetrix-Datensatz mit dem *Hybridization Control* BioB aus Affymetrix-Datensatz und aus den hier untersuchten U95A-Datensätzen dargestellt.

Seit Herbst 2003 bietet Affymetrix ein „*PolyA RNA Control Kit*“ an[16], bei dem die *PolyA Controls* in unterschiedlichen Konzentrationen vorliegen. Mit diesem Kit besteht die Hoffnung, dass in zukünftigen Experimenten die *PolyA Controls* standardmäßig eingesetzt werden.

Abbildung 29 stellt die Bereiche der BioB-Durchschnitts-*Signal*-Werte von MS1, Klein, MS2 und Affymetrix neben denen der *PolyA Controls* des Affymetrix-Datensatzes dar.

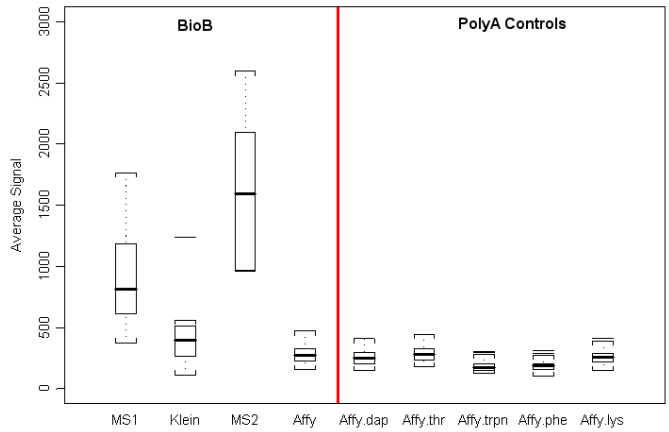


Abbildung 29: Vergleich der BioB-Durchschnitts-*Signal*-Werte mit den PolyA-Durchschnitts-*Signal*-Werten des Affymetrix-Datensatzes

Die Aufweitung der Verteilung ist bei Affymetrix konsistent geringer. Das mag zum einen an der generell niedrigeren Lage der Verteilung liegen oder am homogeneren Sample. Man beachte jedoch auch die Unstimmigkeit, dass BioB im Hybridisierungscocktail in einer Konzentration von 1.5 pM vorliegen soll, die *PolyA Controls* aber mit 10 pM. Dieser Unterschied findet sich in den Daten nicht wieder. Wie bereits im vorangegangenen Abschnitt erwähnt, bedürfen diese Auffälligkeiten weiterer Klärung.

Housekeeping Controls

Um Varianzquellen zu identifizieren, die noch vor dem Isolierungsschritt liegen, ist es erforderlich, Gene zu verwenden, deren RNA-Expression schon in der lebenden Zelle bzw. im lebenden Gewebe möglichst konstant ist. Traditionelle Kandidaten hierfür sind z. B. GAPDH (Glyceraldehyd-3-Phosphat-Dehydrogenase), ein Enzym aus dem Glucose-Stoffwechsel und ß-Actin, ein Bestandteil des Zytoskeletts. Daher befinden sich *probe sets* für die zwei entsprechenden Gene als *Housekeeping Controls* auf jedem Array (siehe Tabelle 18).

Abkürzung	*probe set*-Bezeichner	Beschreibung
GAPDH	**Mensch-Arrays:** AFFX-HUMGAPDH/M33197_5_at AFFX-HUMGAPDH/M33197_M_at AFFX-HUMGAPDH/M33197_3_at	Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
	Maus-Arrays: AFFX-GapdhMur/M32599_5_at AFFX-GapdhMur/M32599_M_at AFFX-GapdhMur/M32599_3_at	M32599 Mouse glyceraldehyde-3-phosphate dehydrogenase mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
β-Actin	**Mensch-Arrays:** AFFX-HSAC07/X00351_5_at AFFX-HSAC07/X00351_M_at AFFX-HSAC07/X00351_3_at	Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
	Maus-Arrays: AFFX-b-ActinMur/M12481_5_at AFFX-b-ActinMur/M12481_M_at AFFX-b-ActinMur/M12481_3_at	M12481 Mouse cytoplasmic beta-actin mRNA (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)

Tabelle 18: Bezeichner und Beschreibungen der *Housekeeping Controls*

Mittlerweile wurde für einige Zellen und Gewebe gezeigt, dass GAPDH und ß-Actin einer erheblichen Regulation unterliegen können. Anhand quantitativer RT-PCR wurde beispielsweise nachgewiesen, dass die RNA-Expression von GAPDH während der Zellkultur von differenzierenden T-Helferzellen signifikanten Änderungen unterliegt (Hamalainen et al.[41]), dass bezogen auf normale Mucosa die GAPDH-Transkription signifikant höher ist in Colon-Adenomen und -Krebs und die ß-Actin-Transkription signifikant höher ist in Colon-Krebs (Tsuji et al.[88]), und dass in serumstimulierten Fibroblasten eine neunfach erhöhte ß-Actin-Expression und dreifach erhöhte GAPDH-Expression vorliegt (Schmittgen and Zakrajsek[79]). Die Identifikation und Quantifizierung von technischer Varianz ist über die Expressionshöhe der *Housekeeping Controls* also nur unter Vorbehalt möglich.

Die Lage der 3'-, M'- und 5'-Profile entspricht für den Datensatz MS1 relativ gut den schon weiter oben dargelegten Erwartungen: Die 3'-Intensitäten liegen bei den meisten Experimenten über den M'-, und diese über den 5'-Intensitäten (siehe Abbildung 30).

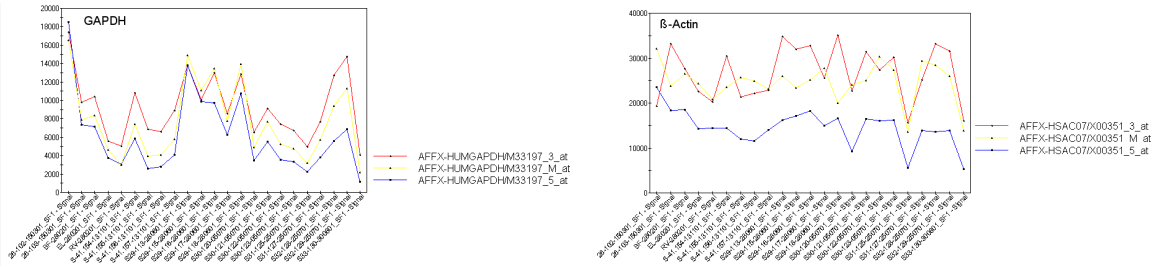


Abbildung 30: *Housekeeping Controls*: 3'-, M'- und 5'-*Signal*-Profile (MS1)

Die anderen Datensätze zeigen ein mehr oder weniger erwartungsgemäßes Bild, beim Klein-Datensatz ist auffällig, dass das 5'-Profil bei der Mehrheit der Experimente über dem 3'-Profil liegt (siehe Abbildung 31).

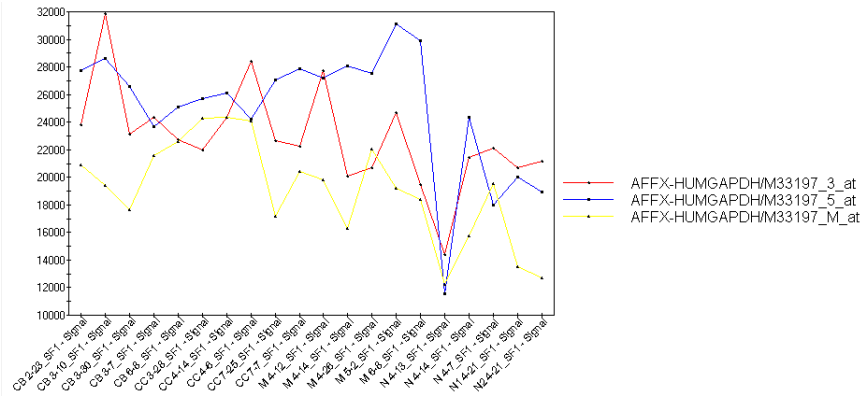


Abbildung 31: *Housekeeping Controls*: Auffälligkeiten des Klein-Datensatzes

Der Grund hierfür ist im Nachhinein nicht mehr feststellbar. Vor weiter gehenden statistischen Auswertungen sollte das nach unten ausreißende Experiment N4_13 idealerweise entfernt oder wiederholt werden.

Abbildung 32 fasst die Verteilungen der Durchschnitts-*Signal*-Werte der *Housekeeping Controls* zusammen. Man beachte, dass bei den MS_Mu-Datensätzen mit Maus-*Housekeeping Controls probe sets* gemessen werden, die nicht direkt vergleichbar mit den *probe sets* der menschlichen *Housekeeping Controls* sind.

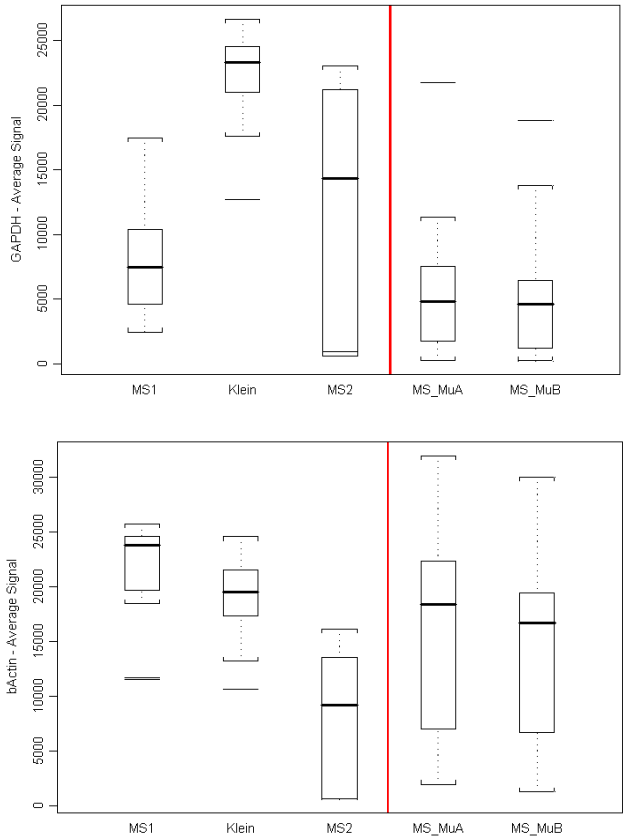


Abbildung 32: *Housekeeping Controls*: Bereiche der Durchschnitts-*Signal*-Werte für alle Datensätze

Wie bei den *Hybridization Controls* zeigt sich ein relativ großer Unterschied zwischen den Medianen und eine relativ große Varianz der *Signal*-Werte. Beides kann seine Ursache in den oben formulierten Vorbehalten bezüglich der Konstanz der *Housekeeping Controls* haben, also in biologischer Varianz. Wie bei den *Hybridization Controls* kommt jedoch auch technische Varianz als Grund in Frage. Darüber hinaus ist zu beachten, dass noch keine Form der Skalierung stattgefunden hat. Eine Diskussion der Varianz der *Housekeeping Controls* in Abhängigkeit der Skalierung findet sich in Kapitel 5.

Da die *Housekeeping Controls per definition* von Anfang an im Sample sind, können mit ihrer Hilfe größere Fehler in der Aufarbeitung identifiziert werden. Im Idealfall würden die beiden *probe sets* für das 3'- und das 5'-Ende eines *Housekeeping Controls* ein gleich großes Signal liefern und der 3'/5'-Quotient wäre 1. Im Realfall wird es jedoch wegen unterschiedlicher Hybridisierungsperformance und vor allem bei frühzeitigem Abbruch der RT oder der IVT vor Erreichen des 5'-Endes eine Erhöhung des 3'-*Signal*-

Wertes gegenüber dem 5'-*Signal*-Wert geben. Affymetrix empfiehlt einen 3'/5'-Quotienten von höchstens 3, jedoch können je nach Zell- oder Gewebetyp auch größere Quotienten noch verwendbare Experimente anzeigen. Affymetrix empfiehlt, die 3'/5'-Quotienten der eigenen Experimente zu quantifizieren und Experimente zu verwerfen, deren Quotienten eine starke Abweichung zeigen (siehe Affymetrix-Website[3]). Abbildung 33 visualisiert die 3'/5'-Quotienten der *Housekeeping Controls* für alle Datensätze.

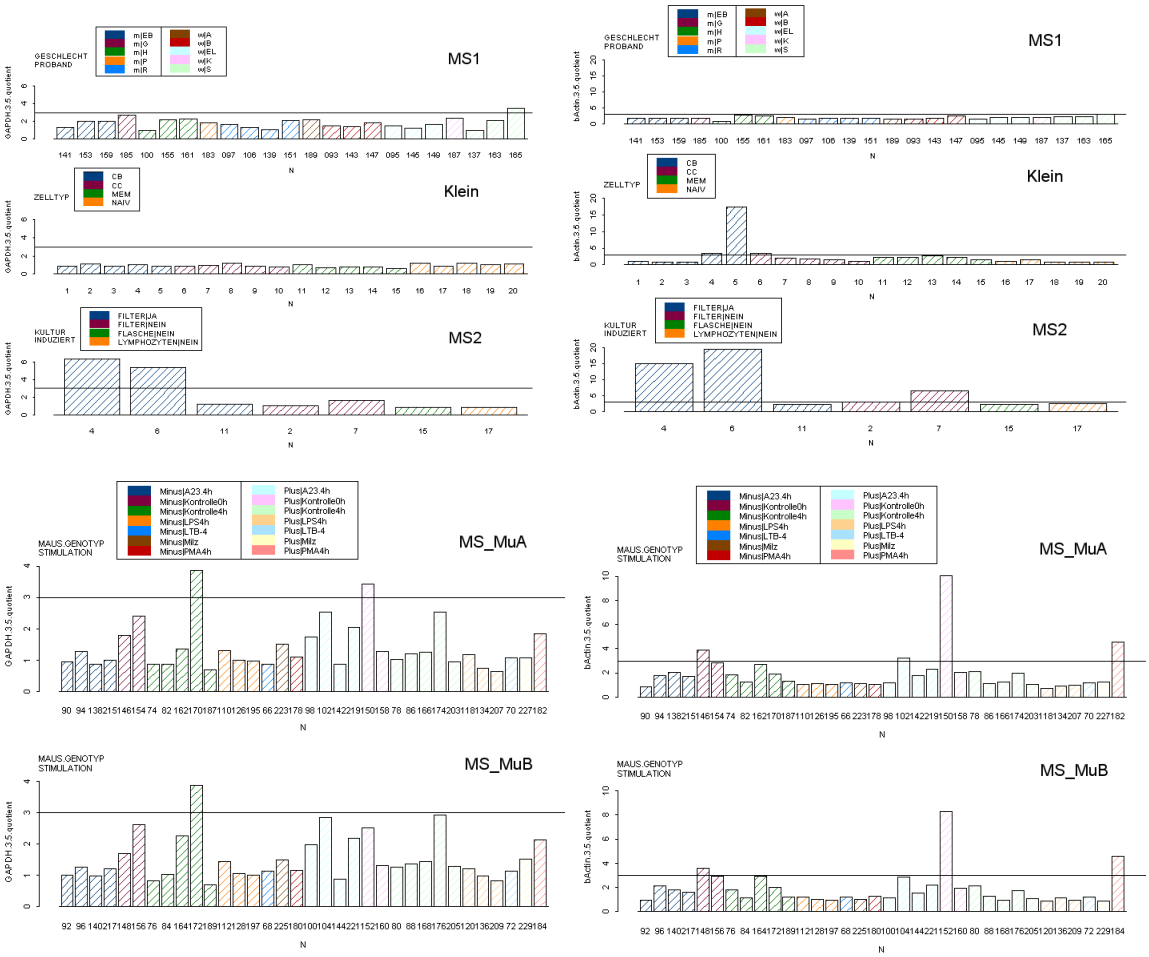


Abbildung 33: *Housekeeping Controls*: 3'/5'-Quotienten aller Datensätze (horizontale Linie: empfohlene Obergrenze 3)

Bei den meisten Experimenten liegt der Quotient innerhalb des empfohlenen Rahmens (horizontale Linie). Vereinzelte Ausreißer lassen sich bei Klein identifizieren. Der MS2-Datensatz zeigt mit den Experimenten 4 und 6 sowohl einen schlechten GAPDH-, als auch β-Actin-Quotienten. Hier ist eine Wiederholung der Experimente angezeigt, um abschätzen zu können, ob lediglich ein einmaliger Fehler in der Aufarbeitung auftrat oder

ob der Quotient durch Einflüsse aus dem Sample in dieser Gruppe systematisch höher liegt.

Die MS_Mu-Datensätze bieten die Möglichkeit zu weiter gehenden Betrachtungen. Die technischen Replikate in MS_MuA und der MS_MuB sind bezüglich der 3'/5'-Quotienten von GAPDH bzw. β-Actin sehr gut korreliert (r=0,94 bzw. 0,99, ohne Abbildung).

Die Betrachtung der Quotienten von GAPDH und β-Actin innerhalb eines Experimentes (siehe Abbildung 34) ergibt ein anderes Bild.

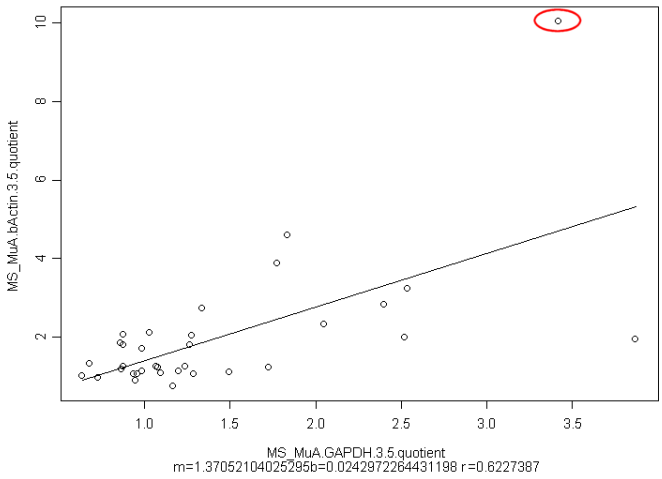


Abbildung 34: Korrelation zwischen GAPDH und β-Actin (MS_MuA-Datensatz)

Insgesamt liegt ein nur mittelmäßiger Korrelationskoeffizient vor (r=0,62). Es kommt in der Tat vor, dass der β-Actin-Quotient sehr viel höher ist als der GAPDH-Quotient (extremes Beispiel: 10,03 zu 3,42, rot markiert). Hieran lässt sich ablesen, dass Probleme bei der Aufarbeitung eines Gens nicht automatisch auf Probleme bei der Aufarbeitung eines anderen Gens schließen lassen.

Normalization Controls

Beim Array-Set HG-U133 hat Affymetrix 100 *probe sets* als ausgewiesene *Normalization Controls* hinzugefügt. Die RNA-Konzentration der entsprechenden Gene liegt in verschiedenen Geweben auf einem vergleichbaren Level (Affymetrix, Inc.[13]). Das Skalieren auf den Durchschnitt dieser *Normalization Controls* soll eine bessere

Vergleichbarkeit der Experimente ermöglichen als das globale Skalieren auf einen Ziel-*Signal*-Wert. Über die Betrachtung der Varianz dieser *Normalization Controls* könnten auch Qualitätsaussagen getroffen werden. Diese vielversprechende Möglichkeit wird hier aufgrund der Beschränkung auf U95A-Datensätze nicht betrachtet.

4.2.3 Anteil *Present Calls*

Im Report zu einer Primäranalyse wird der Anteil der *Present Calls* unter allen *probe sets* angegeben. Über den *Detection p-value* für ein *probe set* und den daraus abgeleiteten *Detection Call* kann eine Aussage zur Detektionsverlässlichkeit des Gens gemacht werden. Sind die Effekte lokaler Varianzquellen auf einem Chip größer oder liegt der Einfluss einer globalen Varianz vor, wird sich die Detektionsverlässlichkeit für viele Gene verändern. Insbesondere wenn die Intensitäten generell dunkler werden, also in Richtung Rauschgrenze verschoben sind, werden weniger Gene verlässlich detektiert. Über den Anteil der *Present Calls* ist also eine Aussage über die Qualität des Experimentes innerhalb eines Datensatzes möglich.

Man beachte, dass über den *Detection Call* allein keine Aussage zu vorhandener oder nicht vorhandener Expression („Ein- / Ausschalten“) eines Gens getroffen werden kann, da eben auch technische Varianz der Grund für eine Veränderung sein kann. Entgegen gängiger Praxis kann also bei Mehrfachexperimenten selbst unter Missachtung des *Detection Call* (also unter Hinzunahme auch der *Absent-probe sets*) eine verlässliche Einschätzung vom Aktivitätsunterschied der Genexpression zwischen verschiedenen Gruppen gewonnen werden.

Abbildung 35 gibt die Anteile der *Present Calls* für die verwendeten Datensätze wieder.

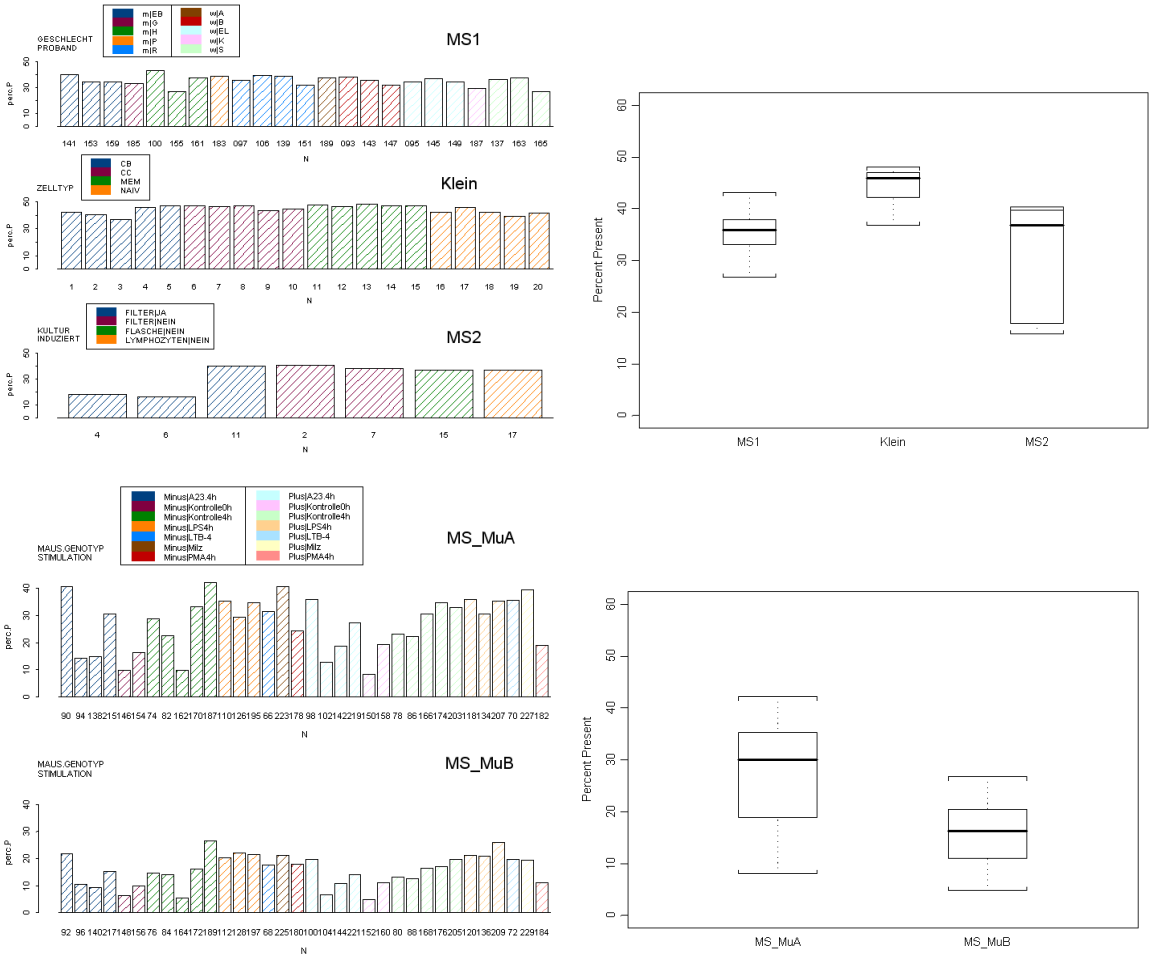


Abbildung 35: Anteil der *Present Calls* (*Percent Present*, *perc.P*) für alle Datensätze

Während die Schwankungen bei MS1 und Klein relativ gering sind, fallen sie bei MS2 und den MS_Mu-Datensätzen stark ins Gewicht. Beim MS2-Datensatz ist auffällig, dass gerade die Experimente einen kleineren Anteil an *Present Calls* zeigen, bei denen der 3'/5'-Quotient von GAPDH größer war. Beide Qualitätskriterien weisen also konsistent darauf hin, dass bei diesen Experimenten Qualitätsprobleme vorliegen.

Wiederum interessant ist das Verhalten dieses Qualitätskriteriums für die Replikate der MS_Mu-Datensätze (siehe Abbildung 36).

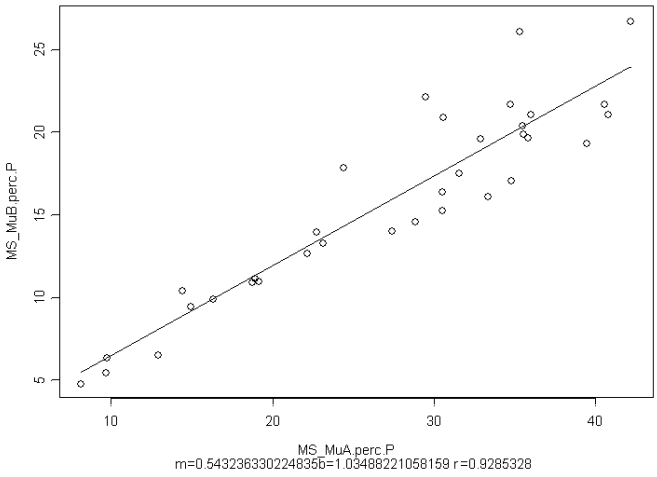


Abbildung 36: Korrelationen von *perc.P* zwischen Replikatpaaren in MS_MuA und MS_MuB

Der Anteil der *Present Calls* ist bei den B-Arrays nur ungefähr halb so groß (m=0.543) wie bei den A-Arrays, ist jedoch mit r=0,93 sehr gut korreliert. Der geringere Anteil kann entweder dadurch erklärt werden, dass die Mehrheit der gemessenen Gene in den durchgeführten Experimenten weniger stark exprimiert war oder wahrscheinlicher dadurch, dass der B-Array im Gegensatz zum A-Array mehr *probe sets* für weniger gut definierte Gene enthält und dadurch die Detektionsgenauigkeit eingeschränkt ist. Die sehr gute Korrelation ist ein deutlicher Hinweis darauf, dass ein globaler Effekt für den schwankenden Anteil der *Present Calls* verantwortlich ist, der nicht in der technischen Varianz von Aufarbeitungsschritten begründet ist, die nach „Befüllen des Chips" einschließlich liegen.

4.2.4 Zusammenfassung

Die betrachteten Qualitätskriterien werden in der Regel einzeln oder in Kombination dazu verwendet, problematische Experimente zu identifizieren. Wie im weiteren Verlauf mit diesen Experimenten zu verfahren ist, hängt beispielsweise ab von der Anzahl der durchgeführten Experimente innerhalb des Datensatzes, von der Art der Untersuchung (Vor- oder Hauptstudie), von der Möglichkeit zu weiterführenden Experimenten und von der Möglichkeit der Bestätigung der Ergebnisse mit anderen Methoden. Je nach Anspruch

können/müssen problematische Experimente aus dem Datensatz ausgeschlossen werden oder ihre Verwendung mit einem „cave“ markiert werden, sodass nachträgliche Bewertungen darauf Rücksicht nehmen. Ist bereits eine Vorstudie abgeschlossen, welche gezeigt hat, dass bei strengen Ausschlusskriterien wenige Experimente für die Auswertung verbleiben, dann würden in Nachfolgestudien einige der Bedingungen relaxiert werden. Je nach Summe der biologischen und technischen Varianz kann dies jedoch auch zur Aufweitung von Verteilungen führen und damit zu stärker verrauschten und weniger klaren Ergebnissen.

Über den reinen Aspekt der Qualitätsbeurteilung hinaus geben die Qualitätskriterien, wie in diesem Kapitel gezeigt, möglicherweise auch Hinweise auf Varianzquellen. Die Folgerungen können nie ganz eindeutig sein, da sich *per definition* alle Varianzquellen überlagern. Erst eine Validierung der Vermutung durch eine systematisch erfolgreiche Verringerung der Varianz könnte eine letztendliche Bestätigung der Hypothesen liefern.

Die Essenz der Beobachtungen dieses Unterkapitels ist einerseits die Tatsache, dass die beobachtete Varianz zum großen Teil multiplikativ ist und andererseits, dass die größte Quelle der beobachteten Varianz nicht in den Aufarbeitungsschritten „Befüllen des Chips“, „Hybridisierung“, „Färben“ und „Scannen“ liegen kann. Außerdem ist der Zusammenhang zwischen mRNA-Konzentration und *Signal*-Wert nicht *a priori* linear und Qualitätsprobleme eines *probe sets* lassen keine Rückschlüsse auf Qualitätsprobleme eines anderen *probe sets* zu.

Darüber hinaus resultierten die bisherigen Ergebnisse in der Idee zu einem neuem Qualitätskriterium (siehe folgendes Unterkapitel 4.3), welches Chip-globale Effekte als Ganzes zusammenfasst und mit den *probe cell*-Intensitäten Messgrößen eines Chips nutzt, die in der Aufarbeitung früher als *Signal*-Werte liegen und allgemeiner als 3’/5’-Quotienten sind. Über diesen Weg führen sie letztlich auch zu neuen Ideen bezüglich Skalierungsmöglichkeiten (siehe Kapitel 5).

4.3 Ein neues Qualitätskriterium: Gesamtintensität

4.3.1 Motivation

Die obigen Betrachtungen der Qualitätskriterien weisen auf eine relativ große technische Varianz hin, deren Ursprung nicht vollständig geklärt werden konnte. Vor allem die Untersuchungen bezüglich des Anteils der *Present Calls* legen nahe, dass es sich um einen globalen Effekt handelt. Schon bei der Betrachtung der CEL-Dateien zweier Experimente mit der MAS kann in der Tat eine globale Intensitätsschwankung ausgemacht werden (siehe Abbildung 37).

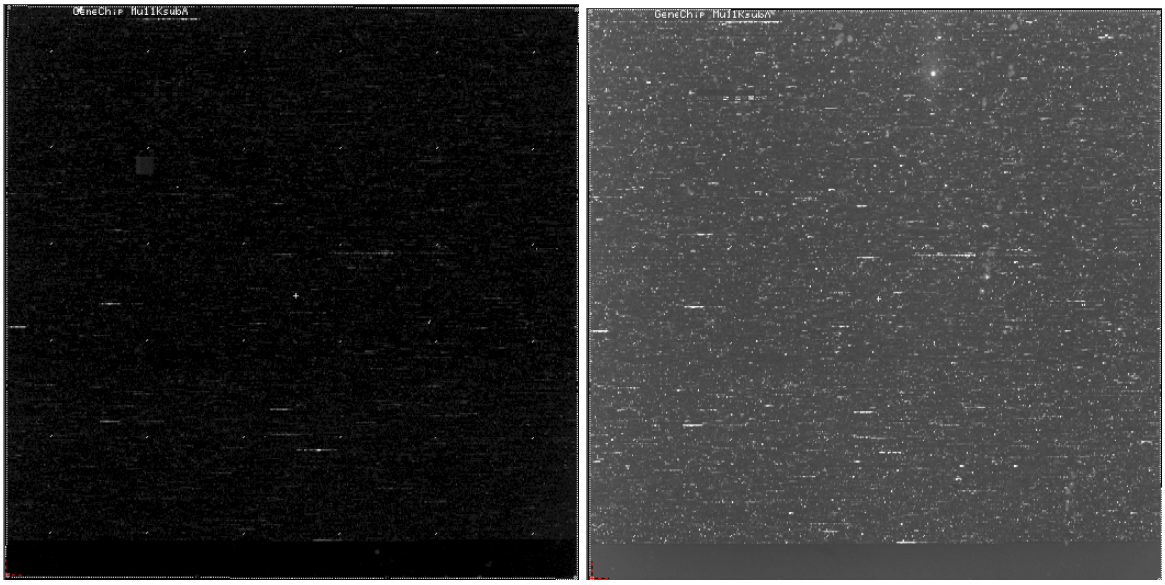


Abbildung 37: Intensitätsbilder (CEL-Datei) zweier MS_MuA-Experimente

Die Betrachtung der Intensitäten ist nicht zuletzt deshalb sinnvoll, weil sie zu den frühesten generierten Daten eines *probe sets* gehören. Da sich der globale Effekt, wie aus den bisherigen Beobachtungen geschlossen, vor dem Aufarbeitungsschritt „Befüllen des Chips“ auswirkt, ist klar, dass er nicht erst in den Maßzahlen, sondern schon in den Intensitäten zu finden ist.

Unter der Voraussetzung, dass der Zusammenhang zwischen der Menge an mRNA in der Probe und der Menge an gebundenem und detektiertem Farbstoff linear ist, bietet es sich an, zunächst die Summe der Intensitäten aller *probe sets* zu betrachten. Unter idealen Bedingungen würden perfekte Replikate dieselbe Gesamtintensität aufweisen. Eine Probe, die zwar eine andere mRNA-Zusammensetzung hat, aber bei der protokollgemäß dieselbe

Menge cRNA auf den Chip gegeben wird, müsste ebenfalls dieselbe Gesamtintensität aufweisen. Die eventuell vorhandenen Unterschiede im globalen Expressionsniveau würden erst durch eine Skalierung berücksichtigt werden. Damit läge also ein Qualitätskriterium vor, das prinzipiell die Möglichkeit bietet, globale technische Varianz zu erfassen.

4.3.2 Verhalten des Qualitätskriteriums „Gesamtintensität“

Für jedes Experiment lässt sich mit den in Unterkapitel 3.4 aufgeführten SPLUS-Funktionen die Gesamtintensität der *probe cells* aller *probe sets* berechnen. Die *probe cells* von ebenfalls auf jedem Array vorhandenen *Control Features* wie z. B. dem Rand oder leere *probe cells* werden dabei nicht berücksichtigt.

Die Gesamtintensitäten für die vorhandenen Datensätze verhalten sich wie in Abbildung 38 dargestellt.

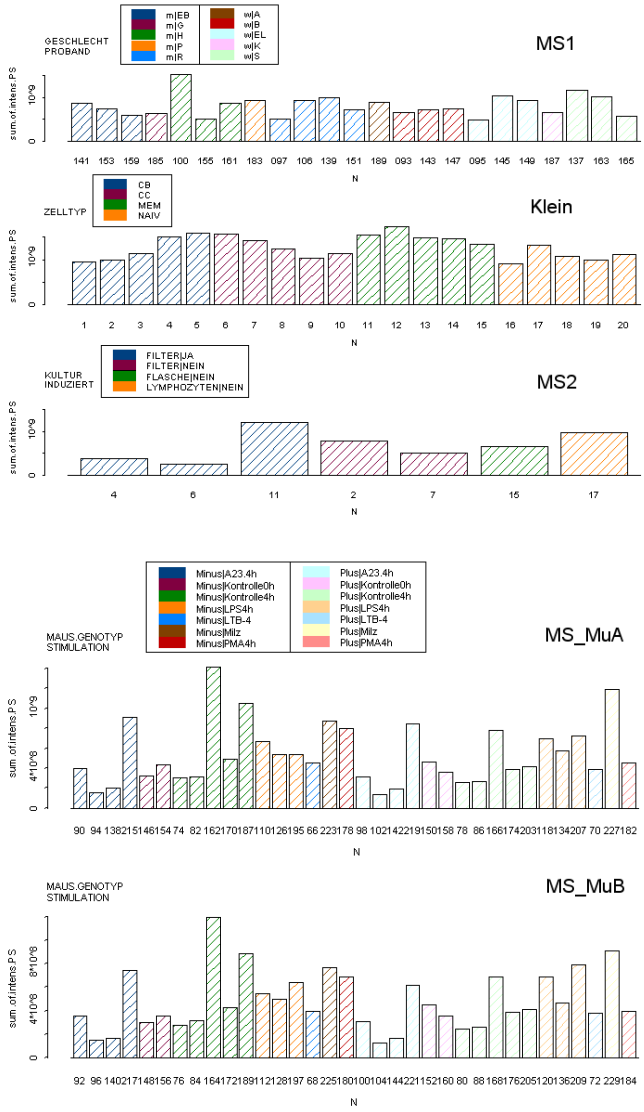


Abbildung 38: Gesamtintensitäten aller Datensätze

Abbildung 39 fasst die Verteilungen der Gesamtintensitäten innerhalb der Datensätze zusammen. Man beachte, dass sie zwischen den U95A- und MS_Mu-Datensätzen aufgrund der unterschiedlichen Array-Typen nicht vergleichbar sind, da unterschiedliche Array-Typen prinzipiell ein unterschiedliches Verhalten der Gesamtintensitäten erwarten lassen.

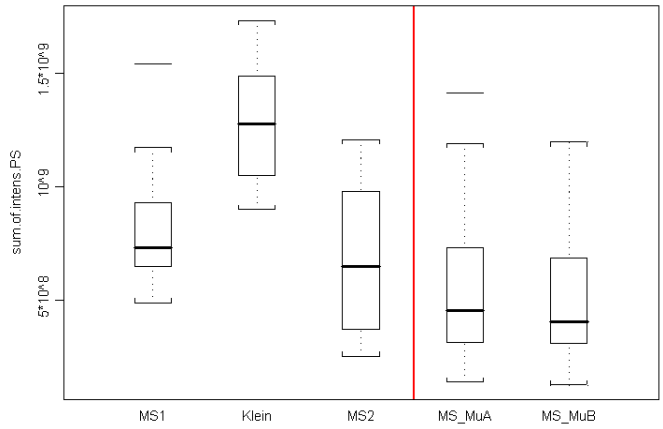


Abbildung 39: Verteilung der Gesamtintensität (alle Datensätze)

Die Gesamtintensitäten der Experimente unterliegen also wie die im vorangegangenen Unterkapitel betrachteten Qualitätskriterien großen Schwankungen. Dies zeichnete sich durch die Beobachtung einer Helligkeitsschwankung im Intensitätsbild und mit der Annahme einer globalen Varianz ja bereits ab, konnte aber aus den bisherigen Betrachtungen nicht direkt abgeleitet werden. Die Replikatexperimente der MS_Mu-Datensätze sind wie schon bei den anderen Qualitätskriterien mit r=0.98 sehr gut korreliert (siehe Abbildung 40), was sich mit den bisherigen Beobachtungen deckt.

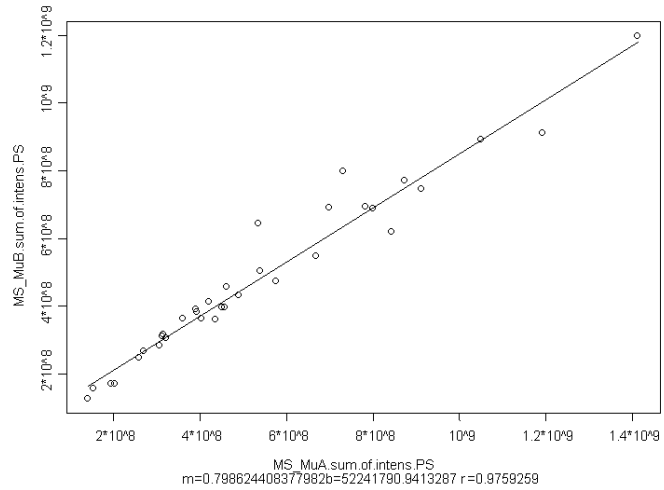


Abbildung 40: Korrelation der Gesamtintensität zwischen Replikatpaaren in MS_MuA und MS_MuB

Nicht nur die Gesamtintensitäten der Experimente unterscheiden sich, sondern auch die Verteilung der *probe cell*-Intensitäten variiert, was in Abbildung 41 exemplarisch für

den MS2-Datensatz dargestellt ist. Auch Histogramme der Intensitätsverteilungen getrennt nach *Perfect Match*- und *Mismatch-probe cells* weisen dieses Verhalten auf.

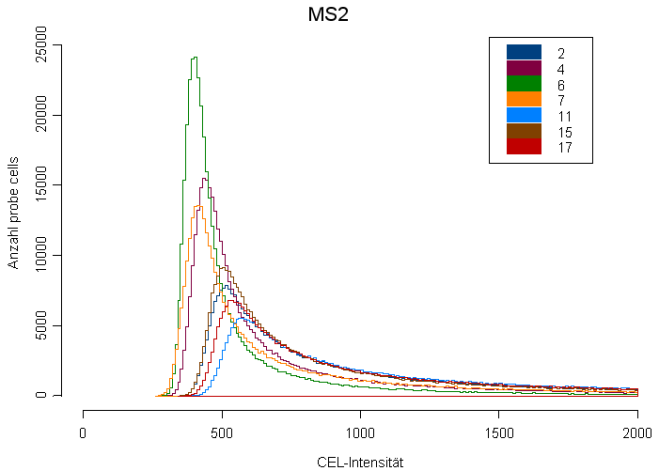


Abbildung 41: Histogramme der *probe cell*-Intensitäten (MS2-Datensatz)

Je größer die Gesamtintensität eines Experiments ist, desto mehr verschiebt sich der Peak in Richtung höherer Intensitäten und desto mehr weitet sich die Verteilung auf.

Die anderen Datensätze verhalten sich tendenziell ähnlich, die Histogramme werden jedoch aus Gründen der Übersichtlichkeit hier nicht aufgeführt. Die Box Plots der Abbildung 42 geben ebenfalls Aufschluss über die Verteilungen. Sie sind nach Gesamtintensität aufsteigend sortiert.

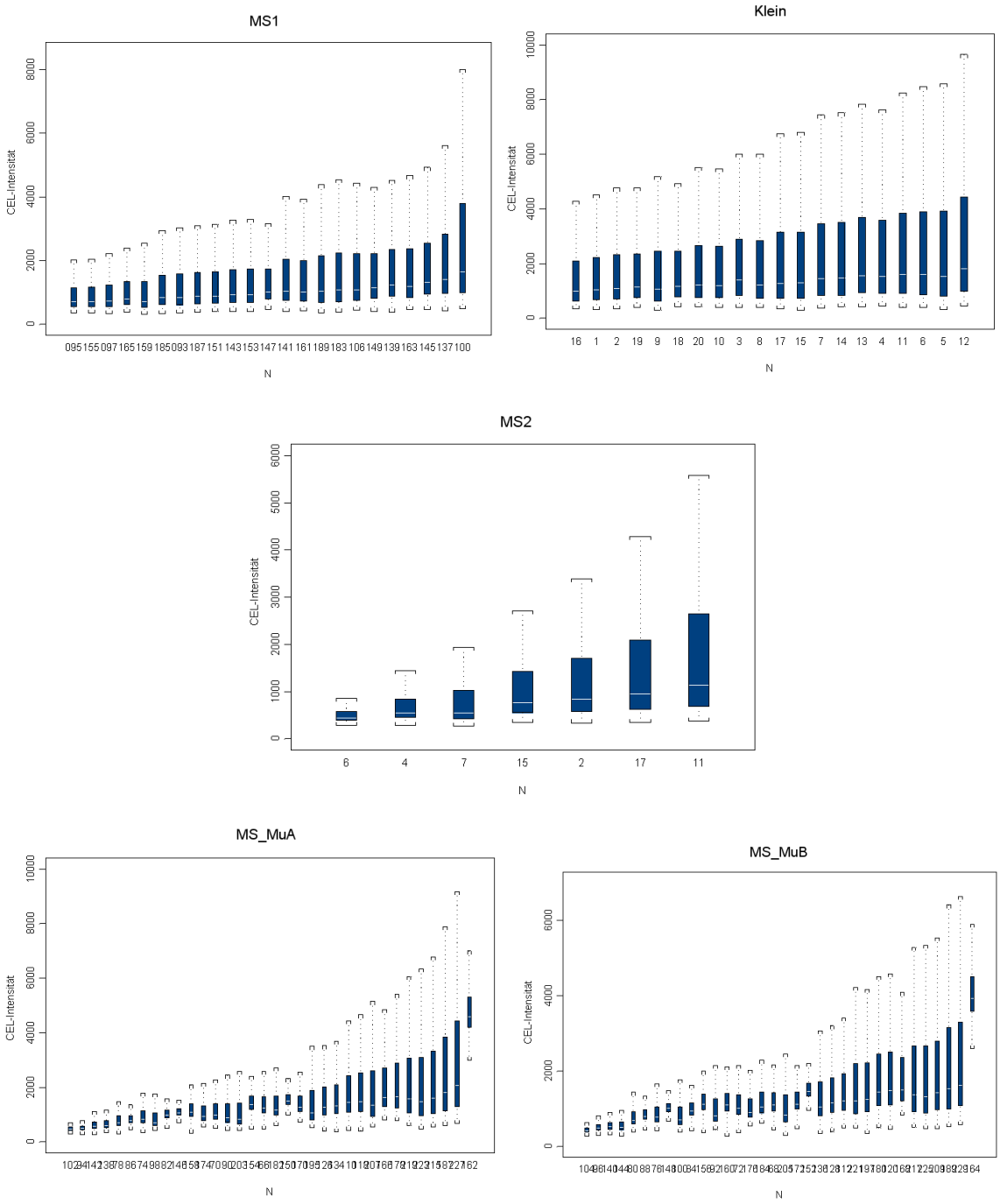


Abbildung 42: Verteilungen der *probe cell*-Intensitäten aller Datensätze (nach Gesamtintensität aufsteigend sortiert)

Mit mehr oder weniger großen Ausnahmen gilt also tendenziell bei allen Datensätzen: Je größer die Gesamtintensität, desto weiter nach oben verschoben ist der Median und eine desto größere Aufweitung liegt vor.

Diese Beobachtung führt zu der folgenden hypothetischen Erklärung der wichtigsten Varianzquelle: Ein Experiment, in dem in allen *probe cells* eine nahezu gleiche Intensität gemessen wird, zeigt eine Intensitätsverteilung mit einem sehr schmalen, hohen Peak. Dieser liegt zusätzlich eher im unteren Intensitätsbereich, wenn generell wenig Farbstoff pro *probe cell* bindet. Die obigen Beobachtungen können als Zwischenzustände zwischen dem beschriebenen fiktiven Experiment und einem Experiment angesehen werden, in welchem die *probe cells* sehr differenzierte Intensitäten im höheren Intensitätsbereich zeigen (z. B. MS2, Experiment 11).

Ein möglicher Grund für das Vorliegen eines schmalen hohen Peaks im unteren Intensitätsbereich ist ein größerer Anteil schlecht oder nicht markierter RNA-Fragmente im Hybridisierungscocktail. Im Gegensatz dazu verbleibt bei einem Experiment mit aufgeweiteter Intensitätsverteilung eine größere Menge des nach der Aufarbeitung eigentlich im Hybridisierungscocktail vorhandenen Farbstoffs auf dem Chip. Nach den weiter oben beschriebenen Folgerungen aus den technischen Replikaten des MS_Mu-Datensatzes kann die Ursache für den beobachteten Effekt allerdings nur vor dem Aufarbeitungsschritt „Befüllen des Chips“ liegen, also nicht im Färbe- und Wasch-Schritt. Eine konkrete Varianzquelle lässt sich ohne weitere Experimente nicht ausmachen, jedoch scheint ein wichtiger Grund die Variabilität in der Menge an gefärbter RNA zu sein, die dem Hybridisierungsschritt zugeführt wird. Diese Variabilität erwächst in der Regel aus Pipettierfehlern und Messungenauigkeiten bei der Quantifizierung der mRNA. Die Mengenbestimmung der mRNA weist auch deshalb Varianzen auf, weil eine experimentell nicht bestimmbare Menge an nicht-mRNA während der gesamten Aufarbeitung nicht aus dem System entfernt werden kann und lediglich durch einen Korrekturfaktor berücksichtigt wird, der von einem konstanten Verhältnis zwischen mRNA und nicht-mRNA im ursprünglichen RNA-Isolat ausgeht.

4.3.3 Korrelation mit den bisherigen Qualitätskriterien

Um das neue Qualitätskriterium Gesamtintensität mit den bisherigen Qualitätskriterien in Bezug zu setzen, kann ein Scatter Plot mit der dazugehörigen Regressionsgerade und dem Korrelationskoeffizienten *r* zu Hilfe genommen werden. Diese werden in Abbildung 43 exemplarisch für den relativ kleinen MS2-Datensatz und die Chip-globalen Qualitätskriterien *noise*, *background*, *Noise(Q)* und *percent Present*, sowie den Qualitätskriterien der *Housekeeping Controls* „GAPDH 3'/5'-Quotient", „GAPDH *Signal*", „β-Actin 3'/5'-Quotient" und „β-Actin *Signal*" aufgeführt. Für die anderen Datensätze werden diese Korrelationskoeffizienten in Tabelle 19 aufgelistet.

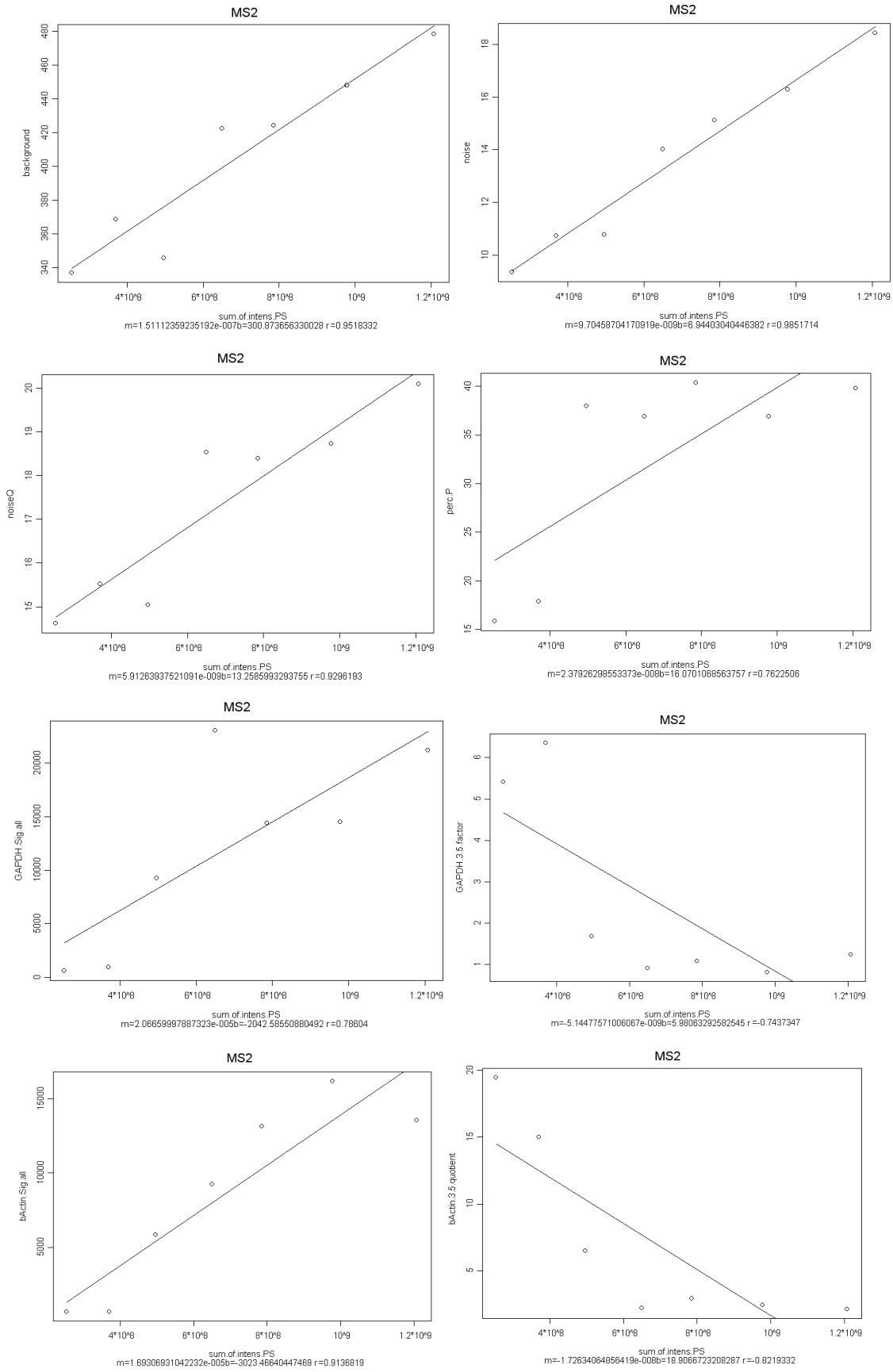


Abbildung 43: Korrelationen der bisherigen Qualitätskriterien mit der Gesamtintensität (MS2-Datensatz)

Datensatz	***noise***	***back-ground***	***Noise(Q)***	***percent Present***	**GAPDH 3'/5'**	**GAPDH *Signal***	**β-Actin 3'/5'**	**β-Actin *Signal***
MS1	0,85	0,84	0,81	0,67	-0,55	0,91	-0,30	0,60
Klein	0,83	0,55	0,66	0,76	-0,48	0,41	0,46	-0,84
MS2	0,99	0,95	0,93	0,76	-0,74	0,79	-0,82	0,91
MS_MuA	0,67	0,63	0,61	0,32	-0,14	0,56	-0,13	0,67
MS_MuB	0,67	0,64	0,62	0,37	-0,06	0,56	-0,15	0,65

Tabelle 19: Korrelationskoeffizienten *r* bisheriger Qualitätskriterien mit der Gesamtintensität

Werden zunächst die Chip-globalen Qualitätskriterien *noise*, *background*, *Noise(Q)* und *percent Present* betrachtet, so sind beim MS2-Datensatz gute bis sehr gute Korrelationen erkennbar. Der MS1-Datensatz zeigt eine vergleichbar gute Korrelation. Diese Tatsache weist darauf hin, dass Schwankungen in diesen Qualitätskriterien dieselbe Ursache haben könnten. Die Vermutung, dass es sich bei dieser Ursache um einen veränderlichen Anteil globaler unspezifischer Hybridisierung mit zusätzlich beeinflusster Farbstoffbindung handelt, erwuchs ja aus der beobachteten Schwankung im Anteil der *Present Calls* und aus der obigen Erklärung für die Veränderungen der Intensitätsverteilung. Unterstützt wird diese Vermutung von der guten positiven Korrelation der Gesamtintensität mit dem *background*, da bei geringerer Gesamtintensität zwar ein größerer Anteil unspezifischer Hybridisierung, aber mit geringeren Intensitäten vorliegen müsste, also auch ein geringerer *background*. Da mit der Verringerung der Gesamtintensität sowohl die Verteilung der Intensitäten über den ganzen Chip, als auch die Verteilung der Intensitäten innerhalb der *background-probe cells* einen höheren, schmaleren Peak erhält, steht auch die gute Korrelation mit *noise* – der Standardabweichung des *background* – der vermuteten Ursache nicht entgegen.

Beim Klein-Datensatz sind die Korrelationen generell schlechter. Dies könnte ein Hinweis auf zusätzliche Ursachen sein, die in anderen Labors ein größeres Gewicht haben als in Münster. Die MS_Mu-Datensätze sind ebenfalls generell schlechter korreliert, was am anderen Array-Typ liegen kann oder ein Hinweis darauf sein kann, dass doch nicht das Labor, sondern eher der untersuchte Zell- oder Gewebetyp stärkere Varianzen hervorruft.

Das Bild bei den Qualitätskriterien, die mithilfe der *Housekeeping Controls* gebildet werden, ist relativ uneinheitlich. Zwar ist für den MS1-Datensatz die Korrelation mit dem

Signal-Wert von GAPDH sehr gut, jedoch mit β-Actin schlecht bis mittelmäßig. Würde sich die Varianz der *Signal*-Werte von den als konstant angenommenen *Housekeeping Controls* einzig mit der Erhöhung der Gesamtintensität erklären, müsste auch β-Actin gut korreliert sein. Hier liegt also ein Hinweis auf andere Varianzquellen vor oder auf eine Regulation der β-Actin-RNA im Sample.

Die (außer bei MS2) relativ schlechten Korrelationen mit den 3'/5'-Quotienten entsprechen der Erwartung, dass über diese Qualitätskriterien Probleme aus gänzlich anderen Aufarbeitungsschritten (RT und IVT) überwacht werden.

Beim Klein-Datensatz fallen die Korrelationen der Gesamtintensität mit β-Actin gegenüber den anderen Beobachtungen generell aus dem Rahmen. Der Betrag von *r* zeigt zwar eine gute Korrelation an, jedoch ist diese entgegen der Erwartung negativ: Je höher die Gesamtintensität, desto geringer der β-Actin-*Signal*-Wert. Eine ähnlich unerwartete Korrelation kommt mit dem 3'/5'-Quotienten zustande. Eine Erklärung für diese Beobachtung kann mit den vorliegenden Daten nicht ohne weiteres gegeben werden.

4.3.4 Zusammenfassung

Unabhängig von obigen Überlegungen bezüglich der Varianzquelle bietet die Erfassung der Gesamtintensität die Möglichkeit, ein neues Qualitätskriterium zu formulieren. Dazu müsste zunächst aus einem ausreichend großen Datensatz ein „Normalbereich" ermittelt werden, innerhalb dessen sich die Gesamtintensität bewegen darf. Zusammengefasst ergeben die Gesamtintensitäten beispielsweise des MS1-Datensatzes ein Histogramm wie in Abbildung 44.

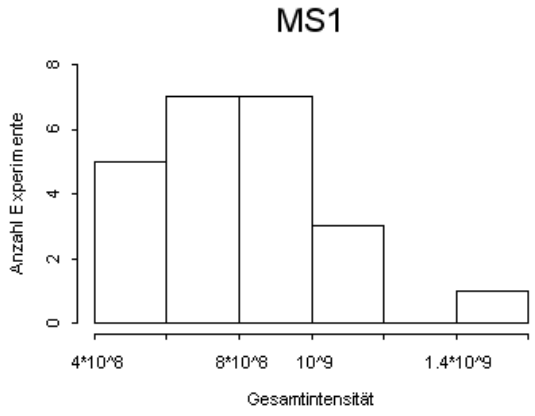


Abbildung 44: Histogramm der Gesamtintensitäten des MS1-Datensatzes

Hieran zeichnet sich ab, dass sich die Verteilung der Gesamtintensitäten bei genügend vielen Experimenten einer Normalverteilung annähern könnte. Eine Forderung an zukünftige Experimente wäre dann beispielsweise, dass ihre Gesamtintensität nicht mehr als drei Standardabweichungen vom Mittelwert entfernt liegt. Ob der „Normalbereich“ einmalig definiert wird oder sich dynamisch mit der Menge an verfügbaren Experimenten ändert, bliebe zu entscheiden, wenn klar ist, ob die Verteilung ab einer gewissen Anzahl von Experimenten stabil bleibt oder sich immer weiter aufweitet.

Die im vorhergehenden Abschnitt durchgeführten Betrachtungen zeigen deutlich, dass das Qualitätskriterium Gesamtintensität zu einem großen Anteil Varianz anzeigt, die auch schon anhand anderer Qualitätskriterien erkannt werden kann. Jedoch ist es durch seine Berücksichtigung des globalen Chip-Zustands (wie auch „Anteil der *Present Calls*“) und der gleichzeitigen Berücksichtigung der früher im Workflow liegenden Intensitäten (im Gegensatz zu den kondensierten Maßzahlen der *Housekeeping Controls*) näher an der Ursache dieser Varianz. Diese Tatsache macht die Nutzung der *probe cell*-Intensitäten und des Merkmals Gesamtintensität zusätzlich interessant für eigene Skalierungsansätze. Diese werden im folgenden Kapitel 5, Abschnitt 5.5.2 vorgestellt und diskutiert.

5 Skalierung von GeneChip-Experimenten

Mit dem Begriff Skalierung werden Verfahren bezeichnet, die die Auswirkungen der technischen Varianz vor der Ermittlung und Bewertung der biologischen Varianz möglichst weitgehend korrigieren. Im Idealfall realisiert also eine Skalierung eine Umkehrfunktion der technischen Varianz. Auf den Vorgang der Skalierung muss sehr viel Sorgfalt verwandt werden, da durch sie die Qualität einer Messmethodik entscheidend beeinflusst wird. Hocquette und Brandstetter[45] zeigen auf, dass selbst Skalierungsvorgänge etablierter Methoden einer eingehenderen Betrachtung bedürfen.

Bei der GeneChip-Technologie findet mit der Affymetrix-Software eine Skalierung ausschließlich auf der Ebene der *probe set*-Maßzahlen statt. Die *probe set*-Maßzahlen entstehen durch Verwendung eines Kondensierungsalgorithmus, dem ein Modell über den Zusammenhang zwischen *probe cell*-Intensitäten und mRNA-Konzentration für ein *probe set* zugrunde liegt. Dieses Modell wurde aus dem Verhalten der RNA-Hybridisierung einer relativ geringen Anzahl bekannter Gene unter der Annahme erstellt, dass sich alle anderen untersuchten mRNAs ebenso verhalten.

Um zunächst von diesen Annahmen zu abstrahieren, werden hier Skalierungsmethoden auf der Ebene der *probe cell*-Intensitäten vorgestellt. Dabei besteht auch die Hoffnung, systematische Varianz besser erkennen und herausrechnen zu können, die durch den Kondensierungsalgorithmus bereits bis zu einem gewissen Grad verwischt wird.

Unterkapitel 5.1 diskutiert zunächst verschiedene Bewertungsmöglichkeiten von Skalierungsmethoden. Unterkapitel 5.2 stellt die wichtigsten Aspekte des Skalierungsverfahrens von Affymetrix vor. Nach Betrachtung des Verhaltens der Affymetrix-Skalierung auf den vorhandenen Datensätzen in Unterkapitel 5.3 wird dann in Unterkapitel 5.4 der Versuch einer Bewertung unternommen. In Unterkapitel 5.5 werden die bisher entwickelten eigenen Skalierungsmethoden auf Intensitätsebene vorgestellt und diskutiert.

5.1 Bewertungsmöglichkeiten von Skalierungsmethoden

Nach den Betrachtungen in Kapitel 4 bezüglich der technischen Varianz ist offenkundig, dass eine Skalierung stattfinden muss, um eine Vergleichbarkeit zwischen Experimenten zu schaffen. Bei Hoffmann et al.[46] wird zusätzlich gezeigt, dass die Wahl

der Skalierungsmethode essenzielle Auswirkungen auf die Ergebnisse der weiter gehenden statistischen Auswertungen hat. Es lässt sich jedoch nicht *a priori* entscheiden, welche Skalierungsmethode vorzuziehen ist.

Forderungen an eine ideale Skalierung könnten beispielsweise wie folgt lauten:

1. Ein Gen, das nachgewiesenermaßen im untersuchten Material verschiedener Experimente in derselben mRNA-Konzentration vorliegt, sollte nach der Skalierung in beiden Experimenten denselben *Signal*-Wert erhalten.
2. Quantitative Verhältnisse zwischen den mRNA-Konzentrationen zweier unterschiedlicher Gene sollten sich in demselben Maße in den *Signal*-Werten wiederfinden. Sie sollten sich ebenso in den Verhältnissen anderer quantitativer Messmethoden wiederfinden.

Diese idealen Forderungen können prinzipiell nur dann erfüllt sein, wenn die erwähnten Messmethoden eine lineare Proportionalität zwischen mRNA-Konzentration und Messwert als Messfunktion verwirklichen.

Zur Bewertung einer Skalierungsmethode und zur Beurteilung, ob obige Forderungen erfüllt sind, können verschiedene Aspekte in Betracht gezogen werden. Im Folgenden werden bereits bekannte und neue Aspekte aufgeführt:

1. Je näher (im Ablauf des gesamten Prozesses) eine Skalierung an der Ursache der technischen Varianz liegt, desto besser der Skalierungseffekt. Überlagerungen von Nicht-Linearitäten und Effekte an den Grenzen von Definitions- und Wertebereich der algorithmischen Transformationen einzelner Schritte werden auf diese Weise vermieden.

2. Unter der Annahme, dass sich das untersuchte Material nicht oder nur wenig voneinander unterscheidet, ist eine Skalierung vorteilhafter, die weniger unterschiedliche Wertebereiche der *probe sets* zur Folge hat.

3. Die Expressionslevels so genannter *Housekeeping*-Gene (z. B. GAPDH) sollten in verschiedenen Experimenten einen gleichen oder zumindest wenig unterschiedlichen *Signal*-Wert erhalten. Wie in Unterkapitel 4.2 bereits diskutiert, wurden dabei Probleme bezüglich unterschiedlicher bzw. überkreuzender 3'-, M'- und 5'-Niveaus beobachtet sowie prinzipielle Probleme bezüglich der *de facto*-Regulation dieser vermeintlichen *Housekeeping*-Gene. Zusätzlich zu den erwähnten *Housekeeping*-Genen kann jedes andere Gen als solches angesehen werden, für welches nachgewiesen werden kann, dass es über die betrachteten Experimente ein stabiles Expressionslevel aufweist.

4. Einen prinzipiell ähnlichen Ansatz stellen die 100 *Normalization Controls* auf den U133-Chips dar, die beim Skalieren selbst eingesetzt werden, aber auch eine Qualitätsbeurteilung anderer Skalierungsmethoden ermöglichen können. Die RNA-Konzentration der entsprechenden Gene liegt laut Affymetrix in unterschiedlichen Geweben auf einem vergleichbar hohen Level (Affymetrix, Inc.[13]).

5. Quantitative RT-PCR, *Serial Analysis of Gene Expression* (SAGE; Vorstellung der Methode: Velculescu et al.[91]; Vergleich GeneChips und SAGE: Evans et al.[34]) oder Protein-Quantifizierungsmethoden (beispielsweise Protein-Microarrays, siehe Talapatra et al.[86]) können herangezogen werden, um zu prüfen, ob sich Verhältnisse in den *Signal*-Werten in den Verhältnissen der Messwerte dieser Methoden wiederfinden.

6. Die bereits in Abschnitt 4.2.2 vorgestellten *PolyA Controls* können zur Skalierung selbst verwendet werden, aber auch zur Bewertung anderer Skalierungsmethoden. Zusätzlich zu den von Affymetrix vorgesehenen *PolyA Controls* können eigene *PolyA Controls* eingesetzt werden. Auf diesen erstmals in dieser Arbeit vorgestellten Vorschlag kann zurückgegriffen werden, wenn nach der Durchführung mehrerer erster Experimente weitere Chips mit gleichem oder ähnlichem experimentellen Design gescannt werden sollen. Dazu werden in den ersten Experimenten *probe sets* / *probe cells* identifiziert, die stets die niedrigsten

Signal-Werte / Intensitäten aufweisen. Es kann vorkommen, dass keine solchen identifiziert werden können. Im Erfolgsfall jedoch können entsprechende RNA-Moleküle hergestellt werden. Diese werden wie die *PolyA Controls* in bekannter Konzentration vor der Reversen Transkription zum Sample jedes weiteren Experimentes hinzugegeben.

Bei den in dieser Arbeit verwendeten Datensätzen können die Aspekte 1 und 2 problemlos zur Bewertung einer Skalierung herangezogen werden. Die aussagekräftigeren Aspekte 3-6 stehen nicht oder nur bedingt zur Verfügung.

Für Aspekt 3 fehlt der Nachweis, dass sich GAPDH und β-Actin in den Proben stabil verhalten. Sollten also unterschiedliche *Signal*-Werte auftreten, kann nicht entschieden werden, ob diese auf Skalierungsartefakte zurückzuführen sind oder auf eine Regulation der entsprechenden RNAs im Ausgangsmaterial. Nichtsdestotrotz können Betrachtungen dieses dritten Aspektes einen tendenziellen Hinweis auf die Qualität einer Skalierungsmethode geben.

Aspekt 4 ist nur bei U133- und neueren Arrays anwendbar. Daten anderer Quantifizierungsmethoden zur Anwendung des Aspekts 5 standen für diese Arbeit nicht zur Verfügung. Wie bereits in Unterkapitel 4.2 erwähnt, wurden weder vorgefertigte noch eigene *PolyA Controls* in der Probenaufarbeitung der verwendeten Datensätze verwendet. Daher kann auch Aspekt 6 nicht zur Bewertung herangezogen werden.

5.2 Die Affymetrix-Skalierung

Die Affymetrix-Skalierung wird als letzter Schritt des Kondensierungsalgorithmus durchgeführt. Eine Skizze des Kondensierungsalgorithmus findet sich in Abschnitt 2.2.1. Analog zu „Skalierung“ wird in der Affymetrix-Terminologie der Begriff „Normalisierung“ verwendet. Hierbei werden dieselben Berechnungen durchgeführt wie bei der Skalierung, allerdings wird der Ziel-*Signal*-Wert nicht vorgegeben, sondern aus einem Baseline-Experiment als randbereinigter Durchschnitt aller *Signal*-Werte errechnet.

Sei $SignalLogValue_i$ der robuste Durchschnitt der logarithmierten *Perfect Match* / *Ideal Mismatch*-Differenzen aller *probe pairs* für jedes *probe set i*. Sei *TGT* der Ziel-*Signal*-Wert. Vor einer Umstellung des dynamischen Messbereiches des GeneChip-

Scanners lag der empfohlene Wert hierfür bei 1000, nach der Umstellung bei 100. Zunächst wird ein randbereinigter Durchschnitt (*trimmed mean*) der delogarithmierten $SignalLogValue_i$ für alle *probe sets i* berechnet. Dabei werden die oberen und unteren zwei Prozent der vorliegenden Werte nicht berücksichtigt, was den Einfluss von Ausreißern verhindern soll. Der Skalierungsfaktor *SF* ergibt sich dann als Quotient aus Ziel-*Signal*-Wert und randbereinigtem Durchschnitt:

$$SF = \frac{TGT}{trimMean(2^{SignalLogValue_i}, 0.02, 0.98)}$$

Der skalierte *Signal*-Wert ergibt sich für jedes *probe set i* als Produkt aus delogarithmiertem $SignalLogValue_i$ und *SF*:

$$Signal_{probe\,set\,i} = SF \cdot 2^{SignalLogValue_i}$$

Zusammenfassend werden also die Wertebereiche (*Signal*-Bereiche) der unskalierten Experimente durch einen konstanten Faktor verschoben und so die randbereinigten Durchschnitte aufeinander gelegt. Neben der wunschgemäßen Verschiebung der Wertebereiche findet dabei jedoch auch eine Dehnung oder Streckung statt. Eine ausführliche Betrachtung der *Signal*-Verteilungen vor und nach der Skalierung für die betrachteten Datensätze findet sich in Unterkapitel 5.3. Um Dehnung und Streckung so weit wie möglich einzugrenzen, empfiehlt Affymetrix, nur Experimente miteinander zu vergleichen, deren Skalierungsfaktoren sich nicht um mehr als einen Faktor drei voneinander unterscheiden (siehe Anhang des *Expression Manuals*[15]).

Aus dieser Argumentation entspringt zwangsläufig der Verbesserungsvorschlag, keinen fest vorgegebenen Ziel-*Signal*-Wert (wie 1000) für die Skalierung zugrunde zu legen, sondern den Durchschnitt der randbereinigten Durchschnitte $trimMean_j$ aller an einer Auswertung beteiligten Experimente $j=1,...,k$ zu verwenden:

$$TGT = \frac{1}{k}\sum_{j=1}^{k} trimMean_j$$

Damit werden nicht nur die Skalierungsfaktoren, sondern etwaige Verschiebungen, Dehnungen und Streckungen der Wertebereiche möglichst gering gehalten. Diese Bestimmung des Ziel-*Signal*-Wertes berücksichtigt im Gegensatz zu dem von Affymetrix empfohlenen festen Wert die spezifische Lage der Wertebereiche in den durchgeführten Experimenten eines Datensatzes.

Werden die Skalierungsfaktoren für einzelne *probe sets* aus einem nicht-skalierten Experiment (*SF*=1) und einem skalierten Experiment berechnet, so weichen sie in der Praxis aufgrund von Rundungsungenauigkeiten, die offenbar gerade während des Logarithmierens und Delogarithmierens entstehen, leicht voneinander ab. In einem Beispiel variierten die Skalierungsfaktoren aller *probe sets* zwischen 5.07 und 5.28 (Verteilung siehe Abbildung 45), während der theoretische Skalierungsfaktor 5.15 betrug.

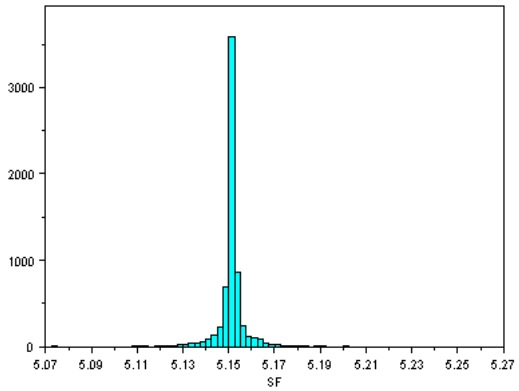


Abbildung 45: Histogramm der Skalierungsfaktoren aller *probe sets* zwischen unskaliertem und skaliertem Experiment; theoretischer Skalierungsfaktor *SF* = 5.15

Dies entspricht einem relativen Fehler von maximal 2.52 %. Die Rundungsungenauigkeiten bei der Skalierung tragen also bereits zur technischen Varianz bei. Je größer der theoretische Skalierungsfaktor, desto weiter ist das Histogramm der Skalierungsfaktoren aller *probe sets* und desto größer ist dementsprechend der maximale relative Fehler. Auch diese Tatsache ist ein Argument für obigen Vorschlag, die Skalierungsfaktoren aller an einer Auswertung beteiligten Experimente möglichst nahe bei eins zu halten.

5.3 Verhalten der Affymetrix-Skalierung

In diesem Unterkapitel wird das Verhalten der Affymetrix-Skalierung bei Anwendung auf die vorhandenen Datensätze beschrieben. Dazu werden jeweils die *Signal*-Verteilungen und die *Signal*-Profile der unskalierten und der auf einen Ziel-*Signal*-Wert von 1000 skalierten Experimente („TGT1000“) gegenübergestellt.

5.3.1 *Signal*-Verteilungen

Im Folgenden findet sich zunächst eine Gegenüberstellung der Wertebereiche der unskalierten und der skalierten Experimente für alle Datensätze. Dazu werden zunächst Box Plots verwendet, bei denen die Outlier nicht eingezeichnet werden (mittlere *Signal*-Bereiche; siehe Abbildung 46).

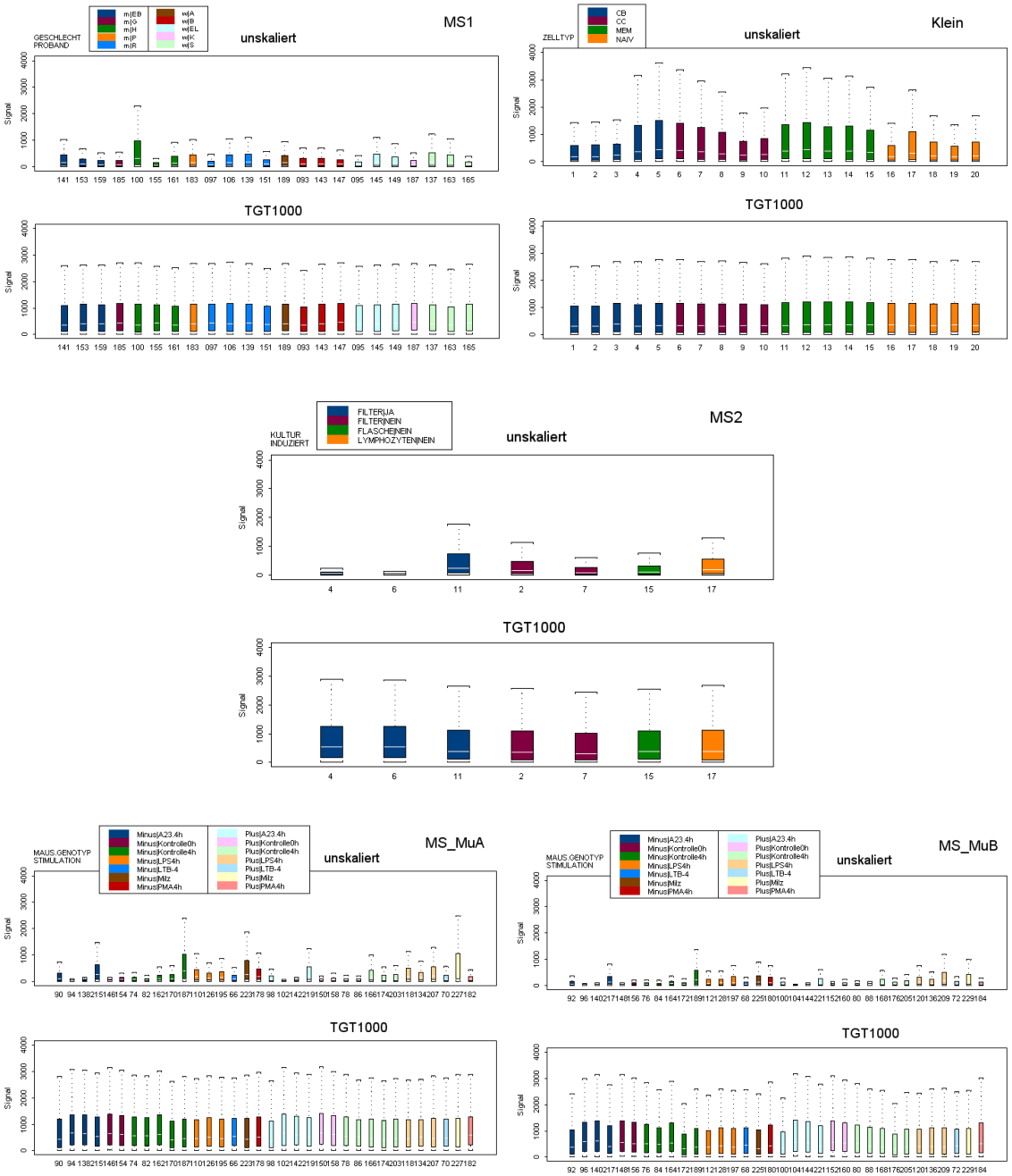


Abbildung 46: Mittlere *Signal*-Bereiche (Box Plots ohne Outlier)
(jeweils oben: unskaliert, unten: TGT1000)

Während sich die in den Box Plots gezeigten *Signal*-Bereiche der unskalierten Experimente in Lage und Streuung relativ stark voneinander unterscheiden, liegen sie bei den auf 1000 skalierten Experimenten näher beieinander. Bei einigen Datensätzen funktioniert die Angleichung der *Signal*-Bereiche nahezu perfekt (z. B. Klein), bei anderen sind immer noch klare Unterschiede festzustellen. Dabei kommt es einerseits vor, dass Experimente, die sich im unskalierten Zustand deutlich von den anderen unterscheiden, im skalierten Zustand keine Auffälligkeiten zeigen (z. B. Experiment 100 des MS1-Datensatzes). Andererseits weisen einige Experimente, die unskaliert einen relativ geringen *Signal*-Bereich zeigen, skaliert einen verhältnismäßig hohen *Signal*-Bereich auf (z. B. 184 des MS_MuB-Datensatzes).

Wird nun zusätzlich die Lage der Outlier betrachtet, die oberhalb und unterhalb der in obigem Box Plot gezeigten Bereiche liegen (vollständige *Signal*-Bereiche), ergibt sich ein weniger homogenes Bild (siehe Abbildung 47).

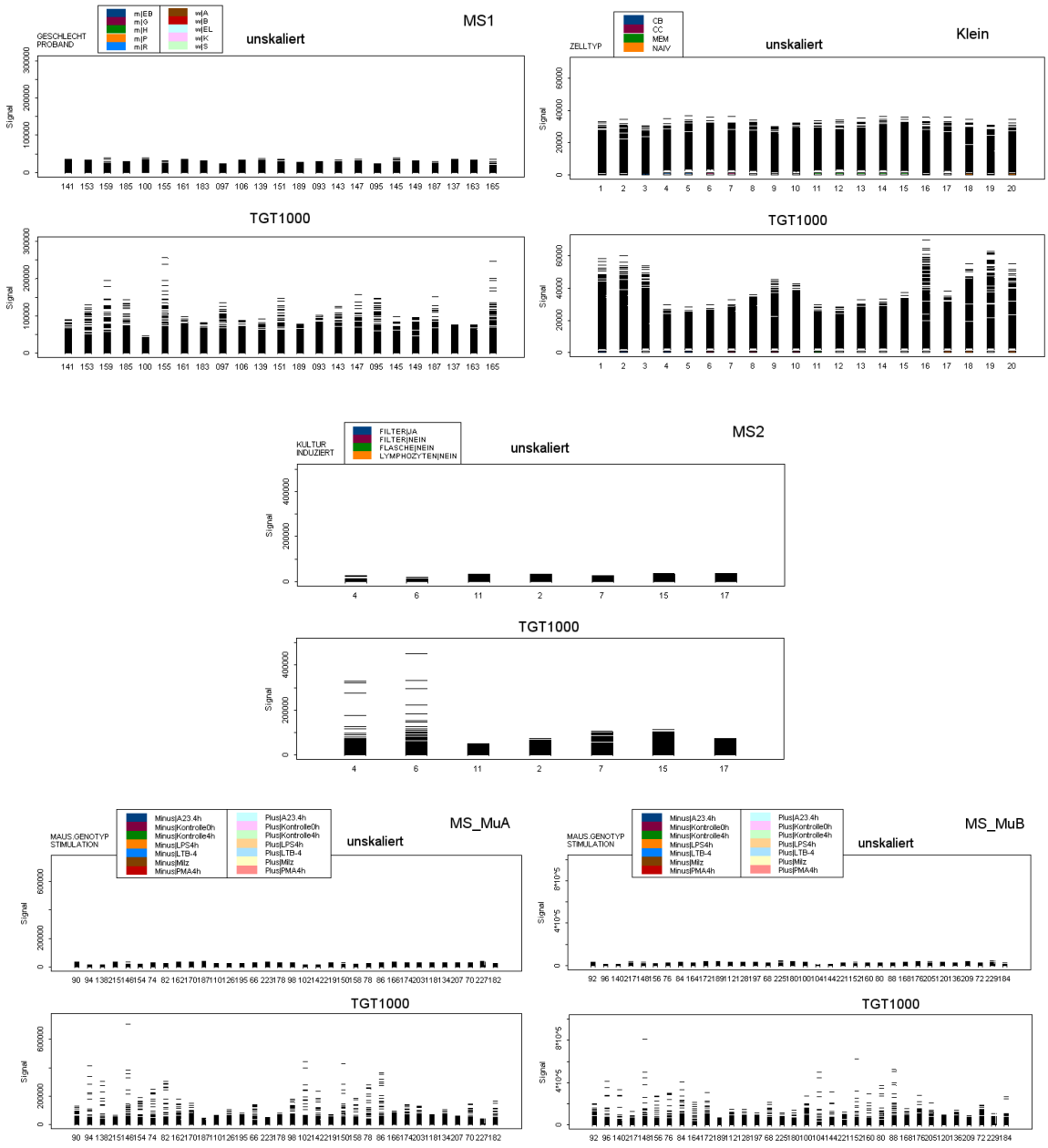


Abbildung 47: Vollständige *Signal*-Bereiche (Box Plots mit Outliern) (jeweils oben: unskaliert, unten: TGT1000)

Im Gegensatz zu den vorhergehenden Betrachtungen haben die *Signal*-Werte der Outlier in den unskalierten Experimenten eine größere Ähnlichkeit als in den skalierten. Scanner und Detektor bilden denselben Intensitätsbereich auf die gleiche numerische Skala ab, auch wenn die Durchschnittsintensität eines Experimentes variiert. *Probe cells*, die eigentlich eine Intensität über der Detektionsgrenze aufweisen, werden auf das Maximum der Detektionsgrenze heruntergebrochen. Diese Tatsache findet sich tendenziell in den *Signal*-Werten wieder, wenn auch nicht mit einem linearen Zusammenhang. Daher

liegen die *Signal*-Werte der Outlier der unskalierten Experimente in einem ähnlichen numerischen Bereich. Erwartungsgemäß werden die Outlier während der Skalierung in Abhängigkeit des linearen Skalierungsfaktors mehr oder weniger verschoben (durch Dehnung bzw. Streckung). Unterscheiden sich die Skalierungsfaktoren stark voneinander, so werden auch die *Signal*-Werte der Outlier nach der Skalierung stark voneinander abweichen.

Die Abhängigkeit vom Skalierungsfaktor lässt sich beispielsweise am Experiment 155 des MS1-Datensatzes gut veranschaulichen. Wie in Abbildung 46 ersichtlich, liegt der mittlere *Signal*-Bereich des unskalierten Experimentes vergleichsweise niedrig. Der Skalierungsfaktor für dieses Experiment (7,89) ist deutlich größer als die der anderen (1,18 - 6,78). Daher liegen die *Signal*-Werte der Outlier des skalierten Experimentes 155 im Vergleich zu den anderen skalierten Experimenten deutlich höher (siehe Abbildung 47).

Die Scatter Plots in Abbildung 48 zeigen am Beispiel des MS1-Datensatzes, dass nach Skalierung ein systematischer Zusammenhang zwischen Skalierungsfaktor und *Signal*-Maximum besteht.

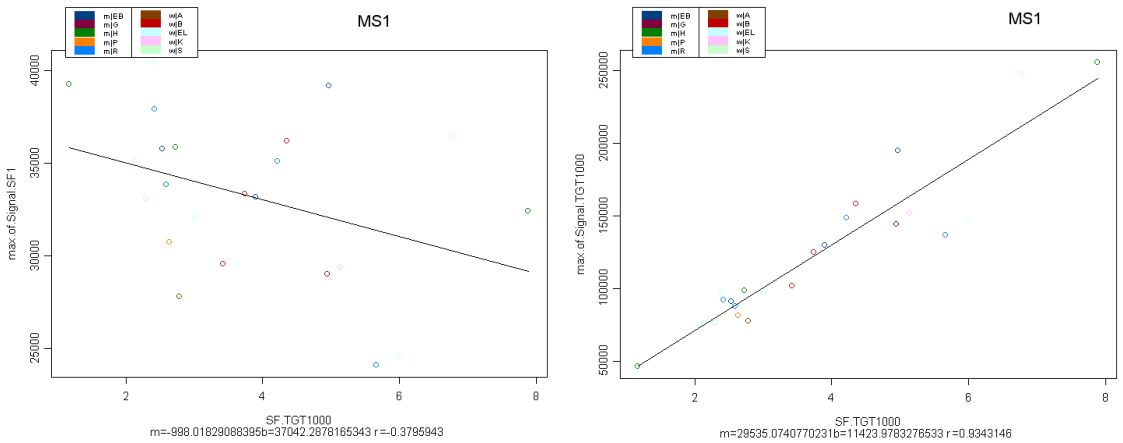


Abbildung 48: Korrelationen zwischen Skalierungsfaktor und *Signal*-Maximum (links: unskaliert, rechts: TGT1000)

Im skalierten Experiment ist die Korrelation zwischen Skalierungsfaktor und *Signal*-Maximum sehr groß, während im unskalierten eine weitaus schwächere Korrelation

besteht. Für die anderen Datensätze gelten ähnliche Beobachtungen (Korrelationskoeffizienten *r* siehe Tabelle 20).

Datensatz	Korrelation mit *Signal*-Maximum (unskaliert)	Korrelation mit *Signal*-Maximum (skaliert)
MS1	-0,380	0,934
Klein	-0,568	0,983
MS2	-0,888	0,985
MS_MuA	-0,756	0,819
MS_MuB	-0,787	0,701
MS_MuA (ohne Ausreißer)	-0,294	0,969
MS_MuB (ohne Ausreißer)	-0,520	0,960

Tabelle 20: Korrelationskoeffizienten *r* zwischen Skalierungsfaktor und *Signal*-Maximum

Die vollständigen MS_Mu-Datensätze bilden bei dieser Beobachtung zunächst eine Ausnahme. Werden die Scatter Plots dann aber im Detail betrachtet (MS_MuA-Datensatz siehe Abbildung 49), zeigt sich, dass jeweils vier der A23.4h-Experimente Ausreißer darstellen. Ohne diese Ausreißerexperimente (MS_MuA: 94, 102, 138, 142; MS_MuB: 96, 104, 140, 144) verhalten sich auch diese Datensätze ähnlich wie die anderen. Insbesondere die Ausreißerexperimente sind diejenigen mit sehr geringen Gesamtintensitäten (siehe Abbildung 38).

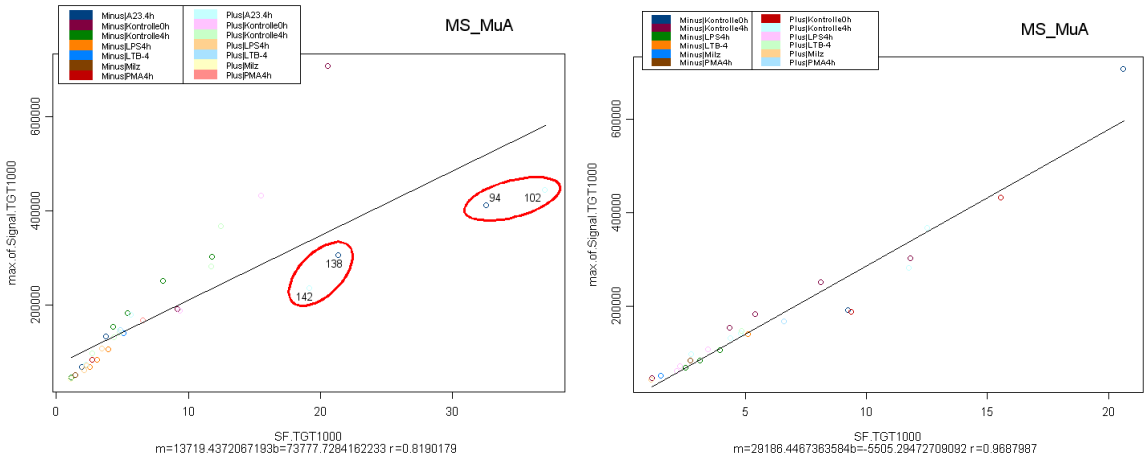


Abbildung 49: Korrelationen zwischen Skalierungsfaktor und *Signal*-Maximum: Ausreißer des MS_MuA-Datensatzes

Mit der Affymetrix-Skalierung werden die randbereinigten Durchschnittswerte der *Signal*-Bereiche aufeinander gelegt. Auch der „mittlere“ Bereich der *Signal*-Verteilung

vom ersten bis zum dritten Quartil wird zwischen zwei Experimenten gut vergleichbar gemacht. Die Ränder der *Signal*-Bereiche werden jedoch bei großen Skalierungsfaktoren unter Umständen sehr weit verschoben. Dies führt beispielsweise zu einer systematischen Überbewertung der *probe sets*, die am Rand des *Signal*-Bereiches liegen. Ergebnisse weiter gehender statistischer Auswertungen, so z. B. eines t-Tests, können daher systematisch falsch sein.

Während der Untersuchungen in Kapitel 4 entstand die Vermutung, dass sich gerade dann große Skalierungsfaktoren ergeben, wenn geringe Gesamtintensitäten im unskalierten Experiment vorliegen. Diese Vermutung wird bestätigt durch die guten (Anti-)Korrelationen von Skalierungsfaktor und Gesamtintensität (siehe Tabelle 21). Damit besteht nun ein Zusammenhang zwischen beobachteter Varianz und Auswirkungen auf weiter gehende Auswertungen trotz (bzw. gerade wegen) der Anwendung einer Skalierung.

Datensatz	Korrelation Gesamtintensität und SF
MS1	-0,869
Klein	-0,927
MS2	-0,779
MS_MuA	-0,631
MS_MuB	-0,652

Tabelle 21: Korrelationskoeffizienten *r* zwischen Gesamtintensität und Skalierungsfaktor

Vermutlich auch aus diesem Grund empfiehlt Affymetrix, keine Experimente miteinander zu vergleichen, deren Skalierungsfaktoren sich um mehr als einen Faktor drei voneinander unterscheiden (siehe Anhang des *Expression Manual*[15]).

5.3.2 *Signal*-Profile

Neben den Betrachtungen zu den *Signal*-Verteilungen jeweils eines Experiments im vorhergehenden Abschnitt sind auch die *Signal*-Profile, also die Verläufe der *Signal*-Werte der *probe sets* über alle Experimente im Vergleich zwischen unskaliertem und skaliertem Experiment von Interesse. Für die folgenden Betrachtungen wird exemplarisch der MS1-Datensatz verwendet.

Werden zunächst die *Signal*-Profile ohne Skalierung betrachtet, so überlagern sich im *line plot* die Profile aller *probe sets*, wodurch nur ein genereller Trend ausgemacht werden kann. Dieser Trend folgt erwartungsgemäß klar dem Verlauf der Gesamtintensitäten der Experimente (siehe Abbildung 50).

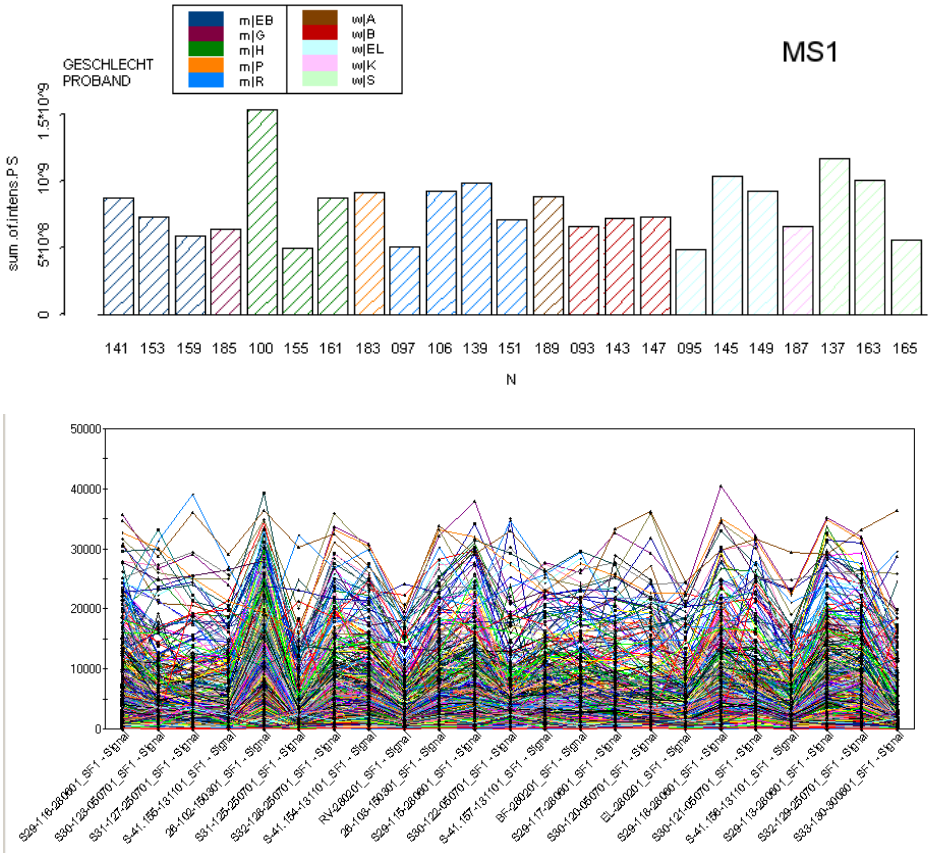


Abbildung 50: MS1, unskaliert: Tendenzielle Übereinstimmung der Gesamtintensitäten mit den *Signal*-Profilen

Dieser beobachtete Trend findet sich in den Clustern eines SOM-Clustering wieder. Bei diesem Verfahren zur Ähnlichkeitsfindung wird die Anzahl *c* der Cluster vorgegeben und dann die *Signal*-Werte der *probe sets* in einen *k*-dimensionalen Raum transformiert (bei *k* Experimenten). Mithilfe von *c* Startknoten werden die *probe sets* in Abhängigkeit ihres Abstandes zu den Clustern in einem iterativen Verfahren sukzessive den Clustern zugeordnet. Abbildung 51 zeigt in Klammern die Anzahl der *probe sets* im jeweiligen Cluster, als rote Linie den Durchschnitt dieser *probe sets* und als blaue Linien die Standardabweichungen.

Der überwiegende Anteil der *probe sets* findet sich hier in Cluster 1 wieder. Die Nummerierung und auch die Zusammensetzung der Cluster kann sich zwischen zwei

Auswertungen unterscheiden, da der Algorithmus teilweise randomisiert abläuft. Die Durchschnittslinie dieses Clusters folgt nahezu exakt dem Verlauf der Gesamtintensitäten mit einem Maximum bei Experiment 100 (fünftes von links) und beispielsweise der absteigenden Folge der Experimente 137, 163, 165 (rechte Seite). Die Standardabweichung ist in diesem Cluster größer als in den anderen Clustern (besonders nach unten). Die Durchschnittslinien der anderen Cluster folgen dem Verlauf der Gesamtintensitäten ebenfalls, allerdings mit mehr oder weniger großen Unterschieden. Ihre Standardabweichungen sind tendenziell geringer, was auch an der geringeren Anzahl von *probe sets* innerhalb dieser Cluster liegt.

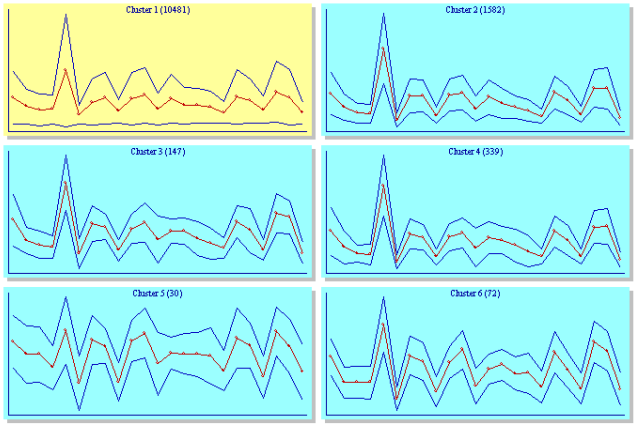


Abbildung 51: MS1, unskaliert: SOM-Clustering der *Signal*-Profile

Nach der Affymetrix-Skalierung auf den Ziel-*Signal*-Wert 1000 ergibt sich ein verändertes Bild der *Signal*-Profile (siehe Abbildung 52). Wie schon bei den Betrachtungen des vorangegangenen Abschnitts wird auch hier augenfällig, dass durch größere Skalierungsfaktoren einige *probe sets* stark betont werden (in diesem Beispiel sehr ausgeprägt in Experiment 155).

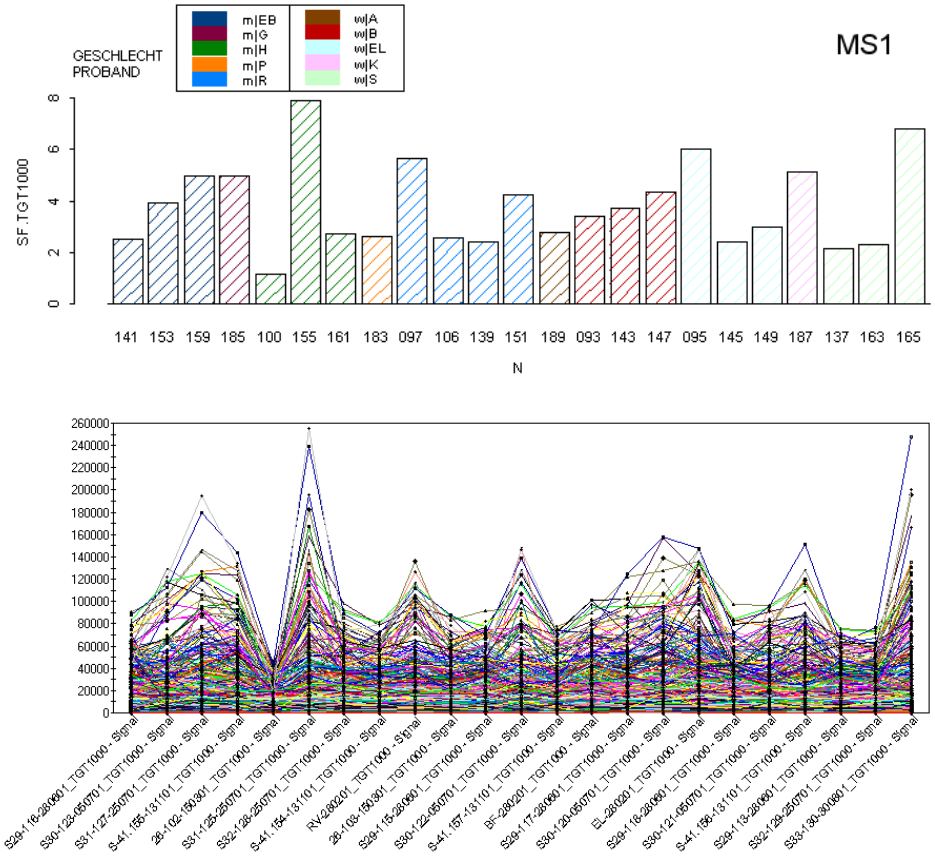


Abbildung 52: MS1, TGT1000: *Signal*-Profile

Nach einem SOM-Clustering der skalierten Experimente ergibt sich gegenüber dem vorherigen SOM-Clustering ein verändertes Bild (siehe Abbildung 53).

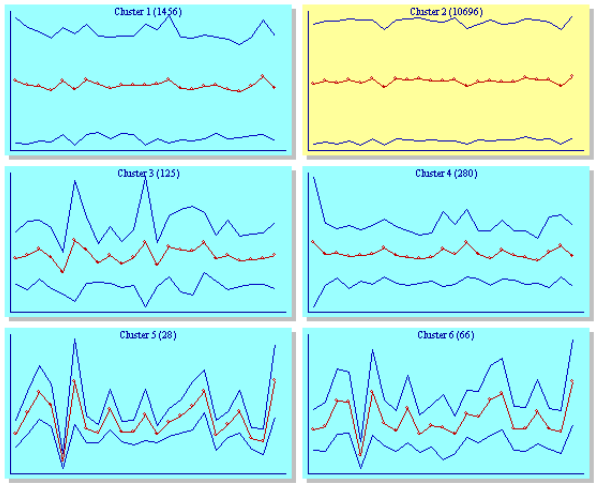


Abbildung 53: MS1, TGT1000: SOM-Clustering der *Signal*-Profile

Cluster 2 enthält hier den größten Anteil der *probe sets*. Seine Durchschnittslinie verläuft relativ gerade auf einem bestimmtem Level. Auch die Standardabweichungen

liegen auf einem vergleichbaren Niveau. Die Affymetrix-Skalierung hat also für den Großteil der *probe sets* den erwarteten Zweck erfüllt: Die Durchschnitts-*Signal*-Werte sind auf ein vergleichbares Niveau gebracht worden.

Cluster 5 ist dasjenige Cluster, welches den am stärksten unterschiedlichen Verlauf zeigt. An einer Gegenüberstellung der Skalierungsfaktoren und der *Signal*-Profile der *probe sets* dieses Clusters wird deutlich, dass die meisten der *probe sets* nach dem Skalieren tendenziell dem Verlauf der Skalierungsfaktoren folgen (siehe Abbildung 54). Es besteht die begründete Annahme, dass die vermeintliche Regulation dieser *probe sets* ein Skalierungsartefakt darstellt.

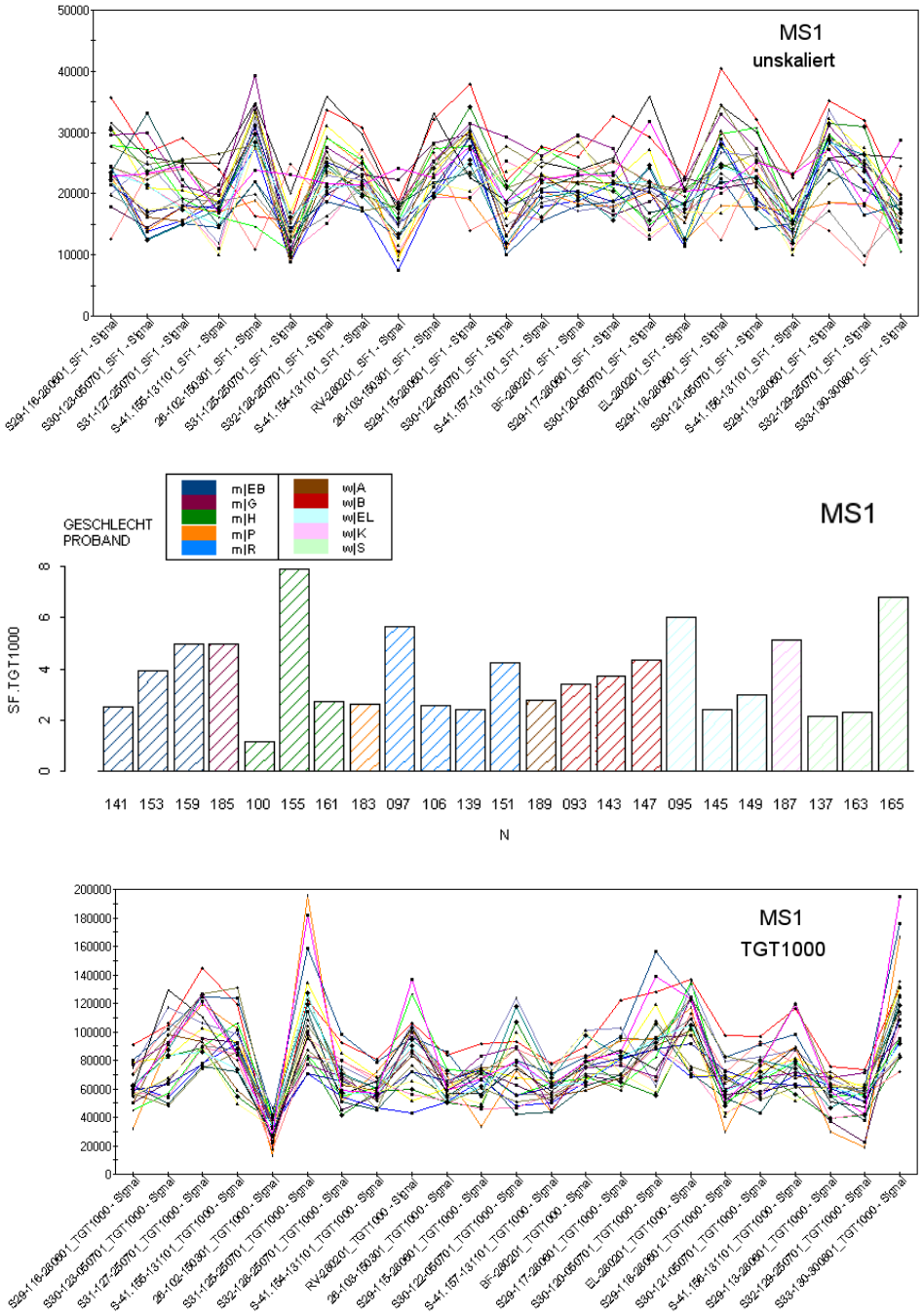


Abbildung 54: MS1: Potenzielle Skalierungsartefakte
(oben: *Signal*-Profile unskaliert, Mitte: Skalierungsfaktoren TGT1000, unten: *Signal*-Profile TGT1000)

5.4 Bewertung der Affymetrix-Skalierung

Die Betrachtungen im weiteren Verlauf dieses Kapitels werden aus Gründen der Übersichtlichkeit auf den hier durchgeführten MS1- und den externen Klein-Datensatz beschränkt. Sie bestehen beide aus HG-U95A-Arrays und sind mit 23 bzw. 20 Experimenten ungefähr gleich groß.

Die Aspekte zur Bewertung einer Skalierungsmethode werden in Unterkapitel 5.1 entwickelt. Nach Aspekt 1 zur Bewertung einer Skalierungsmethode ist die Affymetrix-Skalierung als nicht-optimal zu bewerten, da sie erst auf den kondensierten Maßzahlen ansetzt und also Verfahren denkbar sind, die im Ablauf des gesamten Prozesses näher an der Ursache technischer Varianz ansetzen.

Die Abbildungen 55 und 56 (Wiederholungen aus Abbildungen 46 und 47) untersuchen den Aspekt 2. Sie visualisieren zwei Sichtweisen auf die Wertebereiche der *probe set*-Maßzahlen nach der Kondensierung. Zum einen werden Box Plots der *Signal*-Bereiche mit Outliern (vollständige *Signal*-Bereiche) und zum anderen Box Plots ohne Outlier (mittlere *Signal*-Bereiche) für die unskalierten und die TGT1000-skalierten Experimente gezeigt.

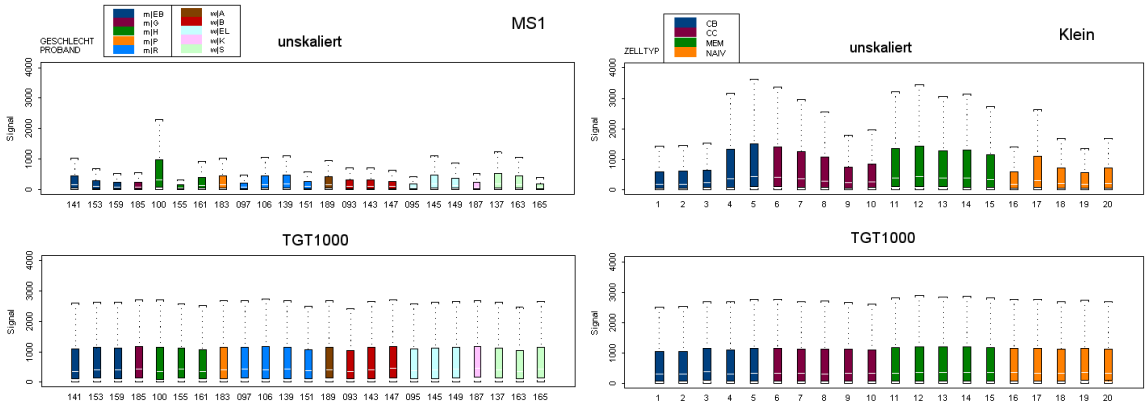


Abbildung 55: Mittlere *Signal*-Bereiche (unskaliert und TGT1000) (MS1- und Klein-Datensatz)

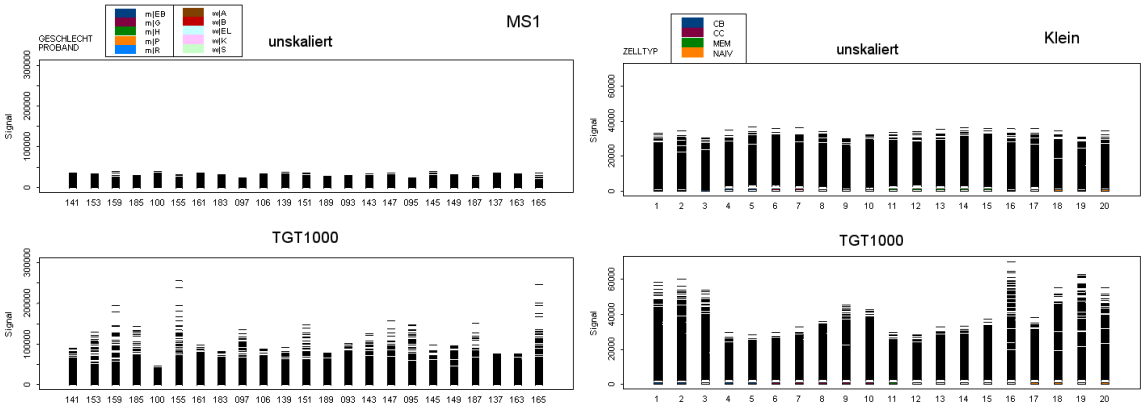


Abbildung 56: Vollständige *Signal*-Bereiche (unskaliert und TGT1000) (MS1- und Klein-Datensatz)

In der Tat gelingt der TGT1000-Skalierung eine Angleichung der mittleren *Signal*-Bereiche (ohne Outlier). Besonders am Experiment 155 des MS1-Datensatzes wird jedoch deutlich, dass die über den mittleren *Signal*-Bereichen liegenden *probe sets* von zuvor schwachen Experimenten durch die Skalierung überdurchschnittlich stark „hochgezogen“ werden (potenzielle Skalierungsartefakte). Während die *Signal*-Maxima vor der Skalierung auf einem vergleichbaren Niveau liegen, unterscheiden sie sich nach der Skalierung in einigen Experimenten stark voneinander.

Aspekt 3 zur Bewertung einer Skalierungsmethode zieht die Expressionslevel der *Housekeeping*-Gene heran. Abbildung 57 zeigt den Verlauf der β-Actin- und GAPDH-*probe sets* in den unskalierten und den TGT1000-skalierten Experimenten.

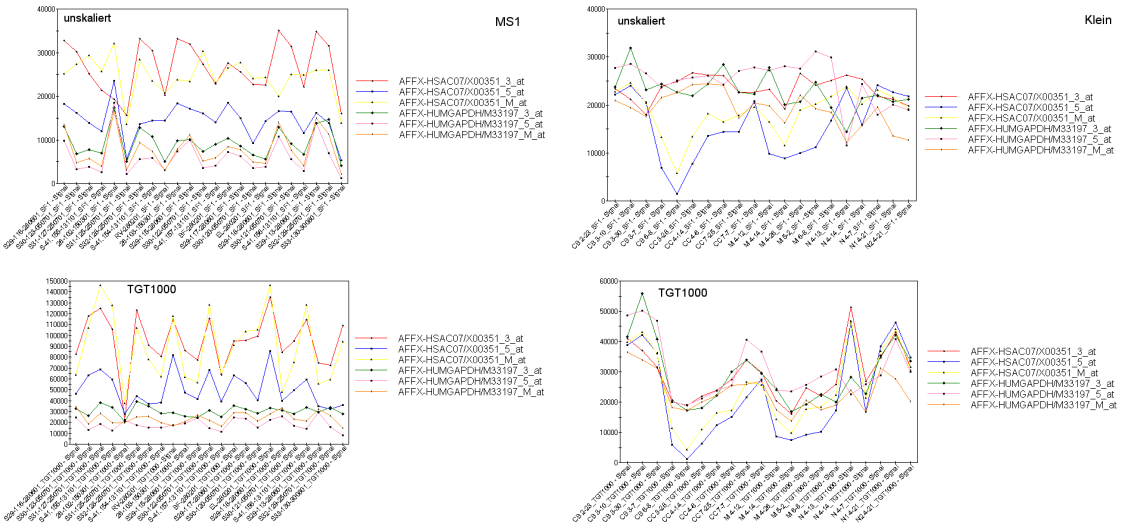


Abbildung 57: *Signal*-Profile der *Housekeeping Controls* (unskaliert und TGT1000)
<u>Achtung</u>: Unterschiedliche y-Achsenbereiche!
(links: MS1, rechts: Klein)

Beim MS1-Datensatz ist eine deutliche Intensivierung der Schwankungen im Expressionslevel der β-Actin-*probe sets* abzulesen (man beachte dazu auch die unterschiedlichen y-Achsenbereiche). Bei den GAPDH-*probe sets* scheint es durch die Skalierung zu einer Stabilisierung zu kommen. Beim Klein-Datensatz weisen β-Actin-5’ (blau) und β-Actin-M’ (gelb) einen besonders auffälligen Verlauf auf. Die Stabilität aller *Housekeeping-probe sets* verringert sich augenscheinlich. Zur genaueren Quantifizierung dieser Beobachtungen wird im Folgenden der Variationskoeffizient verwendet. Dieser wird berechnet, indem die Standardabweichung mittels Division mit dem Durchschnitt

normiert wird. Je kleiner also der Variationskoeffizient, desto stabiler liegen die betrachteten Werte beieinander. Die Tabellen 22 und 23 enthalten die Variationskoeffizienten für die unskalierten Experimente und die TGT1000-Skalierung. Zusätzlich sind auch die durchschnittlichen Variationskoeffizienten der sechs einzelnen *Housekeeping-probe sets* und aller *probe sets* auf dem Array angegeben.

	unskaliert	**TGT1000**
AFFX-HSAC07/X00351_3_at	0,22	0,26
AFFX-HSAC07/X00351_5_at	0,28	0,31
AFFX-HSAC07/X00351_M_at	0,18	0,36
AFFX-HUMGAPDH/M33197_3_at	0,38	0,15
AFFX-HUMGAPDH/M33197_5_at	0,67	0,29
AFFX-HUMGAPDH/M33197_M_at	0,53	0,23
Durchschnitt über die *Housekeeping-probe sets*:	**0,38**	**0,27**
Durchschnitt über alle *probe sets*:	**0,67**	**0,50**

Tabelle 22: Variationskoeffizienten der MS1-Experimente (unskaliert und TGT1000)
rot: besser als bei unskaliert

	unskaliert	**TGT1000**
AFFX-HSAC07/X00351_3_at	0,11	0,31
AFFX-HSAC07/X00351_5_at	0,44	0,68
AFFX-HSAC07/X00351_M_at	0,25	0,52
AFFX-HUMGAPDH/M33197_3_at	0,16	0,36
AFFX-HUMGAPDH/M33197_5_at	0,19	0,32
AFFX-HUMGAPDH/M33197_M_at	0,19	0,27
Durchschnitt über die *Housekeeping-probe sets*:	**0,22**	**0,41**
Durchschnitt über alle *probe sets*:	**0,53**	**0,44**

Tabelle 23: Variationskoeffizienten der Klein-Experimente (unskaliert und TGT1000)
rot: besser als bei unskaliert

Die aus den vorangegangenen Abbildungen abgelesenen Beobachtungen finden sich mit diesen Zahlen bestätigt. Unter der Annahme, dass die Expression der *Housekeeping*-Gene in diesen Experimenten wirklich stabil ist, kann die TGT1000-Skalierung beim Klein-Datensatz als erfolglos und beim MS1-Datensatz als teilweise erfolgreich bezeichnet werden. Lässt man diese Annahme außer Acht und zieht sich darauf zurück, dass sich innerhalb eines Datensatzes nur wenige Gene ändern, ist TGT1000 in beiden

Fällen erfolgreich: die durchschnittlichen Variationskoeffizienten aller *probe sets* werden reduziert.

5.5 Andere Skalierungsansätze

In den vorherigen Unterkapiteln wurden die Eigenschaften der Affymetrix-Skalierung betrachtet und bewertet sowie die Problematik der Dehnung bzw. Streckung des *Signal*-Bereichs diskutiert. Obwohl die letztendliche Beurteilung mangels vielversprechender Kriterien wie beispielsweise der *PolyA Controls* nur indirekt möglich ist, liegt mit den Betrachtungen in Kapitel 4 eine derartige Einsicht in die Varianzquellen vor, dass die Entwicklung eigener Skalierungsmethoden naheliegt.

Hier bieten sich zunächst lokale bzw. semi-lokale Verfahren an, die die *Signal*- oder Intensitätswerte einzelner bzw. mehrerer *probe sets* nutzen, deren konstante mRNA-Konzentration nachgewiesen ist. So stellen Hill et al.[43] eine Methode vor, die die kondensierten Werte der Vorgängerversion der MAS (Version 4) mithilfe von „*spiked cRNA controls*“ zur besseren Vergleichbarkeit skaliert.

Globale Verfahren nutzen, ebenso wie die Affymetrix-Skalierung, die Messwerte aller gemessenen *probe sets* oder *probe cells*. Geller et al.[37] übertragen beispielsweise eine Methode für DNA-Microarrays auf GeneChips und verbinden die Skalierung der Messwerte mit einer gleichzeitigen Transformation, unter anderem zur Stabilisierung der Varianz (ebenfalls mit Werten der MAS, Version 4).

Hier werden globale Verfahren vorgestellt, die aufgrund der vermuteten größeren Nähe zur eigentlichen Ursache der technischen Varianz auf Intensitätsebene ansetzen. Ebenfalls auf Intensitätsebene arbeiten drei Methoden von Bolstad et al.[22] (Implementierungen verfügbar auf der Website des Bioconductor-Projektes[21]); eine nicht-lineare Methode auf Intensitätsebene („*contrast normalization*“) beschreibt Åstrand[19].

5.5.1 Lokale und semi-lokale Verfahren

Lokale bzw. semi-lokale Skalierungen nutzen Messwerte einzelner bzw. mehrerer *probe sets*. Hierfür eignen sich *Housekeeping Controls*, *PolyA Controls*, *Normalization Controls* (ab HG-U133A) und eigene *Housekeeping-probe sets*, für die nachgewiesen ist, dass sie eine stabile Genexpression aufweisen. Eine solche Skalierung

resultiert in einer minimalen Varianz der (durchschnittlichen) *Signal*- oder Intensitätswerte. Gerade bei GeneChips, die mit einer Untermenge von *probe sets* erstellt wurden (*custom made arrays*), sind Skalierungen erforderlich, die auf *Housekeeping*-Genen beruhen und nicht auf Annahmen über die Gesamtheit aller *probe sets*. Für DNA-Microarrays wird ein solcher Ansatz beispielsweise von Wilson et al.[97] vorgestellt.

PolyA Controls, *Normalization Controls* und eigene *Housekeeping-probe sets* standen in den betrachteten Datensätzen nicht zur Verfügung. Für die *Housekeeping Controls* wurde eine stabile Expression ebenfalls nicht geprüft. Daher können hier lokale und semi-lokale Verfahren nicht untersucht werden.

5.5.2 Globale intensitätsbasierte Verfahren

Globale intensitätsbasierte Verfahren zur Skalierung nutzen Messwerte aller *probe sets*. Das Skalieren findet im Gegensatz zur Affymetrix-Skalierung jedoch nicht auf *Signal*-Ebene statt, sondern aufgrund der vermuteten größeren Nähe zur eigentlichen Ursache der technischen Varianz auf Intensitätsebene. Allen hier vorgestellten Skalierungsmethoden liegt die Annahme zugrunde, dass sich die Genexpression des größten Teils der *probe sets* zwischen den Experimenten, die untereinander vergleichbar gemacht werden sollen, nicht wesentlich ändert. Drei Methoden wurden implementiert und werden im Folgenden vorgestellt.

Die erste Methode nutzt die Gesamtintensität als zusammenfassendes Merkmal (**sum.of.intens**-Skalierung), die zweite legt die häufigste gemessene *probe cell*-Intensität zugrunde (**intens.peak**-Skalierung) und die dritte verwendet den häufigsten paarweisen Intensitätsquotienten aller *probe cells* (**pairwise intens.div**-Skalierung).

Skalierung mithilfe der Gesamtintensität (sum.of.intens)

Wie bereits in Unterkapitel 4.3 diskutiert, würden unter idealen Bedingungen perfekte Replikate dieselbe Gesamtintensität aufweisen. Ein Sample, das eine andere RNA-Zusammensetzung hat, aber bei dem ja protokollgemäß dieselbe RNA-Menge auf den Chip gegeben wurde, müsste ebenfalls – abgesehen von den Auswirkungen der ungleichen Hybridisierungs-Performance der unterschiedlich exprimierten Gene – dieselbe Gesamtintensität aufweisen. Unter der Annahme, dass sich die Expression der meisten *probe sets* im Hinblick auf Zusammensetzung und Expressionsniveau zwischen den zu

skalierenden Experimenten nur wenig verändert hat, ist die folgende sum.of.intens-Skalierung anwendbar:

1. Berechne für die zu skalierenden Experimente *i* die Gesamtintensitäten sum.of.intens(*i*).
2. Berechne den Durchschnitt *D* der sum.of.intens(*i*).
3. Berechne für alle Experimente *i* die Skalierungsfaktoren $SF_i := \frac{D}{\text{sum.of.intens(i)}}$
4. Wende auf alle *probe cell*-Intensitäten der CEL-Datei des Experimentes *i* den Skalierungsfaktor SF_i an.
5. Wende den Kondensierungsalgorithmus von Affymetrix an.

Dieser Algorithmus lässt sich nicht mit der MAS, sondern nur mithilfe eigener SPLUS-Funktionen durchführen (siehe Anhang, Funktionen `export.CEL.file`, `add.sum.of.intens.col`, `apply.SF.to.intens`, `perform.LIMS.CEL.-file.import`). Um den Kondensierungsalgorithmus anzuwenden, muss eine modifizierte CEL-Datei zusätzlich zur bereits vorhandenen in das LIMS-System importiert werden. Sie erhält ein Postfix `_SI` zur Kennzeichnung der sum.of.intens-Skalierung. Danach existieren pro skaliertem Experiment zwei CEL-Dateien. Nach Kondensierung der zweiten CEL-Datei mit dem Affymetrix-Skalierungsfaktor eins entsteht eine Primäranalyse mit demselben Postfix. Sie kann nun wieder nach SPLUS importiert werden (`import.LIMS.Analyses`), um dort weitere Betrachtungen durchzuführen.

Vor der Durchführung des sum.of.intens-Skalierungsalgorithmus variieren die Gesamtintensitäten. Nach seiner Durchführung sind die Gesamtintensitäten aller Experimente definitionsgemäß gleich *D* (siehe Abbildung 58).

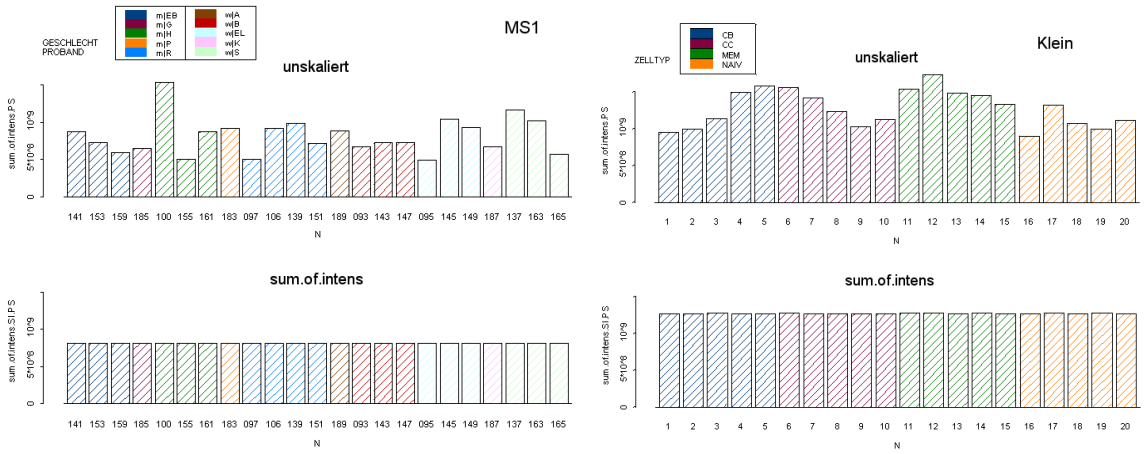


Abbildung 58: Gesamtintensitäten vor und nach der sum.of.intens-Skalierung

Das Verhalten und die Bewertung der sum.of.intens-Skalierung werden in Abschnitt 5.5.3 diskutiert.

Intensity Peak-Skalierung (intens.peak)

Die sum.of.intens-Skalierung nutzt die summierten *probe cell*-Intensitäten, welche bei idealen Replikaten gleich wären. Da der Messbereich der Laser/Detektor-Einheit prinzipiell nach oben (*intensity outlier*) und unten (Rauschgrenze) begrenzt ist, stellen große Differenzen in der Gesamtintensität ein prinzipielles Problem dieser Skalierung dar.

Die schon dort formulierte Annahme, dass sich die Expression der meisten *probe sets* im Hinblick auf Zusammensetzung und Expressionsniveau zwischen den zu skalierenden Experimenten nur wenig verändert, ist Grundlage für die folgende Intensity Peak-Skalierung. Sie ist unabhängig von den oberen und unteren Grenzen des Messbereiches. Zur Ermittlung des linearen Skalierungsfaktors für die Intensitäten der einzelnen *probe cells* wird die Position des Peaks eines Histogramms der *probe cell*-Intensitäten verwendet.

1. Berechne für die zu skalierenden Experimente *i* die Peak-Position intens.peak(*i*) des Histogramms der *probe cell*-Intensitäten.
2. Berechne den Durchschnitt *D* der intens.peak(*i*).
3. Berechne für alle Experimente *i* die Skalierungsfaktoren $SF_i := \frac{D}{\text{intens.peak(i)}}$

4. Wende auf alle *probe cell*-Intensitäten der CEL-Datei des Experimentes *i* den Skalierungsfaktor SF_i an.
5. Wende den Kondensierungsalgorithmus von Affymetrix an.

Dieser Algorithmus lässt sich nicht mit der MAS durchführen (Bemerkungen zur praktischen Durchführung siehe im vorhergehenden sum.of.intens-Abschnitt). Die modifizierten CEL-Dateien dieser Skalierung sind mit dem Postfix `_IP` versehen.

Nach Durchführung des intens.peak-Skalierungsalgorithmus sind die Verteilungen der *probe cell*-Intensitäten verschoben und die Peaks aller Experimente liegen an Position *D* (siehe Abbildung 59).

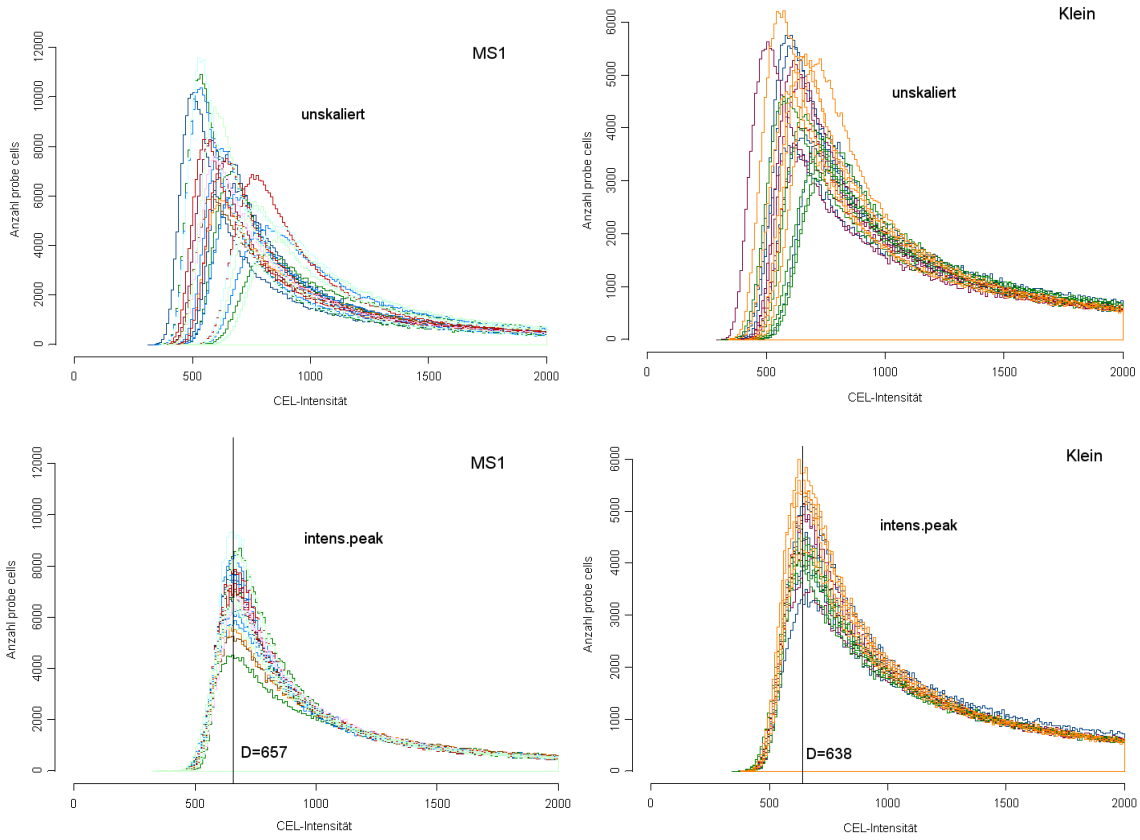


Abbildung 59: Intensitätsverteilungen vor und nach der intens.peak-Skalierung

Das Verhalten und die Bewertung der intens.peak-Skalierung werden in Abschnitt 5.5.3 diskutiert.

Skalierung mithilfe des Intensitätsquotienten (pairwise intens.div)

Bei der intens.peak-Skalierung werden die Histogramme der *probe cell*-Intensitäten verwendet und mit einem linearen Skalierungsfaktor die Peaks übereinander gelegt. Mathematisch möglich wäre, dass die *probe cell*-Intensitäten zwischen zwei Experimenten völlig unterschiedlich, die Histogramme jedoch identisch sind. Die Motivation für Skalierungsmethoden auf Intensitätsebene entstand aus der Folgerung, dass globale technische Varianz alle *probe cells* in gleichem bzw. sehr ähnlichem Maße betrifft. Eine direktere Methode als intens.peak berechnet also das Verhältnis der Intensitäten für jede *probe cell* einzeln und ermittelt unter diesen Intensitätsverhältnissen das häufigste als Skalierungsfaktor. Eine solche Betrachtung kann nur für zwei Experimente gleichzeitig durchgeführt werden. Daher wird vorab ein Experiment mit mittlerer Gesamtintensität bestimmt und für alle übrigen Experimente nacheinander der Skalierungsfaktor zu diesem Experiment berechnet.

1. Bestimme ein Experiment E_{mittel}, welches einen mittleren Intensity Peak aufweist.
2. Berechne für alle anderen zu skalierenden Experimente *i* die Intensitätsverhältnisse *IV* für alle *probe cells j*. $IV_j := \frac{\text{intensity}_{E_{mittel}}(j)}{\text{intensity}_i(j)}$
3. Bestimme für jedes Experiment *i* den Peak eines Histogramms der Intensitätsverhältnisse IV_j als Skalierungsfaktor SF_i für dieses Experiment.
4. Wende auf alle *probe cell*-Intensitäten der CEL-Datei des Experimentes *i* den Skalierungsfaktor SF_i an.
5. Wende den Kondensierungsalgorithmus von Affymetrix an.

Dieser Algorithmus lässt sich nicht mit der MAS durchführen (Bemerkungen zur praktischen Durchführung siehe im Abschnitt zur sum.of.intens-Skalierung). Die modifizierten CEL-Dateien dieser Skalierung sind mit dem Postfix `_ID` versehen.

Für den MS1-Datensatz ergibt sich Experiment 143 als E_{mittel}. Abbildung 60 zeigt das Histogramm der Intensitätsverhältnisse aller *probe cells* zwischen den Experimenten 159 und 143 mit einem Peak bei 1.24. Damit ist auch der Skalierungsfaktor SF_{159}=1.24.

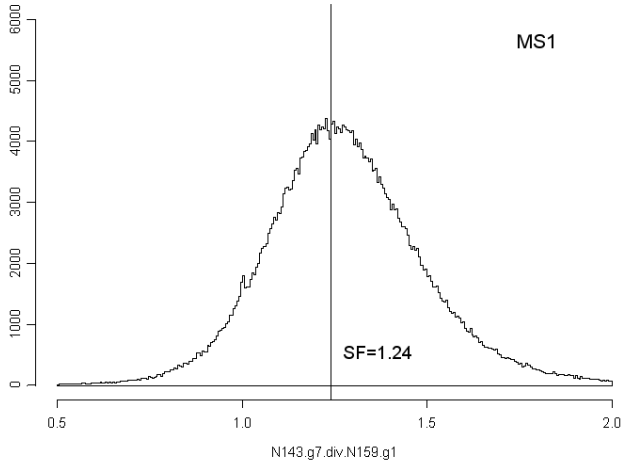


Abbildung 60: Histogramm der Intensitätsverhältnisse (vor intens.div-Skalierung) (*probe cells* der Experimente 143 und 159 des MS1-Datensatzes)

Nach Durchführung der intens.div-Skalierung ergibt sich erwartungsgemäß ein Histogramm der Intensitätsverhältnisse wie in Abbildung 61. Der Peak ist auf die Position 1.0 verschoben. Der größte Teil der *probe cells* weist nun in beiden Experimenten dieselbe Intensität auf.

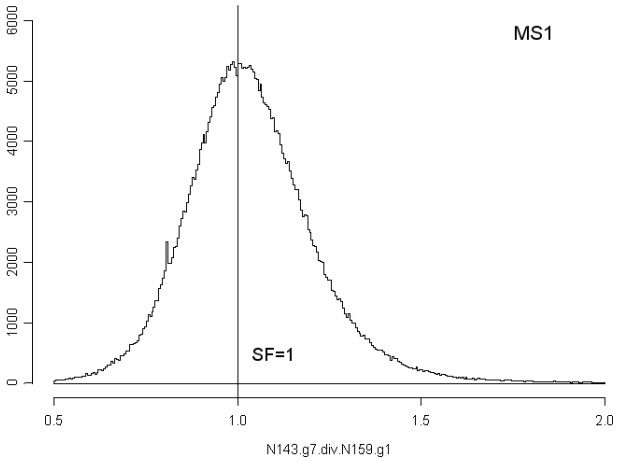


Abbildung 61: Histogramm der Intensitätsverhältnisse (nach intens.div-Skalierung) (*probe cells* der Experimente 143 und 159 des MS1-Datensatzes)

Alle anderen Experimente werden analog behandelt. Das Verhalten und die Bewertung der intens.div-Skalierung werden im folgenden Abschnitt 5.5.3 diskutiert.

5.5.3 Bewertung der sum.of.intens-, intens.peak- und intens.div-Skalierungen

Die Aspekte zur Bewertung einer Skalierungsmethode werden in Unterkapitel 5.1 vorgestellt.

Aspekt 1 zur Bewertung einer Skalierungsmethode ist bei den drei hier vorgestellten anderen Skalierungsansätzen positiv zu bewerten. Diese setzen alle auf der *probe cell*- bzw. Intensitätsebene an, also näher an der Ursache technischer Varianz als die Affymetrix-Skalierung.

Abbildung 62 ermöglicht einen Vergleich der Skalierungsfaktoren der TGT1000-Skalierung und der anderen Verfahren.

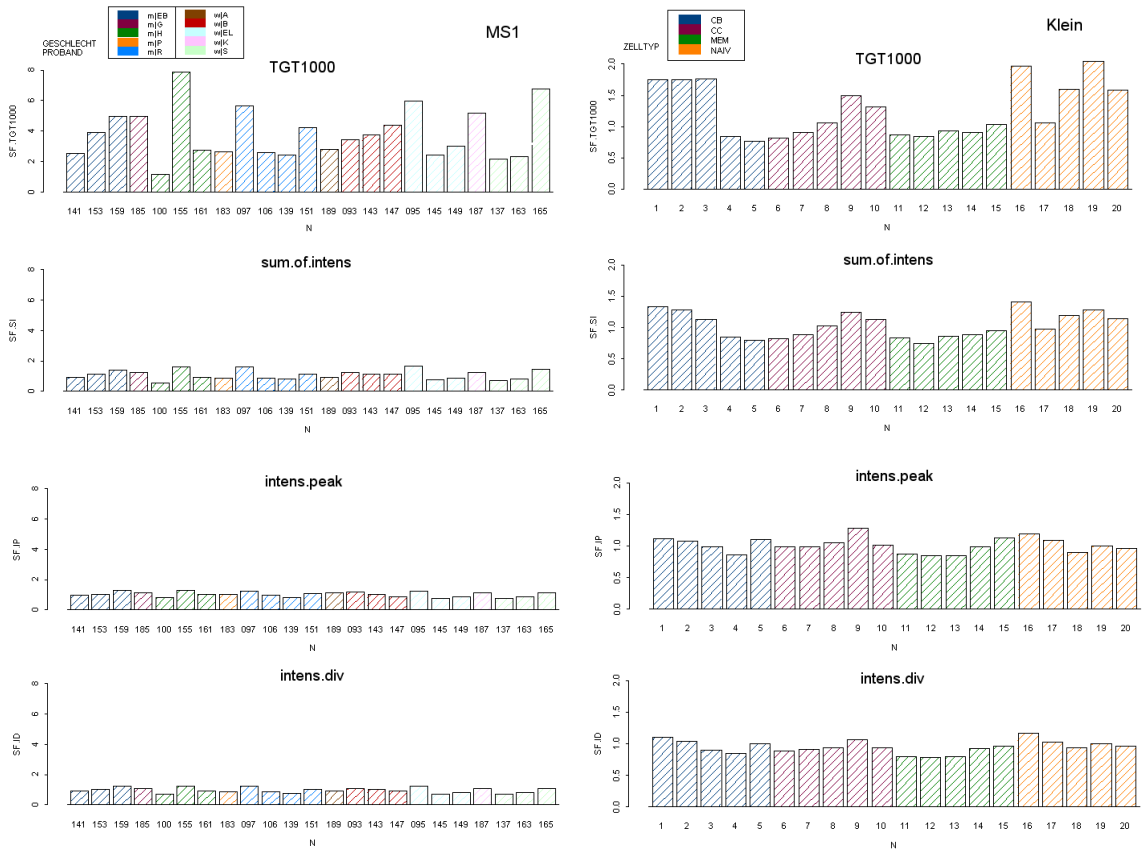


Abbildung 62: Skalierungsfaktoren (alle Skalierungen)
(links: MS1, rechts: Klein)

Die absolute Größe der Skalierungsfaktoren ist zwischen TGT1000 und den Skalierungsmethoden nicht direkt vergleichbar, da sie mit unterschiedlichen Algorithmen auf die Daten angewendet werden. Die Skalierungsfaktoren der TGT1000-Skalierung finden sich tendenziell in den anderen Skalierungen wieder, jedoch sind dort die

Unterschiede der Beträge kleiner. Insbesondere die Skalierungsfaktoren der intens.peak- und der intens.div-Skalierung unterscheiden sich nur geringfügig voneinander. Durch die weniger ausgreifenden Korrekturen im Vergleich zu TGT1000 besteht die Hoffnung, dass bei den Skalierungen auf Intensitätsebene die Outlier bei Experimenten niedriger Gesamtintensität nicht so stark betont werden und damit weniger Skalierungsartefakte auftreten. Die Abbildungen 63 und 64 untersuchen diesen Sachverhalt. Sie visualisieren die mittleren *Signal*-Bereiche und die *Signal*-Werte der Outlier der unskalierten, der TGT1000-skalierten und der mithilfe der anderen Verfahren skalierten Experimente.

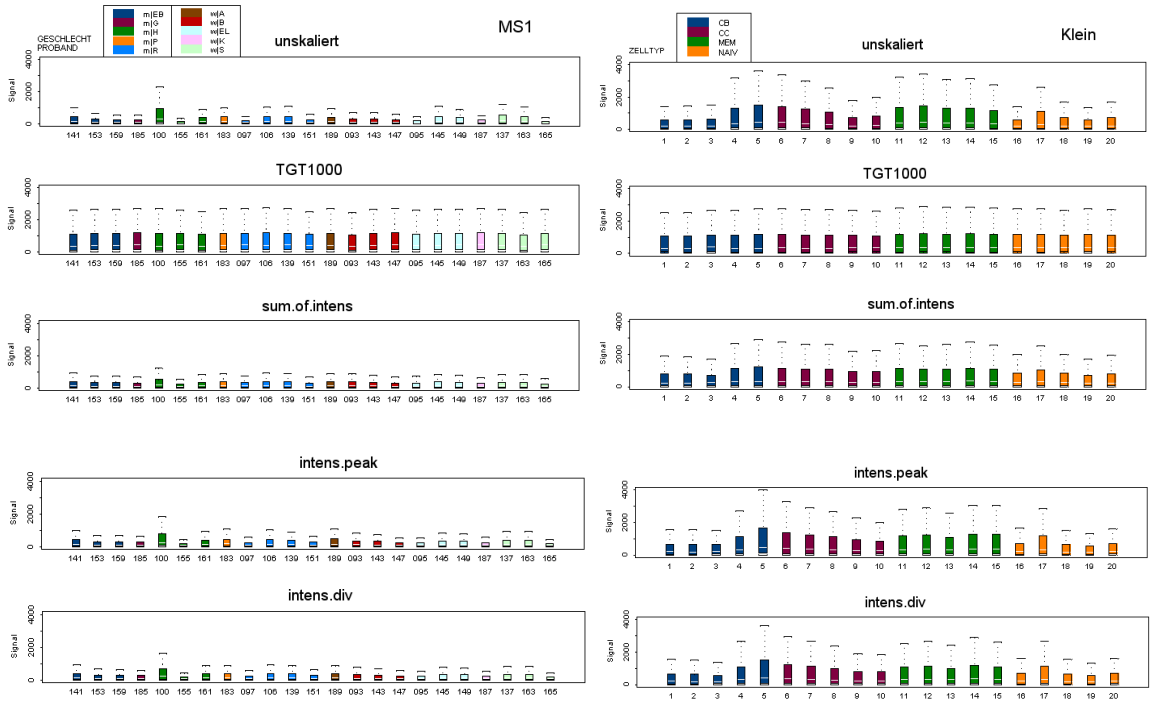


Abbildung 63: Mittlere *Signal*-Bereiche (unskaliert und alle Skalierungen) (links: MS1, rechts: Klein)

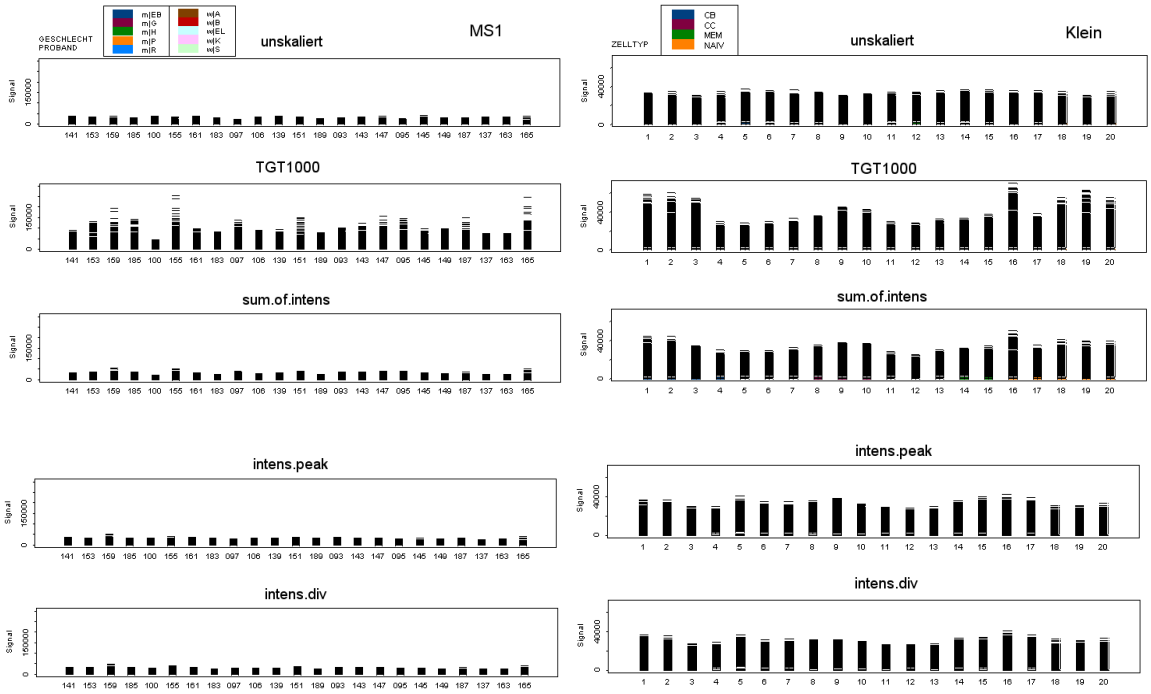


Abbildung 64: Vollständige *Signal*-Bereiche (unskaliert und alle Skalierungen) (links: MS1, rechts: Klein)

Wie erhofft liegen die *Signal*-Maxima bei der sum.of.intens-Skalierung näher beieinander als im Falle der TGT1000-Skalierung. Besonders augenfällig wird das beim MS1-Datensatz (Abbildung 64). Der Preis dafür ist, dass die mittleren *Signal*-Bereiche sich nun ebenfalls deutlich voneinander unterscheiden (siehe Klein-Datensatz, Abbildung 63). Nach Aspekt 2 zur Bewertung einer Skalierungsmethode ist also die sum.of.intens-Skalierung im Hinblick auf die Ränder der *Signal*-Bereiche positiver einzuschätzen als die TGT1000-Skalierung. Sie sollte weniger Gene mit einer falsch positiven Regulation aufgrund von Skalierungsartefakten detektieren. Bei der intens.peak- und der intens.div-Skalierung fallen die Unterschiede der mittleren *Signal*-Bereiche wieder größer aus als bei der sum.of.intens-Skalierung. Die *Signal*-Maxima rücken hingegen näher zusammen. Insgesamt ist sehr anschaulich, dass das Zusammenrücken der mittleren *Signal*-Bereiche mit einem Auseinanderdriften der *Signal*-Maxima erkauft wird. Die intens.peak- und intens.div-Skalierungen skalieren die Experimente nur sehr vorsichtig. Die *Signal*-Bereiche ähneln in hohem Maße den unskalierten Experimenten. Aus Sicht des Aspektes 2 stellt die sum.of.intens-Skalierung einen Kompromiss dar zwischen der sehr invasiven TGT1000-Skalierung und den effektschwachen intens.peak- und intens.div-Skalierungen.

Zur Beurteilung von Aspekt 3 zur Bewertung einer Skalierungsmethode werden zunächst in Abbildung 65 die *Signal*-Profile der *Housekeeping-probe sets* der unskalierten und der skalierten Experimente untereinander aufgeführt.

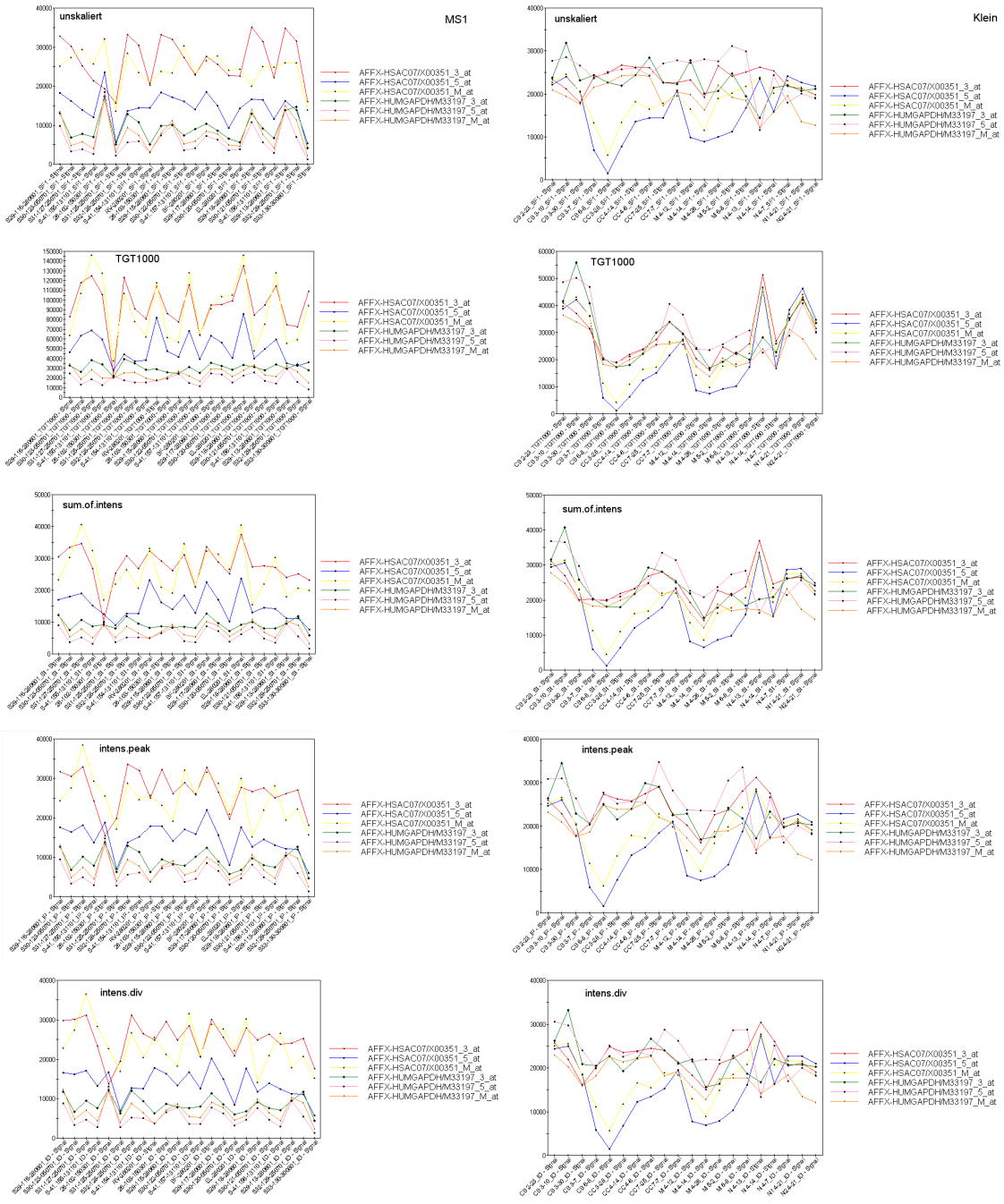


Abbildung 65: *Signal*-Profile der *Housekeeping Controls* (unskaliert und alle Skalierungen)
<u>Achtung</u>: Unterschiedliche y-Achsenbereiche!
(links: MS1, rechts: Klein)

Hierbei ergibt sich kein einheitliches Bild. Beim MS1-Datensatz verstärkt die TGT1000-Skalierung die Schwankungen der β-Actin-*probe sets* (AFFX-HSAC07...),

während die der GAPDH-*probe sets* verringert wird. Die Profile der β-Actin-*probe sets* weisen mit der sum.of.intens-Skalierung weniger Schwankungen als bei der TGT1000-Skalierung auf, während die GAPDH-*probe sets* stärker variieren. Mit den intens.peak- und intens.div-Skalierungen scheinen sich die Schwankungen der β-Actin-*probe sets* weiter zu verringern. Eine klare Beurteilung der GAPDH-Profile ist nicht möglich. Beim Klein-Datensatz scheinen die Profile bei TGT1000 und sum.of.intens eine höhere Variabilität als bei den unskalierten und den intens.peak- und intens.div-skalierten Experimenten aufzuweisen.

Zur genaueren Quantifizierung dieser Beobachtungen wird wie in Unterkapitel 5.4 der Variationskoeffizient verwendet (siehe Tabellen 24 und 25). Die obigen Beobachtungen finden sich in den Zahlen wieder.

	unskaliert	**TGT1000**	**sum.of.intens**	**intens.peak**	**intens.div**
AFFX-HSAC07/X00351_3_at	0,22	0,26	0,20	*0,18*	*0,18*
AFFX-HSAC07/X00351_5_at	0,28	0,31	0,29	0,27	*0,26*
AFFX-HSAC07/X00351_M_at	*0,18*	0,36	0,28	0,23	0,24
AFFX-HUMGAPDH/M33197_3_at	0,38	*0,15*	0,18	0,28	0,25
AFFX-HUMGAPDH/M33197_5_at	0,67	*0,29*	0,40	0,53	0,50
AFFX-HUMGAPDH/M33197_M_at	0,53	*0,23*	0,32	0,41	0,39
Durchschnitt über die *Housekeeping-probe sets*:	**0,38**	***0,27***	**0,28**	**0,32**	**0,30**
Durchschnitt über alle *probe sets*:	**0,67**	***0,50***	**0,51**	**0,58**	**0,56**

Tabelle 24: Variationskoeffizienten der MS1-Experimente (unskaliert und alle Skalierungen)
rot: besser als bei unskaliert, unterstrichen: besser als bei TGT1000, *kursiv*: Minimum der Zeile

	unskaliert	**TGT1000**	**sum.of.intens**	**intens.peak**	**intens.div**
AFFX-HSAC07/X00351_3_at	*0,11*	0,31	0,20	0,17	0,16
AFFX-HSAC07/X00351_5_at	*0,44*	0,68	0,57	0,47	0,50
AFFX-HSAC07/X00351_M_at	*0,25*	0,52	0,40	0,30	0,32
AFFX-HUMGAPDH/M33197_3_at	*0,16*	0,36	0,25	0,19	0,18
AFFX-HUMGAPDH/M33197_5_at	*0,19*	0,32	0,22	0,22	*0,19*
AFFX-HUMGAPDH/M33197_M_at	*0,19*	0,27	0,20	0,20	*0,19*
Durchschnitt über die *Housekeeping-probe sets*:	***0,22***	**0,41**	**0,31**	**0,26**	**0,26**
Durchschnitt über alle *probe sets*:	**0,53**	***0,44***	**0,45**	**0,51**	**0,50**

Tabelle 25: Variationskoeffizienten der Klein-Experimente (unskaliert und alle Skalierungen)
rot: besser als bei unskaliert, unterstrichen: besser als bei TGT1000, *kursiv*: Minimum der Zeile

Zusammenfassend gilt beim MS1-Datensatz: Die Variabilität der β-Actin-*probe sets* wird durch TGT1000 erhöht und durch die neuen Skalierungsansätze teilweise verringert, wobei alle Variationskoeffizienten bei den neuen Skalierungsansätzen geringer sind als bei TGT1000. Die Variabilität der GAPDH-*probe sets* wird von allen Skalierungsmethoden verringert. Dieser Effekt ist bei TGT1000 am stärksten ausgeprägt. Dies trifft auch auf den Durchschnitt aller *Housekeeping-probe sets* und aller *probe sets* insgesamt zu.

Beim Klein-Datensatz vergrößern sämtliche Skalierungen die Variabilität aller *Housekeeping*-Profile. Die neuen Skalierungsansätze resultieren in geringeren Variationskoeffizienten als TGT1000. Dies findet sich im Durchschnitt der Variationskoeffizienten der *Housekeeping*-Profile wieder. Beim Durchschnitt über alle *probe sets* liefert TGT1000 das beste Ergebnis.

Eine sichere Aussage über die Güte einer Skalierungsmethode ist nicht möglich, da keine weiteren Erkenntnisse zur Stabilität der RNA-Konzentration der *Housekeeping*-Gene vorliegen. Insbesondere beim MS1-Datensatz könnte eine Regulation vorliegen, da sich β-Actin und GAPDH beim Skalieren unterschiedlich verhalten. Während TGT1000 unter der Annahme der Stabilität der *Housekeeping-probe sets* bei MS1 als die beste Skalierungsmethode abschneidet, wird sie mit den Ergebnissen des Klein-Datensatzes zur schlechtesten. Wird die Stabilitätsforderung nicht auf die *Housekeeping-probe sets* beschränkt, sondern postuliert, dass sich die Expression der meisten gemessenen Gene innerhalb der Experimente eines Datensatzes nur wenig ändert, ergibt sich bezüglich Aspekt 3 ein konsistentes Bild der Qualität der Skalierungsmethoden: TGT1000 stellt die beste Methode dar, dicht gefolgt von sum.of.intens, auf welche in einigem Abstand intens.div folgt und dann intens.peak (tendenziell findet sich dies auch bei den GAPDH-*probe sets* des MS1-Datensatzes wieder).

Zusammen mit Ergebnissen zu Aspekt 2 (Hinweis auf weniger Skalierungsartefakte) und der grundsätzlich erfüllten Forderung des Aspekts 1 zur Bewertung von Skalierungsmethoden kristallisiert sich heraus, dass die sum.of.intens-Skalierung der TGT1000-Skalierung überlegen sein könnte. Um diese Vermutung bestätigen zu können, sollte mindestens einer der Aspekte 4 bis 6 in weiterführenden Untersuchungen ebenfalls eine Überlegenheit der sum.of.intens-Skalierung zeigen. Diese Aspekte betreffen die

Stabilität der *Normalization Controls* auf den U133-Arrays, den Vergleich von *Signal*-Werten mit den Ergebnissen einer quantitativen PCR und die Stabilität von hinzugefügten *PolyA Controls*.

6 Zusammenfassende Diskussion

Im Folgenden werden die Beobachtungen und Bewertungen der einzelnen Kapitel abschließend im Zusammenhang diskutiert und Empfehlungen bezüglich des weiteren Vorgehens formuliert.

6.1 SPLUS-Schnittstelle und Funktionsbibliothek

Ein erstes Ziel dieser Arbeit war die Modellierung und Implementierung einer Schnittstelle von GeneChip-Daten zu SPLUS und C++. Außerdem sollte eine Funktionsbibliothek zum Gruppieren von Experimenten nach ihren Merkmalen entwickelt werden.

Zu diesem Zweck wurde zunächst das Statistikpaket SPLUS als Entwicklungsumgebung ausgewählt. Es wird mit seiner hierarchischen Datenbankstruktur und seinen vorgefertigten Statistik- und Graphen-Routinen den Anforderungen dieser Aufgabe gerecht. Im Laufe der Zeit wurden jedoch einige Nachteile augenfällig, wie beispielsweise das Fehlen von Pointern, was bei manchen komplexen Operationen zum Überlaufen des Speichers führte. Bei SPLUS handelt es sich um ein kommerzielles Produkt, eine Alternative hierzu bietet das in großen Teilen ähnliche, nicht-kommerzielle Projekt R[73]. Sofern die implementierten Routinen einfach übertragbar sind, sollten sie bald migriert werden, um einer breiteren Forschergemeinde zugänglich gemacht zu werden und von finanziellen Faktoren unabhängig zu sein.

Ausgehend von den elementaren Datenobjekten des LIMS-Systems wurde ein Objektmodell definiert. Auf diesen Objekten wurden grundlegende Funktionen implementiert. Zusätzlich zur Handhabung beliebiger experimenteller Merkmale eines Datensatzes (wie beispielsweise Protokollmodifikationen oder Patienten- / Probanden-Daten) ist der Import sowohl von Intensitätsdaten als auch von Primäranalysen möglich. Damit lassen sich nun prinzipiell beliebige weiter gehende statistische Auswertungen in SPLUS programmieren, die über das hinausgehen, was die Affymetrix-Software bereits bietet. Im Rahmen dieser Arbeit wurden überwiegend Funktionen implementiert, die die Betrachtungen zu Qualitätskriterien und Skalierungsmethoden in Kapitel 4 und 5 ermöglichten. Mit ihnen kann insbesondere das Verhalten der GeneChip-Daten auf Intensitätsebene untersucht werden. Die implementierten Objekte und Funktionen sind prinzipiell erweiterbar, etwa auf neue Generationen von Arrays. Außerdem wird in Zukunft neben der Nutzung des LIMS SDK eine Einbeziehung des File SDK wichtiger

werden, um die binär kodierten CEL-Dateien der von Affymetrix neu entwickelten Front-End-Software GCOS in eigenen Funktionen handhaben zu können.

Zusätzlich zu den implementierten Basisfunktionen entstanden weiter reichende Funktionen zum Gruppieren, Benennen und Kombinieren von Experimenten. Die aus ihnen resultierenden Gruppierungsinformationen können in den Basisfunktionen – gerade auch in den Visualisierungsfunktionen – genutzt werden. Es zeigte sich, dass bei komplexen Datensätzen ein Ausfiltern von Experimenten mit passenden Merkmalen und das Zusammenfassen von Experimenten mit denselben Merkmalen in Gruppen sinnvoll ist, da es die Handhabung großer Experimentmengen vereinfacht. So wurde der MS1-Datensatz aus einer Menge von 196 Experimenten einer Arbeitsgruppe mit vielfältigen Variationen ihrer experimentellen Merkmale extrahiert, deren Handhabung sich ohne die implementierten Funktionen als sperrig und unhandlich gestaltet hatte. In Zukunft wird sich das Gruppierungskonzept bei großen Datensammlungen – beispielsweise bei Expressionsdatenbanken – und in Kombination mit dem MIAME-Standard[23] als unerlässlich erweisen. Durch vorheriges Sortieren nach bestimmten Merkmalen und die Möglichkeit, mehrere Merkmale hierarchisch in den Gruppierungsvorgang einzubeziehen, kann die Gruppierung flexibel auf die Fragestellung abgestimmt werden. Die Benennung der erstellten Gruppen nach Merkmalsausprägungen und die Benennung der Experimente nach Merkmalsausprägungen und Gruppennummern zielt in dieselbe Richtung. Dasselbe Experiment kann so im Kontext unterschiedlicher Fragestellungen (d. h. bei unterschiedlichen Gruppierungen) jeweils anders benannt werden, mit dem Ziel, das gerade interessierende Unterscheidungsmerkmal als Namensbestandteil einzufügen. Eine Zuordnung zum Originalbezeichner ist weiterhin jederzeit möglich. Die implementierten Funktionen erlauben dann die schnelle Spezifikation mehrerer Experimente (beispielsweise aller Replikate eines Datensatzes) durch die Angabe von Gruppennummern. Auch das Konzept zum Kombinieren von Experimenten unter Verwendung der Gruppierungsinformationen konnte prinzipiell erfolgreich umgesetzt werden. Es sieht zurzeit nur einen begrenzten Satz an Kombinationsmöglichkeiten vor, ist jedoch um zusätzliche Kombinationen erweiterbar. Im Laufe der Entwicklungsarbeit zeigte sich allerdings, dass SPLUS an seine oben erwähnten Grenzen bezüglich der Speicherplatzverwaltung stößt, wenn bei großen Datensätzen die Anwendung von Funktionen auf Kombinationen von Experimenten gewünscht wird. Durch einen

gesteigerten Ressourceneinsatz lassen sich zwar größere Datensätze bearbeiten, die prinzipielle Problematik bleibt jedoch bestehen.

Ein Teil der implementierten Funktionalität wurde in C++ realisiert, jedoch eng verzahnt mit den SPLUS-Funktionen. Die C++-Komponenten stellen bei zeitkritischen Operationen wie dem zeilenweisen Einlesen großer Dateien eine gute Ergänzung dar und bieten darüber hinaus die einzige Möglichkeit zur Ansteuerung des LIMS SDK, welches die programmiertechnische Schnittstelle zum LIMS-System liefert. Die implementierte Funktionalität wird durch das im Rahmen dieser Arbeit entwickelte eigenständige Windows-Programm Evaluation Server konsequent ergänzt. Dieses wurde als Client-Server-Lösung konzipiert und ermöglicht (auch plattformübergreifend) die Kommunikation mit anderen Anwendungen. Im Zuge der Untersuchungen für diese Arbeit wurde das Programm zur effizienten Abfrage von CEL-Dateien und Reports genutzt. Parallel dazu diente die Anwendung als Schnittstelle eines Solaris-Webservers zum LIMS-System (hier nicht beschrieben). Eine weitere Rolle könnte sie bei der zukünftigen Vernetzung von LIMS-Servern zur dezentralen Etablierung einer Expressionsdatenbank spielen.

6.2 Varianz, Meta-Analysen und Qualitätskriterien

Das zweite angestrebte Ziel dieser Arbeit war es, durch Meta-Analysen Methoden zur Quantifizierung von Varianz zu entwickeln, die beobachtete Varianz zu beschreiben und Qualitätskriterien für Experimente zu identifizieren und anzuwenden. Dafür wurden neben drei Datensätzen aus dem GeneChip-Standort Münster mit den Array-Typen HG-U95A, Mu11KsubA und Mu11KsubB auch öffentlich zugängliche Experimente anderer Herkunft (HG-U95A) als Testdatensätze ausgewählt. Zunächst wurde eine Definition für die Begriffe „biologische Varianz“ und „technische Varianz“ gegeben. Erstere wird im untersuchten Material aufgrund einer Vielzahl von biologischen Einflussgrößen hervorgerufen und ist unter anderem auch Zielgröße bei der Durchführung vergleichender Genexpressionsexperimente. Letztere ist auf das Messsystem selbst zurückzuführen. Ihr Einfluss soll verringert, im Idealfall sogar ganz vermieden werden. Die Effekte biologischer und technischer Varianz auf die Maßzahlen wurden mithilfe mathematischer Funktionen formuliert, bevor eine Aufzählung verschiedener Varianzarten vorgenommen wurde. Zu deren wichtigsten Vertretern zählen Noise, Bias und lokale Varianz. Unter

Noise wird „zufällige“ Varianz gefasst, die nicht reproduzierbar ist und deren Auswirkungen daher für ein einzelnes Experiment nicht durch Rechenoperationen („Skalierung“) zurückgenommen werden können. Bias stellt eine reproduzierbare Varianz dar, die im günstigsten Fall durch Skalierung herausgerechnet werden kann und sich weiter in additive und multiplikative Varianz aufteilt. Mit lokaler Varianz werden Effekte bezeichnet, die durch räumlich begrenzte Ursachen (z. B. *spots*, Flusen) entstehen. Ihre Folgen können durch Maskieren der in den betroffenen Arealen lokalisierten *probe cells* gemindert werden.

Nach der Definition der Begrifflichkeiten wurden die Quellen biologischer und technischer Varianz diskutiert. Besonderes Augenmerk lag dabei auf der detaillierten Diskussion der technischen Varianzquellen in jedem einzelnen Schritt der Aufarbeitung. Die Auflistung enthält die wichtigsten am hiesigen Standort diskutierten Varianzquellen, ist aber wahrscheinlich nicht erschöpfend, sondern kann um weitere Varianzquellen ergänzt werden. Die erstellte detaillierte Liste dient zum einen der Bewusstmachung einer Vielzahl von Varianzquellen und enthält damit den inhärenten Aufruf zum sorgfältigen Arbeiten. Zum anderen musste klar werden, dass ohne Replikatexperimente an unterschiedlichen Schritten der Aufarbeitung keine genaue Quantifizierung der Varianz der verschiedenen Schritte möglich sein kann, da sich die Varianzen aller Schritte am Ende der Aufarbeitung überlagern. Insbesondere kann der Schritt mit der größten Varianz ohne diese Replikate nicht eindeutig identifiziert werden; die Hauptvarianzquelle muss hypothetisch bleiben.

Im Anschluss an die zunächst theoretischen Betrachtungen folgten Beschreibungen der Varianz von konkreten Experimenten. Diese wurden als Visualisierungen verschiedener Qualitätskriterien anhand von Bar Plots und Box Plots und zusätzlich mithilfe der statistischen Kennzahlen Durchschnitt und Standardabweichung dargestellt. Neben der Untersuchung der von Affymetrix angegebenen Qualitätskriterien wurde das neue Qualitätskriterium „Gesamtintensität“ eingeführt und untersucht.

Die ersten betrachteten Qualitätskriterien waren *noise* und *background*. Es zeigte sich, dass mit ihrer Hilfe eindeutige Ausreißer eines Datensatzes gut identifizierbar sind (MS_Mu: 162 und 164). Die Streuung innerhalb der Datensätze stellte sich als relativ groß heraus (z. B. *background*: MS1: ca. 400 - 700; MS2: ca. 340 - 480). Die Lage der Wertebereiche von *noise* und *background* waren zwischen den Datensätzen unter

Berücksichtigung dieser Streuung (t-Test) in Einzelfällen sogar signifikant unterschiedlich (z. B. MS1 zu MS2: p=0.0001). Aus diesen Ergebnissen leitet sich ab, dass für verschiedene Datensätze in der Regel unterschiedliche Sollbereiche für *noise* und *background* existieren. Damit ist die Identifikation von Ausreißerexperimenten nur innerhalb eines Datensatzes und nur bei Mehrfachexperimenten möglich. Ein Experiment, in dem ein anderes Protokoll oder ein anderes biologisches Material verwendet wird, kann nicht ohne weiteres hinsichtlich der Wertebereiche von *noise* und *background* aus einem bestehenden Datensatz bewertet werden. Diese Beobachtung ist besonders relevant, gerade wenn es um die Qualitätskontrolle von Daten für Expressionsdatenbanken geht. Die Qualitätseinschätzung muss idealerweise vorab mit Erkenntnissen aus ähnlichen experimentellen Aufbauten stattfinden und kann nicht erst nachträglich in der Datenbank stattfinden.

Die Replikatexperimente des MS_Mu-Datensatzes zeigen, dass *noise* und *background* jeweils zwischen zwei Replikatexperimenten gut korreliert sind. Dies führt zu der Folgerung, dass die Hauptvarianzquelle, die sich in diesen Kriterien niederschlägt, nicht in den Schritten „Befüllen des Chips“, „Hybridisierung“, „Färben“ und „Scannen“ liegen kann. Zwischen *noise* und *background* in den Experimenten eines Datensatzes liegt eine gute lineare Korrelation vor. Der *noise* stellt definitionsgemäß die Standardabweichung des *background* dar, daher muss die beobachtete Varianz multiplikativ sein. Wäre sie additiv, würde der *noise*-Wert trotz verändertem *background* gleich bleiben und damit kein linearer Zusammenhang zu beobachten sein. Für das Kriterium *Noise(Q)* gelten ähnliche Beobachtungen. Interessanterweise fanden sich beim Vergleich verschiedener Array-Typen entweder die geringere Pixelzahl einer *probe cell* in diesem Kriterium wieder oder ein Hinweis auf größere Auswirkungen beim Legen des Gitters. Trotz der großen Ähnlichkeit zu *noise* hat also auch *Noise(Q)* eine Existenzberechtigung als Qualitätskriterium, mit dem Probleme beim Legen des Gitters identifiziert werden können. Zuletzt zeigte sich, dass *noise* und *Noise(Q)* gut korreliert sind, mit ihnen wird also vermutlich die gleiche oder eine ähnliche Varianzquelle detektiert.

Zu den Untersuchungen der bisherigen Qualitätskriterien zählt neben der Betrachtung von *noise*, *background* und *Noise(Q)* auch die vergleichende Analyse von vier Gruppen von *Control probe sets*: *Hybridization Controls*, *PolyA Controls*, *Housekeeping Controls* und *Normalization Controls*. Die prokaryotischen *Hybridization Controls* (BioB, BioC,

BioDN und Cre) werden am Ende der Aufarbeitung bei der Herstellung des Hybridisierungscocktails hinzugefügt. Mit ihnen sollen Probleme in den Schritten „Hybridisierung“, „Färben“ und „Scannen“ identifiziert werden. Weisen die *Hybridization Controls* einen *Detection Call* von *Present* auf, so waren diese Schritte erfolgreich. Ist nichtsdestotrotz ein Großteil der anderen *probe sets* des Chips *Absent*, so liegt unter der Voraussetzung, dass gleiche Mengen RNA auf den Chip gegeben wurden, ein Problem in der vorangegangenen Aufarbeitung vor (in der Regel in den Schritten „Entnahme“ oder „Isolierung“) und das Experiment muss aus weitergehenden Auswertungen ausgeschlossen werden. Einzig BioB darf *Absent* sein, da es mit der geringsten Konzentration im Hybridisierungscocktail vorliegt. Die 3’/M’/5’-*Signal*-Werte der *Hybridization Controls* liegen in den untersuchten Datensätzen erwartungsgemäß nicht wie andere *Controls* in einer bestimmten Lage zueinander vor, da sie die Aufarbeitung nicht durchlaufen. Die *Signal*-Werte unterlagen über alle Experimente eines Datensatzes hinweg wider Erwarten großen Schwankungen, obwohl sie ja in konstanten Konzentrationen hinzugegeben werden. Hierfür müssen Einflüsse des Samples oder Pipettierungenauigkeiten verantwortlich sein. Überkreuzungen der 3’/M’/5’-*Signal*-Profile weisen zusätzlich auf Qualitätsprobleme bei der Herstellung des *Control Kits* oder auf Einflüsse des Samples wie z. B. Kreuzhybridisierung hin. Allgemein gültige Kriterien für den Ausschluss von Experimenten aus den Datensätzen können aus den Beobachtungen bezüglich der Überkreuzungen nicht entwickelt werden. Es ist angeraten, die Qualität des *Control Kits* mit unabhängigen Methoden zu überprüfen. Verhalten sich die *Hybridization Controls* aufgrund ihrer Expressionshöhen trotz *Present Call* ohne nachvollziehbaren Grund (wie beispielsweise einer fehlerhafter Mengenberechnung im Labor) deutlich auffällig, sollten die entsprechenden Experimente bei weiter gehenden Auswertungen vorsichtshalber aus dem Datensatz entfernt und durch gleichartige Experimente ersetzt werden (Ausreißer im Klein-Datensatz: Experimente 6, 7 und 8 bzgl. BioB und 7 bzgl. Cre; siehe Seite 93).

Ein interessanter Effekt der eigentlichen Qualitätskontrollfunktion resultierte bei den *Hybridization Controls* aus der Konzentrationsreihe, in der sie im Hybridisierungscocktail vorliegen. Die Verhältnisse der theoretischen Konzentrationen der zugegebenen *Hybridization Controls* (BioB:BioC:BioDN:Cre = 1,5 : 5 : 25 : 100) spiegeln sich in den durchschnittlichen *Signal*-Werten qualitativ sehr gut wider, quantitativ jedoch nicht. Daraus ergibt sich die Folgerung, dass auch für andere Gene allein aus den *Signal*-Werten

nur bedingt quantitative Aussagen hinsichtlich der Verhältnisse ihrer Konzentrationen im Ausgangsmaterial gemacht werden können. Als vollwertig quantitative Methode können GeneChips also höchstens bei horizontalen Vergleichen (*Signal*-Werte eines *probe sets* in verschiedenen Experimenten), nicht bei vertikalen Vergleichen (*Signal*-Werte verschiedener *probe sets* in einem Experiment) gelten. Dies bedeutet auch, dass vergleichende Genexpressionsanalysen mit einer begrenzten Anzahl von Replikaten in der Regel lediglich einen explorativen Charakter haben. Eine Beweisführung für das Vorliegen vermuteter funktioneller oder statistischer Zusammenhänge muss in unabhängigen Experimenten erfolgen. So wird die Bestätigung des Vorhandenseins unterschiedlicher Expressionsniveaus in Fall- und Kontrollprobe in der Regel zusätzlich durch quantitative PCR erfolgen müssen. Die biologische Relevanz muss in eigens hierfür entworfenen Experimenten bestätigt werden. Diagnostisch nutzbare Zusammenhänge werden häufig mithilfe weniger im GeneChip-Experiment als vermutlich diskriminierungsrelevant identifizierter Marker mittels qRT-PCR an größeren Kollektiven bestätigt.

Trotz des Einsatzes vorgefertigter *Hybridization Kits* und konstanter Konzentrationen zeigen sich sowohl eine große Streuung innerhalb der Datensätze, als auch Unterschiede in der Lage zwischen den Datensätzen. Unter der Annahme, dass die anderen *probe sets* eines Arrays einer ebensolchen Streuung selbst bei konstanter Konzentration unterliegen, werden quantitative Auswertungen der Experimente in Expressionsdatenbanken also mit nicht-trivialen Problemen zu kämpfen haben. Wie bereits bei den Kriterien *noise* und *background* zeigt auch das Verhalten der *Hybridization Controls* in den Replikatexperimenten, dass die beobachtete Varianz nicht in den Schritten „Befüllen des Chips“, „Hybridisierung“, „Färben“ und „Scannen“ liegen kann.

PolyA Controls, die nach der Isolierung zum Sample hinzugegeben werden, dienen zur Identifizierung von Problemen in den Schritten „Reverse Transkription“ oder „*in-vitro*-Transkription“. Sie wurden in den betrachteten Datensätzen nicht verwendet. Da mittlerweile ein vorgefertigtes *PolyA Kit* mit unterschiedlichen Konzentrationen der *PolyA Controls* angeboten wird, sollten diese Kontrollen zukünftig eine breitere Anwendung finden. Dies würde neue Einschätzungen von Varianzquellen ermöglichen, die insbesondere in den besagten Schritten vermutet werden. Darüber hinaus könnten

PolyA Controls als Eichstandards in Skalierungsansätzen dienen. Ihre Eignung für diesen Zweck muss sich anhand zukünftiger Untersuchungen erweisen.

Die letzte untersuchte Gruppe von *Control probe sets* waren die *Housekeeping Controls* (GAPDH und β-Actin). Die RNAs der entsprechenden Gene sind bereits im Ausgangsmaterial vorhanden, sie durchlaufen also die gesamte Aufarbeitung. Es wird davon ausgegangen, dass die RNAs dieser Gene in ähnlichen Konzentrationen in allen untersuchten Materialien vorliegen, da sie an zentralen Aufgaben der Zelle beteiligt sind. Mittlerweile wurde für einige Gewebe gezeigt, dass auch die *Housekeeping*-Gene einer Regulation unterliegen können, sodass die Annahme konstanter Expression nur unter Vorbehalt gelten kann. Die Lage der 3'/M'/5'-Profile zueinander entsprach in den betrachteten Datensätzen zum größten Teil den Erwartungen: Der 3'-*Signal*-Wert lag über dem M'- und dieser wiederum über dem 5'-*Signal*-Wert, da im Protokoll sowohl RT als auch IVT am 3'-Ende beginnen und dann nicht immer vollständig bis zum 5'-Ende laufen. Auch die Degradation von RNA beginnt in der Regel am 5'-Ende. Anhand der Expressionsprofile der GAPDH-*probe sets* konnte beispielsweise mit Experiment 16 des Klein-Datensatzes ein eindeutiger Ausreißer identifiziert werden.

Wie bei den anderen Qualitätskriterien waren eine unerwartet große Streuung innerhalb der Datensätze und unerwartet große Unterschiede in der Lage der Wertebereiche zwischen den Datensätzen zu beobachten. Im Gegensatz zu den *Hybridization Controls* kann dies neben technischer Varianz auch in biologischer Varianz begründet liegen (wenn die Annahme konstanter Expression nicht gilt).

Viel wichtiger als die Expressionshöhen der *Housekeeping Controls* sind bei der Beurteilung der Qualität eines Experimentes jedoch die 3'/5'-Quotienten. Sie quantifizieren das Verhältnis vom 3'- zum 5'-*Signal*-Wert, also die Güte der aus den Zellen isolierten RNA und die Qualität der RT bzw. IVT. Wurden letztere systematisch zu früh abgebrochen, liegt ein großer 3'/5'-Quotient (> 3) vor; das Gleiche gilt für eine RNA, die in der präanalytischen Phase degradiert war oder während der Aufarbeitung degradiert wurde. In den betrachteten Datensätzen konnten so einige Ausreißer identifiziert werden (z. B. bezüglich GAPDH: MS1: 165; MS2: 4 und 6; MS_MuA: 170 und 150; MS_MuB: 172). Sind die Quotienten beider *Housekeeping Controls* zu groß, sollte das Experiment verworfen und durch eine Wiederholungsmessung ersetzt werden (MS2: 4 und 6; MS_MuA: 150). Liegt ein systematisch zu großer 3'/5'-Quotient vor, stellt dies einen

eindeutigen Hinweis auf Probleme hinsichtlich des Zustands oder der Aufarbeitung des Untersuchungsmaterials dar. Eine solche systematische Erhöhung wurde in den verwendeten Datensätzen nicht beobachtet. Sie könnte neben präanalytischen Unzulänglichkeiten besonders bei Aufarbeitungen mit einer geringen Menge an mRNA im Ausgangsmaterial und speziellen Aufarbeitungsprotokollen auftreten und müsste in Ermangelung von Alternativen toleriert werden. In diesem Fall wäre es erforderlich, die Anzahl von Mehrfachexperimenten in einer Gruppe zu erhöhen (z. B. von drei auf fünf und mehr), um einen gewissen Ausgleich der Messungenauigkeiten zu schaffen.

Anhand der Experimente des MS_Mu-Datensatzes wurde gezeigt, dass die 3'/5'-Quotienten von GAPDH bzw. β-Actin zwischen Replikatpaaren jeweils sehr gut korreliert sind. Wie schon zuvor ließ sich ableiten, dass der Hauptanteil technischer Varianz dieses Qualitätskriteriums nicht in den Schritten nach „Befüllen des Chips" liegt. Interessanterweise zeigte sich, dass die Korrelation der 3'/5'-Quotienten von GAPDH und β-Actin in einem Experiment nur mittelmäßig ist; diese können sich erheblich voneinander unterscheiden. Hieraus folgt unmittelbar, dass auch die Qualität der Messung der übrigen *probe sets* in einem Experiment erheblich variieren kann und dass selbst bei guten Qualitätskriterien eine schlechte Messung anderer Gene nicht ausgeschlossen werden kann. Die Qualitätseinschätzung eines Experiments auf Basis eines kleinen Ausschnitts aller gemessenen *probe sets* ist also nur bedingt aussagekräftig, jedoch steht ein besseres Qualitätskriterium zur Beurteilung der Integrität der mRNA nicht zur Verfügung.

Normalization Controls existieren erst bei den HG-U133-Arrays und wurden hier folglich nicht untersucht. Es handelt sich um *probe sets*, für die Affymetrix eine relativ konstante Expression in diversen Gewebetypen festgestellt hat. Mit Datensätzen, in denen diese *Control probe sets* gemessen werden, könnte in Zukunft einerseits ihre Eignung als Qualitätskriterium untersucht, und andererseits könnten eigene Skalierungen entwickelt oder die Qualität von Skalierungsmethoden beurteilt werden. Vor ihrer Verwendung ist es angeraten, die konstitutive Expression bestimmter Gene im eigenen experimentellen Aufbau zunächst zu bestätigen.

Zu den bisherigen Qualitätskriterien zählt auch der Anteil an *Present Calls*. Über den *Detection Call* wird die Detektionsverlässlichkeit jedes *probe sets* quantifiziert. Bei Auftreten der hier beobachteten globalen Varianz, die sich in einem insgesamt dunkleren

Intensitätsbild äußert, sinkt die Detektionsverlässlichkeit offensichtlich global ab, sodass auch der Anteil der *probe sets* ohne Detektionsprobleme absinkt, also die Anzahl der *Present Calls* des Experiments. In einigen Fällen traten bei Experimenten mit einem geringeren Anteil an *Present Calls* als bei Vergleichsexperimenten auch schlechtere 3'/5'-Quotienten auf (z. B. MS_MuA: 150; MS2: 4 und 6), was ein Beleg für die grundsätzliche Eignung dieses Qualitätskriteriums ist. Mit seiner Hilfe können weitere Ausreißer-experimente identifiziert werden (z. B. MS_MuA: 146 und 162; MS_MuB: 148 und 164). Der Anteil der *Present Calls* kann analog zu den bereits erwähnten Qualitätskriterien nicht ohne weiteres zwischen zwei Datensätzen verglichen werden. Eine interessante Beobachtung diesbezüglich ermöglichten die Datensätze MS_MuA und MS_MuB. Bei den Mu11KsubB-Arrays gab es im Durchschnitt nur circa halb so viele *Present Calls* wie bei den Replikatexperimenten mit Mu11KsubA-Arrays. Eine mögliche Folgerung hieraus ist, dass im Vergleich zu Mu11KsubA nur halb so viele der Mu11KsubB-Gene oberhalb der Verlässlichkeitsgrenze exprimiert werden. Wahrscheinlicher ist jedoch, dass die *probe sets* dieses Arrays aus den Sequenzen weniger gut definierter Gene erstellt wurden und daher die Detektionsverlässlichkeit beim Mu11KsubB-Array global herabgesetzt ist. Da der Anteil an *Present Calls* zwischen den Replikatexperimenten nichtsdestotrotz einen guten Korrelationskoeffizienten aufweist, muss für die verbleibende Varianz eine Quelle verantwortlich sein, die nicht in den Schritten nach „Befüllen des Chips" liegt.

In der Regel werden Qualitätskriterien einzeln oder in Kombination verwendet, um Experimente fragwürdiger Qualität zu identifizieren. Diese müssen aus allen weiteren Auswertungen ausgeschlossen werden. Aufgrund der hohen Kosten für GeneChip-Experimente wird man – auch abhängig davon, ob es sich sich um eine explorative Untersuchung (Vorstudie) oder eine Klassifikationsstudie (Hauptstudie) handelt oder ob Nachfolgeexperimente einfach möglich sind – in Einzelfällen Kompromisse eingehen und die entsprechenden Experimente trotzdem in weitergehende Auswertungen einbeziehen. Die Ergebnisse dieser Auswertungen müssen mit großen Vorbehalten interpretiert werden.

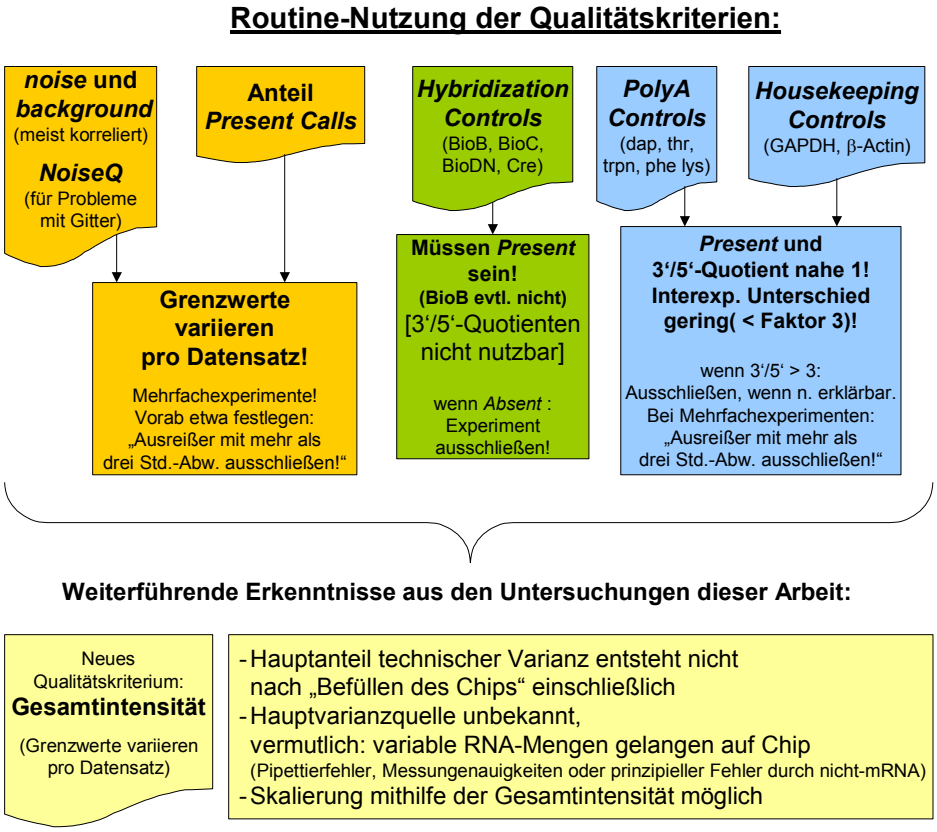
Neben ihrer eigentlichen Funktion zur Identifizierung fragwürdiger Experimente gaben die Qualitätskriterien insbesondere von Replikatexperimenten indirekt Hinweise auf die wichtigsten Quellen technischer Varianz. Mit dem nachfolgend diskutierten, neu entwickelten Kriterium „Gesamtintensität" konnten diese Hinweise weiter konkretisiert werden.

In den vorangegangenen Beobachtungen zeigte sich eine relativ große Varianz, deren Ursprung nicht vollständig geklärt werden konnte. Einige Aufarbeitungsschritte konnten als Quelle ausgeschlossen werden, und es gab Hinweise auf einen globalen Effekt. Bei der Betrachtung der Intensitätsbilder der CEL-Dateien entstand zusätzlich die Idee zu dem neuen Qualitätskriterium „Gesamtintensität". Bei der Untersuchung seines Verhaltens wurden ähnliche Beobachtungen gemacht wie bei den bisherigen Qualitätskriterien: die Streuung innerhalb der Datensätze ist relativ groß und die Lage der Wertebereiche variiert zwischen den Datensätzen relativ stark. Bei den Replikatexperimenten zeigte sich eine gute Korrelation, woraus folgt, dass die Varianzquelle die globale Veränderung der Intensität verursacht. Aus den Betrachtungen der Intensitätsverteilungen der Experimente wurde abgeleitet, dass bei Experimenten mit geringerer Gesamtintensität ein größerer Anteil an unspezifischer Hybridisierung stattfindet, wobei jedoch insgesamt weniger Farbstoff auf dem Chip verbleibt (Näheres siehe Abschnitt 4.3.2). Die Ursache für diesen Effekt muss in Aufarbeitungsschritten vor „Befüllen des Chips" liegen, eine Erklärung für die eigentliche Varianzquelle konnte mit den vorliegenden Experimenten nicht gefunden werden. Als Ursache kommen vor allem unterschiedliche Mengen an RNA in Betracht, die der Hybridisierung mit dem Chip zugeführt wurden. Diese resultieren in der Regel aus Pipettierfehlern und Messungenauigkeiten bei der Quantifizierung der mRNA (bei der eine nicht experimentell bestimmte Menge an nicht-mRNA in Abzug gebracht wird, die während der Aufarbeitung nicht aus dem System entfernt werden kann).

Das in dieser Arbeit neu entwickelte Kriterium Gesamtintensität korreliert gut bis sehr gut mit publizierten Qualitätskriterien; es detektiert also vermutlich keine anderen Varianzquellen als diejenigen, welche auch mit den bekannten Qualitätskriterien erfasst werden. Weil es durch Verwendung der *probe cell*-Intensitäten näher an der Ursache aller beobachteten Varianzquellen liegt als die *Signal*-Werte – und in den Intensitäten die Auswirkungen von Varianzquellen noch nicht durch eine nicht-lineare Beziehung wie den Kondensierungsalgorithmus verkompliziert wurden – und darüber hinaus den globalen Chip-Zustand berücksichtigt, ist das neu entwickelte Kriterium zur Qualitätskontrolle dennoch besser geeignet als die bisherigen Kriterien. Es sollte in Zukunft in den Standardablauf der Qualitätsbeurteilung von GeneChip-Experimenten aufgenommen

werden. Neben diesen Erkenntnissen zeigte sich, dass *probe cell*-Intensitäten und die Gesamtintensität für Skalierungsansätze von Interesse sind (Näheres siehe weiter unten).

Die folgende Übersicht enthält eine Zusammenfassung über die empfohlene Routine-Nutzung der Qualitätskriterien und die weiterführenden Erkenntnisse aus den Betrachtungen dieser Arbeit.



Das zweite Ziel dieser Arbeit, das Verhalten bisher verwendeter Qualitätskriterien zu untersuchen, konnte erreicht werden. Zusätzlich zu der ursprünglich formulierten Aufgabe konnte ein neues Qualitätskriterium definiert und Ideen für neue Skalierungsmethoden entwickelt werden. Ungeklärt ist, welcher der Arbeitsschritte vor dem Befüllen der Hybridisationskassette für den Hauptanteil der beobachteten Varianz verantwortlich ist. Hierzu sind speziell für diesen Untersuchungszweck entworfene Experimente notwendig, die dem Untersucher nicht zugänglich waren. Ein Grund für das Fehlen solcher Experimente sind die damit verknüpften enormen Kosten.

6.3 Skalierungsmethoden

Abschließendes Ziel dieser Arbeit war es, existierende Skalierungsverfahren zu analysieren, um eventuell neue, verbesserte Verfahren zu entwickeln und deren Vorteile darzustellen. Ausgangspunkt hierbei war die Tatsache, dass selbst gut reproduzierte Daten zur Erlangung der Vergleichbarkeit einer Skalierung bedürfen. An anderer Stelle wurde gezeigt, dass die Wahl der Skalierungsmethode Auswirkungen auf die Ergebnisse weiter gehender statistischer Auswertungen hat. Ohne weiteres ist damit jedoch noch nicht geklärt, welche Methode „besser“ oder „schlechter“ ist, d. h. welche die Gegebenheiten im untersuchten Material „besser“ oder „schlechter“ wiedergibt.

Zunächst wurden daher Bewertungsmöglichkeiten für Skalierungsverfahren formalisiert. Forderungen an eine ideale Skalierung sind, dass sich bei Genen mit gleicher mRNA-Konzentration gleiche *Signal*-Werte ergeben und dass sich die quantitativen Verhältnisse zwischen den Konzentrationen zweier Gene in den Verhältnissen ihrer *Signal*-Werte wiederfinden. Aus diesen Forderungen wurden sechs konkrete Aspekte zur Bewertung von Skalierungsmethoden für GeneChip-Experimente entwickelt: Der erste Aspekt verlangt, dass eine gute Skalierungsmethode nahe an der Varianzquelle ansetzt, damit ihre Anwendung nicht durch nachfolgende Transformationen erschwert wird. Der zweite Aspekt beinhaltet den Anspruch, dass sich die Wertebereiche ähnlicher Experimente nach einer erfolgten Skalierung wenig voneinander unterscheiden sollen. Der dritte Aspekt umfasst die Forderung nach wenig varianten *Signal*-Werten der *Housekeeping*-Gene. Diese Stabilität der *Housekeeping*-Gene ist nicht *a priori* klar und muss für die untersuchten Proben zunächst nachgewiesen werden. Die vierte Forderung an eine gute Skalierungsmethode ist die geringe Varianz in den *Normalization Controls* (ab HG-U133). Es folgen als fünfter Aspekt die Forderung nach Reproduzierbarkeit der Ergebnisse mit anderen Messmethoden und als sechster Aspekt eine Konstanz der Messwerte für vorgefertigte oder eigens identifizierte *PolyA Controls*.

Im weiteren Verlauf dieser Arbeit wurde dann die Affymetrix-Skalierung vorgestellt. Bei dieser Methode wird für ein Experiment aufgrund des randbereinigten Durchschnitts aller Messwerte und der Zielintensität ein Skalierungsfaktor errechnet. Dieser Skalierungsfaktor wird auf alle *probe sets* angewendet. Es zeigte sich, dass aufgrund von Rundungsungenauigkeiten beim Logarithmieren und Delogarithmieren die konkreten Skalierungsfaktoren aller *probe sets* einer Verteilung um den errechneten

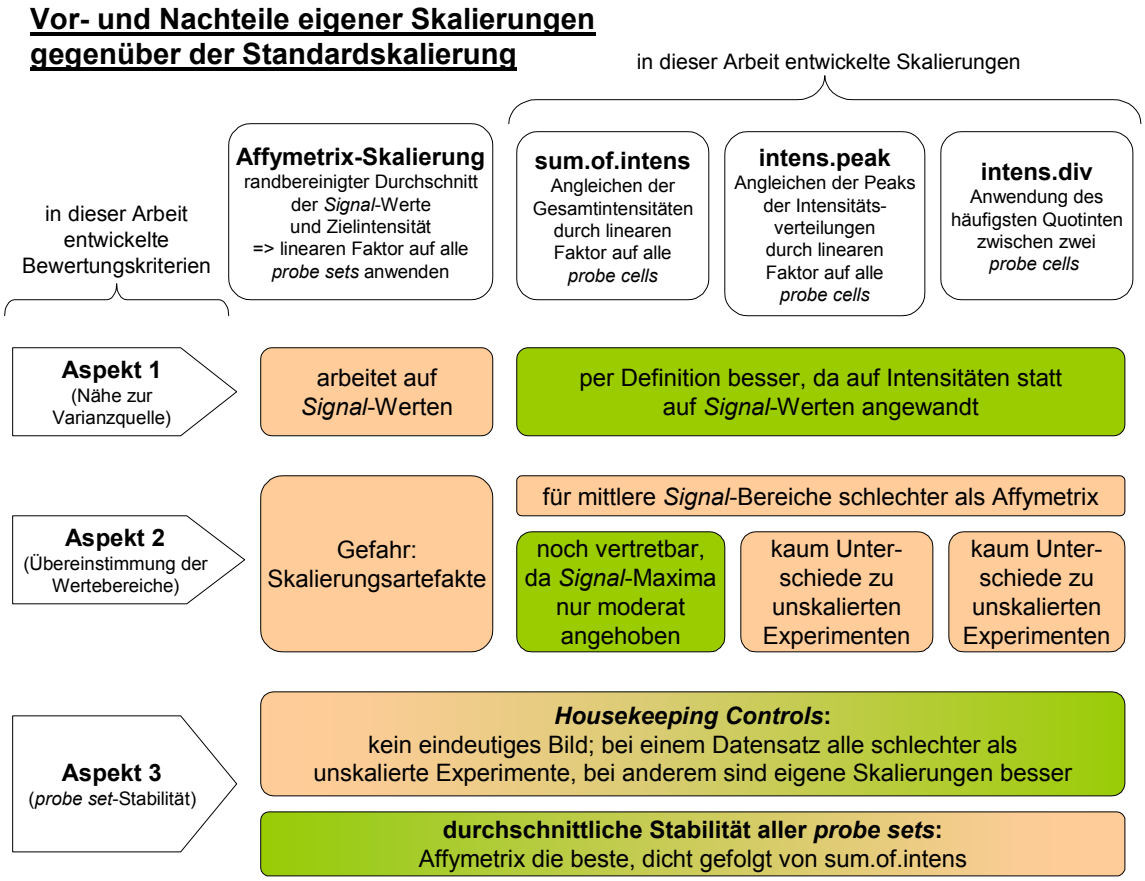
Skalierungsfaktor des Chips folgen. Dies trägt mit einem relativen Fehler von 2,52 % (siehe Beispiel in Unterkapitel 5.2) nicht unerheblich zur technischen Varianz bei. Dieses Problem lässt sich beim Arbeiten mit der Affymetrix-Software nicht umgehen, da der Kondensierungsalgorithmus fest im System integriert ist. Im Anschluss an diese Untersuchung wurde das Verhalten der Affymetrix-Skalierung beschrieben: Bei der Betrachtung der *Signal*-Verteilungen zeigte sich, dass die mittleren *Signal*-Bereiche zwar gut synchronisiert werden, die *Signal*-Maxima jedoch tendenziell den Wert des Skalierungsfaktors widerspiegeln. Die Skalierungsfaktoren korrelieren wiederum mit der Gesamtintensität, sodass Unterschiede in der globalen Varianz auf diesem Weg die Ergebnisse eines t-Tests verfälschen können. Hierdurch wurde die Empfehlung von Affymetrix verständlich, dass sich Skalierungsfaktoren von miteinander verglichenen Experimenten nicht um mehr als einen Faktor drei unterscheiden sollen. Bei der Betrachtung der *Signal*-Profile zeigte sich, dass diese vor der Skalierung tendenziell dem Verlauf der Gesamtintensitäten folgen und nach der Skalierung dem der Skalierungsfaktoren. Zusammen mit den Ergebnissen eines SOM-Clusterings vor und nach der Skalierung verdichtete sich die Befürchtung, dass durch das Auftreten von Skalierungsartefakten eine vorgespiegelte Regulation von Genen möglich wird.

Nach der Beschreibung des Verhaltens wurde der Versuch einer Bewertung der Affymetrix-Skalierung unternommen. Von den zuvor definierten Bewertungsmöglichkeiten sind nur die ersten beiden Aspekte voll anwendbar: In Bezug auf die Nähe zur Varianzquelle muss die Affymetrix-Skalierung als suboptimal gelten; sie setzt erst bei den kondensierten Maßzahlen an. Wie bei der Beschreibung des Verhaltens schon diskutiert wurde, muss der zweite Aspekt differenziert beurteilt werden: zwar werden die mittleren *Signal*-Bereiche einander gut angeglichen, die Ränder der *Signal*-Bereiche (*Signal*-Maxima) jedoch nicht. Die Möglichkeit zur Anwendung des dritten Aspektes ist, wie bereits erwähnt, durch die Konstanz der *Housekeeping*-Gene bedingt, was in Zukunft zunächst bestätigt werden muss. Unter der Annahme, dass die *Housekeeping*-Gene in den betrachteten Datensätzen stabil sind, war die Affymetrix-Skalierung bei einem der betrachteten Datensätze erfolglos und bei dem anderen teilweise erfolgreich. Wird die Annahme dahin gehend modifiziert, dass sich die Expression der meisten Gene zwischen den Experimenten eines Datensatzes nicht ändert, dann ist die Affymetrix-Skalierung bei beiden Datensätzen als erfolgreich anzusehen. Die Aspekte 4-6 konnten aufgrund fehlender Voraussetzungen nicht beurteilt werden.

Nach der Affymetrix-Skalierung wurden andere Skalierungsansätze diskutiert. Ideen für lokale und semi-lokale Verfahren bedienen sich etwa der *Housekeeping Controls*, *PolyA Controls* und / oder *Normalization Controls* als „Eichstandards“. Diese Ideen konnten wegen unklarer oder fehlender Voraussetzungen nicht realisiert werden. Diese Ansätze bergen also noch ein großes Potenzial für zukünftige Forschungen. Realisiert wurden drei globale intensitätsbasierte Skalierungsmethoden: sum.of.intens-Skalierung (Angleichung der Gesamtintensitäten), intens.peak-Skalierung (Angleichung der Intensitäts-Peaks), intens.div-Skalierung (paarweise Anwendung des häufigsten *probe cell*-Intensitätsverhältnisses). Auch diese wurden wie die Affymetrix-Skalierung im Hinblick auf die objektiven Bewertungsaspekte beurteilt: In Bezug auf den ersten Aspekt („Nähe zur Varianzquelle“) sind alle drei Skalierungsmethoden *per definition* besser als die Affymetrix-Skalierung, da sie ja gerade auf den *probe cell*-Intensitäten basieren, welche näher an den Quellen technischer Varianz liegen als die kondensierten Maßzahlen. Der zweite Aspekt betrifft die Ähnlichkeit der Wertebereiche. Er fällt für alle drei Skalierungen hinsichtlich der mittleren *Signal*-Bereiche schlechter aus als die Affymetrix-Skalierung, da sie in sich stärker unterscheidenden mittleren *Signal*-Bereichen resultieren als die Affymetrix-Skalierung. Die Unterschiede sind bei sum.of.intens allerdings noch vertretbar, zumal die Anhebung der *Signal*-Maxima moderater ausfällt als bei Affymetrix. Aspekt 3 bezüglich der Nicht-Varianz der *Housekeeping*-Gene liefert bei den betrachteten Datensätzen kein einheitliches Bild: Bei einem Datensatz sind alle Skalierungen einschließlich Affymetrix schlechter als der unskalierte Datensatz, die drei neuen Skalierungen jedoch besser als die Affymetrix-Skalierung. Für den anderen Datensatz unterscheidet sich die Bewertung zusätzlich nach der Art des *Housekeeping*-Gens: Betreffend GAPDH ist Affymetrix die beste, betreffend β-Actin ist jeweils eine der neuen Skalierungen die beste Methode. Wird die Annahme stabiler *Housekeeping*-Gene fallen gelassen und stattdessen angenommen, dass bei den Experimenten eines Datensatzes die meisten zu messenden *probe sets* nur einer geringen Regulation unterliegen, dann muss die Affymetrix-Skalierung als die beste gelten, allerdings sehr dicht gefolgt von sum.of.intens. Alle drei Aspekte zusammen genommen ergeben die Gesamteinschätzung, dass sum.of.intens der Affymetrix-Skalierung überlegen sein könnte, da sie zwar im Hinblick auf die Genstabilität geringfügig schlechter ist, aber weniger Skalierungsartefakte aufweist. Zukünftige zusätzliche Bewertungen der Aspekte 4-6

könnten dieses unklare Bild verbessern. Sofern sich die sum.of.intens-Skalierung als überlegen erweist, sollte sie statt der Affymetrix-Skalierung in den Standard-Workflow aufgenommen werden.

Die beschriebenen Erkenntnisse sind in der folgenden Übersicht zusammenfassend aufgeführt.



Insgesamt wurde auch das dritte und letzte formulierte Ziel dieser Arbeit erfüllt: die Standardskalierung wurde untersucht, und neue Skalierungen wurden entwickelt, wobei sich Hinweise für eine Verbesserung ergaben. Zusätzlich wurde ein Satz von Bewertungskriterien für Skalierungsmethoden ermittelt. Zukünftige Untersuchungen müssen sich dem Verhalten der *PolyA* und der *Normalization Controls* widmen. Diese könnten sowohl für lokale bzw. semi-lokale Skalierungen dienen, als auch die Bewertung der bereits realisierten Methoden präzisieren. Außerdem müssen sich künftige Arbeiten mit den Auswirkungen der vorgestellten Skalierungen bei Verwendung anderer

Kondensierungsalgorithmen (z. B. Li und Wong[59], weitere siehe Abschnitt 2.2.1) beschäftigen.

6.4 Ausblick

Während der Implementierungen und Auswertungen im Rahmen dieser Arbeit konnte ein stetiger Fortschritt der GeneChip-Technologie verzeichnet werden. Festmachen lässt sich dies beispielsweise an den humanen Array-Typen, von denen der zu Beginn der Arbeit aktuellste (HG-U133) nicht in die Betrachtungen eingeflossen ist. Zum jetzigen Zeitpunkt (Dezember 2003) sind bereits neue Array-Typen eingeführt worden (HG-U133 Plus 2.0), die nun nahezu das gesamte bislang charakterisierte menschliche Genom auf einem Chip abdecken. Hierfür wurde ein neuer Scanner mit höherer Auflösung notwendig. Auch die Software schritt mehrere Produktzyklen voran, auf die jeweils mit einem gewissen Anpassungsaufwand bezüglich der implementierten Funktionsbibliothek reagiert werden musste. Die einschneidendste Änderung betrifft die zukünftigen CEL-Dateien des GCOS-Systems, welche nicht mehr im Textformat, sondern binär gespeichert werden, so dass die entsprechende für diese Arbeit implementierte Funktionalität vollständig überarbeitet werden muss und nur noch nach Lizenzierung des File SDK realisiert werden kann. Im klinischen Umfeld erfordert das Schritthalten mit den rein technischen Modifikationen einen wenig fruchtbaren Ressourcenaufwand. Diese Fakten müssen berücksichtigt werden, wenn man sich auf eher methodische Themen wie die hier bearbeiteten konzentriert. Zur ausführlichen Charakterisierung der Quellen technischer Varianz muss eine Vielzahl grundlegender Experimente durchgeführt werden (z. B. Replikatexperimente, bei denen das Sample an allen möglichen Schritten im Protokoll geteilt wird), was in der Regel jedoch den Rahmen eines Projektes mit konkreter klinischer Fragestellung sprengt. Daher bietet es sich als Konsequenz an, in Zukunft eine unabhängige Instanz in der Art eines TÜVs mit diesen Untersuchungen zu betrauen, zu denen der Hersteller aufgrund der hohen Kosten die Chips beisteuern müsste.

Eine andere Konsequenz aus den Beobachtungen ist, dass die in einem Chip-Experiment gefundenen Regulationsunterschiede zunächst sorgfältig auf technische Ursachen überprüft werden müssen (z. B. mit RT-PCR). Können diese ausgeschlossen werden, muss eine funktionelle Plausibilitätsprüfung erfolgen.

Die Fokussierung auf Fragestellungen der weiter gehenden Auswertung (z. B. Clustering, *Support Vector Machines* oder verfeinerte Methoden der Klassifizierung) hat nicht in einer solchen Unmittelbarkeit mit den geschilderten Problemen zu kämpfen. Die Produktion reiner Kandidatengenlisten allein genügt den Ansprüchen an die GeneChip-Technologie heute nicht mehr. Wichtiger ist die automatisierte funktionelle Eingruppierung der identifizierten regulierten Gene, beispielsweise über die Annotation der *probe sets* mit GeneOntology-Bezeichnern, einer hierarchischen Sammlung von Funktionsbegriffen (Erläuterungen und Tools auf der Website *Gene Ontology Consortium*[38]). Eine für die funktionelle Charakterisierung regulierter Gene essenziell wichtige weiterführende Arbeit stellt die Verknüpfung von Expressionsdaten und Pathway-Informationen dar (frei verfügbare Tools: GenMAPP[29] und MAPPFinder[31]) oder die Verknüpfung von Expressionsdaten mit Protein/Protein-Interaktionsdaten (Segal et al.[81]), die über das „yeast two-hybrid"-System (Uetz et al.[89]) gewonnen wurden.

Das Aufdecken von Regulationszentren innerhalb der zellulären Prozesse ist dabei nur der erste Schritt; die eigentliche Herausforderung besteht in der Aufdeckung neuer Pathway-Komponenten durch Expressionsdaten, also die Einordnung von Genen oder Proteinen bislang unbekannter Funktion in bestehende oder gänzlich neue Pathway-Zusammenhänge und deren hierarchische Architektur. Diese Arbeit wird durch die Modellierung von Pathways als Graphen mit Knoten und Kanten auch näher an der Informatik liegen als die Expressionsanalyse, die ihrem Wesen nach eher der Biostatistik zuzuordnen ist. Der Erfolg eines solchen Ansatzes kann allerdings nur in der parallelen Nutzung vieler Datensätze liegen. Dafür werden gut gepflegte, umfangreiche Expressions-datenbanken mit Experimenten benötigt, die sinnvoll skaliert sind und deren technische Varianz gering ist, da die erwähnten Bioinformatikverfahren grundsätzlich nur so gut sein können wie die zugrunde liegenden Daten. Auch die Einbindung von Datensätzen aus der Messung von Normalgewebe (siehe HugeIndex[49], Haverty et al.[42]) ist obligat. Die Erfahrungen im Umgang mit dem LIMS SDK könnten zu diesem Zwecke umgesetzt werden in die Konzeption einer dezentralen Expressionsdatenbank unter Zuhilfenahme vernetzter LIMS-Server, so dass die Datensätze physikalisch verteilt abgelegt wären, die Auswertung jedoch zentral erfolgen würde.

Literatur

1. Affymetrix Website - Affymetrix Analysis Data Model (AADM). (*http://www.affymetrix.com/support/developer/aadm/content.affx*, zuletzt aufgerufen: 22.6.2004)

2. Affymetrix Website - Array Manufacturing. (*http://www.affymetrix.com/technology/manufacturing/index.affx*, zuletzt aufgerufen: 22.6.2004)

3. Affymetrix Website - FAQ 3'/5' factors. (*http://www.affymetrix.com/support/help/faqs/ge_assays/index.jsp*, zuletzt aufgerufen: 22.6.2004)

4. Affymetrix Website - GDAC Files SDK. (*http://www.affymetrix.com/support/developer/filesdk/GDACFiles/Pages/GDACFiles.affx*, zuletzt aufgerufen: 22.6.2004)

5. Affymetrix Website - Latin square data set. (*http://www.affymetrix.com/support/technical/sample_data/datasets.affx*, zuletzt aufgerufen: 22.6.2004)

6. Affymetrix, Inc. (*Technical Report*): Affymetrix Microarray Suite - User Guide, Version 4.0, (2000).

7. Affymetrix, Inc. (*Technical Report*): Affymetrix Data Mining Tool - User's Guide, Version 3.0, (2001).

8. Affymetrix, Inc. (*Technical Report*): Affymetrix Laboratory Information Management System (LIMS) - Installation and Administration Guide, Version 3.0, (2001).

9. Affymetrix, Inc. (*Technical Report*): Affymetrix Microarray Suite - User's Guide, version 5.0, (2001).

10. Affymetrix, Inc. (*Technical Report*): Array Design for the GeneChip Human Genome U133 Set, (2001).

11. Affymetrix, Inc. (*Technical Report*): New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays (Technical Note), (2001).

12. Affymetrix, Inc. (*Technical Report*): GeneChip® Eukaryotic Small Sample Target Labeling Assay Version II, (2002).

13. Affymetrix, Inc. (*Technical Report*): Performance and Validation of the GeneChip Human Genome U133 Set, (2002).

14. Affymetrix, Inc. (*Technical Report*): Statistical Algorithms Description Document, (2002).

15. Affymetrix, Inc. (*Technical Report*): GeneChip Expression Analysis 2003 (Technical Manual), (2003).

16. Affymetrix, Inc. (*Technical Report*): New Positive Controls for GeneChip Microarray Expression Profiling (data sheet), (2003).

17. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M., & . Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

18. Applications Support Europe. e-mail from Affymetrix Applications Lab Manager from Applications Support Europe (CALL # 44358). 2003.

19. Astrand, M. Contrast normalization of oligonucleotide arrays. *J.Comput.Biol.* **10**, 95-102 (2003).

20. Baugh, L. R., Hill, A. A., Brown, E. L., & Hunter, C. P. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res.* **29**, E29 (2001).

21. BioConductor: Open Source Software For Bioinformatics. (*http://www.bioconductor.org/*, zuletzt aufgerufen: 22.6.2004)

22. Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* **19**, 185-193 (2003).

23. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., & Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat.Genet.* **29**, 365-371 (2001).

24. Chismar, J. D., Mondala, T., Fox, H. S., Roberts, E., Langford, D., Masliah, E., Salomon, D. R., & Head, S. R. Analysis of result variability from high-density oligonucleotide arrays comparing same-species and cross-species hybridizations. *Biotechniques* **33**, 516-8, 520, 522 (2002).

25. Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., & Lockhart, D. J. Transcriptional regulation and function during the human cell cycle. *Nat.Genet.* **27**, 48-54 (2001).

26. Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K., & Kreder, D. E. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* **3**, RESEARCH0005 (2002).

27. Churchill, G. A. & Oliver, B. Sex, flies and microarrays. *Nat.Genet.* **29**, 355-356 (2001).

28. Clohessy, P. A. & Golden, B. E. The mechanism of calprotectin's candidastatic activity appears to involve zinc chelation. *Biochem.Soc.Trans.* **24**, 309S (1996).

29. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., & Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat.Genet.* **31**, 19-20 (2002).

30. Donato, R. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *Int.J.Biochem.Cell Biol.* **33**, 637-668 (2001).

31. Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., & Conklin, B. R. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene- expression profile from microarray data. *Genome Biol.* **4**, R7 (2003).

32. Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., & Coleman, P. Analysis of gene expression in single live neurons. *Proc.Natl.Acad.Sci.U.S.A* **89**, 3010-3014 (1992).

33. El Bahi, S., Caliot, E., Bens, M., Bogdanova, A., Kerneis, S., Kahn, A., Vandewalle, A., & Pringault, E. Lymphoepithelial interactions trigger specific regulation of gene expression in the M cell-containing follicle-associated epithelium of Peyer's patches. *J.Immunol.* **168**, 3713-3720 (2002).

34. Evans, S. J., Datson, N. A., Kabbaj, M., Thompson, R. C., Vreugdenhil, E., De Kloet, E. R., Watson, S. J., & Akil, H. Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression. *Eur.J Neurosci.* **16**, 409-413 (2002).

35. Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., & Adams, C. L. Multiplexed biochemical assays with biological chips. *Nature* **364**, 555-556 (1993).

36. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., & Solas, D. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767-773 (1991).

37. Geller, S. C., Gregg, J. P., Hagerman, P., & Rocke, D. M. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics.* **19**, 1817-1823 (2003).

38. Gene Ontology Consortium. (*http://www.geneontology.org/*, zuletzt aufgerufen: 22.6.2004)

39. Giles, P. J. & Kipling, D. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics.* **19**, 2254-2262 (2003).

40. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

41. Hamalainen, H. K., Tubman, J. C., Vikman, S., Kyrola, T., Ylikoski, E., Warrington, J. A., & Lahesmaa, R. Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Anal.Biochem* **299**, 63-70 (2001).

42. Haverty, P. M., Weng, Z., Best, N. L., Auerbach, K. R., Hsiao, L. L., Jensen, R. V., & Gullans, S. R. HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res.* **30**, 214-217 (2002).

43. Hill, A. A., Brown, E. L., Whitley, M. Z., Tucker-Kellogg, G., Hunter, C. P., & Slonim, D. K. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* **2**, RESEARCH0055 (2001).

44. Hobbs, J. A., May, R., Tanousis, K., McNeill, E., Mathies, M., Gebhardt, C., Henderson, R., Robinson, M. J., & Hogg, N. Myeloid cell function in MRP-14 (S100A9) null mice. *Mol.Cell Biol.* **23**, 2564-2576 (2003).

45. Hocquette, J. F. & Brandstetter, A. M. Common practice in molecular biology may introduce statistical bias and misleading biological interpretation. *J Nutr Biochem* **13**, 370-377 (2002).

46. Hoffmann, R., Seidl, T., & Dugas, M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.* **3**, RESEARCH0033 (2002).

47. Huang, X. & Pan, W. Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Funct.Integr.Genomics* **2**, 126-133 (2002).

48. Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A., & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.* **18 Suppl 1**, S96-S104 (2002).

49. HuGE Index. (*http://www.hugeindex.org*, zuletzt aufgerufen: 22.6.2004)

50. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).

51. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249-264 (2003).

52. Kerkhoff, C., Klempt, M., Kaever, V., & Sorg, C. The two calcium-binding proteins, S100A8 and S100A9, are involved in the metabolism of arachidonic acid in human neutrophils. *J.Biol.Chem.* **274**, 32672-32679 (1999).

53. Kerneis, S., Bogdanova, A., Kraehenbuhl, J. P., & Pringault, E. Conversion by Peyer's patch lymphocytes of human enterocytes into M cells that transport bacteria. *Science* **277**, 949-952 (1997).

54. Klein et. al (2001) J. Exp. Med. 194:1625-1638 - "Gene Expression Profiling of B-cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells" - Source hybridization data (CEL files). (*http://icg.cpmc.columbia.edu/supplement_RDF/Klein_JEM_2001.htm*, zuletzt aufgerufen: 22.6.2004)

55. Klein, U., Tu, Y., Stolovitzky, G. A., Mattioli, M., Cattoretti, G., Husson, H., Freedman, A., Inghirami, G., Cro, L., Baldini, L., Neri, A., Califano, A., & Dalla-Favera, R. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp.Med.* **194**, 1625-1638 (2001).

56. Kraehenbuhl, J. P. & Neutra, M. R. Epithelial M cells: differentiation and function. *Annu.Rev.Cell Dev.Biol.* **16**, 301-332 (2000).

57. Kricka, L. J. Stains, labels and detection strategies for nucleic acids assays. *Ann.Clin.Biochem.* **39**, 114-129 (2002).

58. 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., Raymond, C., 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, B. 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. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

59. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc.Natl.Acad.Sci.U.S.A* **98**, 31-36 (2001).

60. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nat.Genet.* **21**, 20-24 (1999).

61. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., & Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat.Biotechnol.* **14**, 1675-1680 (1996).

62. Lorkowski, S. H. & Cullen, P. H., *Analysing Gene Expression*, Wiley-VCH (2002).

63. Mahadevappa, M. & Warrington, J. A. A high-density probe array sample preparation method using 10- to 100- fold fewer cells. *Nat.Biotechnol.* **17**, 1134-1136 (1999).

64. Manitz, M. P., Horst, B., Seeliger, S., Strey, A., Skryabin, B. V., Gunzer, M., Frings, W., Schonlau, F., Roth, J., Sorg, C., & Nacken, W. Loss of S100A9 (MRP14) results in reduced interleukin-8-induced CD11b surface expression, a polarized microfilament system, and diminished responsiveness to chemoattractants in vitro. *Mol.Cell Biol.* **23**, 1034-1043 (2003).

65. Nacken, W., Roth, J., Sorg, C., & Kerkhoff, C. S100A9/S100A8: Myeloid representatives of the S100 protein family as prominent players in innate immunity. *Microsc.Res.Tech.* **60**, 569-580 (2003).

66. Naef, F., Hacker, C. R., Patil, N., & Magnasco, M. Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* **3**, RESEARCH0018 (2002).

67. Naef, F., Socci, N. D., & Magnasco, M. A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics.* **19**, 178-184 (2003).

68. National Center for Biotechnology Information - Entrez Genome. (*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome*, zuletzt aufgerufen: 22.6.2004)

69. National Center for Biotechnology Information - Entrez Nucleotide. (*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide*, zuletzt aufgerufen: 22.6.2004)

70. Ohyama, H., Zhang, X., Kohno, Y., Alevizos, I., Posner, M., Wong, D. T., & Todd, R. Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization. *Biotechniques* **29**, 530-536 (2000).

71. Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., & Simon, R. Evaluation of normalization methods for microarray data. *BMC.Bioinformatics.* **4**, 33 (2003).

72. Piper, M. D., Daran-Lapujade, P., Bro, C., Regenberg, B., Knudsen, S., Nielsen, J., & Pronk, J. T. Reproducibility of Oligonucleotide Microarray Transcriptome Analyses. AN INTERLABORATORY COMPARISON USING CHEMOSTAT CULTURES OF SACCHAROMYCES CEREVISIAE. *J.Biol.Chem.* **277**, 37001-37008 (2002).

73. The R Project for Statistical Computing. (*http://www.r-project.org/*, zuletzt aufgerufen: 22.6.2004)

74. Rajagopalan, D. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics.* **19**, 1469-1476 (2003).

75. Rammes, A., Roth, J., Goebeler, M., Klempt, M., Hartmann, M., & Sorg, C. Myeloid-related protein (MRP) 8 and MRP14, calcium-binding proteins of the S100 family, are secreted by activated monocytes via a novel, tubulin-dependent pathway. *J.Biol.Chem.* **272**, 9496-9502 (1997).

76. RefSeq. (*http://www.ncbi.nlm.nih.gov/RefSeq/*, zuletzt aufgerufen: 22.6.2004)

77. Sasik, R., Calvo, E., & Corbeil, J. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics.* **18**, 1633-1640 (2002).

78. Schadt, E. E., Li, C., Su, C., & Wong, W. H. Analyzing high-density oligonucleotide gene expression array data. *J.Cell Biochem.* **80**, 192-202 (2000).

79. Schmittgen, T. D. & Zakrajsek, B. A. Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. *J Biochem Biophys.Methods* **46**, 69-81 (2000).

80. Schulze, A. & Downward, J. Navigating gene expression using microarrays--a technology review. *Nat.Cell Biol.* **3**, E190-E195 (2001).

81. Segal, E., Wang, H., & Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics.* **19 Suppl 1**, I264-I272 (2003).

82. Sohnle, P. G., Hunter, M. J., Hahn, B., & Chazin, W. J. Zinc-reversible antimicrobial activity of recombinant calprotectin (migration inhibitory factor-related proteins 8 and 14). *J.Infect.Dis.* **182**, 1272-1275 (2000).

83. Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr., & Brazma, A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, RESEARCH0046 (2002).

84. Strand, A. D., Olson, J. M., & Kooperberg, C. Estimating the statistical significance of gene expression changes observed with oligonucleotide arrays. *Hum.Mol.Genet.* **11**, 2207-2221 (2002).

85. Strupat, K., Rogniaux, H., Van Dorsselaer, A., Roth, J., & Vogl, T. Calcium-induced noncovalently linked tetramers of MRP8 and MRP14 are confirmed by electrospray ionization-mass analysis. *J.Am.Soc.Mass Spectrom.* **11**, 780-788 (2000).

86. Talapatra, A., Rouse, R., & Hardiman, G. Protein microarrays: challenges and promises. *Pharmacogenomics.* **3**, 527-536 (2002).

87. Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P., & Cho, K. W. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.* **30**, e54 (2002).

88. Tsuji, N., Kamagata, C., Furuya, M., Kobayashi, D., Yagihashi, A., Morita, T., Horita, S., & Watanabe, N. Selection of an internal control gene for quantitation of mRNA in colonic tissues. *Anticancer Res.* **22**, 4173-4178 (2002).

89. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

90. UniGene. (*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene*, zuletzt aufgerufen: 22.6.2004)

91. Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).

92. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F., V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., & Nodell, M. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

93. Vogl, T., Roth, J., Sorg, C., Hillenkamp, F., & Strupat, K. Calcium-induced noncovalently linked tetramers of MRP8 and MRP14 detected by ultraviolet

matrix-assisted laser desorption/ionization mass spectrometry. *J.Am.Soc.Mass Spectrom.* **10**, 1124-1130 (1999).

94. Wang, X., Hessner, M. J., Wu, Y., Pati, N., & Ghosh, S. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics.* **19**, 1341-1347 (2003).

95. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., LeVine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C., & Lander, E. S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

96. Welle, S., Brooks, A. I., & Thornton, C. A. Computational method for reducing variance with Affymetrix microarrays. *BMC.Bioinformatics.* **3**, 23 (2002).

97. Wilson, D. L., Buckley, M. J., Helliwell, C. A., & Wilson, I. W. New normalization methods for cDNA microarray data. *Bioinformatics.* **19**, 1325-1332 (2003).

98. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., & Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).

99. Zhou, Y. & Abagyan, R. Algorithms for high-density oligonucleotide array. *Curr.Opin.Drug Discov.Devel.* **6**, 339-345 (2003).

Anhang

Im Folgenden finden sich detaillierte Ausführungen zur Funktionsweise und zu den Schnittstellen und Parameterlisten der implementierten SPLUS-Funktionen. Die Gliederung in Unterkapitel folgt dabei der Gliederung nach Funktionsgruppen in Abschnitt 3.4.2.

A.1 Funktionen zur Handhabung der Merkmale von GeneChip-Experimenten

`import.Experiment.descriptions`: Importieren eines desc.df aus einer Excel-Datei oder einer Access-Datenbank unter Angabe des Namens, unter dem der desc.df in der working-Datenbank abgespeichert wird und unter Angabe von Dateipfad und -name.

`get.experiments`: Auswählen einer Untergruppe von Experimenten durch Übergabe von Ein- und Ausschlusskriterien für bestimmte Merkmale (Parameter `properties, include, values`). Rückgabe eines Merkmals aller Experimente aus der Untergruppe (Parameter `return.property`).

`get.expnames.from.N (hybrid.name,expnumbers)`: Spezialfall von `get.experiments` für `properties = c(„N"), include = c(T), values = list(expnumbers), return.property = „EXPERIMENT.NAME"`: Alle Experimente einschließen, die in der Spalte `N` einen Wert haben, der Element von `expnumbers` ist. Von diesen Experimenten wird der Wert in der Spalte `EXPERIMENT.NAME` zurückgegeben.

`get.experiments.of.groups`: Wie `get.experiments`, aber unter Angabe der Gruppennummern, deren Experimente eingeschlossen werden sollen.

`get.label.of.experiment.in.grouping`: Für die spezifizierten Experimente werden die Experimentbezeichner einer Gruppierung zurückgegeben.

`get.chip.id.of.experiments`: Bei Angabe der Experimentnamen oder -nummern werden über eine C++-Funktion mithilfe des LIMS SDK die `chip.ids` der Experimente zurückgegeben.

`statistic.of.property.of.groups`: Gibt einfache statistische Eigenschaften (Minimum, Mean, Median, Maximum) eines Merkmals aus. Wenn `sep.groups==F`, dann werden alle Experimente der spezifizierten Gruppen berücksichtigt. Wenn `sep.groups==T`, dann wird für jede Gruppe eine eigene Statistik ausgegeben.

`add.property.col.with.Evaluator`: Hinzufügen von Spalten zum desc.df mit Merkmalen von spezifizierten Experimenten. Die Berechnung der Werte findet dabei serverseitig statt clientseitig statt. Eine ausführliche Beschreibung des Client-Server-Konzeptes und der Funktion findet sich im Unterkapitel 3.6.

A.2 Funktionen in Zusammenhang mit Gruppierungen und Kombinationen

`add.grouping`: Hinzufügen einer Gruppierung zu einem desc.df- oder einem groupobj-Objekt. Ausführliche Beschreibung siehe Unterkapitel 3.3.

`remove.grouping`: Löschen einer Gruppierung aus einem groupobj-Objekt. Ausführliche Beschreibung siehe Unterkapitel 3.3.

`get.desc.df (hybrid.name)`: Rückgabe des desc.df mit dem übergebenen Namen oder des erweiterten desc.df des groupobj-Objektes mit dem übergebenen Namen.

`get.df.force.groupobj.name (groupobj.name)`: wie `get.desc.df`, aber nur Angabe des Namens eines groupobj-Objektes erlaubt.

`get.clean.desc.df (groupobj.name)`: Rückgabe des desc.df ohne Gruppierungsspalten des groupobj-Objektes mit dem übergebenen Namen. Fehlermeldung, wenn übergebener Name keinem groupobj-Objekt entspricht.

`get.groupobj (groupobj.name)`: Rückgabe des groupobj-Objektes mit dem übergebenen Namen.

`summary.of.groups`: Ausgabe von Gruppierungsinformationen. Dazu gehören die Anzahl der Gruppen, die Gruppierungsmerkmale, die Gruppenbezeichner und die Ausprägung der Gruppierungsmerkmale innerhalb jeder Gruppe. Darüber hinaus werden die Experimentbezeichner zusammen mit weiteren übergebenen Merkmale des desc.df ausgegeben.

`get.names.of.groups`: Rückgabe der Gruppenbezeichner der spezifizierten Gruppen.

`get.group.prop.label`: Rückgabe der Merkmalsausprägungen der spezifizierten Gruppen als zusammenhängender String.

`create.combinations`: Berechnen einer comb-Struktur, welche Kombinationen innerhalb einer Gruppe speichert. Ausführliche Beschreibung siehe Unterkapitel 3.3.

`apply.to.combinations (comb.struct, groupnos="@ALL", FUN, ...)`: Anwenden der übergebenen Funktion `FUN` auf die Kombinationen der Gruppen mit den Nummern `groupnos`. Ausführliche Beschreibung siehe Unterkapitel 3.3.

A.3 Funktionen zum Umgang mit Intensitäten

`initialize.intens`: Initialisierungen für die Intensitätsalgorithmen, u.a. Festlegung des default-Pfades der CEL-Dateien der *process*-Datenbank des LIMS-Systems.

`evtl.load.intens.by.expname, evtl.load.intens.by.N`: Einladen der Intensitätsdaten des spezifizierten Experimentes (Spalten X, Y, MEAN, STDV, PIXELS der CEL-Datei durch die Unterfunktion `import.intens.byFilename`) und Rückgabe einer Liste mit diesen Komponenten. Ist der logische Parameter `persist==T`, werden neben den eigentlichen Intensitätsdaten auch die Informationen der HEADER-, MASKED-, OUTLIER- und MODIFIED-Bereiche der CEL-Datei importiert (Funktionen `import.header.by.filename` und `import.masked.outlier.modified.by.filename`) und in der intensity-Datenbank von

SPLUS lokal gespeichert. Beim nächsten Aufruf der Funktion wird zunächst geprüft, ob bereits eine lokale Kopie vorliegt und dann aus Geschwindigkeitsgründen diese zur Rückgabe verwendet.

`get.intens.vector`: Rückgabe der MEAN-Spalte einer CEL-Datei eines Experimentes. Dabei ist unerheblich, ob das Experiment bereits lokal vorliegt oder noch aus dem LIMS geladen werden muss. Über den Parameter CEL.postfix können dabei auch durch andere Funktionen (z. B. `apply.SF.to.intens`) modifizierte Intensitätsdaten bzw. CEL-Dateien adressiert werden.

`create.CHIP.DESIGN.table`: Initialisierung des CHIP.DESIGN-Objektes (Näheres siehe Abschnitt 3.2.1). Für die Arrays mit den Chip-IDs 1,…,`max.chip.id` werden die entsprechenden CDF-Dateien im Verzeichnis `CDF.filepath` verwendet, um die Initialisierungsinformationen zusammenzustellen.

`import.chip.layout`: Bei Übergabe der maximal vorkommenden `chip.id` (aktuell: 26 für „HG-U133A“) und dem Pfad mit den CDF-Dateien wird für jeden vorhandenen Array-Typ das Chip-Layout eingelesen und im entsprechenden `CHIP.LAYOUT.`*chip_id*-Objekt in der intensity-Datenbank von SPLUS gespeichert. Neben dem Layout wird auch die Anzahl der *probe sets* des Arrays und die Anzahl der *Quality Features* ermittelt. Vor Aufruf dieser Funktion muss `create.CHIP.DESIGN.table` aufgerufen worden sein, damit die Ausmaße jedes Array-Typs bekannt sind. Nach Hinzufügen eines neuen Array-Typs zum LIMS-System muss die Funktion erneut aufgerufen werden. Über den Aufruf einer C++-Routine (`get_cell_tagsC`) ist das effiziente Einlesen der sehr großen CDF-Datei möglich. Mithilfe der Funktion `convert.chip.id.to.CDF.filename` wird eine `chip.id` in einen CDF-Dateinamen konvertiert. Eine Liste der `chip.ids` mit den zugehörigen Array-Namen findet sich in Tabelle 11.

`get.CEL.positions (chip.id, UNITS)`: Übergabe des numerischen Bezeichners `chip.id` des Array-Typs (Näheres siehe Abschnitt A.7). Mögliche Werte des Parameters `UNITS` siehe Tabelle 12. Zurückgegeben wird ein T/F-Vektor mit gleicher Länge wie die MEAN-Spalte aus der CEL-Datei. Ein Eintrag `T` besagt, dass die entsprechende *probe cell* zur durch `UNITS` spezifizierten Untermenge gehört.

`apply.SF.to.intens (SF, ret.list, new.intens.postfix)`: Benutzerdefiniertes Skalieren auf Intensitätsebene: Anwenden des linearen Faktors `SF` auf die Intensitäten der spezifizierten Experimente. Wenn `ret.list==T`, dann wird das resultierende intens-Objekt zurückgegeben. sonst wird es in der intensity-Datenbank gespeichert. Über die Angabe des Parameters `new.intens.postfix` besteht dabei die Möglichkeit, die Skalierungsmethode und den Skalierungsfaktor im Namen zu kodieren. Im Weiteren können auf diese Weise sowohl in der intensity-Datenbank von SPLUS als auch (nach `export.CEL.file` und `perform.LIMS.CEL.file.import`) in der LIMS-*process*-Datenbank mehrere CEL-Dateien pro Experiment erzeugt und gespeichert werden.

`export.CEL.file`: Export von CEL-Dateien spezifizierter Experimente. Über den Parameter `LIMS.postfix` können dabei die mit anderen Funktionen (z. B. `apply.SF.to.intens`) modifizierten Intensitätsdaten, die mit einem Postfix in der intensity-Datenbank abgelegt wurden, als MEAN-Spalte eingesetzt werden. Der Parameter `CEL.postfix` ermöglicht die Angabe eines Postfix für die CEL-Dateien, die im Verzeichnis `export.path` abgelegt werden.

`perform.LIMS.CEL.file.import`: Für die spezifizierten Experimente werden zusätzlich zu den vorhandenen CEL-Dateien weitere CEL-Dateien mit dem Postfix `CEL.postfix`, die im Verzeichnis `source.CELpath` liegen, in die LIMS-*process*-Datenbank importiert.

`create.intens.div`: Bei Übergabe zweier Experimente wird mithilfe der Funktion `get.intens.div.by.expnames` ein Vektor der komponentenweisen Quotienten der Intensitäten (`intens.div`) berechnet. Zusätzlich hierzu enthält die Rückgabe den Bezeichner <*zaehler*>`.div.`<*nenner*> dieses `intens.div`-Vektors. Das ist gerade dann sinnvoll, wenn

durch den Parameter `positive=T` gefordert wird, dass der Intensitätsvektor mit dem größeren Durchschnittswert den Zähler bildet. Die Übergabe der Experimentbezeichner findet über eine Liste `id` mit zwei Komponenten statt (daher auf Kombinationsobjekten anwendbar). Diese Funktion wird bei der `intens.div`-Skalierung verwendet.

`add.sum.of.intens.col`: Berechnung von Gesamtintensitäten: Hinzufügen einer Spalte `sum.of.intens` zum desc.df oder groupobj-Objekt. Für die spezifizierten Experimente wird die Summe der Intensitäten der mit dem Parameter `UNITS` angegebenen *probe cells* in diese Spalte geschrieben.

`add.mean.of.intens.col`, `add.trimmed.mean.of.intens.col`, `add.median.of.intens.col`: wie `add.sum.of.intens.col`, jedoch Berechnung des Durchschnitts bzw. des Durchschnitts ohne die oberen und unteren zwei Prozent der Werte bzw. des Medians.

`add.sum.of.spechyb.col`: Berechnung der Gesamtintensität der spezifischen Hybridisierung: Hinzufügen einer Spalte `sum.of.spechyb` zum desc.df oder groupobj-Objekt. Berechnung der Summe der PM-MM-Differenzen aller *probe pairs*. Dazu wird als Unterfunktion `create.PM.aligned.MM.vec` aufgerufen.

`add.col.of.combined.properties (col.name, property.names, COMB.FUN)`: Hinzufügen einer neuen Spalte `col.name`, die durch die komponentenweise Anwendung der Funktion COMB.FUN auf die Spalten `property.names` entsteht.

`calc.mean.of.probe.cells.of.group`: Berechnen von Durchschnitt und Varianz der Intensitäten über die Experimente der spezifizierten Gruppen. Funktion kann bei der Suche nach eigenen *PolyA Controls* genutzt werden (Näheres siehe Unterkapitel 5.1).

`calc.intens.distances`: Berechnen der Distanz zwischen den Intensitätsvektoren (nähere Spezifikation über den Parameter `UNITS`) zweier Experimente. Übergabe der Experimentnummern durch eine Liste `id` mit zwei Komponenten (daher auf Kombinationsobjekten anwendbar). Mögliche Distanzfunktionen: "`euclidean`", "`maximum`", "`manhattan`".

A.4 Funktionen zum Import / Export von Primäranalysen

`initialize.LIMS.import`: Initialisierungen für die nachfolgend beschriebenen Import-Algorithmen. Die data.frames mit Informationen zu der mit dem Parameter `Publish.DB.name` spezifizierten *publish*-Datenbank (siehe Abschnitt 3.2.3) werden angelegt bzw. aktualisiert. Nach jeglicher Änderung an zu benutzenden *publish*-Datenbanken muss diese Funktion aufgerufen werden, um die entsprechenden data.frames zu initialisieren.

`names.from.ANALYSIS`: Extrahieren der Bezeichner der Analysen aus der ANALYSIS-Tabelle, Ignorieren der Experimentbezeichner.

`import.LIMS.Analysis.byName`: Unter Angabe des Analysenamens und der *publish*-Datenbank, in der sie gespeichert ist, werden die *probe set*-Bezeichner und die Maßzahlen `Signal`, `Detection p_value`, `Abs.Call`, `Pairs` und `Pairs.Used` einer Primäranalyse importiert und abhängig von einem Parameter `store.in.DB` entweder in der analysis-Datenbank von SPLUS gespeichert oder als data.frame als Funktionsergebnis zurückgegeben. Zusätzlich kann über die Parameter `get.descriptions` und `get.accessions` angegeben werden, ob die *probe set*-

Beschreibungen und die „accession number“ der NCBI-Sequenzdatenbank als Spalte hinzugefügt werden sollen.

`import.LIMS.Analyses`: Ähnlich wie `import.LIMS.Analysis.byName`, jedoch können unter Angabe von mehreren Experimentnamen oder -nummern und unter Angabe eines desc.df oder groupobj-Objekt die Primäranalysen mehrerer Experimente auf einmal importiert werden.

`import.LIMS.Analyses.of.Project`: Unter Angabe des Projektnamens der *process*-Datenbank und des Namens der *publish*-Datenbank, in der die Primäranalysen gespeichert sind, werden die *probe set*-Bezeichner und die durch den Parameter `measures` spezifizierten Maßzahlen aller Primäranalysen des Projekts in einem data.frame in den Spalten *analysis_name*`.`*measure.name* gesammelt und zurückgegeben.

`export.LIMS.Analyses`: Unter Angabe der Analysenamen, eines Dateityps („`EXCEL`“ oder „`ASCII`“) und eines Pfades können Primäranalysen als Excel-Datei oder als tabulatorgetrennte Textdatei ins Dateisystem gespeichert werden. Über den boolschen Parameter `LIMS.names.given` lässt sich angeben, ob dabei die Analysenamen im LIMS-Format oder im SPLUS-Format angegeben werden. Alternativ zu den Analysenamen können auch Experimentnummern, ein desc.df oder groupobj-Objekt und eine *publish*-Datenbank übergeben werden, und damit alle Primäranalysen dieser Experimente in der *publish*-Datenbank spezifiziert werden. Weiterhin kann ein Vektor von Dateinamen übergeben werden, um die Benennung im Dateisystem anzupassen.

A.5 Funktionen zum Umgang mit Primäranalysen

`import.analyses.descriptions`: Unter Angabe eines desc.df oder groupobj-Objekt werden die Namen aller Primäranalysen in der *process*-Datenbank ermittelt und in einem data.frame gespeichert, die von Experimenten in diesem desc.df oder groupobj-Objekt durchgeführt wurden. Zusätzlich zu den Namen werden noch die grundlegenden Parameter „Baseline File“ und „Normalization Factor“ und/oder „Target Intensity“ und „Scale Factor“ einer Primäranalyse in den Spalten `BF`, `NF`, `TGT` und `SF` aufgeführt. Die Spalte `emp.flag` gibt an, ob es sich um eine empirische Primäranalyse (MAS-Version 4) oder um eine statistische Primäranalyse handelt (MAS-Version 5). Über einen Parameter `ret.df` kann gesteuert werden, ob der erzeugte data.frame als Ergebnis der Funktion zurückgegeben oder unter dem Namen `analyses.`*hybridname* in der analysis-Datenbank von SPLUS gespeichert werden soll.

`get.analyses.of.experiment`: Mithilfe des data.frame, der von der Funktion `import.analyses.descriptions` angelegt wurde, wird das durch den Parameter `return.property` angegebene Merkmal (in der Regel `ANALYSIS.NAME`) der Primäranalysen zurückgegeben, die einer Untermenge aller Primäranalysen der durch die Experimentnamen oder Experimentnummern spezifizierten Experimente angehören. Die Untermenge wird dabei durch zwei Vektoren `properties` und `include` und eine Liste aus Vektoren `values` definiert. Eine Analyse gehört zu der Untermenge, wenn ihr Merkmal `properties[i]` eine Bedingung erfüllt, die durch `include[i]` und den Vektor `values[i]` näher spezifiziert ist. Bei `include[i]==T` wird die Analyse in die Untermenge eingeschlossen, wenn ihr Merkmal `properties[i]` eine der Ausprägungen aus `values[i]` annimmt; bei `include[i]==F` wird sie nicht eingeschlossen. Bei `include[i]=='S'` wird sie eingeschlossen, wenn `properties[i]` eine der Komponenten aus `values[i]` als Teil-String enthält.

`get.published.analyses.of.experiment`: Unter Angabe einer *publish*-Datenbank, von Experimentnamen oder -nummern und gegebenenfalls Angabe eines desc.df oder groupobj-Objekt werden die Namen aller in der entsprechenden *publish*-Datenbank befindlichen Primäranalysen dieser Experimente zurückgegeben.

`import.published.analysis.names`, `import.published.experiment.names`: Unter Angabe einer *publish*-Datenbank werden die Namen der in ihr gespeicherten Primäranalysen zurückgegeben bzw. die der Experimente, für die Primäranalysen gespeichert sind.

`map.SF.from.analyses.desc.df.to.hybrid`: Bei Spezifikation des `analyses.desc.dfname` und einer Untermenge der Primäranalysen (mithilfe der Parameter `properties`, `include`, `values`) werden die Skalierungsfaktoren in eine (neue) Spalte des (erweiterten) `desc.df` geschrieben. Das Postfix dieser Spalte wird durch den Parameter `desc.df.col.postfix` angegeben und wird notwendig, wenn die Skalierungsfaktoren verschiedener Skalierungen abgespeichert werden.

`add.min.of.measurements.col`: Bei Angabe eines Maßzahlbezeichners (z. B. „`Signal`“) wird das Minimum dieser Maßzahl für jedes Experiment ermittelt und zum spezifizierten `analysis`-Objekt als Spalte hinzugefügt. Auch ein Präfix des Spaltenbezeichners (`colprefix`) und ein Parameter `postfix` für den Experimentnamen (zur Berücksichtigung eigener Skalierungen) kann übergeben werden. Mit dem Parameter `add.to.groupobj==T` kann die Spalte statt zum `analysis`-Objekt auch zum `desc.df` hinzugefügt werden. Um Mehrdeutigkeiten zu vermeiden, gilt dann das übergebene Postfix auch als Postfix für den Spaltenbezeichner.

`add.max.of.measurements.col`, `add.mean.of.measurements.col`, `add.median.of.measurements.col`, `add.sum.of.measurements.col`, `add.trimmed.mean.of.measurements.col`: Wie `add.min.of.measurements.col`, jedoch statt Minimum Berechnung von Maximum, Durchschnitt, Median, Summe und randbereinigtem Durchschnitt (ohne obere und untere zwei Prozent der Werte).

A.6 Graphenfunktionen

`bar.plot.of.property.of.groups`: Für die Experimente der spezifizierten Gruppen wird ein Balkendiagramm des Merkmals `property.name` aus dem `desc.df` gezeichnet. Mit `sep.groups==T` wird ein Graph pro Gruppe gezeichnet. Bei Angabe der Parameter `properties`, `include`, `values` wird das Merkmal für die spezifizierten Analysen aus dem `analysis.desc.df` gezeichnet. Über den Parameter `yaxs.lim.MAX` kann der y-Achsenabschnitt gesetzt werden oder wird automatisch auf den maximal vorkommenden Wert eingestellt.

`scatter.plot.properties.of.groups`: Für die Experimente der spezifizierten Gruppen wird ein Scatter Plot der zwei anzugebenden Merkmale (`prop1.name`, `prop2.name`) gezeichnet. Bei `sep.groups==T` wird jede Gruppe in einen getrennten Graphen gezeichnet. Wird `fit.line==T` angegeben, wird zusätzlich die lineare Regressionsgerade eingezeichnet und Steigung *m*, y-Achsenabschnitt *b* und Korrelationskoeffizient *r* unterhalb des Graphen angegeben.

`histogram.of.property.of.groups`: Für die Experimente der spezifizierten Gruppen wird ein Histogramm des anzugebenden Merkmals (`property.name`) gezeichnet. Bei `sep.groups==T` wird jede Gruppe in einen getrennten Graphen gezeichnet. Bei `color.alt==T` wird die Farbe pro

Experiment alteriert. Über den Parameter `inside==T`, der über den „...“-Mechanismus an die eigentliche SPLUS-Funktion `histogram` durchgereicht werden kann, kann statt einzelnen Balken nur ein Umriss gezeichnet werden. Dies bietet sich bei Überlagerung mehrerer Histogramme an.

`box.plot.intens.of.groups`: Für die Experimente der spezifizierten Gruppen wird ein Box Plot der *probe cell*-Intensitäten gezeichnet. Mit dem Parameter `UNITS` wird die Untermenge der *probe cells* definiert. Dabei kann zusätzlich zu den in Tabelle 12 vorgestellten Werten auch `@spechyb` verwendet werden, um einen Box Plot der PM-MM-Differenzen (idealerweise die spezifische Hybridisierung) zu erhalten. Über den Parameter `CEL.postfix` können die Intensitäten eigener Skalierungen angesprochen werden. Der durchgereichte Parameter `ylim` kann zur Bestimmung des y-Achsenbereiches verwendet werden.

`histogram.of.intens.of.groups`: Für die Experimente der spezifizierten Gruppen wird ein Histogramm der *probe cell*-Intensitäten gezeichnet. Mit dem Parameter `UNITS` wird die Untermenge der *probe cells* definiert. Dabei kann zusätzlich zu den in Tabelle 12 vorgestellten Werten auch `@spechyb` verwendet werden, um ein Histogramm der PM-MM-Differenzen (idealerweise die spezifische Hybridisierung) zu erhalten. Gilt `apply.col==T`, kann über die Parameter `properties`, `include`, `values` eine Spalte aus einem `analysis`-Objekt spezifiziert werden, mit deren Wert der Intensitätsvektor vorher multipliziert wird. Dies ermöglicht das Zeichnen eines Histogramms mit angewandtem Skalierungsfaktor ohne konkretes Durchführen einer eigenen Skalierung. Über den Parameter `inside==T` kann statt einzelner Balken nur ein Umriss gezeichnet werden. Dies bietet sich bei Überlagerung mehrerer Histogramme an.

`scatter.plot.intensities`: Bei Spezifikation zweier Experimente wird ein Scatter Plot der beiden Intensitätsvektoren gezeichnet. Mit dem Parameter `UNITS` (siehe Tabelle 12) wird die Untermenge der *probe cells* definiert. Wird `fit.line==T` angegeben, wird zusätzlich die lineare Regressionsgerade eingezeichnet und Steigung *m*, y-Achsenabschnitt *b* und Korrelationskoeffizient *r* unterhalb des Graphen angegeben. Die Übergabe der Experimentbezeichner findet über eine Liste `id` mit zwei Komponenten statt, daher ist diese Funktion auf Kombinationsobjekten anwendbar.

`histogram.of.measurement`: Für die Experimente der spezifizierten Gruppen wird ein Histogramm der angegebenen Maßzahl (`analysis.colname`) aus dem `analysis`-Objekt gezeichnet. Spezifikation der Analysen über die Parameter `properties`, `include`, `values`. Über den Parameter `inside==T` kann statt einzelnen Balken nur ein Umriss gezeichnet werden. Dies bietet sich bei Überlagerung mehrerer Histogramme an.

`box.plot.of.measurement`: Für die Experimente der spezifizierten Gruppen wird ein Box Plot der angegebenen Maßzahl (`analysis.colname`) aus dem `analysis`-Objekt gezeichnet. Spezifikation der Analysen über die Parameter `properties`, `include`, `values`. Mit `sep.groups==T` wird ein Graph pro Gruppe gezeichnet.

`intens.div.plot.histograms.body`: Bei Spezifikation zweier Experimente wird mithilfe der Funktion `create.intens.div` ein Histogramm der Intensitätsquotienten gezeichnet. Dabei bildet bei positive==T der Intensitätsvektor mit dem größeren Durchschnittswert den Zähler. Über den Parameter `postfix` können die Intensitäten eigener Skalierungen angesprochen werden.

A.7 Sonstige Funktionen

`.First`: Umschreiben der ursprünglich leeren `.First`-Funktion, die bei jedem Start des SPLUS-Systems aufgerufen wird. Hiermit wird die intensity- bzw. die analysis-Datenbank auf der Hierarchieebene zwei bzw. drei angelegt und die globalen Variablen `default.intens.DB` bzw. `default.analysis.DB` auf zwei bzw. drei gesetzt.

`convert.LIMS.name.to.SPLUS.name (LIMSname)`: Konvertieren von LIMS- oder Dateisystembezeichnern in SPLUS-Bezeichner: Bindestriche und Unterstriche in `LIMSname` werden umgewandelt in Punkte. So wird beispielsweise ein Experimentbezeichner `26_102_150301` (welcher keinen gültigen SPLUS-Bezeichner darstellt) zu `26.102.150301`.

`convert.chip.id.to.ARRAY.TYPE`: Umwandlung einer `chip.id` in den entsprechenden Array-Typ im String-Format (siehe Tabelle 11).

`convert.ARRAY.TYPE.to.chip.id`: Umwandlung eines Array-Typs im String-Format in die entsprechende `chip.id` (siehe Tabelle 11).

`convert.chip.id.to.CDF.filename`: Umwandlung einer `chip.id` in den Bezeichner der entsprechenden CDF-Datei (siehe Tabelle 11 plus Dateierweiterung `.CDF`).

`convert.UNIT.number.to.probe.set.name`: Wandelt den im CDF-Format für ein *probe set* verwendeten UNIT-Bezeichner unter Ausnutzung des `PSI.info`-Objekts in einen *probe set*-Bezeichner um.

`convert.item.id.to.probe.set.name`,
`get.bio.item.id.from.probe.set.name`: Umwandeln von in den *publish*-Datenbanken verwendeten „item.ids“ bzw. „bio.item.ids“ in *probe set*-Bezeichner.

`convert.bio.item.id.to.accession.number`,
`convert.bio.item.id.to.description`: Umwandeln von „bio.item.ids“ in „*accession numbers*“ und *probe set*-Beschreibungen.

MARTIN EISENACHER

Alte Landstr. 21
48161 Münster
Tel.: 0 25 34 / 64 52 08
e-Mail: martin.eisenacher@web.de
Geburtsdatum: 12.5.1973
Geburtsort: Dorsten

1983 - 1992

Heisenberg-Gymnasium, Gladbeck
Abschluss: Abitur (Leistungskurse: Mathematik, Chemie; Abschlussnote: 1.5)

Oktober 1992 - September 1998

Universität Dortmund
Studium der Informatik, Nebenfach: Theoretische Medizin
im Hauptstudium: Betreuung von Softwarepraktikumsgruppen im Rahmen der Lehre
Abschluss: Diplom-Informatiker (Abschlussnote: sehr gut mit Auszeichnung)
Diplomarbeit am Lehrstuhl II („Theoretische Informatik“): „Algorithmen für das Dial-A-Ride-Problem“

Oktober 1998 - September 2000

„Locatech GmbH, Softwareübersetzungen“, Dortmund
Technischer Lektor und Technical Engineering im Bereich „Microsoft CTEC-Schulungen (Certified Technical Education Center)“

seit Oktober 2000

Universitätsklinikum Münster, Medizinische Fakultät

Promotionsstudium der Fächer Medizinische Informatik, Physiologische Chemie und Medizinische Physik

wissenschaftlicher Angestellter mit zwei halben Stellen:

<u>Bioinformatik-Gruppe im Gerhard-Domagk-Institut für Pathologie</u>

Arbeitsgebiete: Bioinformatik der Expressionsanalytik, Nutzer- und Projekt-Betreuung, System-Administration des Affymetrix GeneChip-Systems

<u>Institut für Medizinische Informatik und Biomathematik</u>

Arbeitsgebiete: Auswertung von Studiendaten mithilfe Neuronaler Netze, statistische Beratung, Kooperation mit klinischen Arbeitsgruppen

---

Im **Oktober 2002** Wechsel zur Zentralen Projektgruppe 1 des Interdisziplinären Zentrums für Klinische Forschung (IZKF): <u>„Integrierte Funktionelle Genomik (IFG)“, Abteilung Bioinformatik</u>

Arbeitsgebiete: Forschung, Dienstleistung und Schulungen im Bereich Genexpressionsanalytik

Promotion im Mai 2005

Thema:

„Intensitätsbasierte Qualitätskontrolle und Skalierung von Genexpressionsdaten“