Analyse und Interpretation der Varianz von Genexpressionsdaten

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Diplom-Informatiker Christian Hummert
geboren am 14. März 1979 in Lingen (Ems), Deutschland

Die vorliegende Arbeit wurde am Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie – Hans-Knöll-Institut (HKI) Jena unter der Leitung von Prof. Dr. Reinhard Guthke angefertigt.

Gutachter:

1. Prof. Dr. Stefan Schuster, Jena
2. Prof. Dr. Reinhard Guthke, Jena
3. PD Dr. Sebastian Zellmer, Berlin

Tag der öffentlichen Verteidigung: 21. November 2014

Abstract

This thesis, entitled "Analysis and Interpretation of Variance in Gene Expression Data", summarizes four papers. Initially, the concepts of „technical variance" and „biological variance" are distinguished.

In gene expression profiling with microarrays technical variance refers to the commonly considerable measurement errors. Reasons seem to be manifold. The effect of cross-hybridizations, which means unspecific bindings of RNA-fragments to the probes on the array, is controversially discussed. Some researchers consider this effect the most important source of error while others consider it negligible. The first two studies show that cross-hybridizations are indeed mainly responsible for the measurement errors in microarray experiments. Furthermore, tools for handling unspecific bindings are provided in form of new chip definition files and a manual for the design of new microarrays.

Variance based on real existing biological differences is denoted biological variance. The analysis of gene expression experiments with methods identifying differences in variance yield possible marker transcripts an analysis based on means does not reveal. Mapping the transcripts onto KEGG-pathways excludes false positive results.

In the fourth study an analysis of similarities with the help of correlation coefficients is performed. Hypotheses about the functional pathway in the induced defense of plants can be gathered by analyzing the data with the Kendall coefficient of rank correlation.

Zusammenfassung

Die vorliegende Dissertationsschrift fasst vier Arbeiten unter der Überschrift „Analyse und Interpretation der Varianz von Genexpressionsdaten“ zusammen. Zunächst wird der Begriff der „Technischen Varianz“ von dem der „Biologischen Varianz“ abgegrenzt.

In der Genexpressionsanalyse mit Microarrays wird unter technischer Varianz der traditionell hohe Messfehler verstanden. Die Gründe hierfür scheinen jedoch mannigfaltig zu sein. Höchst umstritten ist hierbei der Effekt von Kreuzhybridisierungen, also unspezifischen Bindungen von RNA-Fragmenten an die Sonden des Arrays. Einige Forscher halten diesen Effekt für die maßgebliche Fehlerquelle, andere beurteilen ihn als vernachlässigbar. In den ersten zwei Arbeiten wird gezeigt, dass Kreuzhybridisierungen in der Tat erheblich für den Messfehler bei Microarray-Experimenten verantwortlich sind. Gleichzeitig werden, mit einem Satz neuer Chip Definition Files und einer Handreichung zum Design neuer Microarrays, Werkzeuge zum Umgang mit unspezifischen Bindungen zur Verfügung gestellt.

Varianz, die auf tatsächlich vorhandenen biologischen Unterschieden basiert, wird biologische Varianz genannt. Bei der Auswertung eines Genexpressionsexperiments werden mittels Analyse der Streuungsparameter mögliche Markertranskripte identifiziert, die bei einer üblichen mittelwertbasierten Auswertung nicht gefunden werden. Durch Mapping der Transkripte auf KEGG-Pathways kann ausgeschlossen werden, dass es sich um falsch positive Treffer handelt.

In der vierten Arbeit wird eine Ähnlichkeitsanalyse mit Hilfe von Korrelationskoeffizienten durchgeführt. Durch Auswertung mit der Korrelation nach Kendall können Hypothesen über den funktionalen Pathway in der induzierten Abwehr von Pflanzen gewonnen werden.

Danksagung

Ich erinnere mich noch gut an mein Vorstellungsgespräch am HKI in Jena. Nach meinem Diplom hatte ich mich in Jena beworben. Prof. Dr. Guthke führte mich durch das Institut und es war wie der Eintritt in eine neue Welt – die ganzen Labore und Geräte hatte ich noch nie gesehen. Mit besonderem Enthusiasmus sprach er von genregulatorischen Netzwerken und von seiner Arbeit hieran. Auch ich sollte aus Daten durch Reverse Engineering solche Netzwerke rekonstruieren.

Am 1. August 2005 trat ich meine neue Arbeit als Doktorand am Leibniz Institut für Naturstoff-Forschung und Infektionsbiologie – Hans-Knöll-Institut (HKI) an, wie oft sollte ich in Zukunft diesen schier endlosen Namen tippen. Nach einer kurzen Einarbeitungsphase bekam ich meinen ersten Datensatz. Es handelte sich um Genexpressionsdaten, gemessen mit Microarrays...

Ich bedanke mich bei Prof. Dr. Reinhard Guthke, der immer sehr geduldig mit mir war und nie aufgehört hat, mich zu unterstützen. Ich danke Prof. Dr. Schuster, der sich sehr für meine Arbeit interessiert hat und mit mir sogar über Batman publizierte. Ich danke der International Leibniz Research School und insbesondere Dr. Dorit Schmidt für die strukturierte Doktorandenausbildung und die Möglichkeit, so viele gute Freundschaften zu schließen. Ich danke Dr. Ulrike Gausmann, die mir vieles von dem beigebracht hat, was ich heute über Transkripte weiß. Ich danke Dr. Martin Hoffmann, der mich ermuntert hat zu zweifeln und in mir das Interesse für Statistik geweckt hat. Und ich bedanke mich bei Folker Dutzmann, der unermüdlich Korrektur gelesen hat.

Nicht zuletzt danke ich meiner lieben Frau und meiner Familie, die meine Launen ertragen haben, als ich diese Arbeit schrieb, und mir so oft Mut gemacht haben.

Christian Hummert

im April 2014

Inhaltsverzeichnis

Kapitel 1

Einleitung

≪The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic.≫

– Sir Ronald Aylmer Fisher[1]

1.1 Fragestellung

Genregulation ist verantwortlich für die Entstehung der Komplexität von Lebewesen und die biologische Diversität im Zuge von Evolution und Adaption (Nguyen and D'haeseleer [2006]). Diese Regulation findet vor allem auf Transkriptebene statt. Aus diesem Grund ist das Verständnis der zugrundeliegenden genregulatorischen Netzwerke von zentraler Bedeutung für die Molekularbiologie (Hache [2009]). Häufig werden solche genregulatorischen Netzwerke aus experimentell gemessenen Daten durch Reverse Engineering rekonstruiert.

Die Qualität eines rekonstruierten Netzwerks hängt dabei direkt von der Qualität der zugrundeliegenden Daten ab (Marseguerra et al. [2005], Guthke et al. [2005]). Bei der Auswertung von Microarray-Daten ist die hohe Streuung der ermittelten Genexpression augenscheinlich (Purdom and Holmes [2005]). Tatsächlich scheinen Messfehler daran einen ganz erheblichen Anteil zu haben.

[1] geäußert von Fisher, einem der Begründer der Varianzanalyse und Erfinder des F-Tests, am 25. Januar 1934 in der anschließenden Diskussion, nachdem John Wishart seine Arbeit „Statistics in agricultural research“ (Wishart [1934]) vor der Industrial and Agricultural Research Section der Royal Statistical Society vorgestellt hatte.

Um vertrauenswürdige genregulatorische Netzwerke rekonstruieren zu können, ist es notwendig, diesen Messfehler genauer zu untersuchen.

Die Gründe für diese hohen Messfehler sind jedoch mannigfaltig, insbesondere ist der Effekt von Kreuzhybridisierungen umstritten. Hierbei handelt es sich um nicht im Arraydesign vorgesehene Bindungen, die aber dennoch systematisch auftreten und nicht eindeutig einem Gen zugeordnet werden können. Einige Forscher sehen hierin die Hauptursache für den hohen Messfehler bei der Bestimmung der Genexpressionen, andere Forscher halten den Effekt für vernachlässigbar. Dieser offenen Frage möchte die vorliegende Arbeit nachgehen und insbesondere folgende Fragen beantworten:

1. Beeinflussen Kreuzhybridisierungen die Messergebnisse von Microarray-Experimenten tatsächlich in einem Maße, dass eine erhebliche Beeinträchtigung der Daten daraus erwächst? Sind Kreuzhybridisierungen also maßgebliche Ursache für den Messfehler und die daraus resultierende Varianz?

2. Wie kann bei der Auswertung vom Microarray-Experimenten und beim Design von Microarrays das Problem von Kreuzhybridisierungen berücksichtigt werden?

Hingegen scheint es falsch zu sein, alle Varianz in den Daten auf Messfehler zurückzuführen. Als die Varianzanalyse zu Beginn des 20. Jahrhunderts entwickelt wurde, wurden damit große Fortschritte erzielt. Bei der Auswertung von biologischen Daten kommen allerdings regelmäßig lediglich Mittelwertvergleiche und die darauf basierende Statistik zum Einsatz, um (etwa durch unterschiedliche Versuchsbedingungen bewirkte) Unterschiede zwischen Gruppen nachzuweisen. In dieser Arbeit wird aber weiter gefragt:

3. Ist es möglich und sinnvoll, bei biologischen Daten die betrachteten Gruppen von Beobachtungen anhand der Intragruppenvarianzen zu unterscheiden? Lassen sich so neue Erkenntnisse gewinnen, die bei einer klassischen Fischerschen Varianzanalyse nicht erkannt werden?

4. Kann die Varianz als Maß für die Homogenität in einer Gruppe auch erweitert werden, um die Ähnlichkeit von biologischen Beobachtungseinheiten zu quantifizieren? Sind dafür von der Korrelation abgeleitete Maße geeignet?

Zunächst soll jedoch eine kurze Einführung in die Genexpression, Microarrays und die Varianzanalyse erfolgen.

1.2 Genexpression

Gene sind Abschnitte auf der **Desoxyribonukleinsäure (DNA)**. Sie tragen die Anleitung für die Herstellung spezifischer Proteine, aber auch anderer Transkripte wie etwa tRNAs oder micro-RNAs. Dabei wird das Protein aber nicht direkt aus dem Gen, sondern über den Umweg der **Ribonukleinsäure (RNA)** synthetisiert. Dieser erste Schritt, vom Gen zur RNA, wird **Transkription** genannt, der zweite Schritt, von der RNA zum Protein, **Translation**. Im Folgenden wird der Begriff **Genexpression** als Vorhandensein von RNA verwendet. Das bedeutet, ein Gen ist exprimiert, wenn zugehörige RNA in der Zelle vorhanden ist. Gene können demnach unterschiedlich stark exprimiert sein. In einigen Fragen der Forschung ist es jedoch ausschließlich von Interesse, ob ein Gen eingeschaltet respektive ausgeschaltet ist, also ob die zugehörige RNA überhaupt vorhanden oder ganz abwesend ist (qualitative Genexpressionsmessung). Bei anderen Fragestellungen ist eine quantitative Expressionsmessung gewünscht.

Die DNA ist ein riesiges Molekül, besteht aber nur aus vier verschiedenen Bausteinen. Ein solcher Baustein heißt **Nukleotid** und besteht aus einem Zucker (der Desoxyribose), einer Phosphatgruppe und einer von vier möglichen stickstoffhaltigen Basen: Adenin (A), Cytosin (C), Guanin (G) oder Thymin (T). Mehrere Nukleotide können sich zu einem DNA-Strang verbinden. Dabei bindet der Zucker eines Nukleotids an die Phosphatgruppe des nächsten. Zwei DNA-Stränge können sich nun antiparallel mit Hilfe der Basen an-

einander binden. Dabei bilden immer paarweise Cytosin und Guanin drei und Thymin und Adenin zwei Wasserstoffbrücken aus. Andere Kombinationen der Basen sind strukturbedingt nicht möglich. In einigen Fällen treten aber beispielsweite bei der tRNA auch sogenannte Nicht-Watson-Crick-Paarungen mit anderen Kombinationen auf (Leontis and Westhof [2001]). Die Gesamtheit der DNA stellt das Erbmaterial eines Organismus dar.

Bei der Transkription wird die DNA mit Hilfe von Enzymen stückweise in ihre Stränge aufgelöst und neue Nukleotidbausteine können sich anlagern und ebenfalls verbinden: Die DNA wird 'abgelesen'. Die 'Ablese'-Bausteine unterscheiden sich kaum von den DNA-Bausteinen. Im Unterschied zur DNA wird bei der RNA Ribose an Stelle von Desoxyribose als Zucker verwendet. Zudem wird als Stickstoffbase Uracil (U) statt Thymin (T) benutzt. Das entstandene Molekül heißt RNA. Diese von den Genen abgelesene RNA wird Boten-RNA, Messenger-RNA oder **mRNA** (engl. messenger: Bote) genannt.

Die Ableserichtung bei der Transkription ist nicht zufällig. Ausgehend vom ersten Nukleotid folgt das Nachbarmolekül, das an den Zucker gebunden ist (und nicht das, das an der Phosphatgruppe bindet). Es wird auch gesagt, die Ableserichtung erfolgt immer vom 5'- zum 3'-Ende des Gens hin. Das 5'-Ende steht also für die Phosphatgruppe, die am 5. Kohlenstoffatom der Desoxyribose bindet. Das 3. Kohlenstoffatom des Zuckers hingegen verbindet sich mit der Phosphatgruppe des Nachbarnukleotids.

Ein Gen ist aus vielen (bis zu mehreren tausend) Nukleotiden zusammengesetzt. Die Reihenfolge, **Sequenz**, der Basen codiert dabei alle wichtigen Erbinformationen. Bei der Transkription wird die gesamte Sequenz eines Gens in RNA übersetzt. In eukaryotischen Zellen wird die RNA verändert, bevor sie den Zellkern verlässt und an den Ribosomen zu Proteinen translatiert wird. Von der ursprünglichen RNA kann ein erheblicher Teil entfernt werden. Dabei werden bestimmte Stücke herausgeschnitten und der Rest wieder zusammengesetzt. Kodierende, also verwendete, Teilsequenzen heißen dabei **Exons** und nichtkodierende **Introns**. Das Weglassen der Introns bei der soge-

nannten RNA-Prozessierung wird **Spleißen** (engl. to splice sth.: etwas verbinden) genannt. Einige Gene in manchen Organismen können durch verschiedene Spleißvarianten verschiedene Transkripte erzeugen. Dieses Phänomen wird **alternatives Spleißen** genannnt und die verschiedenen Transkripte, die aus demselben Gen kodiert wurden, heißen **Isotypen**.

Die Gesamtheit aller Gene einer Zelle wird als **Genom** bezeichnet, die Gesamtheit aller mRNA-Transkripte als **Transkriptom** und die Gesamtheit aller Proteine als **Proteom**. Die Produktion von mRNA aus dem Genom wird als **Genexpression** bezeichnet.

Die Konzentration der Transkripte in einer Zelle ist für die forschende Biologie von Bedeutung. Insbesondere lassen sich in der Transkriptomanalyse Antworten auf die Frage finden, wie Organismen auf Veränderungen ihrer Umwelt reagieren.

Leider ist aber die Relation zwischen Transkriptom und Proteom nicht linear, so dass sich die Proteinkonzentration nicht direkt aus der mRNA-Konzentration vorhersagen lässt. Auf der anderen Seite besteht offensichtlich ein Zusammenhang zwischen Transkriptom und Proteom: Ohne entsprechende RNA kann keine Translation zum Protein stattfinden.

Für die Forschung ist noch von Bedeutung, dass sich das Transkriptom einfacher messen lässt als das Proteom. Auch aus diesem Grund werden bedeutend mehr Transkriptomstudien als Proteomstudien durchgeführt.

1.3 Messen der Genexpression

Zum Messen der Genexpression haben sich verschiedene Verfahren durchgesetzt. Klassischerweise wird die Genexpression mittels des **Northern Blots** bestimmt. Ein anderes gut geeignetes Verfahren ist die quantitative Real-Time Polymerase-Ketten-Reaktion (**qRT-PCR**). Dieses Verfahren hat den Vorteil, dass die Quantität der exprimierten mRNA abgeschätzt werden kann. Andere Verfahren sind Expressed Sequence Tags (**ESTs**) und Serial Analysis of Gene

Expression (**SAGE**) (Malone and Oliver [2011]).

Als Hochdurchsatzverfahren haben sich die sogenannten **Microarrays** etabliert. Das Verfahren wird auführlich in einem Review von Malone und Oliver aus dem Jahr 2011 dargestellt (Malone and Oliver [2011]). Mit sinkenden Kosten von Sequenzern ist auch die Sequenzierung von RNA-Fragmenten und somit eine Messung möglich. Diese Verfahren werden unter dem Begriff Next-generation RNA-Sequencing oder kurz **RNA-Seq** zusammengefasst.

Die Qualität der Messungen bei beiden Hochdurchsatzverfahren ist von hohen Messfehlern geprägt. Als genauestes der oben genannten Verfahren gilt die qRT-PCR, die als Goldstandard für Genexpressionsmessungen akzeptiert ist (Morey et al. [2006], Canales et al. [2006]).

1.4 Microarrays

In der Molekularbiologie und in der forschenden Medizin sind DNA-Chips oder Microarrays als Hochdurchsatztechnologie zur Analyse der Genexpression weit verbreitet. Ein einziger Chip misst die Expression tausender Gene gleichzeitig. Vor allem die Arrays der Marke Affymetrix (GeneChips) haben einen hohen Marktanteil (Ueda et al. [2004]). Solche Affymetrix GeneChips werden auch für die in den Kapiteln 2.1 und 3.1 vorgestellten Untersuchungen verwendet.

Ein Affymetrix GeneChip besteht aus einer Trägerplatte, auf der sehr viele (mehrere Millionen) Sonden aufsitzen. Bei dem weit verbreiteten Affymetrix HG-U133 Plus 2.0 Chip ist die Trägerplatte 1,28 × 1,28 Zentimeter groß. Eine solche Sonde, auch **Probe** oder **Spot** genannt, besteht im Mittel aus 10 Millionen gleichen 25mer langen Oligonukleotiden, also Nukleotidketten aus 25 Bausteinen. Die Spots werden *in situ*, das heißt direkt auf der Trägerplatte, mittels eines photolithografischen Verfahrens synthetisiert.

Die Genexpression wird gemessen, indem mRNA aus dem zu untersuchenden Gewebe oder aus einzelnen Zellen extrahiert und mit fluoreszierenden Farbstoffen markiert wird. Danach werden die einzelnen mRNA-Transkripte

in einem Digestionsschritt in kleinere Fragmente zerlegt und in eine Hybridisierungslösung gegeben. Anschließend wird die markierte mRNA auf das Array gegeben. Die mRNA-Fragmente lagern sich durch Wasserstoffbrückenbindungen zwischen den jeweils komplementären Nukleinbasen an die jeweils passenden Sonden auf dem Chip an. Dieser Prozess heißt **Hybridisierung**. Dann werden nicht gebundene RNA-Stücke abgewaschen. Abschließend wird das Microarray gescannt und jedem Spot durch verschiedene Auswertungsschritte ein numerischer Wert, die sogenannte **Signalintensität**, zugewiesen. Da die RNA-Fragmente fluoreszieren, ist das Signal einer Sonde umso stärker, je mehr RNA daran hybridisiert hat. Die dem Verfahren zugrundeliegende Hypothese ist also, dass sich die gemessene Signalintensität für jedes auf dem Array repräsentierte Transkript relativ zur Stärke der jeweiligen Genexpression verhält (Quackenbush [2002]).

Wie wird nun von den Sondensignalen auf die Expression der Gene rückgeschlossen? Bei den Affymetrix Gene Chip Modellen gehören immer zwei Sonden zusammen, sie bilden das sogenannte **Probepair**. Jedes Probepair besteht aus einer Perfect Match (**PM**) und einer Mismatch (**MM**) Sonde. Bei einem MM-Spot ist das mittlere (13.) Oligonukelotid ausgetauscht, was eine Hybridisierung gleicher RNA-Fragmente sowohl an PM als auch an MM verhindern soll. Ist das MM-Signal gleich Null, so ist das PM-Signal vertrauenswürdig. Geben jedoch beide Sonden Werte an, so deutet das auf eine fehlerhafte Hybridisierung hin.

Die Sequenz eines Transkripts vor der Digestion ist wesentlich länger als 25 Nukleotide. Deshalb wird jedes Transkript, das mit Microarrays untersucht wird, jeweils durch mehrere Probepairs repräsentiert, die zusammengefasst **Probeset** heißen. Für alle Probepairs, die zu einem Probeset gehören, werden die gemessenen Signalintensitäten verrechnet und zu einem einzigen Wert pro Probeset zusammengefasst. Viele Transkripte werden jedoch zusätzlich durch mehrere Probesets repräsentiert (Affymetrix Inc [2003a]). Die Zusammensetzung eines Probesets aus den Probepairs und die entsprechenden Positionen

auf dem Chip sind in einem sogenannten Chip Definition File (**CDF**) definiert.

Ein Standard Affymetrix Probeset besteht aus 11 Probepairs und wird durch einen eindeutigen Bezeichner gekennzeichnet. Dieser besteht aus einer siebenstelligen Zahl, einem _at-Suffix, und dazwischen einer optionalen Zeichenkette _s, _i, _i oder _a (Affymetrix Inc [2003b, 2007]).

Im Jahr 2006 gab es bereits über 12.000 peer-reviewed Veröffentlichungen, die auf der Auswertung von Microarray-Experimenten beruhen (Dalma Weiszhausz et al. [2006]), im Jahr 2011 waren es über 40.000 (Malone and Oliver [2011]). Forschung auf der Basis von Microarray-Daten liefert wichtige Erkenntnisse zum Beispiel zu Signalwegen innerhalb von Zellen (Febbo [2005], Werner [2008], Cheng and Li [2008]) oder im Vergleich von gesundem mit krankem Gewebe. Insbesondere zum Verständnis von Krebserkrankungen (Armstrong et al. [2001], Golub et al. [1999], Guo [2003], Nuyten and van de Vijver [2008]) oder Autoimmunerkrankungen (Centola et al. [2006], Huber et al. [2008]) können Microarray-Experimente beitragen. Aber auch in vielen anderen klinischen Feldern werden so große Fortschritte erzielt. Darüber hinaus spielen Microarrays eine wichtige Rolle bei der Identifikation von Biomarkern (Dieterle and Marrer [2008]). Sie könnten sogar den Weg von einer allgemeinen, für eine ganze Bevölkerung gültigen Medizin hin zu einer personalisierten, das heißt individuellen, Medizin ebnen (Casciano and Woodcock [2006]).

Nichtsdestotrotz erlaubt die Technologie noch Verbesserungen in der Qualität der Messungen (Heber and Sick [2006], Modlich and Munnes [2007]). So sind zum Beispiel die von Microarrays gemessenen Signalintensitäten nur vage Schätzungen für die tatsächlichen Transkriptkonzentrationen. Sie wären nur dann gute Schätzer, wenn die Relation zwischen Intensität und Konzentration in etwa linear wäre. Dies ist allerdings häufig nicht der Fall, weil sich bei hohen Konzentrationen Sättigungseffekte ergeben. Zudem ist die Relation nahe der Detektionsschwelle der Sonden gleichfalls nicht linear (Klipp et al. [2005]).

Obwohl Microarrays in der Forschung weit verbreitet sind, zeigen verschiedene Studien, die die Ergebnisse von Arrays verschiedenen Typs vergleichen,

nur eine geringe Konsistenz (Kuo et al. [2002], Tan et al. [2003], Woo et al. [2004], Järvinen et al. [2004]). Hinzu kommt, dass trotz immenser Arbeit auf diesem Gebiet (Brazma et al. [2001, 2003]) immer noch ein Mangel an standardisierten Protokollen für Microarray-Experimente (Choe et al. [2005]) besteht. Sowohl im experimentellen Vorgang als auch bei der Aufbereitung der Signalintensitäten, dem sogenannten Preprocessing, als auch in der Auswertung fehlen einheitliche Arbeitsweisen. Dies macht einen Vergleich oder das Zusammenführen von Microarray-Daten verschiedener Arbeitsgruppen schwierig.

Mit dem Problem der Vergleichbarkeit von Microarray-Daten aus verschiedenen Quellen beschäftigen sich sogenannte Batch-Correction-Techniken (Johnson et al. [2007], Kupfer et al. [2012]). Dabei werden Modelle, häufig Bayes-Modelle, an die Daten angepasst und eine Standardisierung der Daten durchgeführt. Bei einer solchen Batch-Correction ist es möglich, Unterschiede zwischen Laboren herauszurechnen. Systematische Fehler, die allen Microarray-Experimenten immanent sind, können aber nicht behoben werden.

Ein weiteres Problem ist, dass in hohem Tempo immer neue Erkenntnisse über die bereits sequenzierten Genome bekannt werden. Dies führt dazu, dass die Genomannotationen häufig angepasst werden. Die in den Chip Definition Files festgelegte Zuordnung von Sonden zu Transkripten bleibt aber konstant. So werden neue Erkenntnisse bei der Auswertung nicht berücksichtigt und die Annotationen der Arrays und die Erkenntnisse, die beispielsweise in den großen Datenbanken festgehalten werden, entwickeln sich stetig auseinander (Harrison et al. [2007], Dai et al. [2005]).

Optimalerweise zeigen unterschiedliche Probesets, die aber dasselbe Transkript messen, ähnliche Ergebnisse. Zudem sollten die Expressionswerte der elf Probes eines Probesets konsistent und die Abweichungen zwischen den Probes gering sein. Beides ist leider häufig nicht der Fall (Stalteri and Harrison [2007], Hughes et al. [2001], Harbig et al. [2005]). Um diese Fehler zu vermeiden, schlagen einige Forschungsgruppen eine sondenbasierte Analyse vor, die den Schritt der Verrechnung der Sonden zu Probesets vermeidet (Liu et al. [2006],

Sanguinetti et al. [2005]). Als alternativer Ansatz können Veränderungen an der Zusammensetzung der Probesets vorgenommen werden.

Wie diese Arbeit zeigt, sind unspezifische Bindungen von Transkripten durch Kreuzhybridisierungen ein bedeutendes Problem bei Microarray-Experimenten. Das heißt, dass RNA-Fragmente eines Transkripts an eine Sonde binden, die für ein anderes Transkript entworfen worden ist. Die Gruppe um Wu (Wu et al. [2005]) konnte zeigen, dass nur Fragmente, die länger als 8 Nukleotide sind, hybridisieren können, und vor allem Kreuzhybridisierungen aus Alinierungen zwischen 10 und 16 Nukleotiden entstehen. Zudem scheinen Kreuzhybridisierungen eher an Sonden, die Sequenzen in der Nähe des 5'-Endes, als an Sonden, die Sequenzen am 3'-Endes der Gene abbilden, aufzutreten. Besonders wenig vertrauenswürdig sind solche Sonden, die Ketten von Guaninnukleotiden (sogenannte G-Stacks) enthalten. Beim Affymetrix HG-U133A Chip betrifft das immerhin 16.743 Sonden (Wu et al. [2007]).

Unspezifische Bindungen können sowohl zu falsch positiven wie auch zu falsch negativen Ergebnissen führen, aus denen dann eventuell falsche Hypothesen über die Genexpression abgeleitet werden (Chen et al. [2007], Cambon et al. [2007]). Falsch positive Effekte treten auf, wenn 'falsche' RNA-Fragmente an einer Sonde hybridisieren, deren 'richtige' RNA-Fragmente aber eigentlich nicht oder viel weniger vorhanden sind. Die Sonde gibt also einen positiven Wert, der aber falsch, zu hoch, ist. Außerdem fehlen dann diese RNA-Fragmente, die an einer 'falschen' Sonde gebunden haben, um an der 'eigenen' Sonde zu hybridisieren, die dann einen zu niedrigen Wert anzeigt. Dies wird auch **Stealing-Effekt** (engl. stealing: stehlen) genannt.

In der Praxis hat sich gezeigt, dass falsch positive Ergebnisse häufiger als falsch negative Ergebnisse auftreten. Dies führt dazu, dass der Fehler nicht normalverteilt ist, was eine Auswertung weiter erschwert. Zum Beispiel ist die Normalverteilung der Fehler eine der Vorausetzungen für lineare Regression (Sachs [2004]). Ein Studie aus dem Jahr 2005 (Purdom and Holmes [2005]) zeigt, dass die asymmetrische Laplace-Verteilung eine gute Schätzung für die

Verteilung des Messfehlers bei Microarray-Experimenten ist.

Um den Einfluss von Kreuzhybridisierungen zu minimieren, hat Affymetrix selbst die interne Kontrolle des Probepairs, bestehend aus einer Perfect Match (PM) und einer Mismatch (MM) Sonde, eingeführt. Unglücklicherweise kann dieses Kontrollsystem Kreuzhybridisierungen nicht komplett ausschließen (Zhang et al. [2003]). Die Gruppe um Wu (Wu et al. [2005]) konnte sogar zeigen, dass auch die MM-Sonden selbst kreuzhybridisieren, wenngleich mittels eines anderen Mechanismus als die PM-Sonden. So scheint weitere Forschung an dieser Stelle notwendig (Bolstad et al. [2003]). Bei den neusten Microarray-Modellen von Affymetrix, wie dem MG-430 PM Chip, wird auf Mismatch-Sonden wieder verzichtet (Affymetrix Inc [2009]).

Es existieren bereits verschiedene Ansätze, mit dem Effekt der Kreuzhybridisierung umzugehen. Üblicherweise werden diese in den ersten Schritten der Vorverarbeitung der Microarray-Daten angewandt (Choe et al. [2005], Chen et al. [2006]). Die Gruppe von Haslam (Haslam et al. [2007]) beispielsweise interpretiert Kreuzhybridisierungen als vor allem von Sequenzidentitäten abhängig. Diese werden durch ihren Hamming-Abstand berechnet und es wird eine Wahrscheinlichkeit für eine Kreuzhybridisierung geschätzt.

Andere Ansätze behandeln das Problem der Kreuhybridisierungen durch Definition von alternativen Chip Definition Files (CDFs), die auf verschiedenen Sequenzdatenbanken beruhen. Die Gruppe von Ferrari (Ferrari et al. [2007]) beispielsweise definiert eine Bibliothek von CDFs, die auf der GeneAnnot-Datenbank (Chalifa-Caspi et al. [2004]) beruhen. In diesen CDFs sind alle Probesets, die dasselbe Gen messen sollen, zu einem Probeset zusammengefasst, so dass sich pro Transkript nur ein Messergebnis ergibt.

Eine andere Arbeit an einer CDF-Bibliothek, die auf einem breiten Repertoire von Sequenzdatenbanken, wie RefSeq (Pruitt et al. [2005]) oder Unigene (Pontius et al. [2003]) beruht, wurde von der Gruppe um Dai (Dai et al. [2005]) entwickelt. Auch hier werden Probesets, die dasselbe Gen messen, zusammengefasst. Sie bleiben jedoch getrennt, wenn so verschiedene Isoformen des Tran-

skripts unterschieden werden können. Hier werden auch Sonden ignoriert, die Kreuzhybridisierungen verursachen können, allerdings mit einem sehr laxen Filter.

Aufgrund der verschiedenen oben beschriebenen Probleme bei Microarray-Experimenten, die zu einem erhöhten Messfehler führen können, ist es empfehlenswert, pro biologischem Replikat (zum Beispiel: anderes Individuum unter gleichen Bedingungen) mehrere technische Replikate (Microarrays) anzufertigen. Aus diesen technischen Replikaten wird dann ein Wert pro Probeset für jedes biologische Replikat berechnet. Einige Autoren empfehlen jeweils vier technische Replikate (Churchill [2002], Woo et al. [2004]). Auf der anderen Seite sind viele Autoren zu dem Schluss gekommen, dass biologische Replikate wertvoller als technische sind (Yang and Speed [2002], Yauk et al. [2004]). Da jedes Replikat Ressourcen kostet, werden bei beschränkten Mitteln diese eher in biologische Replikate investiert. Allerdings sinken die Kosten pro Replikat, je mehr Replikate angefertigt werden, die Gesamtkosten steigen aber selbstverständlich (Yoo and Cooper [2004]).

Microarrays sind für relative quantitative Genexpressionsmessungen geeignet. Schlüsse über die absolute Quantitẗ der Genexpression sollten hingegen nicht gezogen werden. Bei der Untersuchung zweier verschiedener Bedingungen X und Y werden die numerischen Werte der gemessenen Genexpression $X_1, \ldots, X_m$ beziehungsweise $Y_1, \ldots, Y_m$ zu sogenannten Fold-Changes verrechnet. Für ein Transkript k, das unter beiden Bedingungen gemessen wird, lautet der Fold-Change:

$$Foldchange(X_k, Y_k) = \begin{cases} X_k \geq Y_k : X_k/Y_k \\ X_k < Y_k : \text{-}(Y_k/X_k) \end{cases} . \tag{1.1}$$

Ist der Fold-Change von Bedingung X nach Bedingung Y positiv, so ist das Transkript in Bedingung X **upregulated**, ist der Fold-Change negativ, so ist das Transkript in Bedingung X **downregulated**. Als Ergebnis einer Microarray-Studie wird in der Regel mit ebendiesen Fold-Changes gerechnet.

1.5 RNA-Sequenzierung

Der erste DNA-Sequenzierautomat wurde im Jahr 1986 am California Institute of Technology entwickelt. Diese Automaten sind in der Lage, die Nukleotidabfolge in einem DNA-Molekül zu bestimmen. Traditionell werden sie eingesetzt, um ganze Genome zu entschlüsseln. Moderne Verfahren bieten allerdings die Möglichkeit der beschleunigten Sequenzierung durch hochparallele Methoden. Diese Automaten werden Next-Generation-Sequenzierautomaten (NGS) genannt. Seit dem Jahr 2004 werden diese verstärkt für die Sequenzierung von mRNAs und anderen, nicht codierenden, RNA-Molekülen verwendet (RNA-Seq). Damit wurde eine konkurrierende Technologie zu den Microarrays etabliert (Zhou et al. [2010]).

Aktuell, im Jahr 2014, dominieren drei verschiedene kommerzielle NGS-Technologien den Markt: die Sequenzierung mit Brückensynthese von Illumina, die Zwei-Basen-Sequenzierung (SOLiD Sequenzer) von Applied Biosystems und die auf Pyrosequenzierung basierenden Sequenzer von Roche. Den drei Verfahren liegen unterschiedliche Prinzipien der biochemischen Sequenzierung und Amplifikation zugrunde. Bei allen aber findet der Prozess der Sequenzierung massiv-parallel und mit niedrigeren Kosten als bei Sequenzierautomaten der ersten Generation statt (Tucker et al. [2009]). Es existieren noch weitere Verfahren, die sich derzeit jedoch nicht durchgesetzt haben.

Eine sehr moderne Technik ist der Pacific Biosciences PacBio RS II Sequenzierautomat. Er erlaubt derzeit als einzige Hochdurchsatztechnologie die Sequenzierung langer Transkripte bis zu 1000 bp Länge. Dies ermöglicht neue Anwendungen für die *de-novo*-Assemblierung und die Analyse von kompletten Transkripten, sofern diese kurz genug sind (Stangier and Hegele [2011]).

Der Hauptvorteil von RNA-Seq gegenüber Microarrays ist, dass die Technologie die Analyse des Transkriptoms ohne vorheriges Wissen über die Sequenz der Transkripte erlaubt. Damit erlaubt RNA-Seq eine *de-novo*-Transkriptomanalyse von Organismen, deren Genomsequenz zuvor nicht sequenziert wurde.

Darüberhinaus können sogar neue, bisher unbekannte, Transkripte entdeckt werden (Hanriot et al. [2008]).

Bei ausreichend hoher Genauigkeit der Sequenzierautomaten, der sogenannten Tiefensequenzierung (Deep Sequencing), können auch sehr kleine Mengen von Transkripten (niedrig exprimierte Transkripte) gemessen werden. Desweiteren lassen sich neue Spleißstellen für alternatives Spleißen (Mortazavi et al. [2008]) und Einzelnukleotid-Polymorphismen (SNPs) feststellen (Wang et al. [2009]).

Die Ergebnisse von NGS-Experimenten werden in sehr großen Dateien gespeichert, die zeilenweise die Sequenzen der sequenzierten Transkripte und mehrere Kennzahlen zur Qualität der jeweiligen Sequenz enthalten. Tatsächlich bestehen diese Dateien oft aus mehreren hundertmillionen Einträgen und sind hunderte von Gigabytes bis mehrere Terrabytes groß (Fox et al. [2009]).

Die wohl größte Herausforderung in der Datenanalyse von RNA-Seq besteht darin, die gelesenen Fragmente (Reads) dem Referenzgenom zuzuordnen. In diesem Schritt muss die Alinierung (Alignment) der gemessenen RNA-Seq-Transkripte erfolgen. Alinierung von DNA-Sequenzen ist ein klassisches Problem der Bioinformatik. Der Standardalgorithmus für diese Aufgabenstellung ist BLAST (Basic Local Alignment Search Tool) (Altschul et al. [1990]). Aufgrund der kurzen Länge, der sehr großen Anzahl und der hohen Fehlerrate der gelesenen RNA-Transkripte sowie den großen Sequenzlücken (Gaps) sind konventionelle Algorithmen für diese Aufgabestellung aber ungeeignet. Daher kommen häufig heuristische Verfahren zum Einsatz (Polyanovsky et al. [2011]). Bei Anwendung von solchen Heuristiken ergeben sich, wenngleich selten, auch Fehlinterpretationen und damit Fehler (Li and Homer [2010]).

Mit der RNA-Seq-Technologie werden höhere Genauigkeiten erzielt als beim Einsatz von Microarrays. Unter idealen Bedingungen ist die Varianz technischer Replikate sehr gering. In einer Arbeit aus dem Jahr 2008 wurde die Fehlerrate bei technischen Replikaten unter Idealbedingungen auf 0,5 % geschätzt (Spearman-Korrelation zwischen den Replikaten 0,96). Dieser Wert ist dem von

Microarrays deutlich überlegen (Marioni et al. [2008]). Der geschätzte Wert für Microarrays liegt bei etwa 10 % (Love and Carriquiry [2009]). Eine Studie aus dem Jahr 2009 kam zu dem Ergebnis, dass die Wahrscheinlichkeit, dass ein 20mer langes Transkript, gemessen mit einem Illumina-Sequenzierautomaten, einen oder mehrere Sequenzierfehler enthält, 0,0048 % beträgt (Philippe et al. [2009]). Erscheint diese Zahl zunächst als sehr klein, so ist in Betracht zu ziehen, dass sehr viele Transkripte parallel gemessen werden. Eine einzige Gelspur (Lane) enthält schon mehr als 100 Millionen gemessene Transkripte (Reads). Wird diese Zahl nun mit der Fehlerrate multipliziert, ergibt das mehr als 480.000 fehlerhafte Reads pro Lane.

Kreuzhybridisierungen können bei RNA-Seq nicht auftreten. Viele Autoren nehmen dies zum Anlass, das Problem für gelöst zu erklären (Mortazavi et al. [2008], Wang et al. [2009], Müller et al. [2012]). Da aber auch beim RNA-Seq der Digestionsschritt, in dem die extrahierte RNA in kleinere Fragmente zerlegt wird, stattfindet, ergibt sich bei Genen mit gleichen Abschnitten wieder das Problem der Zuordnung, analog zum Problem der Kreuzhybridisierung bei Microarrays (Mardis [2008]). Beim RNA-Seq wird dieser Effekt mit *Cross-Alignment* analog zu *Cross-Hybridization* bezeichnet (Valdes et al. [2013]).

Daneben ist die Durchführung von RNA-Seq-Experimenten deutlich aufwändiger als von Microarray-Experimenten. Das bedeutet nicht nur, dass mehr Ressourcen aufgewendet werden müssen, sondern auch, dass die längere Bearbeitungskette anfälliger für mögliche Fehler ist.

In der RNA-Seq-Technologie sehen viele Autoren jedoch die Zukunft der Transkriptomanalyse (Shendure [2008]).

1.6 Varianz

Varianz (vom lateinischen *variantia*) bedeutet ursprünglich Verschiedenheit. In der Statistik wird hiermit der Effekt bezeichnet, dass sich bei der mehrfachen Durchführung ein und desselben Experimentes der Messwert X von

einer Durchführung zur anderen verändern kann. Als Ergebnis liegen dann verschiedene konkrete Messwerte $x_1, \ldots, x_n$ desselben Experiments vor. Natürlich können auch unterschiedliche Versuchsbedingungen verschiedene Werte hervorbringen. Um quantitative Aussagen über diese Verschiedenheiten zu machen, benutzt die Statistik sogenannte Streuungsmaße oder Schwankungsmaße. Die Schwankung erfolgt dabei um einen noch festzulegenden Referenzwert. Das gebräuchlichste Streuungsmaß ist die Varianz beziehungsweise die Standardabweichung als deren Quadratwurzel. Die Varianz ist also hier die erwartete quadratische Abweichung der Messwerte vom erwarteten Wert als Referenzwert.

Definition: Die Varianz ist definiert als

$$Var(X) = E((X - E(X))^2). \tag{1.2}$$

Da der Erwartungswert $E(X)$ aus einer Stichprobe $x = (x_1, \ldots, x_n)$ vom Umfang n durch das arithmetische Mittel $\bar{x}$ geschätzt wird, kann die empirische Varianz

$$s^2(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.3}$$

als erwartungstreuer Schätzwert für die Varianz berechnet werden.

Bei Genexpression wird häufig das arithmetische Mittel als Referenzwert benutzt, mitunter auch der Median, der eine größere Robustheit gegenüber grob fehlerhaften Messwerten besitzt. In diesem Fall wird in der Formel (1.3) $\bar{x}$ durch den Median ersetzt und von der Medianvarianz gesprochen.

Statistische Tests werden eingesetzt, um Hypothesen über Varianzen zu prüfen. Insbesondere ist häufig zu prüfen, ob für unterschiedliche Beobachtungsgruppen Varianzhomogenität vorliegt oder umgekehrt festgestellt werden kann, dass sich die Streuungen innerhalb unterschiedlicher Beobachtungsgruppen signifikant unterscheiden (Sachs [2004]).

In der klassischen Varianzanalyse nach R. A. Fisher werden für Daten aus

Messungen mit unterschiedlichen Versuchsbedingungen zwei Varianzen unterscheiden:

1. Die Varianz zwischen Gruppen, die **Intergruppenvarianz**.
   Werden in einem Experiment Daten unter unterschiedlichen Bedingungen erhoben und haben diese Bedingungen tatsächlich Einfluss auf die Daten, dann wird die Wirkung der Bedingungen in der Varianz zwischen den Gruppen sichtbar. Diese Form der Varianz, die Intergruppenvarianz, ist vom Experimentator in der Regel erwünscht (Workman et al. [2002]).

2. Die Varianz in einer Gruppe, die **Intragruppenvarianz**.
   Die Varianz zwischen den Gruppen ist von der Varianz der Daten innerhalb einer einzelnen Gruppe zu unterscheiden. Letztere beschreibt die Abweichungen der Daten bei gleichen Bedingungen. Diese Form der Varianz ist in der Regel vom Experimentator nicht erwünscht (Sartor et al. [2003]).

Die Intragruppenvarianz im biologischen Experiment lässt sich im Wesentlichen auf zwei Ursachen zurück führen:

1. Individualität biologischer Daten. Biologische Individuen entsprechen sich auch unter exakt gleichen Bedingungen nicht vollkommen. Werden beispielsweise menschliche Zellen betrachtet, ist die Genexpression in jedem Individuum auch unter noch so gleichen Bedingungen natürlich verschieden.

2. Der technische Messfehler. In der eingesetzten Technik können Ungenauigkeiten auftreten, beim Microarray-Experiment unter anderem verursacht durch Kreuzhybridisierungen.

Bei der Varianzanalyse dient die Unterscheidung von Inter- und Intragruppenvarianz dem Nachweis der Wirkung unterschiedlicher Versuchsbedingungen auf Lageparameter der Messwerte in den Gruppen. Zusätzlich erweist es sich aber als sinnvoll, auch die Streuungsparameter der Gruppen zu vergleichen.

Signifikante Unterschiede zwischen den Varianzen zweier Gruppen können mit verschiedenen statistischen Tests bestimmt werden (Sachs [2004]). Zu nennen sind insbesondere der Levene-Test (Levene [1960]), der Bartlett-Test (Bartlett [1937]), der Cochran-Test (Cochran [1941]) und die Gruppe der F-Tests. Dabei ist zu beachten, dass das Ergebnis des klassischen F-Tests nach R. A. Fisher (Fisher [1922]) auch durch kleine Abweichungen von der Normalverteilung stark beeinflusst ist. Bei experimentellen Daten, bei denen nicht sicher von einer Normalverteilung ausgegangen werden kann, ist aus diesem Grund ein nichtparametrisches Verfahren, wie etwa der Brown-Forsythe-Test, vorzuziehen (Brown and Forsythe [1974]).

Da der Brown-Forsythe-Test in der im Kapitel 3.1 angeführten Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) verwendet wird, soll diese Methode an dieser Stelle kurz dargestellt werden (Brown and Forsythe [1974]).

Sei $x_{ij} = \mu_i + \varepsilon_{ij}$ die j-te ($j = 1, \ldots, n_i$) Beobachtung in Gruppe i ($i = 1, \ldots, g$), wobei die Erwartungswerte μ_i nicht bekannt sind. Auch sei nicht vorausgesetzt, dass die μ_i gleich sind. Die Differenzen ε_{ij} der Beobachtungen zum Erwartungswert seien stochastisch unabhängige und ähnlich verteilte Zufallsgrößen mit Erwartungswert 0 und möglichweise verschiedenen Varianzen. Für jede Gruppe wird der Median $\tilde{x}_i$ und die empirische Medianvarianz $\tilde{s}_i^2$ berechnet.

Mit $z_{ij} = x_{ij} - \tilde{x}_i$ wird nun eine einfaktorielle ANOVA analog zum bekannteren Levene-Test (Levene [1960]) berechnet:

$$W_{50} = \frac{\sum_i n_i(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2/(g-1)}{\sum_i \sum_j (\bar{z}_{ij} - \bar{z}_{i\cdot})^2 / \sum_i (n_i - 1)}, \tag{1.4}$$

wobei

$$\bar{z}_{i\cdot} = \sum_j \frac{z_{ij}}{n_i} \quad \text{und} \quad \bar{z}_{\cdot\cdot} = \sum_i \sum_j \frac{z_{ij}}{\sum_i n_i}. \tag{1.5}$$

Der kritische Wert für W_{50} wird nun aus der Snedecor-F-Tabelle (Snedecor [1934]) mit $g-1$ und $\sum_i (n_i - 1)$ Freiheitsgraden abgelesen. Selbstverständlich lässt sich auch ein p-Wert nach Fisher (Fisher [1973]) berechnen.

Um Unterschiede in den Streuungsparametern zweier Gruppen X und Y quantifizierbar zu machen, hat der Autor der vorliegenden Arbeit den Varianzfold definiert:

$$VarFold(X,Y) = \begin{cases} Var_X \geq Var_Y : Var_X / Var_Y \\ Var_X < Var_Y : \text{-}(Var_Y / Var_X) \end{cases} . \tag{1.6}$$

Dieses Maß wird in der als Kapitel 3.1 angeführten Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) verwendet.

Bei biologischen Experimenten ist es häufig sinnvoll, die Varianz nicht nach Gruppen, sondern nach ihrer Ursache zu unterscheiden:

1. Die **technische Varianz** entsteht durch Ungenauigkeiten der eingesetzten Verfahren oder durch ungewollte Unterschiede bei den Versuchsbedingungen. Sie wird häufig unter dem Begriff **Messfehler** zusammengefasst.

2. Die **biologische Varianz** entsteht durch biologisch gegebene Unterschiede, dies schließt die Reaktion von Organismen auf verschiedene Bedingungen ebenso ein wie die biologische Individualität.

Bei der Analyse biologischer Experimente ist die Gleichsetzung der Intergruppenvarianz mit der biologischen Varianz und die der Intragruppenvarianz mit der technischen Varianz nicht immer korrekt. In der als Kapitel 3.1 angeführten Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) wird beispielsweise gezeigt, dass es sich auch bei Intragruppenvarianz um biologische Varianz handeln kann. Auf

der anderen Seite kann es vorkommen, dass sich ungewollte Gruppen durch Messfehler oder Unterschiede bei der Datenerhebung ergeben und damit eine technische Intergruppenvarianz. Dies kann beispielsweise der Fall sein, wenn ein Experiment von verschiedenen Experimentatoren durchgeführt wird. Solche Effekte werden als Batch-Effekte bezeichnet (Kupfer et al. [2012]).

Während sich die Intergruppenvarianz von der Intragruppenvarianz sehr gut trennen lässt, ist es nicht möglich, biologische Varianz und technische Varianz zu trennen. Auer und Doerge formulierten dazu:

> «It is essentially impossible to partition biological variation from technical variation. When these two sources of variation are confounded, there is no way of knowing which source is driving the observed results. No amount of statistical sophistication can separate confounded factors after data have been collected.» (Auer and Doerge [2010])

Auch ist eine Wertung: „Biologische Varianz ist gute Varianz“ und: „Technische Varianz ist schlechte Varianz“ nicht immer korrekt. Abhängig von der Fragestellung ist in einigen Fällen bei Experimenten die biologische Varianz genauso unerwünscht wie die technische Varianz (Churchill [2002]).

1.7 Kovarianz und Korrelation

Eine Verallgemeinerung des Varianzbegriffs für mehrere Messgrößen ist die Kovarianz. Die Kovarianz ist ein nichtstandardisiertes Zusammenhangsmaß für den monotonen Zusammenhang zwischen zwei Zufallsvariablen. Sie ist ein Maß für die gemeinsame Variation zweier Zufallsvariablen (Sachs [2004]).

Definition: Die Kovarianz ist definiert als

$$Cov(X, Y) = E((X - E(X) \cdot (Y - E(Y)). \tag{1.7}$$

Damit ist die Varianz ein Spezialfall der Kovarianz, denn es gilt:

$$Var(X) = Cov(X, X). \tag{1.8}$$

Die Varianz ist demnach die Kovarianz einer Zufallsvariablen mit sich selbst.

Analog zur empirischen Varianz kann für eine zweidimensionale Stichprobe $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ vom Umfang n mit den arithmetischen Mitteln $\bar{x}$ und $\bar{y}$ die empirische Kovarianz als erwartungstreuer Schätzwert für die Kovarianz berechnet werden:

$$s(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{1.9}$$

Die Kovarianz wird wie folgt interpretiert:

- Die Kovarianz ist positiv, wenn tendenziell ein monoton wachsender Zusammenhang zwischen X und Y besteht. Das bedeutet: Hohe Werte von X gehen mit hohen Werten von Y einher und niedrige Werte von X treffen auf niedrige Werte von Y. Es besteht also ein positiver Zusammenhang zwischen X und Y.
- Die Kovarianz ist hingegen negativ, wenn entsprechend ein entgegengerichteter monotoner Zusammenhang zwischen X und Y besteht. Das bedeutet: Hohe Werte von X gehen mit niedrigen Werten von Y einher und niedrige Werte von X treffen auf hohe Werte von Y. Es besteht also ein negativer Zusammenhang zwischen X und Y.
- Für unabhängige Zufallsgrößen ist die Kovarianz Null. Umgekehrt wird aus dem Verschwinden der Kovarianz meist auf die Unabhängigkeit geschlossen. Tatsächlich besteht dann zwar kein monotoner Zusammenhang zwischen X und Y, nicht auszuschließen sind allerdings nichtmonotone Beziehungen.

Der Wert der Kovarianz gibt zwar die Richtung einer monotonen Beziehung

zwischen den Zufallsvariablen an, die Stärke des Zusammenhangs lässt sich aber nicht an ihm ablesen. Dies liegt in der Linearität der Kovarianz begründet, das heißt, sie hängt vom Maßstab der Zufallsvariablen ab.

Um das Maß des Zusammenhangs vergleichbar zu machen, ist es notwendig die Kovarianz zu normieren. Die Normierung der Kovarianz führt zum Korrelationskoeffizienten.

Wird von Korrelation gesprochen, ist zumeist die von Karl Pearson entworfene Pearson-Korrelation gemeint (Pearson [1895]).

Definition: Der Korrelationskoeffizient ρ ist definiert als

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}, \tag{1.10}$$

die Kovarianz wird mit den Standardabweichungen der Zufallsvariablen normiert. Werden in der Formel (1.10) die Varianzen $Var(X)$ und $Var(Y)$ durch die empirischen Varianzen $s^2(x)$ und $s^2(y)$ sowie die Kovarianz durch die empirische Kovarianz $s(x,y)$ ersetzt, ergibt sich der empirische Maßkorrelationskoeffizient nach Bravais und Pearson $\rho(x,y)$ als erwartungstreuer Schätzer für die Korrelation.

Der Korrelationskoeffizient ist ein dimensionsloses Maß für den linearen Zusammenhang zwischen zwei Zufallsvariablen. Er kann Werte zwischen -1 und +1 annehmen. Es ergibt sich folgende Interpretation:

- $\rho = 1$: Es besteht ein vollständiger, linearer Zusammenhang zwischen den Merkmalen.
- $\rho =$ -1: Es besteht ein vollständiger, negativer, linearer Zusammenhang zwischen den Merkmalen.
- $\rho = 0$: Es besteht kein linearer Zusammenhang zwischen den Merkmalen. Es wird jedoch keine Aussage über nichtlinearere Zusammenhänge getroffen.

Anders als die Kovarianz gibt Pearsons ρ nicht nur die Richtung einer monotonen Beziehung zwischen den Zufallsvariablen, sondern auch die Stärke

des Zusammenhangs an. Im Gegensatz zur Kovarianz ist ρ invariant gegenüber streng monoton steigenden linearen Transformationen.

Andere Maße für die Korrelation sind Spearmans ρ (Spearman [1904]) und Kendalls τ (Kendall [1938]). Beide Maße sind nichtparametrische Rangkorrelationskoeffizienten, die auch für ordinale Daten definiert sind. Da die Kendall-Korrelation in der im Kapitel 3.2 angeführten Arbeit *Quantification of growth-defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover* (Ullmann-Zeunert et al. [2013]) verwendet wird, soll diese Methode an dieser Stelle kurz dargestellt werden.

Definition: Die Kendall-Korrelation ist wie folgt definiert:
Seien $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ eine nach der ersten Variable sortierte Menge von Realisierungen der verbundenen Zufallsvariablen X und Y. Ein Paar $(x_i, y_i), (x_j, y_j)$, mit $i < j \leq n$, heißt konkordant, wenn gilt: $x_i < x_j$ und $y_i < y_j$ oder $x_i > x_j$ und $y_i > y_j$. Die Anzahl der konkordanten Paare wird mit C bezeichnet. Ein Paar heißt diskordant, wenn gilt: $x_i < x_j$ und $y_i > y_j$ oder $x_i > x_j$ und $y_i < y_j$. Die Anzahl der diskordanten Paare wird mit D bezeichnet. Dann ist

$$\tau(X, Y) = \frac{C - D}{\frac{1}{2}n(n-1)}. \tag{1.11}$$

Dabei wurde von Bindungen, das heißt von übereinstimmenden Realisierungen, abgesehen. Wie für Pearsons ρ gilt für Kendalls τ:

$$\text{-1} \leq \tau \leq +1. \tag{1.12}$$

Kendalls τ ist vor allem für Daten geeignet, die nicht normalverteilt sind oder ungleiche Skalenteilungen aufweisen. Auch bei kleinen Stichprobengrößen ist Kendalls τ stets vorzuziehen. Kendalls τ ist leicht konservativer als Pearsons ρ oder Spearmans ρ, das heißt, dass die Werte in der Regel etwas kleiner als die der anderen Korrelationskoeffizienten sind.

Der Korrelationskoeffizient lässt sich jedoch nicht nur als Zusammenhangsmaß, sondern auch als Ähnlichkeitsmaß oder nach einer geeigneten Transformation als Abstandsmaß interpretieren (Egghe and Leydesdorff [2009]).

Werden etwa die Versuchsergebnisse eines Experimentes als (Zufalls-) Vektor $X = (X_1, \ldots, X_m)$ und die entsprechenden Ergebnisse unter veränderten Versuchbedingungen als (Zufalls-) Vektor $Y = (Y_1, \ldots, Y_m)$ zusammengefasst, so entspricht der Pearson-Korrelationskoeffizient ρ dem Cosinus des zwischen beiden Vektoren X und Y eingeschlossenen Winkels ϕ, also:

$$\cos(\phi) = \rho(X, Y). \tag{1.13}$$

Zeigen beide Vektoren in dieselbe Richtung, ist $\phi = 0$ und damit $\cos(\phi) = 1$. Sind die Vektoren orthogonal zueinander, so ist $\cos(\phi) = 0$, zeigen sie in entgegengesetzte Richtungen, so ist $\cos(\phi) = -1$. Diese für $m = 2$ oder $m = 3$ geometrische Deutung der Beziehung zwischen X und Y entspricht der dargelegten Interpretation des Korrelationskoeffizienten.

Die Euklidische Distanz $d(X, Y)$ ist mit der Pearson-Korrelation eng verwandt. Für zwei normierte Zufallsvektoren mit $X_1^2 + \ldots + X_m^2 = 1$ und $Y_1^2 + \ldots + Y_m^2 = 1$ gilt:

$$\rho(X, Y) = 1 - \frac{d^2(X, Y)}{2} \qquad \text{und} \qquad d(X, Y) = \sqrt{2(1 - \rho(X, Y))}. \tag{1.14}$$

Die nichtparametrischen Korrelationskoeffizienten nach Spearman und Kendall lassen sich analog als Ähnlichkeitsmaße oder Abstandsmaße interpretieren. In der vorliegenden Arbeit wird im Kapitel 3.2 Kendalls τ in diesem Sinne verwendet.

In den nachfolgenden zwei Kapiteln werden insgesamt vier Arbeiten vorgestellt, in denen der Autor zunächst zeigt, dass Kreuzhybridisierungen in erheblichem Maße zum Messfehler bei Genexpressionsmessungen mit Microarrays beitragen. Dazu entwickelt er zunächst neue Chip Definition Files, die Kreuzhybridisierungen ausschließen. Zwei Datensätze werden einmal mit den

neuen CDFs und einmal mit den Original-CDFs ausgewertet. Danach wird in der zweiten Arbeit ein komplett kreuzhybridisierungsfreies Array geschaffen und gezeigt, dass die technische Varianz signifikant abnimmt. In der dritten Arbeit wird der Varianzfold definiert und ein Genexpressionsdatensatz mit diesem Werkzeug und anderen Methoden zum Vergleich der Streuungsparameter ausgewertet. Dabei werden neue Markertranskripte für die rheumatoide Arthritis identifiziert. In einer weiteren Arbeit wird eine Ähnlichkeitsanalyse der Verteilung von Stickstoffmetaboliten in Pflanzenteilen durchgeführt. Dabei wird Kendalls τ als Ähnlichkeitsmaß interpretiert.

Kapitel 2

Technische Varianz in Microarray-Experimenten

Die als Kapitel 2.1 angeführte Arbeit *Creation and comparison of different chip definition files for Affymetrix microarrays* (Hummert et al. [2011]) analysiert das Problem der Kreuzhybridisierung aufgrund unspezifischer Bindungen. Es werden vier verschiedene Affymetrix GeneChip Arrays untersucht, drei Human Genome Arrays, das HG-U133A, das HG-U133B und das HG-U133 Plus 2.0, sowie das Mouse Genome 430 2.0 Array. Mit Hilfe von BLAST wird gezeigt, dass Kreuzhybridisierungen für alle vier untersuchten Arrays sehr häufig sind.

Für den HG-U133A Chip können beispielsweise 45 % der Sonden des Arrays kreuzhybridisieren. Außerdem spiegelt ein beträchtlicher Teil der Sonden die angegebenen Transkripte nicht korrekt wider.

Um den tatsächlichen Einfluss von Kreuzhybridisierungen zu untersuchen, werden in der Arbeit neue Chip Definition Files (CDFs) zur Verfügung gestellt, die alle kreuzhybridisierenden oder unpassenden Sonden ausschließen. Die neuen CDFs werden mit Hilfe von Korrelation zwischen Microarray- und qRT-PCR-Ergebnissen mit drei anderen CDFs verglichen, den originalen Affymetrix CDFs, denen von Dai *et al.* (Dai et al. [2005]) sowie denen von Ferrari *et al.* (Ferrari et al. [2007]).

Es wird gezeigt, dass die neuen CDFs ohne Kreuzhybridisierungen bessere

Korrelation zur qRT-PCR zeigen als die originalen Affymetrix CDFs. Dies ist insofern besonders bemerkenswert, da für die neuen CDFs deutlich weniger Sonden zur Verfügung stehen. Für den HG-U133A Chip werden beispielsweise fast 50 % der Sonden in dem neuen CDF unberücksichtigt gelassen. Insgesamt zeigt sich, dass das Ergebnis umso besser ist, je mehr Sonden pro Probeset zur Verfügung stehen. In der Arbeit wird also gezeigt, dass Kreuzhybridisierungen tatsächlich ein relevantes Problem bei Microarray-Experimenten sind und die Berücksichtigung von kreuzhybridisierenden Sonden das Ergebnis verschlechtert und den Messfehler somit erhöht.

Die dem Kapitel 2.2 entsprechende Arbeit *Optimization of a microarray probe design focusing on the minimization of cross-hybridization* (Horn et al. [2011]) beschäftigt sich ebenfalls mit dem Problem der Kreuzhybridisierung aufgrund unspezifischer Bindungen. In dieser Arbeit wird ein neues Microarray-Chipdesign entworfen, das von Anfang an Kreuzhybridisierungen konsequent vermeidet. Es wird eine neue Methode vorgestellt, die bereits existierende Sonden bewertet und dann auf bestimmte Kriterien hin (hier Vermeidung von Kreuzhybridisierungen) optimiert. Eine solche Optimierung eines Microarray-Sondendesigns, das auf die Vermeidung von Kreuzhybridisierung fokussiert, ist exemplarisch für *Aspergillus nidulans* durchgeführt worden.

Nachdem das neue Chipdesign vorlag, wurde der Chip tatsächlich gespottet. Ein Experiment wurde sowohl mit einem alten Chipdesign (mit möglichen Kreuzhybridisierungen) als auch dem neuen (kreuzhybidisierungsfreien) Array durchgeführt.

Das neue Sondendesign wird mittels Medianvarianz interner technischer Replikate experimentell evaluiert. Im Ergebnis wird gezeigt, dass das neue Design dem alten signifikant überlegen ist. Auch diese Arbeit belegt also, dass Kreuzhybridisierungen tatsächlich ein relevantes Problem bei Microarray-Experimenten sind und der konsequente Ausschluss von kreuzhybridisierenden Sonden das Ergebnis verbessert sowie den Messfehler reduziert.

2.1 C. Hummert, F. Mech, F. Horn, M. Weber, S. Drynda, U. Gausmann, R. Guthke. *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11)*, Vol.I, 16-22. CSREA Press, Las Vegas, USA, July 18-21, 2011.

Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert[1], F. Mech[1], F. Horn[1], M. Weber[1], S. Drynda[2], U. Gausmann[3], R. Guthke[1]
[1]Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena
[2]Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty
[3]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Abstract— *Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. Nevertheless, the quality of the technology is still capable for further improvements. One of the main problems is cross-hybridization of the transcripts to non-corresponding probes on the array by unspecific binding.*

Four different Affymetrix GeneChip arrays are analyzed, namely the Human Genome arrays HG-U133A, HG-U133B, HG-U133 Plus 2.0 and the Mouse Genome 430 2.0 array. It is shown that putative cross-hybridizations are common for the examined arrays (e.g., 45 % of all probes for the U133A). Furthermore, a considerable amount of probes does not match the annotated transcript correctly. A new set of CDFs is created avoiding putative cross-hybridization completely. It is compared with three other CDFs (Affymetrix, Dai et al.*, Ferrari* et al.*) with the help of correlation between microarray and qRT-PCR results for two datasets. The newly created and the Ferrari CDFs perform significantly better than the original Affymetrix CDFs. The new CDFs are available as R-packages at http://www.sysbio.hki-jena.de/software and have been submitted to BioConductor.*

Keywords: microarrays, unspecific binding, cross-hybridization, Chip Definition Files

1. Background

Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. They are applied to measure changes in expression levels for thousands of genes simultaneously. Until 2011, more than 20,000 measurement series based on microarray technology have been published in public repositories like NCBI's Gene Expression Omnibus.

Nevertheless, the quality of the technology is still capable for further improvements [1], [2]. Several studies tried to compare data derived from different types of arrays and showed a rather poor consistency [3], [4]. Although microarrays are commonly used, this is a daunting problem. In addition, although there has been extended work on this field [5], there is still a lack of standardized experimental protocols among different laboratories [6].

The main problem of microarray analysis is unspecific binding of transcripts by cross-hybridization. This means that RNA fragments hybridize to a probe which is not designed for this gene. It was shown that fragments longer than 8 nucleotides are able to hybridize and that cross-hybridization can emerge from alignments ranging from 10 to 16 nucleotides. Further, the 5'-ends were found to cross-hybridize more likely than the 3'-ends [7].

Unspecific binding may lead to false-positive and false-negative results following in incorrect hypotheses about gene expression [8], [9]. Affymetrix, a technology widely used [10], accounts for the influence of cross-hybridization by introducing internal controls: each probepair comprises a Perfect Match (PM) and a Mismatch (MM) probe which are statistically evaluated [11]. Unfortunately, this procedure cannot solve the problem of cross-hybridization completely [12] and further refinements are suggested [13]. For example, Wu *et al.* [7] stated that the MM probes can also cross-hybridize, even though by another mechanism as the PM probes. Therefore, they recommended ignoring the MM probes.

Generally, expressed transcripts are represented on the array by a series of probepairs called probesets. The signal intensities are summarized to a single value per probeset. A large number of single transcripts are represented by multiple probesets. Multiple probesets representing the same gene are expected to show similar fold changes calculated from the signal intensities of the hybridized samples. However, this is in fact not the case [14], [15], [16]. This problem arises from single probes in the probeset which are capable of cross-hybridization. Ways to deal with this problem is either a probe-based analysis, leaving out the probe-to-probeset summarization step [17], [18], or the composition of the probesets could be improved by setting up alternative Chip Definition Files (CDFs) based on information contained in different sequence databases. For example, the group of Ferrari *et al.* [19] created a set of custom CDFs based on the GeneAnnot database [20]. In these CDFs the probesets that match the same gene were merged into one probeset. Hence, the existence of more than one probeset per gene was eliminated, avoiding discordant expression signals for the same transcript.

Another set of custom CDFs relying on a broad repertoire

of databases like RefSeq or Unigene has been created by the group of Dai *et al.* [21]. Probesets matching the same gene were merged, but remained divided if they were able to discriminate different isoforms of a gene. Probes causing cross-hybridizations were removed from the new probesets, but the filter had been not very strict.

Several groups dealt with the question of the minimum probeset size [19], [21]. For example, the group of Lu *et al.* [22] sets the minimum probeset size to 4 probes because smaller probesets result in high error rates. In this study the minimum probeset size was set to 4 [19], [21]. From these new probesets custom CDFs and the corresponding Bioconductor libraries for Affymetrix GeneChips were created.

In the work presented here, a new set of CDFs is introduced avoiding putative cross-hybridization completely. These CDFs are compared with those from Affymetrix, Ferrari, and Dai by validation of the respective microarray results using qRT-PCR for two different datasets.

2. Results

Four different Affymetrix GeneChip arrays are analyzed, namely the HG-U133A, HG-U133B, HG-U133 Plus 2.0 designed for human samples, and the Mouse Genome 430 2.0 array. For the detection of putative cross-hybridizations, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23] as described in the methods section.

The GeneChip HG-U133A consists of 22,283 probesets, each of 11–20 probepairs and 247,937 probepairs in total. Additional 1,155 probepairs are controls and are furthermore ignored. About 44 % of the PM probes (109,245) match exactly one single gene. 11 % of the probes (26,159) do not match any annotated gene. 45 % of the probes (112,533) match more than one gene and thus have cross-hybridization potential.

Furthermore, the direction of the probes was analyzed. Normally, sense strand RNA fragments are expected, although there are some loci in the human genome [24], as well as in the mouse genome [25], where both sense and antisense strands are transcribed. However, mixing up probes detecting sense or antisense strands in one single probeset could cause wrong expression results. Here, only probes matching the sense strand are considered as correct. For the U133A microarray all probes match the sense strand.

The GeneChip HG-U133B consists of 22,645 probesets, each of 11–20 probepairs and 249,491 probepairs in total. Again, there are additional probesets containing more than 11 probes as controls and are ignored (1,100). About 35 % of the probes (87,067) are found to match exactly one gene. 2 % of the probes (5,453) match more than one gene, so they possibly cross-hybridize, 5 % of the probes (12,805) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 58 % of the probes (144,166) do not match any annotated gene.

The GeneChip HG-U133 Plus 2.0 consists of 54,675 probesets and 604,247 probepairs. Like in the other arrays, additional probesets containing more than 11 probes are controls and are discarded. Here, 37 % of the remaining probes (221,821) match exactly one gene, 23 % of the probes (141,146) match more than one gene, 11 % of the probes (65,327) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 29 % of the probes (175,953) do not match any annotated gene.

The Mouse Genome 430 2.0 array consists of 45,036 probesets and 496,457 probepairs. About 52 % of the counted probes (257,331) match exactly one gene and 5 % of the probes (27,112) match more than one gene. About 1 % of the probes (4,661) match genes only in the wrong direction and 42 % of the probes (207,353) do not match any annotated gene.

Nearly all Affymetrix probesets contain at least one probe which has cross-hybridization potential. In fact, for the HG-U133 Plus 2.0 Chip about 65 % of all probesets include more cross-hybridizing probes than non-ambiguous ones.

All probes matching exactly one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes, that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. Only the good probes are used to create the new CDFs as described in the methods chapter. Accordingly, for the HG-U133A microarray originally measuring 14,500 genes by 22,283 probesets the newly created CDF contains 12,400 probesets representing 12,400 genes. For the HG-U133 Plus 2.0 the number of probesets is reduced from 54,675 (representing 38,500 genes) to 18,800 (representing 18,800 genes). The HG-U133B comprises 22,645 probesets measuring the expression of 18,400 genes. Here, the number of probesets is reduced to 6,500 matching 6,500 transcripts. The Mouse 430 2.0 microarray consists of 45,036 probesets for 39,000 genes. With the new CDF there are 16,400 probesets matching 16,400 genes. Hence, the number of identifiable genes is reduced in order to achieve a higher specificity of the probesets. The result for the HG-U133 Plus 2.0 is in good agreement to the results of Barnes *et al.* [26], who used BLAT and the Golden Path database and achieved a number of 17,143 genes that can be measured.

Small probesets lead to higher error rates and result in lower statistical significance. In the Affymetrix CDFs the size is 11 for nearly all probesets, but in the newly created probesets the size is not fixed. Some probesets are smaller than those from Affymetrix due to the removal of the problematic probes. However, many probesets increase in size due to useful probes on the array that have not been used for the matching gene before and probesets measuring

the same gene beeing merged. For example, for the HG-U133 Plus 2.0 the mean probeset size increases from 11 to 17.

For the validation of all CDFs two test datasets are chosen: (i) the Etanercept (ETC) and (ii) the MAQC dataset. The first of the two datasets is derived from a study analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points [27]. It is a typical dataset that arises in medical studies and is rather representative. One Affymetrix HG-U133A array experiment was performed for each time point. The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. The subset of the 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here.

qRT-PCR results are considered to reflect the real transcript concentrations with higher reliability than those determined by microarrays. Therefore, qRT-PCR experiments are regarded as a 'gold standard' for chip analyses [29], [30]. The Pearson correlation coefficient (PCC) of the microarray and the qRT-PCR data is computed for each gene using the different CDFs.

For the Etanercept dataset we performed qRT-PCR experiments for 16 genes. In total, this dataset now contains results from 51 microarrays and 816 qRT-PCR experiments. In addition, the genes with qRT-PCR data in both records are analyzed in more detail.

The perfomance of these CDFs were compared: the original Affymetrix CDFs (A), the two alternative CDFs of Ferrari *et al.* (F) [19] and Dai *et al.* (D) [21], and the new CDFs (H) presented here. The CDFs from Ferrari, using the GeneAnnot database, contain merged probesests (see background chapter), and cross-hybridization was not considered. The group of Dai offers a broad spectrum of different CDFs based on different databases. The one using RefSeq is chosen for comparison because it corresponds best to the new CDFs, using RefSeq as well. In the Dai CDFs different probesets matching a single gene are combined, although there are exceptions for genes comprising different isoforms. A check for cross-hybridization is also included. However, it applies a different algorithm than the new CDFs and the filter is much less strict.

For the probe to probeset summarization step two algorithms are used as described in the methods section: (i) the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and (ii) the Affymetrix Microarray Suite MAS5 [32]. These were compared repeatedly, but it is difficult or even impossible to decide which of the both algorithms performs better in any case [33], [34], [35].

For the Etanercept dataset, the mean correlation coefficient of all 16 genes for the Affymetrix CDF is 0.61 using the robust multi-array analysis algorithm (RMA) and 0.60 using the Affymetrix Microarray Suite MAS5. These values include 31 probesets in total matching these 16 genes according to the Affymetrix annotation file. If only the best correlating probeset for each gene is considered, the average correlation coefficient increases to 0.73 for RMA and 0.71 for MAS5. However, this value is more of theoretical interest because the knowledge which probeset will perform best is gained not until the qRT-PCR experiments and correlation analysis is finished. On average, the incorporated probesets contain 5.58 putative cross-hybridizations calculated by BLAST (4.47 including only the best performing probesets).

The Dai CDF contains 23 probesets for the 16 genes of the Etanercept dataset. Their mean correlation coefficient increases to 0.67 for both RMA and MAS5 compared to the 0.60 using the Affymetrix CDF. Considering the best correlating Dai probesets only, the values further increase to 0.73 for RMA and 0.69 using MAS5. The mean size of the Dai probesets increases to 20.59 probes containing 8.82 putative cross-hybridizations. This number changes to 4.71 if normalized to a probeset size of 11. Here, normalization means the number of putative cross-hybridizations calculated for a hypothetical Dai probeset size of 11. Considering only the best Dai probesets, the number of putative cross-hybridizations decreases to 7.88 on average.

For the Ferrari CDF, the mean correlation coefficient equals 0.73 for RMA and 0.69 using MAS5 on average. The mean probeset size increases to 19.56, harboring 10.81 possible cross-hybridizations (6.07 if normalized).

Using the new CDF the mean correlation coefficient amounts to 0.72 for RMA and 0.68 for MAS5. The mean probeset size decreases to 10.25 with no cross-hybridizations at all. The detailed results are shown in the table below:

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
TNF	A: 207113_s_at	0.88	0.85	N/A	8	11
	D: NM_000594_at	0.88	0.85	N/A	8	11
	F: GC06P031652_at	0.88	0.85	N/A	8	11
	H: gi_25952110	0.86	0.81	N/A	0	3
IL1B	A: 205067_at	0.95	0.90	0.37	6	11
	A: 39402_at	0.95	0.87	0.82	6	16
	D: NM_000576_at	0.96	0.89	0.74	12	27
	F: GC02M113303_at	0.96	0.89	0.74	12	27
	H: gi_27894305	0.95	0.88	0.86	0	15
IL6	A: 205207_at	0.69	0.71	0.81	3	11
	D: NM_000600_at	0.69	0.71	0.81	3	11
	F: GC07P022732_at	0.69	0.71	0.81	3	11
	H: gi_10834983	0.65	0.72	0.71	0	8
IL8	A: 202859_x_at	0.88	0.81	0.90	6	11
	A: 211506_s_at	0.86	0.73	0.98	6	11
	D: NM_000584_at	0.88	0.73	0.96	12	22
	F: GC04P074845_at	0.88	0.73	0.96	12	22
	H: gi_28610153	0.89	0.73	0.95	0	10
IL1RN	A: 212657_s_at	0.75	0.87	N/A	2	11
	A: 212659_s_at	0.77	0.84	N/A	4	11
	A: 216243_s_at	0.75	0.86	N/A	6	11
	A: 216244_s_at	0.13	0.07	N/A	4	11
	A: 216245_at	0.21	0.11	N/A	10	11
	D: NM_173841_at	0.80	0.88	N/A	12	33
	D: NM_000577_at	0.80	0.88	N/A	12	33
	D: NM_173842_at	0.80	0.88	N/A	12	33
	D: NM_173843_at	0.84	0.86	N/A	15	42
	F: GC02P113591_at	0.83	0.86	N/A	16	44
	H: gi_27894315	0.78	0.88	N/A	0	23
ICAM1	A: 202637_s_at	0.63	0.73	0.97	7	11
	A: 202638_s_at	0.62	0.72	0.98	4	11
	A: 215485_s_at	0.71	0.73	0.94	3	11
	D: NM_000201_at	0.70	0.76	0.99	14	33
	F: GC19P010247_at	0.70	0.77	0.99	14	33
	H: gi_4557877	0.72	0.74	0.97	0	20
SOD2	A: 215078_at	0.25	0.35	N/A	10	11

Continued on next page

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
	A: 215223_s_at	0.15	0.28	N/A	7	11
	A: 216841_s_at	0.18	0.39	N/A	3	11
	A: 221477_s_at	0.32	0.44	N/A	10	11
	D: NM_001024466_at	0.16	0.33	N/A	6	12
	D: NM_000636_at	0.19	0.37	N/A	10	22
	D: NM_001024465_at	0.16	0.33	N/A	6	13
	F: GC06M160020_at	0.20	0.36	N/A	20	33
	H: gi_67782304	0.20	0.39	N/A	0	12
TRAF1	A: 205599_at	0.61	0.50	0.88	6	11
	D: NM_005658_at	0.61	0.50	0.88	6	11
	F: GC09M122704_at	0.61	0.50	0.88	6	11
	H: gi_53759116	0.59	0.47	0.89	0	5
ZFP36	A: 201531_at	0.84	0.86	N/A	5	11
	A: 213890_x_at	-0.01	-0.46	N/A	8	11
	D: NM_003407_at	0.84	0.86	N/A	5	11
	F: GC19P044589_at	0.84	0.86	N/A	5	11
	H: gi_141802261	0.85	0.82	N/A	0	6
PTGS2	A: 204748_at	0.91	0.71	0.97	4	11
	D: NM_000963_at	0.91	0.71	0.97	4	11
	F: GC01M184907_at	0.91	0.71	0.97	4	11
	H: gi_4506264	0.89	0.72	0.95	0	9
TNFAIP3	A: 202643_s_at	0.78	0.82	0.97	4	11
	A: 202644_s_at	0.87	0.85	0.93	6	11
	D: NM_006290_at	0.82	0.83	0.96	10	22
	F: GC06P138230_at	0.82	0.83	0.96	10	22
	H: gi_26051241	0.80	0.82	0.98	0	13
DUSP2	A: 204794_at	0.75	0.66	N/A	5	11
	D: NM_004418_at	0.75	0.66	N/A	5	11
	F: GC02M096230_at	0.75	0.66	N/A	5	11
	H: gi_12707563	0.74	0.60	N/A	0	6
ADM	A: 202912_at	0.80	0.67	0.92	5	11
	D: NM_001124_at	0.80	0.67	0.92	5	11
	F: GC11P010283_at	0.80	0.67	0.92	5	11
	H: gi_4501944	0.82	0.67	0.94	0	6
CROP	A: 203804_s_at	0.44	0.56	N/A	5	11
	A: 208835_s_at	0.43	0.36	N/A	5	11
	A: 220044_x_at	0.43	0.44	N/A	4	11
	D: NM_016424_at	0.49	0.50	N/A	13	32
	D: NM_006107_at	0.49	0.45	N/A	13	30
	F: GC17P046151_at	0.48	0.48	N/A	14	33
	H: gi_52426741	0.46	0.47	N/A	0	17
NFκBIA	A: 201502_s_at	0.81	0.73	N/A	4	11
	D: NM_020529_at	0.81	0.73	N/A	4	11
	F: GC14M034940_at	0.81	0.73	N/A	4	11
	H: gi_10092618	0.82	0.77	N/A	0	7
JUNB	A: 201473_at	0.44	0.44	0.94	7	11
	D: NM_002229_at	0.44	0.44	0.94	7	11
	F: GC19P012763_at	0.44	0.44	0.94	7	11
	H: gi_44921611	0.54	0.44	0.73	0	4
Ø	all Affymetrix	0.61	0.59	0.88	5.58	11.16
	best Affymetrix	0.73	0.71	0.92	4.47	11.00
	Dai	0.67	0.67	0.91	8.82	20.59
	best Dai	0.73	0.69	0.91	7.88	18.69
	Ferrari	0.73	0.69	0.91	10.81	19.56
	Hummert	0.72	0.68	0.89	0.00	10.25

Evaluating the PM and MM probes statistically, the MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis. Following this recommendation and using only those results for the correlation analysis that are marked as 'present' the mean correlation coefficient increases from 0.59 to 0.66 (0.74 including only the best performing probesets). Hence, incorporating the Affymetrix detection call indeed improves the correlation, but using alternative CDFs is still better than using the Affymetrix probesets and the detection call.

Analyzing the MAQC reference dataset using the RMA suite, the results are almost in accordance with those of the Etanercept data described above. The mean correlation coefficient for all 1,000 genes is 0.47 for the Affymetrix CDF (0.71 incorporating only the best probeset for each gene). Using the Dai CDF, the mean correlation increases to 0.63 (0.64 for the best probesets). With the Ferrari and the new CDF the mean correlations are 0.63 and 0.58, respectively. The detailed results for all MAQC genes can be downloaded.

Discussion

Results from microarray experiments contain considerably high error rates [36]. Due to error propagation, it is of particular importance to minimize errors in the beginning of the analysis chain [37]. Therefore, especially the pre-processing of the chip data has to be done as accurate as possible. Many efforts were spent on these problems before [38], such as the notable results of the 'Golden Spike Project' [6]. The question which statistical method should be adequately chosen is even more complicated if experimental data from different laboratories are incorporated in one single analysis [39].

For microarray analyses algorithms are essential which combine the 11-20 probepair intensities for a given gene and define a measure of expression that represents the amount of the corresponding mRNA species. In this study, two of these algorithms are compared, the robust multi-array analysis algorithm (RMA) and the Affymetrix Microarray Suite MAS5. Applying both algorithms to the Etanercept dataset RMA outperforms MAS5 on average. Other studies revealed similar results. However, their performance is assumed to be dependent on the actual dataset [40]. In fact, normalisation steps are applied after the probe to probeset summarization. Some of these steps depend on global parameters (e.g. mean of total gene expression) which depend on the total set of probesets. Therefore, identical probesets within different CDFs vary slightly in the final gene expression values.

Analyzing the probes of the Affymetrix microarrays discloses many inaccuracies. A large number of problematic probes are based on the fact that Affymetrix had to rely on genome annotation available at the time the chips were designed (U133A and U133B: 2001; U133 Plus 2.0 and Mouse 430 2.0: 2003). Because genome annotation improves permanently, the chip design does not properly match the present annotations anymore. Due to compatibility reasons, Affymetrix is not able to keep the design of their microarrays up to date.

The problem of cross-hybridization is well known. The first work on custom CDFs examining this error source was published by the group of Dai in 2005 [21]. They created a large amount of high quality custom CDFs related to different reference databases. Some probes, causing cross-hybridizations, are deleted from the probesets, but the filter is quite loose, so the number of problematic probes decreased but did not vanish. The use of the new CDFs can avoid full length, i.e., 25 mer long, cross-hybridizations completely. Cross-hybridization of shorter fragments are very difficult to handle due to the fact that the number of putative bindings grows exponentially the shorter the considered fragments are. Hence, if all putatively cross-hybridizing probes are excluded the amount of measurable genes will be reduced extremely.

The underlying gene annotation which is used for sequence alignment has a big impact on the number of cross-hybridizations. Manually curated mRNA sequences have a high chance of missing transcripts. Therefore, the inclusion of computational proposed gene annotations decreases the

number of false negative predicted cross-hybridizations. The drawback is that a number of false positive hybridizations increases. A more strict approach should be preferred, because it does not significantly decrease the number of covered transcripts as there is a high amount of availabe probes. In this study, the exclusion of XM-RefSeq-accessions results in smaller differences between the different CDFs in the number of putative cross-hybridzing transcripts. Interestingly, the correlation coeefficents of the newly created probesets do not change significantly.

Evaluating the four different CDFs, we figured out that the usage of the original Affymetrix CDFs leads to poorer results than the usage of the custom CDFs, although the best Affymetrix probesets give equally good or even better results than the other CDFs. However, as already mentioned, this cannot be taken into account, because it is not known which probeset will perform best before the correlation analysis is completed. The Dai probesets perform better, but the problem of several probesets representing a single gene had not been solved. Although multiple probesets representing the same gene are expected to show similar signal intensities, this is in fact not the case [14], [15]. Thus, it is difficult to decide which of the probesets matching the same gene is the most reliable. The Ferrari and the new CDFs comprise only one probeset per gene, which is of great advantage. The Ferrari CDFs perform slightly better on the Etanercept dataset and both CDFs perform equally well on the MAQC data.

The analysis of the genes for which qRT-PCR results are available in the Etanercept dataset as well as in the MAQC dataset clearly shows higher correlation coefficients in the MAQC dataset. This is most likely due to the fact that the U133 Plus 2.0 arrays which were used in the MAQC dataset outperform the older U133A microarrays.

The results show that probesets consisting of more probes, i.e., larger probesets, lead to better correlation results in general, whereas smaller probesets perform poorer. This finding correlates to the results of the study of Cui *et al.* [14] that merges probesets matching the same transcript. Interestingly, probesets containing many putative cross-hybridizations do not considerably perform poorer than probesets containing only a few. This result is very surprising, because it is obvious that cross-hybridization is one of the main error sources in microarray experiments [8], [9]. The normalization step in the two summarizing algorithms RMA and MAS5 may explain for that because they possibly eliminate some cross-hybridization effects. Another explanation is that leaving out the problematic probes does not compensate the influence of cross-hybridization. Unspecific binding leads to two types of error: (i) false-positives because RNA fragments bind to problematic probes of the probeset, and (ii) gene expression events are missed or underestimated, leading to a false-negative error if the RNA fragments are already bound to problematic probes of other probesets (competitive binding). Custom CDFs can only account for the first type of error by leaving out the problematic probes, the second effect could only be overcome by better array design.

The newly created CDFs perform slightly poorer than the Ferrari probesets (0.72 vs. 0.73) on the Etanercept dataset and equally well on the much larger MAQC dataset. On the one hand, the Ferrari CDFs can obviously countervail the negative effect by their much larger probesets in comparison to the new CDFs. On the other hand, using the new CDFs, putative cross-hybridizations are systematically excluded whereas using the Ferrari CDFs, the negative effect vanishes for statistical reasons due to the larger probesets. For exact studies, it is better to avoid a putative error source instead of averaging the cross-hybridization effects out as the Ferrari CDFs do. In addition, it has to be mentioned that the new CDFs provide as good or better results as the other CDFs using only about half the amount of probes (HG-U133A: 44 %, HG-U133B: 35 %, HG-U133 Plus 2.0: 37 %, Mouse Genome 430 2.0 Array: 52 %). Hence, designing new microarrays without the problematic probes, the dimension can be reduced by half without loosing any information and minimize the costs of the technology tremendously. Future microarray design using only the good probes and incorporating probesets of large sizes like in the Ferrari CDFs will certainly provide optimal solutions.

Methods

Probe Analysis

For the detection of putative cross-hybridizations by sequence alignment, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23]. For the U133A and the U133 Plus 2.0 the RefSeq release from 05/14/07 was used (download from ftp://ftp.ncbi.nih.gov/-refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz), for the U133B the realease from 01/10/08, and for the Mouse 430 2.0 microarray the release from 05/09/08 (∼M_musculus/mRNA_Prot/mouse.rna.fna.gz) was used. These parameters were applied: ValW = 7, ValE = 1000, ValHspmax = 1.

In this work all those RefSeq accession numbers beginning with XM or NM are used. The XM-identifiers indicate mRNA-RefSeq-accessions which are produced by computationally annotated genome submissions. The NM-identifier show that the RefSeq records are subsequently curated. Using both accessions in our model leads to more predicted cross-hybridizations which increases the reliability of the specificity of the probes.

The strand direction of the probes is analyzed. For each probe it is counted how many genes match and checked whether the match has the correct direction, i.e., the sense direction.

All BLAST hits for different transcript isoforms are merged, i.e., if the probe hybridizes to alternative splice variants of one gene but not to another gene, it is considered as unambiguous. Different gene isoforms of one gene are identified by screening the gene descriptions of the RefSeq database.

All probes matching only one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. For the creation of the new CDFs only the good probes are used. The probe sequences are annotated with GeneIDs derived from RefSeq. The GeneID is a database cross-reference qualifier, which supports access to the Entrez Gene database and provides a distinct tracking identifier for a gene or locus. Probes sharing the same GeneID are grouped together into a new probeset. The intersection between two different probesets is therefore always empty for all probesets. The size of the newly created probesets is variable and not fixed to 11 like in the Affymetrix CDFs.

Datasets

Two datasets were chosen for the validation of the different CDFs. The first of the two datasets chosen is derived from a study published by Koczan *et al.* [27] analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points. One Affymetrix HG-U133A array was performed for each time point. The data are available at the Array Express archive [41] with the accession number E-MTAB-11.

Expression levels of 16 genes were measured by quantitative real-time RT-PCR (qRT-PCR) performed with TaqMan assay reagents according to the manufacturer's instructions on a 7900 High Throughput Sequence Detection System (Applied Biosystems, Foster City, CA, USA) using predesigned primers and probes (GAPDH Hs99999905_m1, ICAM1 Hs00164932_m1, TNFAIP3 Hs00234713_m1, IL1B Hs00174097_m1, NFκBIA Hs00153283_m1, IL8 Hs00174103_m1, ADM Hs00181605_m1, TNF Hs00174128_m1, IL6 Hs00174131_m1, IL1RN Hs00277299_m1, SOD2 Hs00167309_m1, TRAF1 Hs00194638_m1, ZFP36 Hs00185658_m1, PTGS2 Hs00153133_m1, DUSP2 Hs00358879_m1, CROP Hs00538879_s1, JUNB HS00357891_s1).

The threshold cycle values (C_T) for specific mRNA expression in each sample were normalized to the C_T values of GAPDH mRNA in the same sample. This provides ΔC_T values that were used for the correlation analysis. In total, 816 qRT-PCR experiments were performed and complement the 51 microarray experiments (17 patients, 3 time points) described in [27]. The results of the qRT-PCR experiments can be downloaded.

The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. All available 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here. The MAQC data discussed in this publication are available in NCBI's Gene Expression Omnibus with accession number GSE5350. In addition, the nine genes for which qRT-PCR results are available in both datasets, are analyzed in more detail.

Comparison of the CDFs

For the comparison of different CDFs, the correlation between the microarray and the qRT-PCR experiments is used [29], [30]. As a performance index the Pearson correlation coefficient of the microarray results and the qRT-PCR experiments is calculated. Calculation of the Spearman correlation coefficient showed very similar results (data available at http://sysbio.hki-jena.de/software).

The raw chip data (CEL Files) are analyzed using the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and the Affymetrix Microarray Suite MAS5 [32] in combination with the different CDFs.

The MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis [32]. For an additional correlation analysis only the 'present' probesets are used to check if the calculated detection call from MAS5 gives a good prediction for the probeset quality.

Availability

The newly created CDFs as R-packages and additional files are available for download at http://www.sysbio.hki-jena.de/software. Using the CDFs does not interfere with all further steps of microarray analysis.

Acknowledgements

This work was supported by the ILRS - International Leibniz Research School for Microbial and Molecular Interactions (CH, FH) and by the ERASysBio+ project Linconet (MW).

References

[1] S. Heber and B. Sick, "Quality assessment of Affymetrix GeneChip data," *OMICS A Journal of Integrative Biology*, vol. 10, no. 3, pp. 358–368, Fall 2006.

[2] O. Modlich and M. Munnes, "Statistical framework for gene expression data analysis," *Methods in Molecular Biology*, vol. 377, pp. 111–130, May 2007.

[3] P. K. Tan, T. J. Downey *et al.*, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5676–5684, October 2003.

[4] A.-K. Järvinen, S. Hautaniemi *et al.*, "Are data from different gene expression microarray platforms comparable?" *Genomics*, vol. 83, no. 6, pp. 1164–1168, June 2004.

2.1 C. Hummert, F. Mech, F. Horn, M. Weber, S. Drynda, U. Gausmann, R. Guthke. *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11)*, Vol.I, 16-22. CSREA Press, Las Vegas, USA, July 18-21, 2011.

[5] A. Brazma, P. Hingamp *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, December 2001.

[6] S. E. Choe, M. Boutros *et al.*, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biology*, vol. 6, no. 2, p. R16, January 2005.

[7] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, no. 9, p. e84, May 2005.

[8] Z. Chen, M. McGee *et al.*, "A distribution free summarization method for Affymetrix GeneChip® arrays," *Bioinformatics*, vol. 23, no. 3, pp. 321–327, February 2007.

[9] A. C. Cambon, A. Khalyfa *et al.*, "Analysis of probe level patterns in Affymetrix microarray data," *BMC Bioinformatics*, vol. 8, no. 146, May 2007.

[10] H. R. Ueda, S. Hayashi *et al.*, "Universality and flexibility in gene expression from bacteria to human," *The Proceedings of the National Academy of Sciences (US)*, vol. 101, no. 11, pp. 3765–3769, March 2004.

[11] Affymetrix Inc, "GeneChip custom express array design guide. part no. 700506 rev. 4," Tech. Rep., 2003.

[12] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–821, July 2003.

[13] B. M. Bolstad, R. A. Irizarry *et al.*, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 3, pp. 185–193, January 2003.

[14] X. Cui and A. E. Loraine, "Consistency analysis of redundant probe sets on Affymetrix three-prime expression arrays and applications to differential mRNA processing," *PLoS One*, vol. 4, no. 1, p. 4229, January 2009.

[15] T. R. Hughes, M. Mao *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nature Biotechnology*, vol. 19, no. 4, pp. 342–347, April 2001.

[16] M. A. Stalteri and A. P. Harrison, "Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips," *BMC Bioinformatics*, vol. 8, no. 13, January 2007.

[17] X. Liu, M. Milo *et al.*, "Probe-level measurement error improves accuracy in detecting differential gene expression," *Bioinformatics*, vol. 22, no. 17, pp. 2107–2113, September 2006.

[18] G. Sanguinetti, M. Milo *et al.*, "Accounting for probe-level noise in principal component analysis of microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3748–3754, October 2005.

[19] F. Ferrari, S. Bortoluzzi *et al.*, "Novel definition files for human GeneChips based on GeneAnnot," *BMC Bioinformatics*, vol. 8, no. 446, November 2007.

[20] V. Chalifa-Caspi, I. Yanai *et al.*, "GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes," *Bioinformatics*, vol. 20, no. 9, pp. 1457–1458, June 2004.

[21] M. Dai, P. Wang *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, no. 20, p. e175, November 2005.

[22] J. Lu, J. C. Lee *et al.*, "Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays," *BMC Bioinformatics*, vol. 8, no. 108, March 2007.

[23] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, pp. W20–W25, July 2004.

[24] R. Yelin, D. Dahary *et al.*, "Widespread occurrence of antisense transcription in the human genome," *Nature Biotechnology*, vol. 21, no. 4, pp. 379–386, April 2003.

[25] H. Kiyosawa, N. Mise *et al.*, "Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized," *Genome Research*, vol. 15, no. 4, pp. 463–474, April 2005.

[26] M. Barnes, J. Freudenberg *et al.*, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5914–5923, October 2005.

[27] D. Koczan, S. Drynda *et al.*, "Molecular discrimination of responders and nonresponders to anti-TNFalpha in rheumatoid arthritis therapy by Etanercept," *Arthritis Research & Therapy*, vol. 10, p. R50, May 2008.

[28] L. Shi, L. H. Reid *et al.*, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, September 2006.

[29] J. S. Moray, J. C. Ryan, and F. M. Van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR," *Biological Procedures Online*, vol. 8, no. 1, pp. 175–193, December 2006.

[30] R. D. Canales, Y. Luo *et al.*, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, September 2006.

[31] R. A. Irizarry, B. Hobbs *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, April 2003.

[32] Affymetrix Inc, "Statistical algorithms description document. whitepaper. part no. 701137 rev. 3," Tech. Rep., 2002.

[33] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 22, no. 7, pp. 789–794, July 2006.

[34] J. Seo and E. P. Hoffman, "Probe set algorithms: is there a rational best bet?" *BMC Bioinformatics*, vol. 7, no. 395, August 2006.

[35] S. D. Pepper, E. K. Saunders *et al.*, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 273, July 2007.

[36] M. Eisenstein, "Microarrays: Quality control," *Nature*, vol. 442, pp. 1067–1070, August 2006.

[37] M. Grabe, *Measurement Uncertainties in Science and Technology*. New York: Springer Press, 2005.

[38] P. Boutros, "Systematic evaluation of the microarray analysis pipeline," in *Proceedings of the First 11th MGED Meeting: 1-4 September 2008; Riva del Garda*, G. Sherlock, Ed. MGED, 2008, pp. 16–27.

[39] H.-C. Liu, C.-Y. Chen *et al.*, "Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 570–579, August 2008.

[40] K. Shedden, W. Chen *et al.*, "Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data," *BMC Bioinformatics*, vol. 6, no. 26, 2005.

[41] H. Parkinson, M. Kapushesky *et al.*, "ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D868–D872, January 2009.

2.2 F. Horn, H.-W. Nützmann, V. Schroeckh, R. Guthke, C. Hummert.
***Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11)*, Vol.I, 3-9. CSREA Press, Las Vegas, USA, July 18-21, 2011.**

Optimization of a microarray probe design focusing on the minimization of cross-hybridization

Fabian Horn*, Hans-Wilhelm Nützmann†, Volker Schroeckh†, Reinhard Guthke*, Christian Hummert*
*Research Group Systems Biology / Bioinformatics
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI)
D-07745 Jena, Germany
Email: fabian.horn@hki-jena.de
†Department of Molecular and Applied Microbiology
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI)
D-07745 Jena, Germany

***Abstract*—Microarrays are extensively used for high-throughput gene expression analyses in molecular biology and medicine. Microarray analysis is reliable if the probe binds specifically to the intended target transcript. Cross-hybridizations of microarray probes is one of the main systematic errors which is influenced by microarray probe design. Newly released genome annotations make it possible and necessary to improve given probe designs in order to reduce this source of error.**

We present a new method which evaluates and optimizes existing probe designs in a modular way. The workflow can include existing software and it can be adapted to additionally required probe design criteria. A microarray probe design optimization which focuses on the avoidance of cross-hybridization was exemplarily done for *Aspergillus nidulans*. We show the high impact of the underlying structural genome annotation on the probe design process. The new probe design was experimentally evaluated with the help of the mean variance of internal technical replicates.

We recommend to check existing microarray probe designs for cross-hybridization and if possible to optimize them based on the preferred version of genome annotation.

INTRODUCTION

Microarray technique represents one of the most common methods to carry out genome-wide research based on sequenced genomes. A microarray experiment consists of many different steps which are all vulnerable to errors. The experiments do not necessarily satisfy the underlying assumption that probes representing the same gene show the same signal intensity [1]. Signal intensities depend strongly on the probe sequence. Different sequences generate varying physical properties, which are important for hybridization [1]. Certain probe sequences are also vulnerable of building secondary structures which inhibit hybridization with the transcript (target) [2]. These properties are used for the design of microarray probes [3].

The main objective of this process is to increase the reliability of signal intensities by reducing systematic errors caused by the probe sequences. Among other criteria, the hybridization process itself is modeled with the help of criteria, like melting temperature uniformity [4]–[6], GC-content [7], and Free Gibbs energy [8].

In order to guarantee a high discrimination between between targets and non-targets the probe design is checked for cross-hybridization. Cross-hybridization is a non-target binding between a probe and a transcript fragment which is not intended to match the probe. In fact, cross-hybridizations are one of the main sources of systematic error that even affect the well-established microarrays from Affymetrix [9] and also affect tiling arrays [10], [11]. Several studies have shown that nucleotide sequences are capable of hybridization, even when the complementary region between probe and transcript has only a 70% identity [1], [12], [13]. Besides this identity threshold, non-specific bindings additionally need a longest continuous complementary substring of a certain minimum length [4], [7], [12]. Signal intensities in the data may therefore result from unspecific bindings and may lead to false-positively detected target genes.

There are approaches to cope with cross-hybridizations by creating new alternative Chip Definition Files (CDFs) of existing custom microarray probe designs [14], [15]. These methods correct and avoid the impact of cross-hybridizations by disregarding a certain fraction of the probes during data analysis. It is evident that the same level of information can be obtained with less probes spotted onto the microarray. The reannotation of oligonucleotide libraries is therefor the first step in order to obtain up-to-date microarray probe designs [16], [17]. It is preferable to exclude existing cross-hybridizing oligonucleotides during the process of microarray probe design [18]. The removal of unspecific probes in existing probe designs leads to a reduced production cost for each utilized data point. New alternative probes can be spotted onto the microarray which leads to a higher genome coverage rate or a higher number of replicates per gene.

Many different algorithms have been proposed for designing microarray probes [3]. Each algorithm has a different scope of application and consequently utilizes different probe design criteria and, as a consequence, perform differently. The different foci make it difficult to directly evaluate and compare the quality of the proposed algorithms with a theoretical optimization criterion. In fact, the limitations of the applied experimental protocol determine suitable probe design criteria

and narrow down the set of available methods. It is favorable to use an extendable und adjustable general framework where different probe design criteria can be integrated [19], [20]. This allows to adjust for application-specific design criteria and enables the reuse of existing modular software.

In this work, we present a workflow which evaluates and optimizes an already given reference probe design concerning the avoidance of cross-hybridization. The optimization of the probe design is exemplarily done for a microarray for *Aspergillus nidulans* which is the model organism of filamentous fungi [21]. The obtained probe design minimizes unspecific bindings. We show that this design yields more reliable results. In addition to the avoidance of cross-hybridizations, it is possible to include different design criteria which are applied due to experimental constraints.

RESULTS

Evaluation of reference probe design

The mapping of a given full-genome probe design for *Aspergillus nidulans* was examined by aligning the probe sequences against three structural genome annotations: two different versions available from the Broad institute and one version from the Central Aspergillus Data REpository (CADRE). (The annotations are referred to as BROAD (2008), BROAD (2010) and CADRE (2009), respectively.) For further information see methods and figure 1.

The given reference probe design contains 342 and 377 probes that cross-hybridize with BROAD (2008) and CADRE (2009) annotation, respectively (see table I). Regarding the newer BROAD (2010) annotation, only 148 probes are considered as cross-hybridizing.

Using the BROAD (2008) annotation and the CADRE (2009) annotation respectively, 317 and 313 probes in the reference probe design do not match any transcript with a perfect sequence identity.

The reference probe design contains probes that do not match any transcript in the given annotation: 74 probes using BROAD (2008), 204 probes using CADRE (2009), and 993 probes using the newer BROAD (2010).

The reference probe design does not cover a number of predicted transcripts in each annotation: 442 transcripts in BROAD (2008), 478 transcripts in CADRE (2009), and as much as 968 transcripts in BROAD (2010).

The evaluation also calculated the thermodynamic properties of the probe sequences. The result reveals that the melting temperatures of the probes are in a narrow range between $80°C$ and $90°C$. This desirable property is achieved with the help of a uniform GC content of 48%.

In summary, the reference probe design is not optimized for any of the used annotations. Depending on the used annotation version, 7...11% of all probes do not match a transcript unambiguously. The current annotation causes a poorer performance which can be seen explicitly at the decreased number of perfect probes (see table I).

Probe design optimization

A large fraction of the reference probe design is not optimized for any genome annotation and needs improvement. The objective of the optimization was to get 50 nucleotides long optimized oligonucleotides which use the BROAD (2008) annotation. The probes should be placed at the 5'-end because cDNA is used in the hybridization protocol.

The workflow of the proposed probe design method can be separated into three consecutive steps (see figure 2). In the first step new probe candidates are generated with the help of ArrayOligoSelector [22]. In the second step, probe candidates are evaluated with the help of evaluation tool to exclude cross-hybridizations (see above). The evaluation also calculates thermodynamic properties that are used in a following third step - a further selection. The selection step is necessary because only one probe sequence per gene is spotted.

The optimization showed that it was not possible to find a valid unique probe sequence for every transcript. In order to achieve a higher gene coverage, design criteria have to be mitigated. New probe candidates are iteratively generated from intervals of elongated transcript sequences. 1,303 probes were found in the smallest interval of 600 basepairs (see table II). In the next two steps the interval is extended to 1,200 and 2,000 basepairs which only led to 30 and 24 additional probes, respectively. In a last step, probes that are capable of cross-hybridization are exceptionally allowed. The relaxation of this last criterion increased gene coverage with 53 additional probes. In total, the softening of the design criteria leads to 107 additionally covered genes in the presented study.

Finally, there are 188 genes without a valid probe sequence which leads to a transcript coverage rate of 98,2%.

TABLE II
Composition of the gene coverage

	Number of genes
Reference probe design (validated probes)	9,103
Probe design optimization:	
Sequence range: 0...600 bp	1,303
Sequence range: 0...1,200 bp	30
Sequence range: 0...2,000 bp	24
Ignoring cross-hybridizations	53
Uncovered genes	188
Total	10,701

The gene coverage of the probe design results from different steps. A high number of genes are covered by validated probes from the reference probe design. The probe design optimization leads to an additional number of covered genes which are obtained by iteratively mitigating the probe design criteria. First, the transcript sequences are extended and at last the cross-hybridization criterion is relaxed. In the end, some genes remain that are not covered by any valid probe.

The comparison of the resulting new probe design with the given reference probe design shows that the new probe design is optimized for the BROAD (2008) annotation (see table I). The new design consists of 10,512 probes (99.5%) which match perfectly and do not show any cross-hybridization. Notably, the comparison with the reference probe design

TABLE I
Results of probe classification and gene coverage

Annotation	BROAD (2008)		CADRE (2009)		BROAD (2010)	
Probe design	old	new	old	new	old	new
Number of probes	10,676	10,566	10,676	10,566	10,676	10,566
Perfect probes	9,943 (93.1%)	10,513 (99.5%)	9,782 (91.6%)	10,287 (97,4%)	9,535 (89.3%)	9,535 (90.2%)
Cross-hybridizing	342 (3.2%)	53 (0.5%)	377 (3.5%)	133 (1.3%)	148 (1.4%)	63 (0.6%)
Not identical match	317 (3.0%)	0 (0.0%)	313 (2.9%)	3 (0.0%)	0 (0.0%)	0 (0.0%)
Not matching	74 (0.7%)	0 (0.0%)	204 (1.9%)	143 (1.4%)	993 (9.3%)	968 (9.2%)
Total number of genes	10,701	10,701	10,546	10,546	10,560	10,560
Covered genes	10,259 (95,9%)	10,513 (98.2%)	10,068 (95.5%)	10,282 (97.5%)	9,592 (90.8%)	9,561 (90.5%)
Uncovered genes	442 (4.1%)	188 (1.8%)	478 (4.5%)	264 (2.5%)	968 (9.2%)	999 (9.5%)

Probes from the reference probe design (old) and the optimized (new) probe design have been mapped to different genome annotations. Probes either show no systematic error (perfect probes), hybridize with multiple genes (cross-hybridizing), match one gene without total sequence identity (not identical match), or do not match any transcript at all (not matching). The lower part of the table shows how many genes of the annotation are perfectly covered by the corresponding probe design.

demonstrates that 254 genes are additionally covered in the optimized design while avoiding systematic errors.

Remarkably, there are also 214 extra covered genes if the CADRE (2009) annotation is used as basis. This result is achieved by a lower number of genes with systematic errors. The number of potentially cross-hybridizing probes is only 133 in comparison to 377 probes in the reference probe design. Only three specific probes match a transcript without a total sequence identity whereas this number is much higher in the reference probe design with 313 probes. Changes in the annotation lead to 143 probes that do not match any given transcript in contrast to 204 probes in the reference probe design. The number of uncovered genes is 264 which corresponds to a gene coverage rate of 97,5%.

For the current BROAD (2010) annotation the gene coverage of the probe design is reduced to 90,5% and the number of covered genes (9,561 vs. 9,592) is comparable between both versions of the probe design. Nevertheless, the new probe design still minimizes systematic errors. 63 probes are prone to cross-hybridizations in contrast to 148 probes in the reference probe design. A high number of 968 probes do not match any transcript at all which is again comparable to the performance of the reference probe design.

In summary, the new probe design reduces systematic errors regardless of the structural annotation used. Concerning the cross-hybridizations, the improvements become apparent. For BROAD (2008) and CADRE (2009) the gene coverage of the optimized probe design is higher as compared to the reference probe design.

Impact of genome annotation

The evaluation of different probe designs clearly highlights the big impact of the underlying structural genome annotation on the results (see table I).

The new probe design was optimized for the BROAD (2008) annotation and the gene coverage could be increased to 98.2%. The optimization also takes effect for the CADRE (2009) annotation with a gene coverage rate of 97.5%. In comparison to the current BROAD (2010) annotation, the gene coverage rate is dramatically decreased to 90.5% which is comparable with the coverage rate of the reference probe design. The same trend for gene coverage can be seen for the reference probe design where the gene coverage rate also decreases to 90.8% if the BROAD (2010) annotation is used.

The differences in gene coverage result from probes which are vulnerable to systematic errors. The new probe design shows only a small fraction of probes that are prone to cross-hybridization in the BROAD (2008) annotation. This number doubles if the CADRE (2009) annotation is used. In the BROAD (2010) annotation only a few cross-hybridizing probes occur. This results from the increased number of error prone probes that do not match any transcript at all. The number of unmatched probes constitutes the largest error source which is affected by the change in genome annotation.

In the probe design optimized for BROAD (2008), the number of probes that are not classified as perfect increases from 54 (0.6%) over 279 (2,7%) to 1031 (9.8%) for the BROAD (2008), CADRE (2009), and BROAD (2010) annotation, respectively. The same trend holds for the non-perfect probes from the reference probe design which increases from 733 (6.9%) over 894 (8.4%) to 1141 (10.7%). It is noteworthy that a change in the annotation basis can cause almost 10% of all probes to be classified as invalid.

Experimental Validation

The new probe design is optimized for the minimization of systematic errors in respect to the BROAD (2008) annotation. Especially, the avoidance of cross-hybridization should significantly increase the reliability of experimental data. An indicator for improved reliability is a lower mean variance of internal technical replicates over each array. For this purpose, a highly reproducible experiment with the reference and the new probe design was performed (see methods). Microarray raw data was obtained from *Aspergillus nidulans - Streptomyces rapamycinicus* interaction experiments. The co-cultivation was performed because most of the secondary metabolite gene clusters are silent under laboratory conditions and the fungal-bacterial interaction leads to specific activations [23], [24]. (Microarray data is available at Gene Expression Omnibus - GSE25266.)

First, a microarray experiment using the reference probe design was performed. The following second experiment used the same experimental setup except that the new optimized probe design was used. It is not possible to compare the variance of probes for each single gene individually because an altered probe sequence has an essential impact on the signal intensities. Probes with the same nucleotide sequences have a high Pearson correlation coefficient of 0.928 whereas altered probe sequences result in a low correlation coefficient of 0.554.

Overall, the internal technical replicates should however show the desirable property of a lower mean variance over each array. The first experiment with the reference probe design used 4,148 internal technical replicates for 164 genes whereas the second experiment with the new probe design had 1,368 internal technical replicates for 157 genes. The mean variance of the internal technical replicates for the reference probe design range from 4.27...4.7 for the biological sample of the *A. nidulans-S. rapamycinicus* interaction and *A. nidulans* wildtype, respectively (see table III). The new probe design shows a lower mean variance of internal replicates, namely 3.55 for the wildtype and 3.69 for the interaction sample. This change corresponds to an reduction of the mean variance with a ratio of 0.76...0.86. The application of a Shapiro-Wilk test indicated a normal distribution of signal intensities with a p-value < 0.05. An F-test with a subsequent Holm-correction confirmed the significance of the change in variance. All adjusted p-values are below 0.05. The lower mean variance over each array of the new probe design is significant. In summary, the statistical analysis of experimental results obtained from technical replicates supports the applied method and shows that the new probe design yields more reliable results.

TABLE III
Mean variance of technical replicates over each array

Sample/Replicates	Old design	New design	ratio
A. nidulans rep1	4.79	3.73	0.78
A. nidulans rep2	4.69	3.85	0.82
A. nidulans mean	4.70	3.55	0.76
A. nidulans+S. rapamycinicus rep1	4.00	3.76	0.94
A. nidulans+S. rapamycinicus rep2	4.51	4.12	0.91
A. nidulans+S. rapamycinicus mean	4.27	3.69	0.86

Mean variance of internal technical replicates which were included in the first microarray experiment using the reference probe design and in the second experiment using the optimized probe design. Two technical replicates were used for each of the biological samples (*A. nidulans* and *A. nidulans + S. rapamycinicus*). Mean variances and the ratio between both experiments are given for each replicate and for the mean of each biological sample.

DISCUSSION

Probe Design Optimization

The reliability of used probe designs need to be checked whenever new genome annotations are available [16]–[18]. For *A. nidulans* the evaluation of the given reference probe design showed this necessity as it contains many systematic errors and the possibility to cover a higher number of transcripts is not fully exploited. The approach combines both steps - the evaluation of reference probe designs and the design of new probes. Frequently, a probe design already exists and probe sequences that satisfy the design criteria do not need to be recalculated.

It is challenging to find the right software which applies all probe design criteria described above. The usage of a modular workflow which allows for the flexible integration of different design criteria helps to adjust the oligonucleotide design to the specific experimental requirements. This approach allows the integration of own probe design criteria and existing software. A similar workflow with different steps has been proposed and implemented in the tool Teolenn [19]. This framework was not considered due to the missing integration of re-evaluation of existing probe designs.

For the generation of probe candidates many different software tools have been proposed. In the proposed workflow we decided to use ArrayOligoSelector [22] which applies a large fraction of required design criteria and was recommended in an evaluation of custom microarray applications [3]. The tool chosen is interchangable and should be orientated at the specific probe design requirements.

In this working example, hybridization are only considered if the alignment has a minimum sequence identity of 90% (see methods). This way, cross-hybridization can not be fully excluded because it was shown that it already occurs at a identity of 70% [12]. If the evaluation tool uses a more stringent cut-off, more probes are classified as invalid and more genes are not covered by any probe. The setting of this threshold is always a trade-off because the aim is to cover as many genes as possible while excluding cross-hybridizations. Hybridization with *S. rapamycinicus* transcripts was not checked because poly-dT-priming ensures that only eukaryotic RNA is amplified.

Due to the experimental objectives, the position of the probe and the GC content range were used as design criteria. The filtering for a narrow GC content range is a fast calculabe filter criterion and effectively obtains a close melting temperature uniformity. The computational costly application of the Nearest-Neighbor Model [25] gives a more precise estimation of the melting temperature. A direct application of this methods for probe design is limited because it assumes that both nucleotide strands interact freely in a solution which is not the case for microarrays.

Generally, if more probe design criteria are applied more probe candidates are excluded leading to a lower number of valid probe sequences. Overall, the used approach utilizes only a small set of all possible probe design criteria. Despite that, it was not possible to find a valid probe for 188 genes. Several factors contribute to this number of uncovered genes: If the gene annotation allows for transcripts which are shorter than the desired probe length or consist of highly repetitive sequence stretches, it is apparently not possible to find a valid probe sequence for them. In addition, a few transcripts share the same 3'-end, represent different splice variants, or are positioned within the same locus but on different strands. Finally, some sequences are at different loci, but have a high

sequence similarity which may result from gene homology.

It is a challenge to select non-unique probes to identify the presence of targets in a sample. This problem especially arises when designing microarrays for the study of host-parasites or host-pathogen interactions. In this exemplarily application, we manually chose 8 non-unique probes for genes of high interest. The manual selection of non-unique probes is very time-consuming and not standardized. It is desirable to integrate proposed approaches which select non-unique probes [10], [11], [18], [26], [27].

Impact of annotation databases

It is crucial to decide what structural genome annotation should be used as reference for the probe design. The reason are new genome assemblies and differences in the formal definition of the characteristics of a gene. Large fractions of the annotation of *Aspergillus nidulans* are done automatically with the help of bioinformatic tools. It is evident that with ongoing research the annotation of transcripts is subject to change. A large fraction of the oligonucleotide libraries can not be unambiguously matched to existing structural genome annotations [16], [17]. The progress in laboratory research and, consequently, the related manual curation of genome annotations lead to more robust genome annotations.

Experimental Validation

The quality of the designed probes, and therefore the quality of the proposed approach, is eventually assessed by experimental validation. Probe sequences may be evaluated with spike-in experiments [28], self-hybridization experiments with the analysis of gene coverage [19], correlation of experimental data with probe design criteria [19], [20], experimental selection of probes [20], and the usage of internal technical replicates [29]. Without a transcriptome golden standard the impact of modifications can not be directly linked to the overall improvement of the array design. Spike-in experiments, Northern Blots, and qRT-PCR can only focus on a selection of chosen transcripts and are therefore not suited to assess a whole microarray probe design. Furthermore, it is not distinguishable which specific probe design criterion has an effect on the results because the criteria are mutually dependent. An altered probe sequence, for instance, does not only change the sequence similarity but also the physical properties of the probe and the hybridization. Nevertheless, it is necessary for an improvement of the design process.

In this study we used internal replicates to assess the quality of the new probes. Internal technical replicates allow to check for the performance of probes regardless of the experimental influences. If the mean variance of internal technical replicates is low, the reproducibility of the probe signal is high. The results are more reliable which is also the goal of the proposed evaluation method which aims at the avoidance of systematic errors. A significant decrease of mean variances of internal replicates over each array was observed. This shows that the probes have a higher signal reproducibility. The optimized microarray probe design is more reliable as it has been shown with the help of statistically significant lower mean variance of the internal technical replicates.

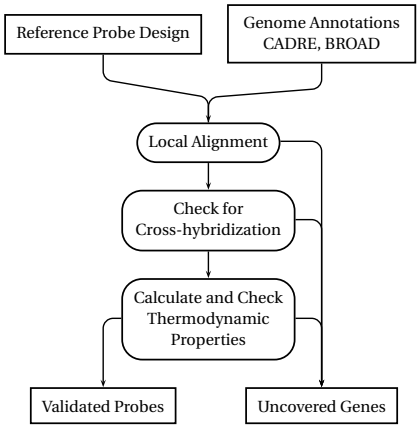


Fig. 1. **Schematic overview of evaluation process.** A reference probe design is locally aligned to selected genome annotation databases. Probes that cross-hybridize are filtered and thermodynamic properties of the hybridization are calculated for further assessment.

Materials and Methods

Material

The probe design from febit biomed GmbH (Heidelberg, Germany) was as used as 'reference probe design' (see GSE25266 and [23]). It was analyzed regarding the structural genome annotations from BROAD institute [30] (two different versions downloaded October, 10th 2008 and February, 18th 2010) and from CADRE [31] (downloaded February, 16th 2009). The annotation versions are referred to as 'BROAD2008', 'BROAD2010', and 'CADRE2009', respectively.

Probe design evaluation

Probe sequences were aligned locally to the known corresponding transcripts with the help of FASTA (Parameters: expectation value 1.0, alignment type 0) [32]. The thermodynamic properties of each probe and the hybridization were calculated with the nearest-neighbor model [25], which is implemented in the freely available software MELTING (Parameters: '-Hdnadna -N0.2 -P0.0001 -Ksan98a') [33]. A probe is considered to match a transcript if there is at least one 16 basepairs long common subsequence and if both sequences share a sequence identity not less than 90%. Although literature suggests that hybridization already occurs at 70% sequence identity [12], a less stringent cut-off was applied. A stricter constraint dramatically decreases the number of valid probe sequences and prevents a full-genome probe design. All probes are finally classified into four classes. Probes that i) match perfectly, ii) cross-hybridize, iii) do not match any transcript, and iv) hybridize, but are not fully identical with the target sequence.

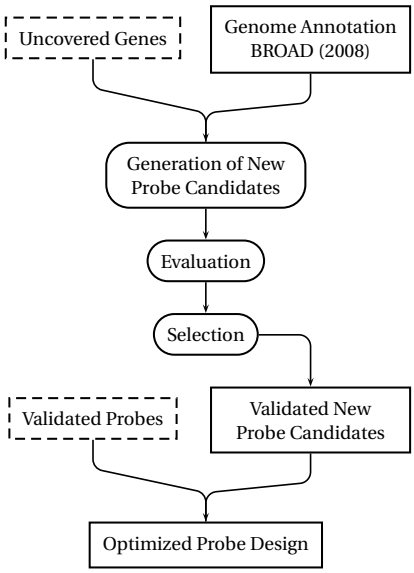


Fig. 2. **Workflow of probe design optimization.** New probe candidates are generated for the genes where there are no current valid probe sequences. Probe candidates are evaluated with the evaluation tool. If more than one probe candidate is valid, different selection criteria are applied to select the best optimized probe. The final new optimized probe design is obtained by the combination of these probe candidates with the validated probes from the reference probe design. (Dashed lines represent results from the evaluation of the reference probe design.)

Generation of new probe candidates

New probe candidates were generated for genes where no perfect matching probe is given in the reference probe design. Different available algorithms could be applied for this step. In this study, we integrated the public available tool ArrayOligoSelector (Parameters: target GC percentage 48.0, length of oligonucleotides 50), number of oligos per gene 5) [22], which utilizes sequence similarity, a given GC content range, tests for low-complexity regions, and recognition of self-complementary sequences. The transcript sequence was trimmed to the first 600 basepairs to reduce computational time and to meet the probe design objective of placing the probe near the 3'-end. The generated probe candidates were checked with the help of the evaluation tool described above. This guarantees that new probe candidates meet the given cross-hybridization criterion and that systematic errors are avoided.

Selection of validated probes

The aim of covering the full genome of *Aspergillus nidulans* allows only to spot one oligonucleotide for each gene considering the given spotting density constraint. Validated newly generated probe candidates are preferred if they are positioned at the 3'-end of the transcript. If several probes exist within an overlapping close interval of 50bp, the following second design criterion is applied: Probes with a GC content closest to the mean GC content of the reference probe design are chosen if the difference to the mean is below 8%. This ensures similar thermodynamic properties of all probes. After the application of these criteria, at most one single probe candidate per gene remains.

Iterative softening of design criteria

We start with a transcript sequence ranging from the 3'-end to 600 basepairs. In order to get a better gene coverage, the used transcript sequence range was iteratively extended to 1,200 and 2,000 basepairs for the remaining uncovered genes. Finally, the stringent cross-hybridization criterion was relaxed for the remaining uncovered genes. Hence, probe candidates are even considered if they are vulnerable to cross-hybridization. Probe sequences were chosen manually for genes of high biological interest and without a valid probe candidate. The manually chosen sequences minimize the number of cross-hybridizations and fall within the narrow range of the desired mean GC content ($\pm$ 8%).

Merging the valid probes from the reference probe design with the selected new probe candidates resulted in the new and optimized probe design (see GSE25266 and figure 2).

In summary, in this study the following probe design criteria have been applied: cross-hybridization, sequence complexity, lack of self-binding, GC content, and position on reverse strand.

Experimental validation

Microarray raw data was obtained from *Aspergillus nidulans* - *Streptomyces rapamycinicus* interaction experiments [23]. The fungus was incubated over night in liquid Aspergillus minimal media (AMM) and shifted into new media. Actinomycetes were cultivated in M79 medium and 5 ml of the culture was added to 100ml AMM and both organisms were further incubated at 37°C. The reference culture is incubated without bacteria. After 3 h, each sample was split into two identical technical replicates and total-RNA was isolated using RiboPure-Yeast Kit (Applied Biosystems) according to the manufacturers instructions. cDNA synthesis, labeling and microarray measurements were done by febit biomed GmbH. In the first experiment, the reference probe design was used. The same samples were used for the second experiment where the new probe design was utilized (see figure 3). All microarray data is compliant to the MIAME standard and can be accessed at GEO (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE25266.

Both microarrays contain several internal technical replicates which can be used to assess the quality of microarray design. The comparability of both experiments is shown with the help of Pearson correlation coefficients of the signal intensities. The mean variance of the internal technical replicates were calculated over each array. The application of a Shapiro-Wilk tests for a normal distribution of signal intensities. The significance of the change in variances are evaluated by an F-test and a subsequent Holm-correction.

2.2 F. Horn, H.-W. Nützmann, V. Schroeckh, R. Guthke, C. Hummert. *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11)*, Vol.I, 3-9. CSREA Press, Las Vegas, USA, July 18-21, 2011.

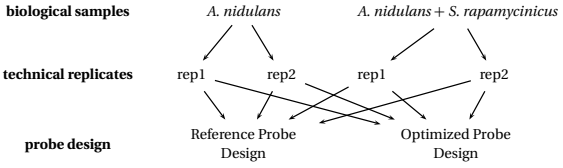


Fig. 3. **Schematic overview of Experimental Design.** In the first sample *A. nidulans* is cultivated without *S. rapamycinicus* and in the second sample it is co-cultivated with *S. rapamycinicus*. Each sample was split in two identical technical replicates. For each replicate a microarray experiment is performed with the reference and the new optimized probe design. The microarrays contain internal technical replicates that are used for the experimental validation.

CONCLUSION

We proposed a worflow for the evaluation and optimization of existing microarray probe designs. This workflow is capable of integrating existing software and adjusting the probe design according to the experimental requirements. Exemplarily, this approach has been applied for a full-genome microarray for *Aspergillus nidulans* with the focus on avoiding systematic errors, especially cross-hybridizations. The reduction of cross-hybridization improves the reliability of the probe design which can be seen in a reduced mean variance of internal technical replicates over each array. We showed the high influence of different structural genome annotations on the design process. It is recommended to check for cross-hybridizations based on a current version of genome annotation prior to microarray data analysis.

ACKNOWLEDGMENT

This work was supported by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS) and the Jena School for Microbial Communication (JSMC).

REFERENCES

[1] T. R. Hughes *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." *Nat Biotechnol*, vol. 19, no. 4, pp. 342–347, Apr 2001.

[2] R. T. Koehler and N. Peyret, "Effects of DNA secondary structure on oligonucleotide probe binding efficiency." *Comput Biol Chem*, vol. 29, no. 6, pp. 393–397, Dec 2005.

[3] S. Lemoine *et al.*, "An evaluation of custom microarray applications: the oligonucleotide design challenge." *Nucleic Acids Res*, vol. 37, no. 6, pp. 1726–1739, Apr 2009.

[4] Z. He *et al.*, "Empirical establishment of oligonucleotide probe design criteria." *Appl Environ Microbiol*, vol. 71, no. 7, pp. 3753–3760, Jul 2005.

[5] A. Halperin *et al.*, "On the hybridization isotherms of DNA microarrays: the Langmuir model and its extension," *J. Phys.: Condens. Matter*, vol. 18, pp. S463–S490, 2006.

[6] N. Ono *et al.*, "An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays." *Bioinformatics*, vol. 24, no. 10, pp. 1278–1285, May 2008.

[7] Y. A. Chen *et al.*, "A multivariate prediction model for microarray cross-hybridization." *BMC Bioinformatics*, vol. 7, p. 101, 2006.

[8] H. Binder and S. Preibisch, "Specific and nonspecific hybridization of oligonucleotide probes on microarrays." *Biophys J*, vol. 89, no. 1, pp. 337–352, Jul 2005.

[9] J. Mieczkowski *et al.*, "Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements." *BMC Bioinformatics*, vol. 11, p. 104, 2010.

[10] S. Graf *et al.*, "Optimized design and assessment of whole genome tiling arrays." *Bioinformatics*, vol. 23, no. 13, pp. i195–i204, Jul 2007.

[11] P. Bertone *et al.*, "Design optimization methods for genomic DNA tiling arrays." *Genome Res*, vol. 16, no. 2, pp. 271–281, Feb 2006.

[12] M. Kane *et al.*, "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays," *Nucleic Acids Res.*, vol. 28, no. 22, pp. 4552–7, Nov 2000.

[13] S.-K. Rhee *et al.*, "Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays." *Appl Environ Microbiol*, vol. 70, no. 7, pp. 4303–4317, Jul 2004.

[14] F. Ferrari *et al.*, "Novel definition files for human GeneChips based on GeneAnnot." *BMC Bioinformatics*, vol. 8, p. 446, 2007.

[15] M. Dai *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." *Nucleic Acids Res*, vol. 33, no. 20, p. e175, 2005.

[16] P. Casel *et al.*, "sigReannot: an oligo-set re-annotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters." *BMC Proc*, vol. 3 Suppl 4, p. S3, 2009.

[17] P. B. T. Neerincx *et al.*, "Oligorap - an oligo re-annotation pipeline to improve annotation and estimate target specificity." *BMC Proc*, vol. 3 Suppl 4, p. S4, 2009.

[18] H.-H. Chou, "Shared probe design and existing microarray reanalysis using PICKY." *BMC Bioinformatics*, vol. 11, p. 196, 2010.

[19] L. Jourdren *et al.*, "Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments." *Nucleic Acids Res*, vol. 38, no. 10, p. e117, Jun 2010.

[20] F. Bidard *et al.*, "A general framework for optimization of probes for gene expression microarray and its application to the fungus Podospora anserina." *BMC Res Notes*, vol. 3, p. 171, 2010.

[21] W. Vongsangnak and J. Nielsen, *Aspergillus: Molecular Biology and Genomics*. Caister Academic Press, Jan 2010, ch. Bioinformatics and Systems Biology of Aspergillus, pp. 61–84.

[22] Z. Bozdech *et al.*, "Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray." *Genome Biol*, vol. 4, no. 2, p. R9, 2003.

[23] V. Schroeckh *et al.*, "Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans." *Proc Natl Acad Sci U S A*, vol. 106, no. 34, pp. 14 558–14 563, Aug 2009.

[24] A. A. Brakhage and V. Schroeckh, "Fungal secondary metabolites - strategies to activate silent gene clusters." *Fungal Genet Biol*, Apr 2010.

[25] J. SantaLucia and D. Hicks, "The thermodynamics of DNA structural motifs." *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–440, 2004.

[26] I. Lee *et al.*, "Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray," *Nucleic Acids Res.*, vol. 32, no. 2, pp. 681–90, Feb 2004.

[27] S. Rimour *et al.*, "GoArrays: highly dynamic and efficient microarray probe design," *Bioinformatics*, vol. 21, no. 7, pp. 1094–103, Apr 2005.

[28] I. V. Yang, "Use of external controls in microarray experiments." *Methods Enzymol*, vol. 411, pp. 50–63, 2006.

[29] D. L. Leiske *et al.*, "A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray." *BMC Genomics*, vol. 7, p. 72, 2006.

[30] J. E. Galagan *et al.*, "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." *Nature*, vol. 438, no. 7071, pp. 1105–1115, Dec 2005.

[31] J. E. Mabey *et al.*, "Cadre: the Central Aspergillus Data REpository." *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D401–D405, Jan 2004.

[32] W. R. Pearson, "Flexible sequence similarity searching with the FASTA3 program package." *Methods Mol Biol*, vol. 132, pp. 185–219, 2000.

[33] N. L. Novère, "MELTING, computing the melting temperature of nucleic acid duplex." *Bioinformatics*, vol. 17, no. 12, pp. 1226–1227, Dec 2001.

Kapitel 3

Analyse der biologischen Varianz

In der als Kapitel 3.1 eingefügten Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) stehen biologische Varianzunterschiede in der Genexpression im Vordergrund. Zentral leistet diese Arbeit die Auswertung eines umfangreichen Microarray-Experiments.

Es wird zwischen gelenkdegenerierenden Entzündungserkrankungen unterschieden, der Osteoarthrose (OA) und der rheumatoiden Arthritis (RA). Die Differentialdiagnose der beiden Erkrankungen ist nicht einfach. Es gibt jedoch Hinweise auf genetische Marker, die eine schnelle Diagnose erlauben würden. Um solche Marker zu identifizeren, wird ein Datensatz mit Affymetrix U133A Microarray-Daten über 31 Patienten ausgewertet. Die Arraydaten stehen im Gene Expression Omnibus zur Verfügung (Edgar et al. [2002]).

Bei der Analyse der Datensätze ist auffällig, dass nicht nur die Mittelwerte der Gruppen mitunter stark voneinander abweichen (hoher Fold-Change), sondern dass es gelegentlich auch große Abweichungen in der Varianz der einzelnen Gruppen gibt. Diese Varianzunterschiede werden mit dem Varianzfold (siehe Seite 19) quantifiziert. Die biologische Erklärung dieser Varianzunterschiede ist, dass in diesem Fall Krankheit (RA respektive OA) sich nicht durch

eine einfache Hoch- oder Herabregulierung widerspiegelt, sondern allgemeiner durch eine Missregulierung in eine beliebige Richtung – also eine falsche Regulierung, die bei einigen Patienten nach unten, bei anderen aber nach oben abweicht.

Die Hypothese, dass es Probesets gibt, deren Varianz zwischen den Patientengruppen signifikant abweicht, wird mit dem Brown-Forsythe-Test (siehe Seite 18) überpüft. Um nur solche Probesets auszuwählen, bei denen auch tatsächlich ein erheblicher Varianzunterschied besteht, wird der neue Begriff des Varianzfolds definiert. Zur Überprüfung der Relevanz der so selektierten Transkripte werden diese auf KEGG-Pathways gematcht. Eine Untersuchung der KEGG-Pathways ergibt tatsächlich, dass solche Pathways getroffen werden, die bereits als mit rheumatischen Erkrankungen assoziiert bekannt sind.

Die Untersuchung zeigt, dass rheumatische Erkrankungen in der Tat nicht nur durch Hoch- beziehungsweise Herabregulierung eines Gens oder von Genen erklärt werden kann, sondern vielmehr allgemeiner eine Missregulierung, das heißt Abweichungen in der Varianz, die Krankheit erklärt. Weiter zeigt die Arbeit, dass eine Analyse der Streuungsparameter bei Genexpressionsexperimenten sinnvoll ist, dass also nicht nur Mittelwertvergleiche in der Auswertung biologischer Phänomene in der Genexpressionsanalyse herangezogen werden können, sondern dass eine Untersuchung von Varianzunterschieden tatsächlich relevante Ergebnisse erbringt.

In der dem Kapitel 3.2 entsprechenden Arbeit *Quantification of growth-defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover* (Ullmann-Zeunert et al. [2013]) wird die Stickstoffverteilung innerhalb von Tabakpflanzen untersucht. Anstelle von Varianzunterschieden werden in dieser Arbeit Ähnlichkeiten, also Abschnitte mit hoher Korrelation, gesucht.

Die Arbeit behandelt die pflanzliche Abwehr von Herbivoren. Eine Strategie der pflanzlichen Abwehr ist die sogenannte induzierte Abwehr. Hierbei sind

die Abwehrmechanismen nicht ständig ausgeprägt, sondern werden als direkte Reaktion auf eine Beschädigung aktiviert. Im Vergleich zur konstitutiven (dauerhaften) Abwehr werden deutlich weniger Ressourcen (Energie und damit Nährstoffe) benötigt. Viele Pflanzen, wie auch der wilde Tabak (*Nicotiana attenuata*), verwenden Jasmonsäure in der induzierten Abwehr. Der wichtigste Stoff für Tabak in der Abwehr von Herbivoren ist jedoch Nikotin.

Im Vergleich von wildem Tabak (*Nicotiana attenuata*) Widtyp (WT) mit zwei Mutanten irLOX3 und irMYB8, deren *Jasmonate Defense Signaling Pathway* gestört ist, zeigt sich eine unterschiedliche Stickstoffverteilung innerhalb der Pflanzen. Stickstoff ist insofern interessant, da Nikotin Stickstoff enthält. Die Verteilung des Stickstoffs in der Pflanze auf verschiedene Verbindungen, wie Nikotin, Caffeoylputrescin oder Dicaffeoylspermidin, ergibt ein bestimmtes Muster für jedes Individuum. Für die zwei Mutanten und den Wildtyp werden unter jeweils zwei Bedingungen (Kontrolle und unter Einfluss von herbivoren Speichelsekret) die Stickstoffmuster aufgenommen. Nun werden jeweils verschiedene Pflanzenteile betrachtet und es wird nach Ähnlichkeiten beziehungsweise Unterschieden zwischen den Stickstoffmustern gesucht.

Für diese Aufgabe wird der Korrelationskoeffizient als Abstand zweier Vektoren interpretiert (Siehe Seite 24). In der Arbeit wird Kendalls τ (siehe Seite 23) für alle Kombinationen berechnet und in einer Heatmap dargestellt. So kann sehr schnell erkannt werden, welche Mutanten und Bedingungen sich gleichmäßig (konkordant) und welche sich verschieden (diskordant) verhalten. Durch die Analysen ergeben sich Hypothesen über den funktionalen Pathway in der induzierten Abwehr.

3.1 R. Huber, C. Hummert, U. Gausmann, D. Pohlers, D. Koczan, R. Guthke, R. W. Kinne. *Arthritis Research & Therapy*, 10:R98, August 2008.

Available online http://arthritis-research.com/content/10/4/R98

Research article

Open Access

Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane

René Huber[1,2], Christian Hummert[3], Ulrike Gausmann[4], Dirk Pohlers[1], Dirk Koczan[5], Reinhard Guthke[3] and Raimund W Kinne[1]

[1]Experimental Rheumatology Unit, Department of Orthopedics, University Hospital Jena, Waldkrankenhaus 'Rudolf Elle', Klosterlausnitzer Str. 81, 07607 Eisenberg, Germany
[2]Institute for Clinical Chemistry, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
[3]Systems Biology/Bioinformatics Group, Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany
[4]Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany
[5]Proteome Center Rostock, University of Rostock, Schillingallee 69, 18055 Rostock, Germany

Corresponding author: Raimund W Kinne, Raimund.W.Kinne@med.uni-jena.de

Received: 25 Oct 2007 Revisions requested: 5 Dec 2007 Revisions received: 16 Jul 2008 Accepted: 22 Aug 2008 Published: 22 Aug 2008

Arthritis Research & Therapy 2008, **10**:R98 (doi:10.1186/ar2485)
This article is online at: http://arthritis-research.com/content/10/4/R98

Abstract

Introduction Rheumatoid arthritis (RA) is a chronic inflammatory and destructive joint disease characterized by overexpression of pro-inflammatory/pro-destructive genes and other activating genes (for example, proto-oncogenes) in the synovial membrane (SM). The gene expression in disease is often characterized by significant inter-individual variances via specific synchronization/desynchronization of gene expression. To elucidate the contribution of the variance to the pathogenesis of disease, expression variances were tested in SM samples of RA patients, osteoarthritis (OA) patients, and normal controls (NCs).

Method Analysis of gene expression in RA, OA, and NC samples was carried out using Affymetrix U133A/B oligonucleotide arrays, and the results were validated by real-time reverse transcription-polymerase chain reaction. For the comparison between RA and NC, 568 genes with significantly different variances in the two groups ($P \leq 0.05$; Bonferroni/Holm corrected Brown-Forsythe version of the Levene test) were selected. For the comparison between RA and OA, 333 genes were selected. By means of the *Kyoto Encyclopedia of Genes and Genomes*, the pathways/complexes significantly affected by higher gene expression variances were identified in each group.

Results Ten pathways/complexes significantly affected by higher gene expression variances were identified in RA compared with NC, including cytokine–cytokine receptor interactions, the transforming growth factor-beta pathway, and anti-apoptosis. Compared with OA, three pathways with significantly higher variances were identified in RA (for example, B-cell receptor signaling and vascular endothelial growth factor signaling). Functionally, the majority of the identified pathways are involved in the regulation of inflammation, proliferation, cell survival, and angiogenesis.

Conclusion In RA, a number of disease-relevant or even disease-specific pathways/complexes are characterized by broad intra-group inter-individual expression variances. Thus, RA pathogenesis in different individuals may depend to a lesser extent on common alterations of the expression of specific key genes, and rather on individual-specific alterations of different genes resulting in common disturbances of key pathways.

Introduction

Human rheumatoid arthritis (RA) is characterized by chronic inflammation and destruction of multiple joints, perpetuated by an abnormally transformed and invasive synovial membrane

ECM: extracellular matrix; IL: interleukin; IL2RG: interleukin 2 receptor gamma; JNK: c-jun kinase; KEGG: *Kyoto Encyclopedia of Genes and Genomes*; MAPK: mitogen-activated protein kinase; MMP: matrix metalloproteinase; NC: normal control; OA: osteoarthritis; PCR: polymerase chain reaction; RA: rheumatoid arthritis; RT-PCR: reverse transcription-polymerase chain reaction; SM: synovial membrane; TGF-β: transforming growth factor-beta; TNF: tumor necrosis factor; VEGF: vascular endothelial growth factor.

(SM), forming the so-called pannus tissue [1]. Many activated cell types contribute to the development and progression of RA. Monocytes/macrophages, dendritic cells, T and B cells, endothelial cells, and synovial fibroblasts are major components of the pannus [2-8] and participate in maintaining joint inflammation, degradation of extracellular matrix (ECM) components, and invasion of cartilage and bone [2,4] as well as fibrosis of the affected joints [9].

The extended analysis of gene expression profiles in RA SM during the last decades has revealed several relevant gene groups affecting development and progression of the disease. Central transcription factors involved as key players in RA pathogenesis are AP-1, NF-κB, Ets-1, and SMADs [10-12]. These factors show binding activity for their cognate recognition sites in the promoters of inflammation-related cytokines (for example, tumor necrosis factor-alpha [TNF-α], interleukin [IL]-1β, and IL-6 [3]) and matrix-degrading enzymes (for example, matrix metalloproteinase [MMP]-1 and MMP-3 [13,14]). The latter contribute to tissue degradation by destruction of ECM components, including aggrecan or collagen type I-IV, X, and XI [15].

The analysis of those comprehensive expression data has become feasible due to the implementation of microarray-based methods [16]. Therefore, a variety of comparisons can be performed, including differences in gene expression among different groups and/or individuals. In contrast to conventional differential gene expression analyses, the determination of inter-individual gene expression variances, often affecting gene expression of members of the same patient/donor group, is generally not considered in rheumatology, although those variances are known to be a characteristic of many diseases. In trisomy 21, for instance, inter-individual expression variances affect a number of tightly regulated genes. In addition, the variances are independent of the respective level of gene expression, and although only a minority of genes are affected, these genes are thought to be involved in the symptoms of trisomy 21 with the highest phenotypical differences [17]. Significant inter-individual expression variances have also been reported to affect the expression of telomerase subunits in malignant glioma [18] as well as protein tyrosine kinases and phosphatases in human basophils in asthma and inflammatory allergy [19]. The latter implies that such alterations may also play an important role within inflammatory diseases, reflected in either synchronization (that is, a loss of inter-individual gene expression variances) or desynchronization (that is, increased inter-individual gene expression variances) of gene expression within a group of different individuals/patients.

In RA, differences in gene expression profiles for specific genes among two subgroups of RA patients have been reported, but within these subgroups, the differences are limited to distinct expression levels without significant intra-subgroup expression variances [12]. To the best of our knowledge, there are as yet no reports on broad intra-group inter-individual gene expression variations among RA patients.

Interestingly, although the majority of reports show expression variances in tissues from patients with different diseases, variances have also been reported in normal tissues (for example, the human retina [20] or human B-lymphoblastoid cells [21]). In contrast to expression variations in diseases, the variations in normal donors are generally limited to a small number of genes (for example, 2.6% in the human retina [20]). To analyze inter-individual mRNA expression variances in RA, the occurrence of gene-specific expression differences in the SM was analyzed using the Bonferroni/Holm corrected Brown-Forsythe version of the Levene test for variance analysis [22-24] on the basis of genome-wide mRNA expression data in RA (n = 12), osteoarthritis (OA) (n = 10), and normal control (NC) (n = 9) synovial tissue.

Materials and methods

Patients and tissue samples

SM samples were obtained within 10 minutes following tissue excision upon joint replacement/synovectomy from RA (n = 12) and OA (n = 10) patients at the Department of Orthopedics, University Hospital Jena, Waldkrankenhaus 'Rudolf Elle' (Eisenberg, Germany). Tissue samples from joint trauma surgery (n = 9) were used as NCs (Table 1). After removal, tissue samples were frozen and stored at -70°C. Informed patient consent was obtained and the study was approved by the Ethics Committee of University Hospital Jena (Jena, Germany). RA patients were classified according to the American College of Rheumatology criteria [25], OA patients according to the respective criteria for OA [26].

Isolation of total RNA

Tissue homogenization, total RNA isolation, treatment with RNase-free DNase I (Qiagen, Hilden, Germany), and cDNA synthesis were performed as described previously [27].

Microarray data analysis

RNA probes were labeled according to the instructions of the supplier (Affymetrix, Santa Clara, CA, USA). Analysis of gene expression was carried out using U133A/B oligonucleotide arrays. Hybridization and washing procedures were performed according to the supplier's instructions and microarrays were analyzed by laser scanning (Hewlett-Packard Gene Scanner; Hewlett-Packard Company, Palo Alto, CA, USA). Background-corrected signal intensities were determined using the MAS 5.0 software (Affymetrix). Subsequently, signal intensities were normalized among arrays to facilitate comparisons between different patients. For this purpose, arrays were grouped according to patient/donor groups (RA, n = 12; OA, n = 10; and NC, n = 9). The arrays in each group were normalized using quantile normalization [28]. Original data from microarray analyses were deposited in the Gene Expression

Available online http://arthritis-research.com/content/10/4/R98

Table 1

Clinical characteristics of the patients at the time of synovectomy/sampling

Patients, total	Gender, male/female	Age, years	Disease duration, years	Rheumatoid factor, +/-	ESR, mm/hour	CRP[a], mg/L	Number of ARA criteria for RA	Concomitant medication (number)
Rheumatoid arthritis								
12	3/9	65.9 ± 2.9	15.8 ± 4.2	10/2	42.6 ± 6.2	31.9 ± 7.2	5.3 ± 2.1	MTX (5)
								Prednis. (10)
								Sulfas. (3)
								NSAIDs (9)
Osteoarthritis								
10	2/8	71.9 ± 2.0	6.2 ± 2.7	1/9	22.9 ± 4.0	7.6 ± 2.9	0.1 ± 0.1	NSAIDs (4)
								None (7)
Normal controls								
9	7/2	49.9 ± 6.7	0.4 ± 0.3	ND	ND	ND	0.0 ± 0.0	None

[a]Normal range: <5 mg/L. For the parameters of age, disease duration, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), and number of American Rheumatism Association (ARA) (now American College of Rheumatology) criteria for rheumatoid arthritis (RA), mean ± standard error of the mean is given. For the remaining parameters, numbers are provided. +/-, positive/negative; MTX, methotrexate; ND, not determined; NSAID, nonsteroidal anti-inflammatory drug; Prednis., prednisolone; Sulfas., sulfasalazine.

Omnibus of the National Center for Biotechnology Information (Bethesda, MD, USA) (accession number GSE12021 [29]).

Real-time reverse transcription-polymerase chain reaction

The data obtained by Affymetrix microarrays were validated for six selected genes (*IL13*, *MAPK8*, *SMAD2*, *IL2RG*, *PLCB1*, and *ATF5*) using real-time reverse transcription-polymerase chain reaction (RT-PCR). PCRs were performed as previously described using a Mastercycler® ep realplex (Eppendorf, Hamburg, Germany) and SYBR-green. To normalize the amount of cDNA in each sample, the expression of the housekeeping gene *GAPDH* (glyceraldehyde 3-phosphate dehydrogenase) was determined [27]. Product specificity was confirmed by (a) melting curve analysis, (b) agarose gel electrophoresis, and (c) cycle sequencing of the PCR products.

Statistical analysis of gene expression variance

This analysis did not concentrate on differently expressed genes, but on genes with different variances in the three patient groups [30]. The assumption of homogeneity of variance can be rejected by a variance analysis according to Levene [22]. The Brown-Forsythe version of this test was used [23]. For independent groups of data, the null hypothesis (that is, variances are equal) was tested.

To control the stability of the variance, the variance calculation was tested for 2, 3, 5, 7, and 10 samples per group. For fewer than 5 samples, the calculation did not reach stable results, but stable results were achieved for more than 5 patients. In addition, the results of the statistical tests were influenced by the number of samples in each group (that is, small groups did not reach statistical significance).

The *P* value can be obtained by calculating the value of the cumulative distribution function at the point *F*. This is equivalent to the integral of the probability density function of the normal distribution over the interval [0, *F*]. To prevent the accumulation of false-positives due to multiple comparisons, the very strict Bonferroni correction was used [31]. Alternatively, the less conservative Holm correction was applied for the correction of the data [24]. The application of the Holm correction yielded results comparable to those obtained by Bonferroni correction and pointed out only very few new genes.

The variance-fold is defined as the quotient of the variance of one group (for example, OA patients) and the variance of another group (for example, RA patients). If the variance in the second group is higher than 1, the result is the multiplicative inverse and the algebraic sign is inverted. This way, all groups can be compared:

$$VarFold = \begin{cases} var_x \geq var_y & : var_x / var_y \\ var_x < var_y & : -1 * (var_y / var_x) \end{cases}$$

The application of a variance filter before testing of the data (excluding variance-fold values between 2.5 and -2.5 from the analysis) yielded equivalent results compared with the initial data analysis including the *a posteriori* application of the Bonferroni or the Holm correction. Following *Kyoto Encyclopedia of Genes and Genomes* (KEGG) analysis (see below), the

same pathways/complexes were indicated and only the ranking of selected pathways/complexes was changed (for example, the ranking of cytokine–cytokine receptor interactions and the mitogen-activated protein kinase [MAPK] pathway were inverted).

Analysis of inter-individual gene expression variances

Relevant genes were selected using different criteria: (a) a significance level of $P \leq 0.05$ (Bonferroni/Holm corrected Brown-Forsythe version of the Levene test) for variance-fold values and (b) a cutoff value for absolute variance-fold levels of greater than 2.5 for higher variances in RA, OA, and NC, respectively. Using these criteria, 568 genes were selected for the comparison between RA and NC (307 with higher variances in RA and 261 with higher variances in NC) while 542 genes were used for the comparison OA versus NC (314 with higher variances in OA and 228 with higher variances in NC). Finally, 333 genes were selected for the comparison between RA and OA (186 with higher variances in RA and 147 with higher variances in OA). All selected genes are presented in Supplementary Table 1 (sorted according to absolute variance-fold values). Inter-individual variances of gene expression among the different groups were analyzed using predefined pathways and functional categories annotated by KEGG [32].

Mapping of probesets onto gene names

Gene names used for KEGG inputs follow the nomenclature of the HUGO Genome Nomenclature Committee [33] and are mostly derived from the Affymetrix annotation feature 'Gene Symbol' for the respective probeset. If required, corresponding RefSeqs were manually inspected.

Statistical KEGG analysis

To ensure that only KEGG pathways with a significant enrichment of more variant genes were obtained for further analyses, the χ^2 test statistic was used. Following the calculation of the expected frequency of affected genes in each pathway, the difference between the expected frequency and the absolute frequency was determined. All pathways with a difference of less than 2 were ignored. As a second criterion of the multilevel test, *P* values of less than or equal to 0.15 were considered statistically significant [34]. Pathways with insignificant *P* values were examined in detail and subdivided into two or more sub-pathways if possible. In some cases, *P* values for selected sub-pathways decreased considerably.

Results

Analysis of inter-individual gene expression variances in rheumatoid arthritis, osteoarthritis, and normal control synovial membrane

For the comparison of inter-individual gene expression variances between RA SM (n = 12) and NC SM (n = 9), 568 genes were used (307 with significantly higher variances in RA and 261 with significantly higher variances in NC; $P \leq 0.05$, Bonferroni/Holm corrected Brown-Forsythe version of the Levene test), resulting in the identification of 129 affected KEGG pathways/complexes in total (Supplementary Table 1a; shown for *IL13* and *CXCL13* in Figure 1). These pathways include 10 pathways significantly affected by higher gene expression variances in RA and 6 pathways significantly affected by higher gene expression variances in NC (in both cases $P \leq 0.15$, χ^2 test).

For the comparison of OA (n = 10) and NC (n = 9) SM, 542 genes were used (314 with significantly higher variances in OA and 228 with significantly higher variances in NC; Supplementary Table 1b). A total of 128 affected KEGG pathways/complexes were identified, including 7 pathways significantly affected by higher gene expression variances in OA and 4 pathways significantly affected by higher gene expression variances in NC.

The comparison of RA (n = 12) and OA (n = 10) SM was performed with 333 genes (186 with significantly higher variances in RA and 147 with significantly higher variances in OA; Supplementary Table 1c). This comparison culminated in the identification of 114 pathways, 3 of which were significantly affected by higher gene expression variances in RA and 4 of which were significantly affected by higher gene expression variances in OA.

Real-time reverse transcription-polymerase chain reaction validation

Validation of the microarray data by real-time RT-PCR was attempted in RA, OA, and NC samples for the genes *IL13*, *MAPK8*, *SMAD2*, *IL2RG*, *PLCB1*, and *ATF5*. In three cases (50%), the results of microarray analyses and real-time RT-PCR were equivalent for RA versus NC *(MAPK8*: variance-fold 9.8 versus 5.2; *IL2RG*: variance-fold 5.6 versus 8.9; *ATF5*: variance-fold 1.7 versus 2.3); in addition, two cases (33%) tended to result in comparable variance-fold values for microarray and real-time RT-PCR (*IL13*: variance-fold 12 versus 1.3; *SMAD2*: variance-fold 5 versus 1.1). In only one case (*PLCB1*; 17%), microarray analyses and real-time RT-PCR validation showed contradictory results (higher variance in NC versus higher variance in RA). For OA versus NC, comparable results were achieved (only *IL2RG* and *ATF5* showed contradictory results).

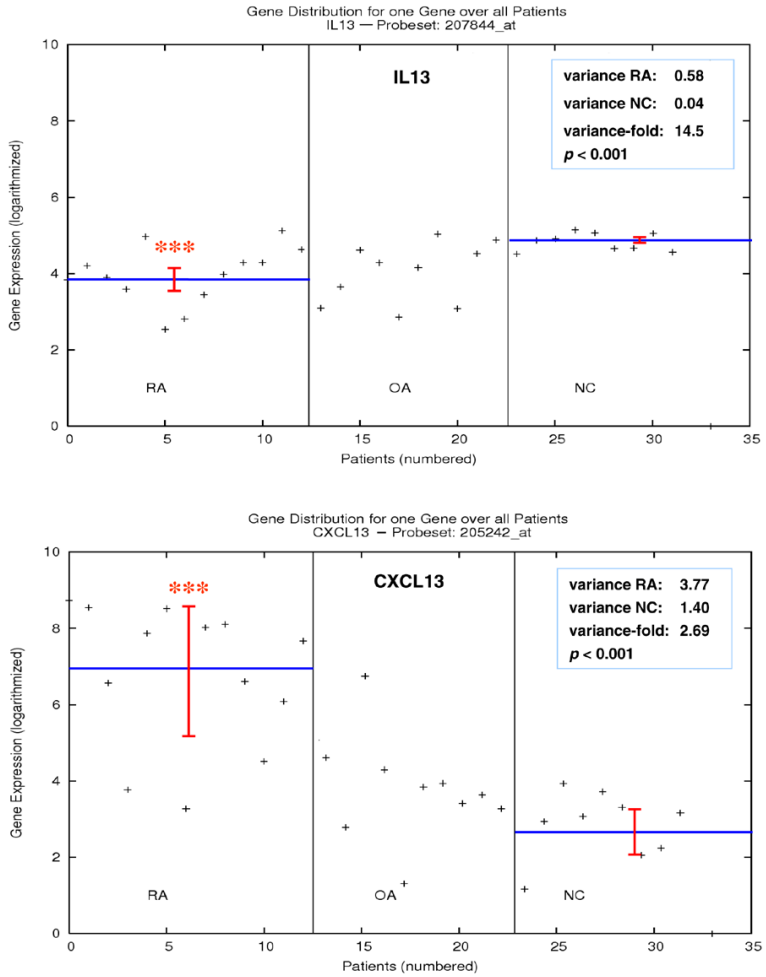
KEGG pathways identified in the comparison between rheumatoid arthritis and normal control

Pathways significantly affected by inter-individual gene expression variances in rheumatoid arthritis

Ten pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in the comparison between RA and NC, 7 of which were specific for RA, that is, did not appear in the comparison between OA and NC (for example, cytokine–cytokine receptor interactions; Figure 2). The occurrence of gene expression variances in the complete MAPK, transforming growth factor-beta (TGF-β), and

Available online http://arthritis-research.com/content/10/4/R98

Figure 1



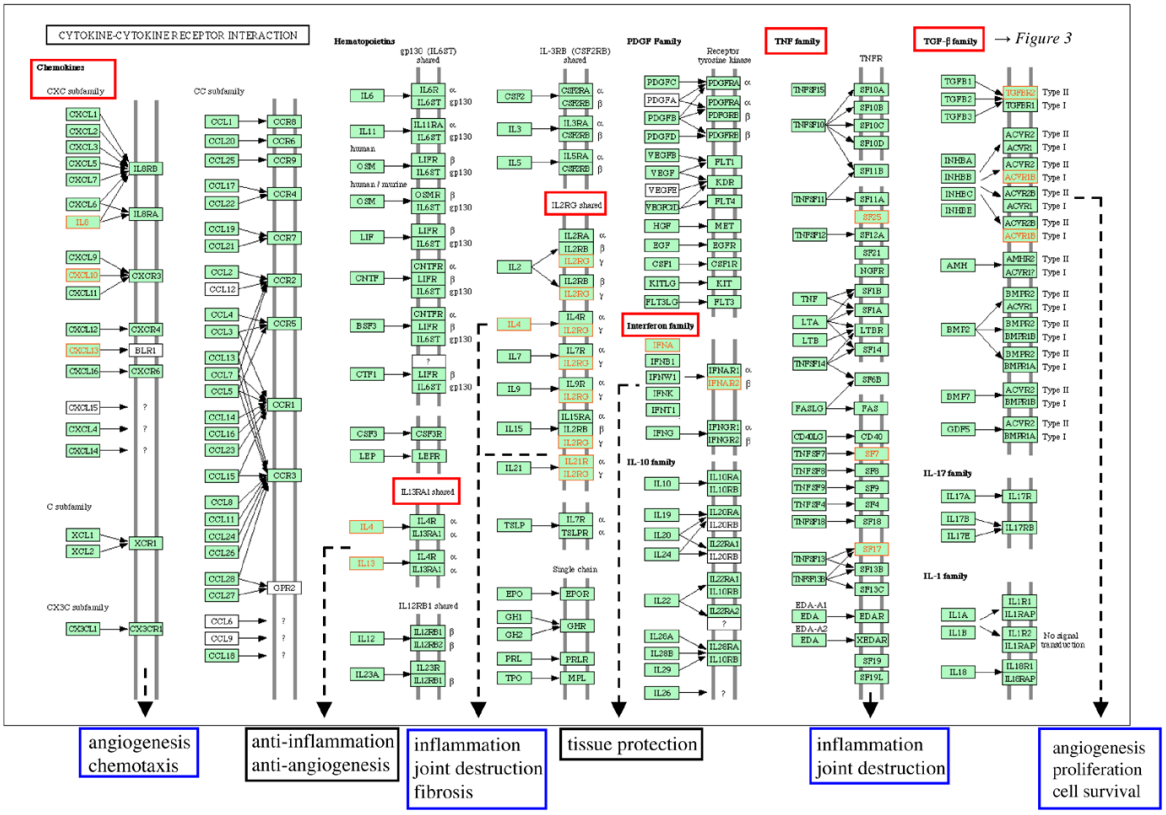
Gene-specific inter-individual gene expression variances. The graph shows the individual gene expression level of rheumatoid arthritis (RA) (n = 12) and osteoarthritis (OA) (n = 10) patients as well as normal control (NC) donors (n = 9) for *IL13* and *CXCL13* (cytokine–cytokine receptor interactions). The mean gene expression (blue line) and the intra-group inter-individual variances in RA and NC synovial membrane (red bar) are indicated, resulting in significantly enhanced variances among patients within the RA group ($P < 0.001$, Bonferroni/Holm corrected Brown-Forsythe version of the Levene test).

apoptosis pathways/complexes did not reach statistical significance. Interestingly, within these pathways, significantly affected sub-pathways/sub-complexes could be identified: the classical TGF-β sub-pathway (Figure 3), the classical and the c-jun kinase (JNK)/p38 MAPK sub-pathway(s) (Figure 4), and the sub-complex of anti-apoptosis (Figure 5). A complete list of significantly affected pathways/complexes is presented in Table 2.

Pathways significantly affected by inter-individual gene expression variances in normal control

Six pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in NC compared with RA, including the cell cycle and the Wnt (wingless-type MMTV integration site family) signaling pathway. All pathways/complexes were specific for NC. A complete list of significantly affected pathways/complexes is presented in Table 3.

Figure 2



Inter-individual mRNA expression variances among cytokine–cytokine receptor interactions in rheumatoid arthritis (RA) compared with normal control (NC). The graph shows genes affected by significant intra-group inter-individual mRNA expression variances in RA compared with NC ($P \leq 0.05$; Bonferroni/Holm corrected Brown-Forsythe version of the Levene test; labeled in red) among *Kyoto Encyclopedia of Genes and Genomes* (KEGG) cytokine–cytokine receptor interactions, including the respective sub-pathways ($P \leq 0.15$, χ^2 test; labeled in red). Cellular processes with potential influence on or relevance for RA pathogenesis (for example, inflammation, proliferation, and cell survival) are labeled in blue, and anti-inflammatory/anti-destructive processes are labeled in black.

KEGG pathways identified in the comparison between osteoarthritis and normal control

Pathways significantly affected by inter-individual gene expression variances in osteoarthritis

Seven pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in OA compared with NC. Among these pathways/complexes, six were specific for OA, including the complexes of apoptosis. A complete list of significantly affected pathways/complexes is presented in Table 4.

Pathways significantly affected by inter-individual gene expression variances in normal control

Four pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in NC compared with OA. Three of those were specific for NC, including the Toll-like receptor signaling pathway. A complete list of significantly affected pathways/complexes is presented in Table 5.

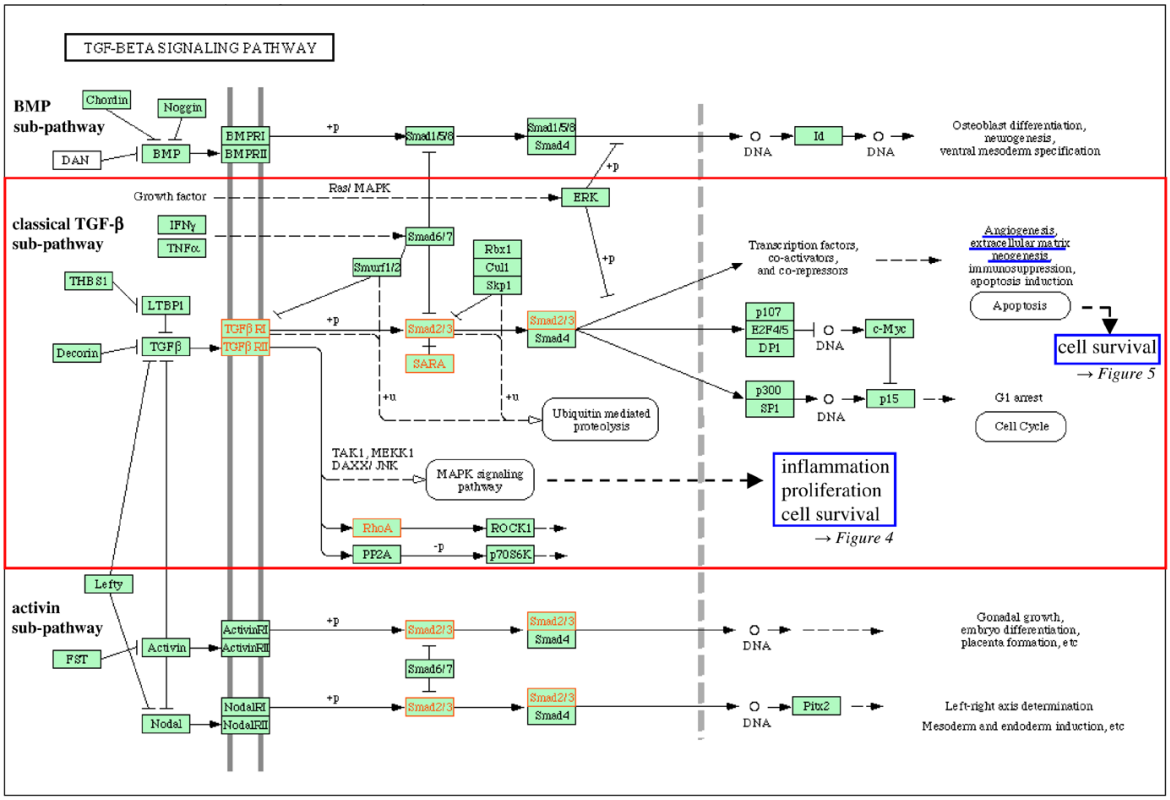
KEGG pathways identified in the comparison between rheumatoid arthritis and osteoarthritis

Pathways significantly affected by inter-individual gene expression variances in rheumatoid arthritis

Three pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in RA compared with OA. All pathways/complexes were specific for RA, including the vascular endothelial growth factor (VEGF) and the B-cell receptor signaling pathways. A complete list of significantly affected pathways/complexes is presented in Table 6.

Available online http://arthritis-research.com/content/10/4/R98

Figure 3



Inter-individual mRNA expression variances in the transforming growth factor-beta (TGF-β) signaling pathway in rheumatoid arthritis (RA) compared with normal control (NC). The graph shows genes affected by significant intra-group inter-individual mRNA expression variances in RA compared with NC ($P \leq 0.05$; Bonferroni/Holm corrected Brown-Forsythe version of the Levene test; labeled in red) in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) TGF-β signaling pathway. Among the three TGF-β family sub-pathways, the classical TGF-β sub-pathway is significantly affected by gene expression variances ($P \leq 0.15$, χ^2 test; indicated in red). TGF-β-regulated cellular processes with potential influence on or relevance for RA pathogenesis (for example, angiogenesis and cell survival) are labeled in blue.

Pathways significantly affected by inter-individual gene expression variances in osteoarthritis

Four pathways/complexes significantly affected by inter-individual mRNA expression variances were identified in OA compared with RA (for example, the complex of oxidative phosphorylation). All of them were specific for OA. A complete list of significantly affected pathways/complexes is presented in Table 7.

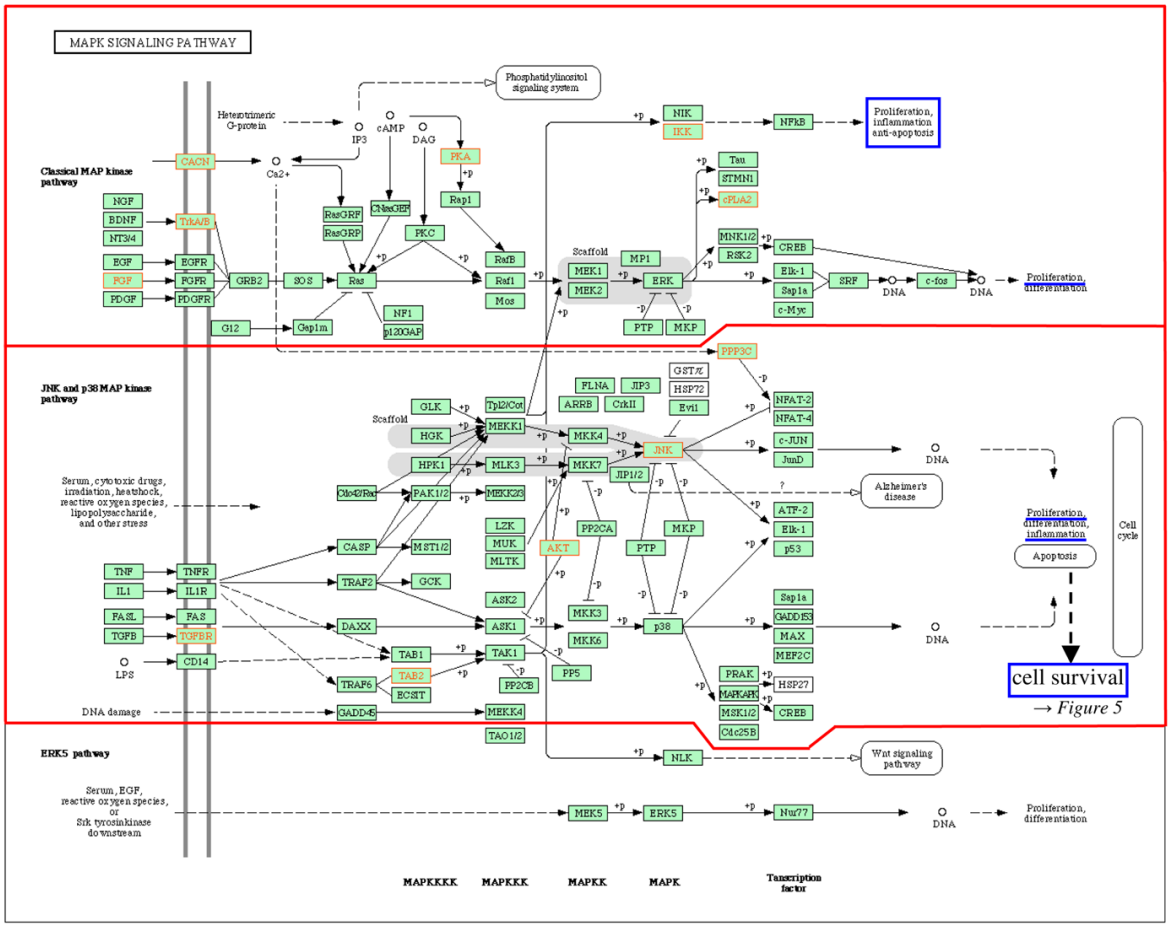
Discussion

The present microarray-based and real-time RT-PCR-validated, genome-wide mRNA expression analysis in RA, OA, and NC SM by KEGG mapping shows that gene-specific, significant, intra-group/inter-individual variances in gene expression profiles occur in RA. These variances affect a variety of genes involved in numerous pathways/complexes potentially relevant for RA pathogenesis. Since significant variance-fold values are observed for many genes with comparable mean expression levels among different patient/donor groups (data not shown), the manifestation of gene expression variances does not necessarily depend on the respective mean mRNA expression level.

To our knowledge, gene expression variances in RA samples have been reported only for distinct subgroup-specific differences in gene expression profiles of RA patients [12]. Consequently, the present data demonstrate for the first time broad intra-group/inter-individual gene expression variances in RA SM samples, previously observed in other severe diseases such as trisomy 21, malignant glioma, and inflammatory allergy [17-19]. It has been hypothesized that expression variances of regulatory key genes contribute to the individual phenotype of the given disease [17], whether independent of or depending on the expression level.

3.1 R. Huber, C. Hummert, U. Gausmann, D. Pohlers, D. Koczan, R. Guthke, R. W. Kinne. *Arthritis Research & Therapy*, 10:R98, August 2008.

Figure 4



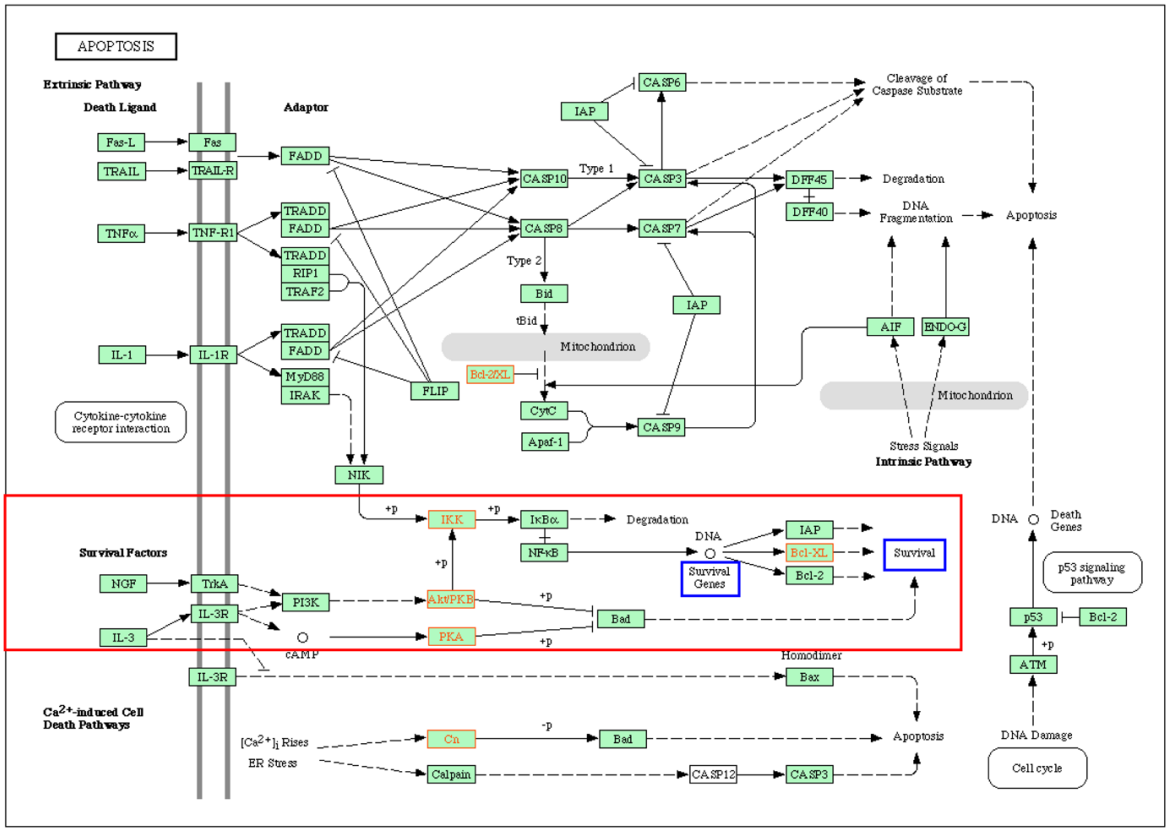
Inter-individual mRNA expression variances in the mitogen-activated protein kinase (MAPK) signaling pathway in rheumatoid arthritis (RA) compared with normal control (NC). The graph shows genes affected by significant intra-group inter-individual mRNA expression variances in RA compared with NC ($P \leq 0.05$; Bonferroni/Holm corrected Brown-Forsythe version of the Levene test; labeled in red) in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) MAPK signaling pathway. Among the three MAPK family sub-pathways, the classical and the c-jun kinase (JNK)/p38 MAPK sub-pathways were significantly affected by gene expression variances ($P \leq 0.15$, χ^2 test; indicated in red). MAPK-regulated cellular processes with potential influence on or relevance for RA pathogenesis (for example, proliferation, inflammation, and anti-apoptosis) are labeled in blue.

Currently, the causes for gene expression variances among RA patients are unknown. Possible external reasons may include the higher average age of the individuals in the RA group as well as medication influencing immunological processes and the expression of immunologically relevant genes (for example, methotrexate, prednisolone, sulfasalazine, and/or nonsteroidal anti-inflammatory drugs [35,36]) or differences in nutrition, with general effects on individual gene expression [37]. The inflammatory status of the respective joint at the time of surgical intervention may also substantially influence gene expression in the RA SM [38]. However, an analysis of the differential gene expression shows that the present RA group is generally characterized by an expression profile highly compatible with previous gene expression studies [39], including the overexpression of several transcription factors (for example, *FOS*, *FOSB*, *JUN*, and *STAT1* [10-12]), cytokines/chemokines (for example, *IL2*, *IL4*, *CCL23*, and *CCL25* [40]), signal transduction molecules (for example, *MAPK9*, *MAP3K2*, *PTPN7*, and *AKT2* [41,42]), cell cycle regulators (for example, *CDC12*, *CCNB2*, and *CCNE2* [43]), and heat shock proteins (DNAJ molecules; [44]; data not shown), indicating that the present RA cohort is representative for RA patients in general.

Regarding internal molecular changes in the individuals, a participation of mutations or single nucleotide polymorphisms in different genes is plausible, either directly [45,46] or via mutated regulators (for example, transcription factors, mRNA

Available online http://arthritis-research.com/content/10/4/R98

Figure 5



Inter-individual mRNA expression variances in the complex of apoptosis in rheumatoid arthritis (RA) compared with normal control (NC). The graph shows genes affected by significant intra-group inter-individual mRNA expression variances in RA compared with NC ($P \leq 0.05$; Bonferroni/Holm corrected Brown-Forsythe version of the Levene test; labeled in red) in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) complex of apoptosis. Among the three apoptosis sub-complexes, the survival factor-dependent sub-complex was significantly affected by gene expression variances ($P \leq 0.15$, χ^2 test; indicated in red). Cellular processes with potential influence on or relevance for RA pathogenesis (expression of survival genes and cell survival) are labeled in blue.

stability modifiers, and so on [47]). This also includes broader genomic rearrangements (for example, chromosomal translocations or polysomies [48,49]) as well as epigenomic modifications (for example, gene/promoter methylation [50]). In addition, the individual composition of cell types in the analyzed SM samples may influence the mRNA expression profile, depending on the inflammatory status and/or cell proliferation, potentially resulting in enhanced immigration/proliferation of T cells, B cells, or synovial fibroblasts [51].

In RA compared with NC, 10 KEGG pathways/complexes are specifically and significantly affected by gene expression variances. As expected, the importance of immunological processes for RA progression [8] is reflected in several pathways directly involved in such networks (Toll-like, T cell, and Fc ε receptor signaling [52-54]). In the SM, alterations in immunological pathways/complexes may contribute to the development of local (and systemic) inflammation, reflecting the highly inflamed status of the joint as one of the major characteristics of RA [2,55].

RA-specific gene expression variances also occur in cytokine–cytokine receptor interactions. Within this complex, a striking involvement of sub-pathways can be observed, with relevance for chemotaxis (CXC family chemokines [56]), angiogenesis, proliferation, and cell survival (TGF-β family [57,58]) as well as inflammation, joint destruction, and fibrosis (TNF family [59,60] and IL2RG shared pathway [9,61]; Figure 2). Sub-pathways influencing tissue protection (interferon family [62]) or anti-inflammation and anti-angiogenesis (IL13RA1 [interleukin-13 receptor alpha-1] shared pathway [63]) are scarcely affected. Therefore, a specific influence of gene expression variances on cytokine-mediated aspects of the RA can be assumed [64].

Table 2

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in rheumatoid arthritis (RA) compared with normal control (that is, higher variances in RA)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa04060	Cytokine–cytokine receptor interaction[a]	14 (8)	4.56	0.12	*CXCL13, IFNA8, FNAR2, IL2RG, IL4, IL8, IL13, CXCL10, IL21R, TNFRSF17, TGFBR2, CD27, TNFRSF25, ACVR1B*
2	hsa04010	MAPK signaling pathway[a]	13 (8)	3.32	0.22	*CHP, AKT2, MAP3K7IP2, PLA2G2D, IKBKB, NTRK2, PRKACA, MAPK8, PRKX, TGFBR2, CACNB1, FGF18, ACVR1B*
2a	hsa04010	MAPK signaling pathway[a] (classical + JNK/p38 MAPK sub-pathway)	13 (7)	4.39	0.13	*CHP, AKT2, MAP3K7IP2, PLA2G2D, IKBKB, NTRK2, PRKACA, MAPK8, PRKX, TGFBR2, CACNB1, FGF18, ACVR1B*
3	hsa05212	Pancreatic cancer[a]	9 (2)	20.29	<0.01	*E2F3, AKT2, IKBKB, SMAD2, MAPK8, BCL2L1, STAT1, TGFBR2, ACVR1B*
4	hsa04620	Toll-like receptor signaling pathway[a]	9 (3)	14.01	<0.01	*AKT2, MAP3K7IP2, IFNA8, IFNAR2, IKBKB, IL8, CXCL10, MAPK8, STAT1*
5	hsa04660	T-cell receptor signaling pathway[a]	7 (3)	5.98	0.05	*CHP, AKT2, IKBKB, IL4, RHOA, PDK1, PLCG1*
6	hsa04664	Fc epsilon receptor I signaling pathway[a]	7 (2)	9.53	0.01	*AKT2, PLA2G2D, IL4, IL13, PDK1, PLCG1, MAPK8*
7	hsa04520	Adherens junction[a]	6 (2)	5.56	0.07	*CSNK2A1, RHOA, SMAD2, TGFBR2, ACVR1B, CDH1*
8	hsa05220	Chronic myeloid leukemia[a]	6 (2)	5.73	0.06	*E2F3, IKBKB, BCL2L1, TGFBR2, ACVR1B, AKT2*
9	hsa04350	TGF-β signaling pathway[a]	5 (3)	1.86	0.38	*RHOA, SMAD2, TGFBR2, ACVR1B, ZFYVE9*
9a	hsa04350	TGF-β signaling pathway[a] (classical TGF-β sub-pathway)	5 (2)	6.7	0.05	*RHOA, SMAD2, TGFBR2, ACVR1B, ZFYVE9*
10	hsa04210	Apoptosis[a]	5 (3)	2.25	0.34	*AKT2, IKBKB, PRKACA, BCL2L1, CHP*
10a	hsa04210	Apoptosis[a] (anti-apoptotic sub-complex)	5 (1)	6.7	0.03	*AKT2, IKBKB, PRKACA, BCL2L1, CHP*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; JNK, c-jun kinase; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; MAPK, mitogen-activated protein kinase; TGF-β, transforming growth factor-beta.

Although the following pathways/complexes are not significantly affected by gene expression variances in total, embedded sub-pathways include the majority of affected genes, thus reaching statistical significance. In the TGF-β pathway, only members of the classical TGF-β sub-pathway are significantly affected, thus potentially influencing angiogenesis [58], cell survival [65], and cell proliferation [66] amongst others (Figure 3). Indeed, this (sub-) pathway appears to occupy a central position for the RA pathogenesis, due to the integration of various RA-relevant cellular functions. This is further underlined by its prominent role within the framework of cytokine–cytokine receptor interactions (Figure 2) and its influence on pro-inflammatory/pro-destructive features, either independent of or via MAPK (Figures 3 and 4). Within the MAPK signaling pathway, the 'classical' and the JNK/p38 MAPK sub-pathways – regulating proliferation, anti-apoptosis, and inflammation – are significantly affected by gene expression variances (Figure 4). This may be an indication of a participation of variable gene expression in inflammatory processes via MAPK variants (especially via *JNK/MAPK8* [67]) and proliferation of activated cells (for example, synovial fibroblasts and T cells) in RA [68,69] and MAPK-mediated anti-apoptosis (Figure 4).

Regarding apoptosis, genes particularly involved in the regulation of cell survival and anti-apoptosis are significantly affected by expression variances (Figure 5) [70]. Interestingly, the respective genes in this particular pathway also show increased expression levels in RA SM (data not shown). Pro-apoptotic genes are not affected in this pathway, corresponding to the absence of gene expression variances within the complex of p53-induced apoptosis (data not shown).

Depending on the individual gene expression level in each patient, gene expression variances in regulatory pathways may lead to enhanced inflammation [53,54], angiogenesis [71,72], enhanced collagen synthesis and secretion [9], and/or a reduced rate of apoptosis [73], thus potentially contributing to

Available online http://arthritis-research.com/content/10/4/R98

Table 3

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in normal control (NC) compared with rheumatoid arthritis (that is, higher variances in NC)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa03010	Ribosome[a]	8 (3)	27.62	<0.01	*RPL7, RPL9, RPL21, RPL27, RPL30, RPS6, RPS10, RPS12*
2	hsa04110	Cell cycle[a]	7 (4)	13.11	<0.01	*CDKN1A, E2F1, GADD45B, ATM, SKP1A, CCNA2, CDC2*
3	hsa04310	Wnt signaling pathway[a]	7 (5)	7.8	0.01	*CACYBP, PPP2R1B, PRKACB, PSEN1, SKP1A, TBL1XR1, FZD1*
4	hsa04640	Hematopoietic cell lineage[a]	4 (3)	4.15	0.15	*CSF1, EPOR, FLT3LG, ITGA4*
5	hsa05010	Alzheimer disease[a]	3 (1)	13.3	<0.01	*GAPDH, LRP1, PSEN1*
6	hsa01510	Neurodegenerative disorders[a]	3 (1)	8.18	0.01	*GAPDH, NR4A2, PSEN1*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; Wnt, wingless-type MMTV integration site family.

hyperplasia of the SM [74], collagen-dependent fibrosis of the joints [64], and a prolonged life span of activated synovial cells in RA [73,75].

Since RA and OA samples share many aspects of their respective mRNA expression profiles [76,77], genes in a number of pathways show comparable variance-fold values in both RA and OA (for example, apoptosis; Tables 2 and 4), thus reflecting basic similarities of joint diseases. However, RA and OA SM samples can be clearly differentiated regarding gene expression variances in other pathways/complexes. In OA, the pathways/complexes affected by higher expression variances than in NC indicate an OA-specific desynchronization of metabolic processes (Table 7). In contrast, RA-specific pathways/complexes are involved in the regulation of VEGF-mediated angiogenesis [74,75] and vascular permeability [78], as well as B cell-dependent auto-immunity and inflammation [79]. The latter represents the elevated activity status of B cells (including cytokine production and T-cell activation) and – in connection with the affection of the anti-apoptotic subpathway – the enhanced survival of self-reactive B cells [5,6,80]. This may result in a pronounced role of B cells for disease development in RA compared with OA, which is also reflected in the increasing impact of B cell-directed treatment in RA [81].

Table 4

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in osteoarthritis (OA) compared with normal control (that is, higher variances in OA)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa04310	Wnt signaling pathway	7 (4)	3.44	0.21	*CSNK2A1, SMAD2, PPP3CB, PRKACA, TBL1X, BTRC, RBX1*
1a	hsa04310	Wnt signaling pathway (canonical subpathway)	6 (3)	4.56	0.12	*CSNK2A1, BTRC, SMAD2, PRKACA, TBL1X, RBX1*
2	hsa04210	Apoptosis[a]	6 (2)	8.13	0.01	*AKT2, IKBKB, PP3CB, PRKACA, RKAR2A, BCL2L*
3	hsa03010	Ribosome[a]	5 (2)	3.99	0.16	*RPL18, RPL35A, RPL38, RPS10, RPL14*
3a	hsa03010	Ribosome[a] (large subunit)	4 (1)	6.49	0.04	*RPL18, RPL35A, RPL38, RPL14*
4	hsa04520	Adherens junction[a]	5 (2)	5.57	0.07	*CSNK2A1, SMAD2, ACP1, TGFBR2, YES1*
5	hsa05212	Pancreatic cancer[a]	5 (1)	6.22	0.04	*AKT2, IKBKB, SMAD2, BCL2L1, TGFBR2*
6	hsa04120	Ubiquitin-mediated proteolysis[a]	4 (2)	8.12	0.01	*ANAPC5, UBE2D2, BTRC, RBX1*
7	hsa05050	Dentatorubropallidoluysian atrophy[a]	3 (1)	19.79	<0.01	*ATN1, RERE, MAGI1*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; Wnt, wingless-type MMTV integration site family.

Table 5

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in normal control (NC) compared with osteoarthritis (that is, higher variances in NC)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa04310	Wnt signaling pathway	8 (3)	6.55	0.04	*CSNK1A1, DKK2, JUN, MYC, PPP2R1B, PRKACB, WNT5B, FZD1*
2	hsa05120	Epithelial cell signaling in *Helicobacter pylori* infection[a]	5 (2)	7.97	0.01	*JUN, NFKBIA, ATP6V1C1, ADAM17, ATP6V0D1*
3	hsa05211	Renal cell carcinoma[a]	5 (2)	7.75	0.01	*AKT2, HGF, JUN, TCEB1, VEGFA*
4	hsa04620	Toll-like receptor signaling pathway[a]	5 (2)	4.43	0.12	*AKT2, JUN, NFKBIA, TLR7, STAT1*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; Wnt, wingless-type MMTV integration site family.

In summary, these pathways indicate limited but distinct molecular/cellular differences between RA and OA and demonstrate a major contribution of inflammation and angiogenesis in RA. It is reasonable to assume that the RA pathogenesis is influenced by broad alterations of gene expression in general. For years, only differential gene expression analyses have been performed, resulting in the identification of some key genes but leading to the disregard of several genes with a more limited influence on RA, whose collective influence may still be as large as that of the already-known key players. Therefore, besides ubiquitous elevated expression levels of exceptional pro-inflammatory/pro-destructive key regulators/mediators like TNF-α, IL-1β [82], or MMP-1 [83], elevated or reduced expression levels of many different genes in various pathways/complexes may also influence RA development and progression. In this process, the affection of pathologically relevant pathways with differentially expressed genes may be more important than the character of the respective genes, resulting in different gene expression profiles among individual RA patients as reflected in the gene expression variances of the present study. As a consequence, synchronized or desynchronized gene expression in RA potentially shifts cellular activity from the normal to an activated status.

Regarding diagnosis and therapy of RA, the present results indicate that a more individualized approach for different patients may represent the future of RA treatment. Thus, the determination of individual gene expression patterns may facilitate the selection of the best medication or, more ambitiously, may allow directed modulation of (individually) selected pathways/complexes instead of broad suppression of inflammation by anti-inflammatory/anti-rheumatic drugs [84]. In addition, the present study helped to identify the TGF-β pathway as an accessory key player in RA, due to its central position within the regulatory networks. This suggestion is strongly supported by an emerging number of publications reporting a decisive impact of TGF-β on RA development/progression [57,58,85,86]. The affected pathways (and the respective genes) reported here may provide the basis for further analyses of the RA pathogenesis and the differences between RA and OA on a cellular and molecular level.

Conclusion

In RA, a number of disease-relevant or even disease-specific KEGG pathways/complexes (for example, TGF-β signaling and anti-apoptosis) are characterized by broad intra-group inter-individual expression variances. This indicates that RA pathogenesis in different individuals may depend to a lesser extent on common alterations of the expression of specific key genes, and rather on individual-specific alterations of different genes resulting in common disturbances of key pathways. Numerous affected pathways, including TGF-β signaling in a central position, are involved in inflammation, angiogenesis, proliferation, and cell survival, thus potentially influencing char-

Table 6

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in rheumatoid arthritis (RA) compared with osteoarthritis (that is, higher variances in RA)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa04916	Melanogenesis[a]	6 (3)	6.53	0.03	*ADCY2, LEF1, PRKCB1, PRKX, TCF7, WNT8B*
2	hsa04662	B-cell receptor signaling pathway[a]	5 (2)	9.72	0.01	*MALT1, PIK3CD, PLCG2, PRKCB1, CD72*
3	hsa04370	VEGF signaling pathway[a]	4 (2)	4.09	0.15	*PLA2G2D, PIK3CD, PLCG2, PRKCB1*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; VEGF, vascular endothelial growth factor.

Available online http://arthritis-research.com/content/10/4/R98

Table 7

KEGG pathways/complexes significantly affected by intra-group inter-individual gene expression variance in osteoarthritis (OA) compared with rheumatoid arthritis (that is, higher variances in OA)

	KEGG identification number	Pathway/complex	B (E)	χ^2	*P* value	Affected genes
1	hsa00190	Oxidative phosphorylation[a]	10 (1)	75.6	<0.01	*COX5B, NDUFA6, NDUFA8, NDUFB2, NDUFB4, SDHC, NDUFB6, aNDUFC1, NDUFA13, ATP5G3*
2	hsa04010	MAPK signaling pathway[a]	5 (2)	3.8	0.17	*DUSP5, RASGRP3, FAS, MAPK11, TAOK1*
2a	hsa04010	MAPK signaling pathway[a] (JNK/p38 MAPK sub-pathway)	4 (1)	6.54	0.03	*DUSP5, FAS, MAPK11, TAOK1*
3	hsa00790	Folate biosynthesis[a]	3 (0)	22.03	<0.01	*ASCC3, SETX, SMARCA5*
4	hsa00500	Starch and sucrose metabolism[a]	3 (1)	7.86	0.01	*ASCC3, SETX, SMARCA5*

[a]Specifically affected in rheumatoid arthritis. B, absolute frequency; E, expected frequency; JNK, c-jun kinase; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; MAPK, mitogen-activated protein kinase.

acteristic features of RA pathology.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RH performed the KEGG analyses, contributed to the real-time RT-PCR analyses, and participated in the writing of the manuscript. CH analyzed the microarray data, performed the bioinformatic analyses, and participated in the writing of the manuscript. RH and CH contributed equally to this work. UG participated in the data analyses. DP performed the real-time RT-PCR analyses. DK performed the Affymetrix microarray experiments. RG participated in the design and coordination of the study, including supervision of the bioinformatic analyses. RWK contributed to the design and coordination of the study and participated in the writing of the manuscript. All authors read and approved the final version of the manuscript.

Additional files

The following Additional files are available online:

Additional file 1
'Supplementary Table 1A: Genes affected by intra-group, inter-individual mRNA expression variances (RA compared to NC)', 'Supplementary Table 1B: Genes affected by intra-group, inter-individual mRNA expression variances (OA compared to NC)', 'Supplementary Table 1C: Genes affected by intra-group, inter-individual mRNA expression variances (RA compared to OA)'. For KEGG analyses, relevant genes were selected according to (i) a significance level of $p \leq 0.05$ (Bonferroni/Holm corrected Brown-Forsythe version of the Levene test) for variance-fold values and (ii) a cutoff value for absolute variance-fold levels of > 2.5 for higher variances in RA, OA, and NC, respectively. (A) 568 genes were selected for the comparison between RA and NC (307 with higher variances in RA, 261 with higher variances in NC), (B) 542 genes were used for the comparison OA versus NC (314 with higher variances in OA, 228 with higher variances in NC), and (C) 333 genes were selected for the comparison between RA and OA (186 with higher variances in RA, 147 with higher variances in OA). All genes are sorted according to absolute variance-fold values.
See http://www.biomedcentral.com/content/supplementary/ar2485-S1.doc

Acknowledgements
We thank Ernesta Palombo-Kinne for critical reading of the manuscript and Bärbel Ukena, Ulrike Körner, and Ildiko Toth for excellent technical assistance. We are grateful to Andreas Roth, Rando Winter, Renée Fuhrmann, and Rudolf-Albrecht Venbrocks (Department of Orthoped-

3.1 R. Huber, C. Hummert, U. Gausmann, D. Pohlers, D. Koczan, R. Guthke, R. W. Kinne. *Arthritis Research & Therapy*, 10:R98, August 2008.

ics, University Hospital Jena, Eisenberg, Germany) as well as Wolfgang Lungershausen (Department of Traumatology, University Hospital Jena, Jena, Germany) for providing patient/donor material. The study was supported by the German Federal Ministry of Education and Research (BMBF) (grant FKZ 010405 to RWK), the Interdisciplinary Center for Clinical Research (IZKF) Jena (grant FKZ 0313652A to RG), and the Jena Centre for Bioinformatics (grant FKZ 0313652B and grant 01GS0413, NGFN-2 to RWK). RH was supported by a grant from the German National Academic Foundation.

References

1. Grassi W, De Angelis R, Lamanna G, Cervini C: **The clinical features of rheumatoid arthritis.** *Eur J Radiol* 1998, **27(Suppl 1):**S18-S24.
2. Kinne RW, Palombo-Kinne E, Emmrich F: **Activation of synovial fibroblasts in rheumatoid arthritis.** *Ann Rheum Dis* 1995, **54:**501-504.
3. Abeles AM, Pillinger MH: **The role of the synovial fibroblast in rheumatoid arthritis: cartilage destruction and the regulation of matrix metalloproteinases.** *Bull NYU Hosp Jt Dis* 2006, **64:**20-24.
4. Karouzakis E, Neidhart M, Gay RE, Gay S: **Molecular and cellular basis of rheumatoid joint destruction.** *Immunol Lett* 2006, **106:**8-13.
5. Weyand CM, Seyler TM, Goronzy JJ: **B cells in rheumatoid synovitis.** *Arthritis Res Ther* 2005, **7(Suppl 3):**S9-12.
6. Keystone E: **B cell targeted therapies.** *Arthritis Res Ther* 2005, **7(Suppl 3):**S13-S18.
7. Firestein GS: **Evolving concepts of rheumatoid arthritis.** *Nature* 2003, **423:**356-361.
8. Firestein GS: **Immunologic mechanisms in the pathogenesis of rheumatoid arthritis.** *J Clin Rheumatol* 2005, **11:**S39-S44.
9. Postlethwaite AE, Holness MA, Katai H, Raghow R: **Human fibroblasts synthesize elevated levels of extracellular matrix proteins in response to interleukin 4.** *J Clin Invest* 1992, **90:**1479-1485.
10. Firestein GS, Manning AM: **Signal transduction and transcription factors in rheumatic disease.** *Arthritis Rheum* 1999, **42:**609-621.
11. Han Z, Boyle DL, Manning AM, Firestein GS: **AP-1 and NF-kappaB regulation in rheumatoid arthritis and murine collagen-induced arthritis.** *Autoimmunity* 1998, **28:**197-208.
12. Pouw Kraan TC van der, van Gaalen FA, Kasperkovitz PV, Verbeet NL, Smeets TJ, Kraan MC, Fero M, Tak PP, Huizinga TW, Pieterman E, Breedveld FC, Alizadeh AA, Verweij CL: **Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues.** *Arthritis Rheum* 2003, **48:**2132-2145.
13. Beer S, Oleszewski M, Gutwein P, Geiger C, Altevogt P: **Metalloproteinase-mediated release of the ectodomain of L1 adhesion molecule.** *J Cell Sci* 1999, **112(Pt 16):**2667-2675.
14. McCachren SS: **Expression of metalloproteinases and metalloproteinase inhibitor in human arthritic synovium.** *Arthritis Rheum* 1991, **34:**1085-1093.
15. Wu JJ, Lark MW, Chun LE, Eyre DR: **Sites of stromelysin cleavage in collagen types II, IX, X, and XI of cartilage.** *J Biol Chem* 1991, **266:**5625-5628.
16. Hardiman G: **Microarrays Technologies 2006: an overview.** *Pharmacogenomics* 2006, **7:**1153-1158.
17. Sultan M, Piccini I, Balzereit D, Herwig R, Saran NG, Lehrach H, Reeves RH, Yaspo ML: **Gene expression variation in Down's syndrome mice allows prioritization of candidate genes.** *Genome Biol* 2007, **8:**R91.
18. Shervington A, Patel R, Lu C, Cruickshanks N, Lea R, Roberts G, Dawson T, Shervington L: **Telomerase subunits expression variation between biopsy samples and cell lines derived from malignant glioma.** *Brain Res* 2007, **1134:**45-52.
19. MacGlashan DW Jr: **Relationship between spleen tyrosine kinase and phosphatidylinositol 5' phosphatase expression and secretion from human basophils in the general population.** *J Allergy Clin Immunol* 2007, **119:**626-633.
20. Chowers I, Liu D, Farkas RH, Gunatilaka TL, Hackam AS, Bernstein SL, Campochiaro PA, Parmigiani G, Zack DJ: **Gene expression variation in the adult human retina.** *Hum Mol Genet* 2003, **12:**2881-2893.
21. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM: **Gene-expression variation within and among human populations.** *Am J Hum Genet* 2007, **80:**502-509.
22. Levene H: **Robust tests for equality of variances.** In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* Edited by: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB. Palo Alto, CA: Stanford University Press; 1960:278-292.
23. Brown MB, Forsythe AB: **Robust tests for equality of variances.** *J Amer Statist Assoc* 1974, **69:**364-367.
24. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Statist* 1979, **6:**65-70.
25. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, Healey LA, Kaplan SR, Liang MH, Luthra HS, Medsger TA Jr, Mitchell DM, Neustadt DH, Pinals RS, Schaller JG, Sharp JT, Wilder RL, Hunder GG: **The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis.** *Arthritis Rheum* 1988, **31:**315-324.
26. Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, Christy W, Cooke TD, Greenwald R, Hochberg M, Howell D, Kaplan D, Koopman W, Longley S III, Mankin H, McShane DJ, Medsger T Jr, Meenan R, Mikkelsen W, Moskowitz R, Murphy W, Rothschild B, Segal M, Sokoloff L, Wolfe F: **Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association.** *Arthritis Rheum* 1986, **29:**1039-1049.
27. Huber R, Kunisch E, Gluck B, Egerer R, Sickinger S, Kinne RW: **Comparison of conventional and real-time RT-PCR for the quantitation of jun protooncogene mRNA and analysis of junB mRNA expression in synovial membranes and isolated synovial fibroblasts from rheumatoid arthritis patients.** *Z Rheumatol* 2003, **62:**378-389.
28. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
29. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo]
30. Fisher RA: **The correlation between relatives on the supposition of mendelian inheritance.** *Trans Roy Soc Edinb* 1918, **52:**399-433.
31. Abdi H: **Bonferroni and Sidak corrections for multiple comparisons.** In *Encyclopedia of Measurement and Statistics* Edited by: Salkind NJ. Thousand Oaks, CA: SAGE Publications; 2007.
32. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27:**29-34.
33. **HUGO Gene Nomenclature Committee** [http://www.genenames.org]
34. Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D: **Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies.** *BMC Genet* 2005, **6:**18.
35. Haupl T, Yahyawi M, Lubke C, Ringe J, Rohrlach T, Burmester GR, Sittinger M, Kaps C: **Gene expression profiling of rheumatoid arthritis synovial cells treated with antirheumatic drugs.** *J Biomol Screen* 2007, **12:**328-340.
36. Ospelt C, Gay S: **Antirheumatic drugs and gene signatures.** *Curr Opin Investig Drugs* 2007, **8:**385-389.
37. Paoloni-Giacobino A, Grimble R, Pichard C: **Genetics and nutrition.** *Clin Nutr* 2003, **22:**429-435.
38. Hahn G, Stuhlmuller B, Hain N, Kalden JR, Pfizenmaier K, Burmester GR: **Modulation of monocyte activation in patients with rheumatoid arthritis by leukapheresis therapy.** *J Clin Invest* 1993, **91:**862-870.
39. Batliwalla FM, Baechler EC, Xiao X, Li W, Balasubramanian S, Khalili H, Damle A, Ortmann WA, Perrone A, Kantor AB, Gulko PS, Kern M, Furie R, Behrens TW, Gregersen PK: **Peripheral blood gene expression profiling in rheumatoid arthritis.** *Genes Immun* 2005, **6:**388-397.
40. Sweeney SE, Firestein GS: **Rheumatoid arthritis: regulation of synovial inflammation.** *Int J Biochem Cell Biol* 2004, **36:**372-378.
41. Hammaker DR, Boyle DL, Chabaud-Riou M, Firestein GS: **Regulation of c-Jun N-terminal kinase by MEKK-2 and mitogen-acti-**

3.1 R. Huber, C. Hummert, U. Gausmann, D. Pohlers, D. Koczan, R. Guthke, R. W. Kinne. *Arthritis Research & Therapy*, 10:R98, August 2008.

vated protein kinase kinase kinases in rheumatoid arthritis. *J Immunol* 2004, **172:**1612-1618.
42. Liagre B, Vergne-Salle P, Leger DY, Beneytout JL: **Inhibition of human rheumatoid arthritis synovial cell survival by hecogenin and tigogenin is associated with increased apoptosis, p38 mitogen-activated protein kinase activity and upregulation of cyclooxygenase-2.** *Int J Mol Med* 2007, **20:**451-460.
43. Taranto E, Leech M: **Expression and function of cell cycle proteins in rheumatoid arthritis synovial tissue.** *Histol Histopathol* 2006, **21:**205-211.
44. Kurzik-Dumke U, Schick C, Rzepka R, Melchers I: **Overexpression of human homologs of the bacterial DnaJ chaperone in the synovial tissue of patients with rheumatoid arthritis.** *Arthritis Rheum* 1999, **42:**210-220.
45. Kim SY, Han SW, Kim GW, Lee JM, Kang YM: **TGF-beta1 polymorphism determines the progression of joint damage in rheumatoid arthritis.** *Scand J Rheumatol* 2004, **33:**389-394.
46. Han SW, Kim GW, Seo JS, Kim SJ, Sa KH, Park JY, Lee J, Kim SY, Goronzy JJ, Weyand CM, Kang YM: **VEGF gene polymorphisms and susceptibility to rheumatoid arthritis.** *Rheumatology (Oxford)* 2004, **43:**1173-1177.
47. Martinez A, Valdivia A, Pascual-Salcedo D, Balsa A, Fernandez-Gutierrez B, De la CE, Urcelay E: **Role of SLC22A4, SLC22A5, and RUNX1 genes in rheumatoid arthritis.** *J Rheumatol* 2006, **33:**842-846.
48. Kinne RW, Liehr T, Beensen V, Kunisch E, Zimmermann T, Holland H, Pfeiffer R, Stahl HD, Lungershausen W, Hein G, Roth A, Emmrich F, Claussen U, Froster UG: **Mosaic chromosomal aberrations in synovial fibroblasts of patients with rheumatoid arthritis, osteoarthritis, and other inflammatory joint diseases.** *Arthritis Res* 2001, **3:**319-330.
49. Kinne RW, Kunisch E, Beensen V, Zimmermann T, Emmrich F, Petrow P, Lungershausen W, Hein G, Braun RK, Foerster M, Kroegel C, Winter R, Liesaus E, Fuhrmann RA, Roth A, Claussen U, Liehr T: **Synovial fibroblasts and synovial macrophages from patients with rheumatoid arthritis and other inflammatory joint diseases show chromosomal aberrations.** *Genes Chromosomes Cancer* 2003, **38:**53-67.
50. Shin HJ, Park HY, Jeong SJ, Park HW, Kim YK, Cho SH, Kim YY, Cho ML, Kim HY, Min KU, Lee CW: **STAT4 expression in human T cells is regulated by DNA methylation but not by promoter polymorphism.** *J Immunol* 2005, **175:**7143-7150.
51. Hoffmann M, Pohlers D, Koczan D, Thiesen HJ, Wolfl S, Kinne RW: **Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions.** *BMC Bioinformatics* 2006, **7:**369.
52. Andreakos E, Sacre S, Foxwell BM, Feldmann M: **The toll-like receptor-nuclear factor kappaB pathway in rheumatoid arthritis.** *Front Biosci* 2005, **10:**2478-2488.
53. Zhang Z, Gorman C, Clark JM, Cope AP: **Rheumatoid arthritis: a disease of chronic, low-amplitude signals transduced through T cell antigen receptors?** *Wien Med Wochenschr* 2006, **156:**2-10.
54. Takai T: **Fc receptors and their role in immune regulation and autoimmunity.** *J Clin Immunol* 2005, **25:**1-18.
55. Huber LC, Distler O, Tarner I, Gay RE, Gay S, Pap T: **Synovial fibroblasts: key players in rheumatoid arthritis.** *Rheumatology (Oxford)* 2006, **45:**669-675.
56. Pierer M, Rethage J, Seibl R, Lauener R, Brentano F, Wagner U, Hantzschel H, Michel BA, Gay RE, Gay S, Kyburz D: **Chemokine secretion of rheumatoid arthritis synovial fibroblasts stimulated by Toll-like receptor 2 ligands.** *J Immunol* 2004, **172:**1256-1265.
57. Pohlers D, Beyer A, Koczan D, Wilhelm T, Thiesen HJ, Kinne RW: **Constitutive upregulation of the TGF-b pathway in rheumatoid arthritis synovial fibroblasts.** *Arthritis Res Ther* 2007, **9:**R59.
58. Maruotti N, Cantatore FP, Crivellato E, Vacca A, Ribatti D: **Angiogenesis in rheumatoid arthritis.** *Histol Histopathol* 2006, **21:**557-566.
59. Kollias G, Douni E, Kassiotis G, Kontoyiannis D: **The function of tumour necrosis factor and receptors in models of multi-organ inflammation, rheumatoid arthritis, multiple sclerosis and inflammatory bowel disease.** *Ann Rheum Dis* 1999, **58(Suppl 1):**I32-I39.
60. Rannou F, Francois M, Corvol MT, Berenbaum F: **Cartilage breakdown in rheumatoid arthritis.** *Joint Bone Spine* 2006, **73:**29-36.
61. Young DA, Hegen M, Ma HL, Whitters MJ, Albert LM, Lowe L, Senices M, Wu PW, Sibley B, Leathurby Y, Brown TP, Nickerson-Nutter C, Keith JC Jr, Collins M: **Blockade of the interleukin-21/interleukin-21 receptor pathway ameliorates disease in animal models of rheumatoid arthritis.** *Arthritis Rheum* 2007, **56:**1152-1163.
62. Tak PP: **IFN-beta in rheumatoid arthritis.** *Front Biosci* 2004, **9:**3242-3247.
63. Haas CS, Amin MA, Ruth JH, Allen BL, Ahmed S, Pakozdi A, Woods JM, Shahrara S, Koch AE: ***In vivo* inhibition of angiogenesis by interleukin-13 gene therapy in a rat model of rheumatoid arthritis.** *Arthritis Rheum* 2007, **56:**2535-2548.
64. Szekanecz Z, Koch AE: **Update on synovitis.** *Curr Rheumatol Rep* 2001, **3:**53-63.
65. Kawakami A, Urayama S, Yamasaki S, Hida A, Miyashita T, Kamachi M, Nakashima K, Tanaka F, Ida H, Kawabe Y, Aoyagi T, Furuichi I, Migita K, Origuchi T, Eguchi K: **Anti-apoptogenic function of TGFbeta1 for human synovial cells: TGFbeta1 protects cultured synovial cells from mitochondrial perturbation induced by several apoptogenic stimuli.** *Ann Rheum Dis* 2004, **63:**95-97.
66. Bira Y, Tani K, Nishioka Y, Miyata J, Sato K, Hayashi A, Nakaya Y, Sone S: **Transforming growth factor beta stimulates rheumatoid synovial fibroblasts via the type II receptor.** *Mod Rheumatol* 2005, **15:**108-113.
67. Han Z, Boyle DL, Aupperle KR, Bennett B, Manning AM, Firestein GS: **Jun N-terminal kinase in rheumatoid arthritis.** *J Pharmacol Exp Ther* 1999, **291:**124-130.
68. Ospelt C, Neidhart M, Gay RE, Gay S: **Synovial activation in rheumatoid arthritis.** *Front Biosci* 2004, **9:**2323-2334.
69. Forre O, Waalen K, Natvig JB, Kjeldsen-Kragh J: **Evidence for activation of rheumatoid synovial T lymphocytes – development of rheumatoid T cell clones.** *Scand J Rheumatol Suppl* 1988, **76:**153-160.
70. Vermeulen K, Berneman ZN, Van Bockstaele DR: **Cell cycle and apoptosis.** *Cell Prolif* 2003, **36:**165-175.
71. Hayes AJ: **Angioneogenesis in rheumatoid arthritis.** *Lancet* 1999, **354:**423-424.
72. Hirohata S, Sakakibara J: **Angioneogenesis as a possible elusive triggering factor in rheumatoid arthritis.** *Lancet* 1999, **353:**1331.
73. Gaur U, Aggarwal BB: **Regulation of proliferation, survival and apoptosis by members of the TNF superfamily.** *Biochem Pharmacol* 2003, **66:**1403-1408.
74. Malemud CJ: **Growth hormone, VEGF and FGF: involvement in rheumatoid arthritis.** *Clin Chim Acta* 2007, **375:**10-19.
75. Byrne AM, Bouchier-Hayes DJ, Harmey JH: **Angiogenic and cell survival functions of vascular endothelial growth factor (VEGF).** *J Cell Mol Med* 2005, **9:**777-794.
76. Firestein GS, Alvaro-Gracia JM, Maki R: **Quantitative analysis of cytokine gene expression in rheumatoid arthritis.** *J Immunol* 1990, **144:**3347-3353.
77. Nakamura Y, Nawata M, Wakitani S: **Expression profiles and functional analyses of Wnt-related genes in human joint disorders.** *Am J Pathol* 2005, **167:**97-105.
78. Middleton J, Americh L, Gayon R, Julien D, Aguilar L, Amalric F, Girard JP: **Endothelial cell phenotypes in the rheumatoid synovium: activated, angiogenic, apoptotic and leaky.** *Arthritis Res Ther* 2004, **6:**60-72.
79. Amu S, Stromberg K, Bokarewa M, Tarkowski A, Brisslert M: **CD25-expressing B-lymphocytes in rheumatic diseases.** *Scand J Immunol* 2007, **65:**182-191.
80. Szodoray P, Alex P, Frank MB, Turner M, Turner S, Knowlton N, Cadwell C, Dozmorov I, Tang Y, Wilson PC, Jonsson R, Centola M: **A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis.** *Rheumatology (Oxford)* 2006, **45:**1466-1476.
81. Anolik JH, Ravikumar R, Barnard J, Owen T, Almudevar A, Milner EC, Miller CH, Dutcher PO, Hadley JA, Sanz I: **Cutting edge: anti-tumor necrosis factor therapy in rheumatoid arthritis inhibits memory B lymphocytes via effects on lymphoid germinal centers and follicular dendritic cell networks.** *J Immunol* 2008, **180:**688-692.
82. Dayer JM: **Interleukin 1 or tumor necrosis factor-alpha: which is the real target in rheumatoid arthritis?** *J Rheumatol Suppl* 2002, **65:**10-15.

3.1 R. Huber, C. Hummert, U. Gausmann, D. Pohlers, D. Koczan, R. Guthke, R. W. Kinne. *Arthritis Research & Therapy*, 10:R98, August 2008.

83. Pardo A, Selman M: **MMP-1: the elder of the family.** *Int J Biochem Cell Biol* 2005, **37:**283-288.
84. Smolen J, Aletaha D: **The burden of rheumatoid arthritis and access to treatment: a medical overview.** *Eur J Health Econ* 2008, **8(Suppl 2):**S39-S47.
85. Hammaker DR, Boyle DL, Inoue T, Firestein GS: **Regulation of the JNK pathway by TGF-beta activated kinase 1 in rheumatoid arthritis synoviocytes.** *Arthritis Res Ther* 2007, **9:**R57.
86. Szekanecz Z, Haines GK, Harlow LA, Shah MR, Fong TW, Fu R, Lin SJ, Rayan G, Koch AE: **Increased synovial expression of transforming growth factor (TGF)-beta receptor endoglin and TGF-beta 1 in rheumatoid arthritis: possible interactions in the pathogenesis of the disease.** *Clin Immunol Immunopathol* 1995, **76:**187-194.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **75,** 417–429 doi: 10.1111/tpj.12210

Quantification of growth–defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover

Lynn Ullmann-Zeunert[1,2,†], Mariana A. Stanton[1,†], Nathalie Wielsch[3], Stefan Bartram[4], Christian Hummert[5], Aleš Svatoš[3], Ian T. Baldwin[1,*] and Karin Groten[1,*]

[1]*Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Knoell-Str. 8, 07745 Jena, Germany,*
[2]*Qiagen, Qiagenstr. 1, 40724 Hilden, Germany,*
[3]*MS Group, Max Planck Institute for Chemical Ecology, Hans-Knoell-Str. 8, 07745 Jena, Germany,*
[4]*Department of Bioorganic Chemistry, Max Planck Institute for Chemical Ecology, Jena, Germany, and*
[5]*Systems Biology/Bioinformatics Research Group, Leibniz Institute for Natural Product Research and Infection Biology, Beutenbergstr. 11a, 07745 Jena, Germany*

Received 11 January 2013; revised 3 April 2013; accepted 11 April 2013; published online 17 May 2013.
*For correspondence (e-mails baldwin@ice.mpg.de; kgroten@ice.mpg.de).
†These authors contributed equally to this work.

SUMMARY

Induced defenses are thought to be economical: growth and fitness-limiting resources are only invested into defenses when needed. To date, this putative growth–defense trade-off has not been quantified in a common currency at the level of individual compounds. Here, a quantification method for ^{15}N–labeled proteins enabled a direct comparison of nitrogen (N) allocation to proteins, specifically, ribulose-1,5-bisposphate carboxylase/oxygenase (RuBisCO), as proxy for growth, with that to small N-containing defense metabolites (nicotine and phenolamides), as proxies for defense after herbivory. After repeated simulated herbivory, total N decreased in the shoots of wild-type (WT) *Nicotiana attenuata* plants, but not in two transgenic lines impaired in jasmonate defense signaling (ir*LOX3*) and phenolamide biosynthesis (ir*MYB8*). N was reallocated among different compounds within elicited rosette leaves: in the WT, a strong decrease in total soluble protein (TSP) and RuBisCO was accompanied by an increase in defense metabolites, ir*LOX3* showed a similar, albeit attenuated, pattern, whereas ir*MYB8* rosette leaves were the least responsive to elicitation, with overall higher levels of RuBisCO. Induced defenses were higher in the older compared with the younger rosette leaves, supporting the hypothesis that tissue developmental stage influences defense investments. We propose that MYB8, probably by regulating the production of phenolamides, indirectly mediates protein pool sizes after herbivory. Although the decrease in absolute N invested in TSP and RuBisCO elicited by simulated herbivory was much larger than the N-requirements of nicotine and phenolamide biosynthesis, ^{15}N flux studies revealed that N for phenolamide synthesis originates from recently assimilated N, rather than from RuBisCO turnover.

Keywords: *Nicotiana attenuata*, caffeoyl-putrescine, dicaffeoyl-spermidine, nicotine, ribulose-1,5-bisphosphate carboxylase/oxygenase, total soluble protein, R2R3-MYB transcription factor, *Manduca sexta*.

INTRODUCTION

Plants have evolved two general direct strategies against herbivory: constitutive and inducible defenses. The biosynthesis of these defenses requires fitness-limiting resources that could otherwise be invested into growth and reproduction. Hence, induced plant defenses are thought to be a cost-saving strategy compared with constitutive defenses, as they are only produced when needed, e.g. after herbivory (Karban and Baldwin, 1997), and this cost-saving model plays a central role in most theoretical treatments of induced defenses (for a review of plant defense hypotheses, see Stamp, 2003). Several studies have quantified the costs of induction by measuring photosynthesis rates, plant bio-

mass, size and/or yield associated with an increase in defense metabolites (Bazzaz *et al.*, 1987; Karban and Baldwin, 1997; Zangerl *et al.*, 2002). Although measurements of the impact of anti-herbivore defenses on plant yield are important for understanding their ultimate fitness costs, measurements of plant biomass do not discriminate among the relative investments into compounds that function in growth, storage and defense processes in the tissues analyzed (Chapin *et al.*, 1990). Therefore, the investment into growth is preferably estimated by measuring components of biomass that directly promote the acquisition of resources for growth, such as photosynthetic proteins (Chapin *et al.*, 1990). Additionally, the costs of defense should be measured in the currency of a fitness-limiting resource (Mole, 1994; Baldwin *et al.*, 1998). Nitrogen (N) is often such a fitness-limiting resource, determining the growth and reproduction of plants, and of the herbivores that eat them. N availability also influences N allocation to defense metabolites (Baldwin *et al.*, 1998; Lou and Baldwin, 2004; Simon *et al.*, 2010). Thus, it is an ideal currency to use for the study of growth–defense trade-offs in plant–herbivore interactions.

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is the most abundant foliar protein in plants, and is essential for the dark reaction of photosynthesis. RuBisCO constitutes 30–50% of the total soluble protein (TSP) in C3 plants (Ellis, 1979; Makino *et al.*, 1984; Imai *et al.*,2008), and may function as a potential N storage protein (Millard, 1988); consequently, it represents a major N sink in plants. Its large and small subunits (LSUs and SSUs, respectively) are synthesized from separate precursor pools that have different metabolic origins (Allen *et al.*, 2012). Although the concentration and activity of RuBisCO are not the only factors controlling growth (Stitt and Schulze, 1994), changes in RuBisCO expression influence growth and lead to complex changes in N metabolism (Stitt and Schulze, 1994; Stitt and Krapp, 1999; Matt *et al.*, 2002), making this enzyme a reasonable proxy for growth parameters.

Nicotiana attenuata is a wild tobacco native to the Great Basin Desert in south-western USA that synchronizes its germination from long-lived seed banks in response to exposure to cues from pyrolized vegetation (Preston and Baldwin, 1999). By timing its germination with the immediate post-fire environment, *N. attenuata* takes advantage of the abundant, yet ephemeral, pools of inorganic N in burned soil (Lynds and Baldwin, 1998), but is subject to high intraspecific competition for this fitness-limiting resource because of its mass-germination behavior. Furthermore, because it is a pioneer species, *N. attenuata* is attacked by a diverse herbivore community, including the specialist tobacco hornworm (*Manduca sexta*). Herbivore attack elicits the jasmonic acid (JA) signaling cascade (Kessler *et al.*, 2004), which activates JA-responsive transcription factors that lead to the biosynthesis of a plethora of induced small metabolites (Figure 1a; Woldemariam *et al.*, 2011), such as the N-intensive alkaloid nicotine, and a variety of phenolamides, which decrease herbivore performance (Baldwin, 1999; Steppuhn *et al.*, 2004; Kaur *et al.*, 2010; Onkokesung *et al.*, 2010). The biosynthesis of nicotine and phenolamides requires the same amino acid precursors (ornithine and arginine for putrescine and spermidine biosynthesis; Kaur *et al.*, 2010; Steppuhn *et al.*, 2004; Takano *et al.*, 2012), but nicotine is produced only in the roots (Hibi *et al.*, 1994), whereas phenolamides are synthesized in the attacked leaf (Kaur *et al.*, 2010).

Nicotine is present constitutively in undamaged *N. attenuata* tissues, and foliar concentrations increase substantially after herbivory (McCloud and Baldwin, 1997; Baldwin, 1999). The two major phenolamides found in *N. attenuata* are the N-acylated polyamines caffeoyl-putrescine (CP) and dicaffeoyl-spermidine (DCS), the biosynthesis of which is regulated by the transcription factor NaMYB8 (hereafter MYB8; Figure 1a). Both CP and DCS accumulate constitutively in reproductive tissues, and are strongly induced in leaves by simulated herbivory (Kaur *et al.*, 2010). Herbivory also causes large-scale changes in *N. attenuata*'s transcriptome and proteome, decreasing the levels of photosynthetic genes and proteins, including RuBisCO (Halitschke *et al.*, 2003; Voelckel and Baldwin, 2004b; Giri *et al.*, 2006). Because of the important constraints imposed by N availability upon both its growth and defense, as well as the wealth of understanding of its anti-herbivore defenses and the availability of isogenic transgenic lines impaired in individual classes of defenses, *N. attenuata* is an ideal model in which to study growth–defense trade-offs in a common N currency.

The induction of defense responses in wild tobacco can be simulated in a standardized and synchronized way by wounding leaves and applying the oral secretions (OS) of *M. sexta* larvae to the wounds (W + OS, Figure 1). The major elicitors in *M. sexta* OS are fatty-acid amino acid conjugates (FACs), which are recognized by the plant, triggering defense responses (Schittko *et al.*, 2001; Halitschke *et al.*, 2003; Giri *et al.*, 2006). The FAC composition of OS, and the resulting gene expression and metabolite induction in the plant, differ between specialist and generalist folivores (Voelckel and Baldwin, 2004a; Diezel *et al.*, 2009; Steinbrenner *et al.*, 2011).

Here, we quantified the N investments into different plant parts and among different N pools within a tissue to compare the investments into growth and defense in the same currency after repeated simulated herbivory by W + OS elicitation from a specialist herbivore. Repeated simulated herbivory, in contrast to single W + OS elicitation, more closely mimics natural herbivore feeding, which varies in duration and timing (Van Dam *et al.*, 2001; Skibbe *et al.*, 2008; Stork *et al.*,2009). A stable isotope labeling technique was used to track N flux among different pools

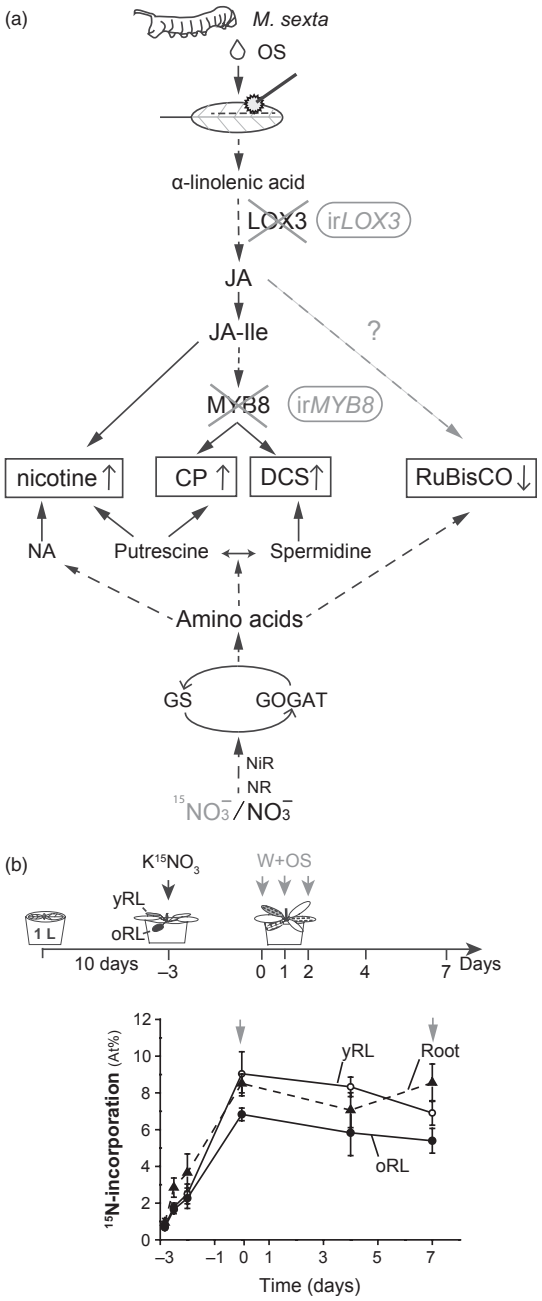


Figure 1. Overview of experimental strategy used to study growth–defense trade-offs in *Nicotiana attenuata* in a common nitrogen (N) currency.
(a) The biosynthesis of nicotine, caffeoyl-putrescine (CP) and dicaffeoylspermidine (DCS) is induced after simulated herbivory in the wild type (WT) by wounding (W) with a pattern wheel and by the application of oral secretions (OS) of *Manduca sexta*, but is impaired in the transgenic plants silenced in the expression of *lipoxygenase 3* (*LOX3*) or *MYB8* by RNAi with inverted-repeat (ir) constructs. The concentration of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) decreases in the WT after W + OS, but the effects of jasmonic acid (JA) on N investment into RuBisCO are unclear. Amino acids serve as precursors for putrescine, spermidine and nicotinic acid (NA), which provide N for the synthesis of these metabolites. Amino acids are derived from nitrate (NO_3^-) reduction, followed by assimilation catalyzed by glutamine synthetase (GS) and glutamate synthase (GOGAT), and are also used as precursors for RuBisCO synthesis. JA-Ile, JA-isoleucine; NiR, nitrite reductase; NR, nitrate reductase.
(b) Incorporation of ^{15}N into roots, and younger (yRL) and older rosette leaves (oRL) following pulse labeling with $K^{15}NO_3$ 27 days after germination was determined by isotope-ratio mass spectrometry (IRMS; $n = 5$). Grey arrows indicate the time points of elicitation in the experiments that followed. During this time frame ^{15}N-incorporation was stable. At%, atomic percentage.

of individual compounds in locally elicited and systemic leaves and seeds. We applied ^{15}N-labeled nitrate to the soil because nitrate is the most common form of N taken up by *N. attenuata* in nature, after the rapid biological nitrification of the ammonium generated by pyrolysis (Figure 1; Lynds and Baldwin, 1998).

The N flux into the three major N-intensive small metabolites of *N. attenuata* (nicotine, CP and DCS) was used as a proxy for defense investment that could be directly compared with the N investment into proteins, and in particular the abundant photosynthetic protein, RuBisCO, as a proxy for growth-related investment. These different molecule classes could not be measured in the past with comparable precision and accuracy because of a lack of suitable methods, especially for proteins. Here, we used a high-throughput LC-MS^E method for the absolute quantitation of proteins and the incorporation of ^{15}N into peptides (Ullmann-Zeunert *et al.*, 2012), which allows for the quantification of single large proteins with the same accuracy as for the small defense metabolites quantified by UPLC/UV/ToF-MS.

To further disentangle the effects of induced defenses on N allocation after herbivory, we compared two previously described transgenic lines, one deficient in JA signaling, ir*LOX*3 (Allmann *et al.*, 2010), and one deficient in the biosynthesis of phenolamides, ir*MYB8* (Kaur *et al.*, 2010), with wild-type (WT) plants (Figure 1a). This design allows for a direct comparison of N flux into specific classes of defense compounds with that into growth-related proteins measured in the same N currency, and an evaluation of the hypothesis that RuBisCO is used as an N-storage compound for defense responses.

RESULTS AND DISCUSSION

Anti-herbivore defense elicitation alters the nitrogen content of the shoot

Herbivory is known to change resource allocation within plants (Bazzaz *et al.*, 1987; Frost and Hunter, 2008; Gomez *et al.*, 2010). To estimate the impact of the biosynthesis of N-containing defense metabolites on N accumulation in *N. attenuata*, we compared the shoot N contents (% dry mass) of the two transgenic lines impaired in defense

responses with WT plants after repeated simulated herbivory with W + OS. The isotope ratio mass spectrometry (IRMS) measurements revealed that repeated elicitation reduced the N content of WT shoots (i.e. the total of N per dry mass of shoots; Welch's two-sample *t*-test, d.f. = 7.24, $P = 0.032$), but not of the transgenic lines (Figure 2), whereas the N pool sizes were slightly reduced after elicitation for all three genotypes (Figure S1b). The changes in the N pool sizes of elicited ir*LOX3* and ir*MYB8* plants were the result of a reduction in shoot dry mass (Figure S1), whereas the elicited WT showed both reduced shoot dry mass and reduced shoot N content, suggesting a possible N reallocation within the plant caused by the biosynthesis of N-containing defense metabolites.

Plants can allocate N to roots to protect this resource from folivores to reduce the nutritional value of the attacked tissues, which, together with increased defenses, can slow herbivore growth and increase their exposure to natural enemies (Trumble *et al.*, 1993). Previous studies with *Solanum lycopersicum* (tomato) demonstrated that N allocation in the form of amino acids from the shoot to the roots was rapidly induced by methyl-jasmonate (MeJA; Gomez *et al.*, 2010) and *M. sexta* feeding (Steinbrenner *et al.*, 2011; Gomez *et al.*, 2012). In *N. attenuata*, OS elicitation has been shown to cause a rapid allocation of carbon from the shoot to the roots, which can later be used for regrowth and reflowering (Schwachtje *et al.*, 2006). The reduced N concentration of WT shoots in our experiment suggests that this species can also allocate N from the shoot to the roots after herbivory. This inference is consistent with the observation that the N contents of WT roots increased after elicitation, as measured in a separate experiment, although the increase was not quite significant (Welch's two-sample *t*-test, d.f. = 4.71, $P = 0.054$; inset Figure 2). Alternatively, the increased N content of roots may have resulted from increased N assimilation, but previous ^{15}N labeling experiments in this species have found no evidence for changes in N assimilation rate after herbivory (Baldwin and Ohnmeiss, 1994; Lynds and Baldwin, 1998). Therefore, we conclude that the induced biosynthesis of N-containing metabolites after OS elicitation alters whole-plant N partitioning.

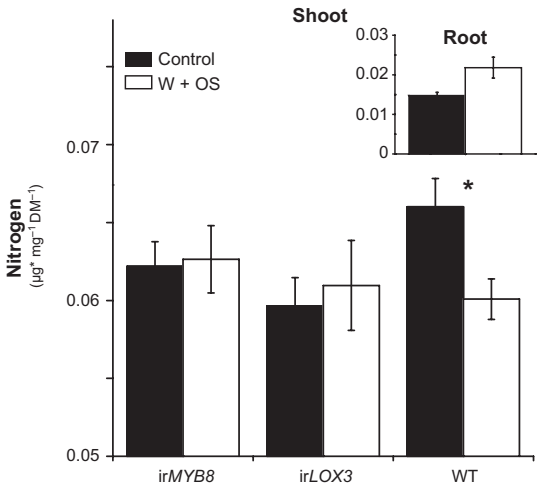


Figure 2. Total nitrogen (N) content in wild-type (WT) shoots decreases after simulated herbivory.
The N content of shoots of ir*LOX3*, ir*MYB8* and WT ($n = 5$) was determined by IRMS 4 days after the first W + OS elicitation. Unelicited plants were controls. Asterisks represent significant differences between treatments ($*P \leq 0.05$; $n = 5$). Inset: N content of WT roots ($n = 5$) was determined in a separate experiment at the same time point. DM, dry mass.

Changes in absolute pool sizes depend on developmental stage

To analyze the influence of anti-herbivore defense induction, especially phenolamide biosynthesis, on within-shoot N allocation, we determined the absolute N pools of different leaf types (hereafter, total N pools) and N allocation to seeds by IRMS. Expressing resource allocation as concentrations reveals proportional allocations within an organ; however, total pools allow for comparisons among organs, as they are a function of both organ size and concentration (Chapin *et al.*, 1990). We analyzed elicited older (oRL) and younger (yRL) rosette leaves to explore the influence of leaf development on N reallocation after elicitation, and the first (unelicited) stem leaf (S1) to examine systemic effects.

Overall, there was no clear effect of genotype or elicitation on the leaf total N pools. Total N pools varied among genotypes only in the S1 leaf (ANOVA, $F1_{,27} = 4.86$, $P = 0.036$), whereas OS elicitation only reduced the total N pool of ir*LOX3* (two-sample *t*-test, d.f. = 8, $P = 0.006$) and WT (Welch's two-sample *t*-test, d.f. = 8, $P = 0.021$) in the yRL (ANOVA, $F1_{,28} = 7.40$, $P = 0.011$). The N pool size in oRL was unaffected by genotype and elicitation (Figure 3). As N pool size correlates with biomass at the whole-plant scale (Baldwin and Hamilton, 2000), we evaluated whether the observed changes in total N pools of single leaves could be explained by changes in growth. Although the leaf size of yRL was reduced after elicitation (ANOVA, $F1_{,24} = 12.33$, $P = 0.002$; Figure S2a), it did not correlate with total N pools (ANCOVA, $P = 0.187$). Similarly, the change in total N pools of S1 leaves was not correlated with changes in leaf size (ANCOVA, $P = 0.406$).

It is possible that changes in the total N pool of a leaf reflect changes in a major pool within the leaf, such as proteins. Although TSP pool size dramatically decreased in the yRL after elicitation, it did not correlate with the total N pool size in this tissue (ANCOVA, $P = 0.122$; Figure 3). Thus, we conclude that although both pools are reduced by elicitation, the total N pool of the rosette leaves does not reflect the changes in TSP pool size or leaf size. This result

3.2 L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, K. Groten. *The Plant Journal*, 75(3):417-429, July 2013.

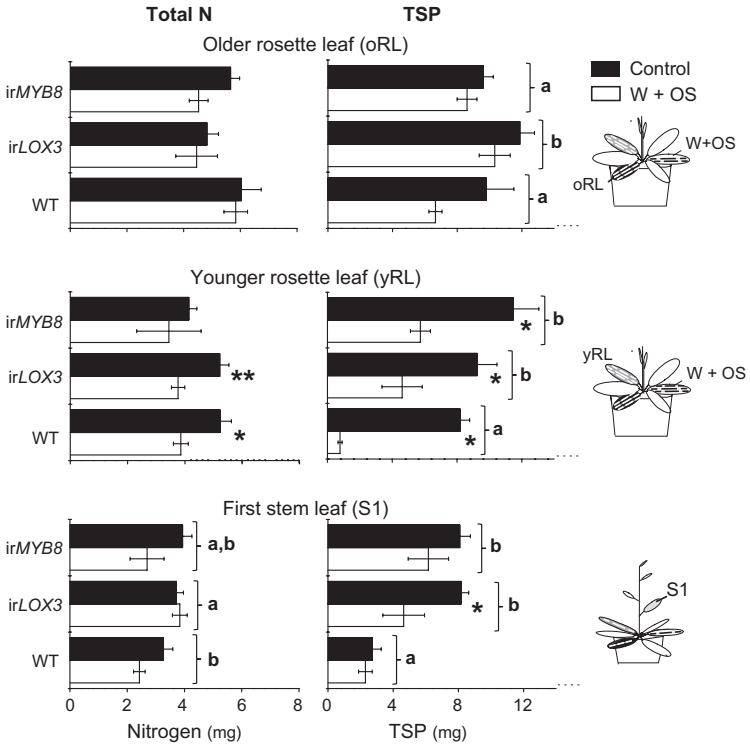


Figure 3. Silencing of *LOX3* and *MYB8* alters the distribution of nitrogen (N) between and within leaves. The N pools and total soluble protein (TSP) of leaves (oRL, older rosette leaf; yRL, younger rosette leaf; S1, first stem leaf) of ir*LOX3*, ir*MYB8* and WT, calculated based on leaf mass. The N content was determined by IRMS and the TSP was measured by the Bradford assay. Plants were elicited as described for Figure 2. yRL and oRL were harvested 4 days after the first W + OS elicitation, and when S1 leaves underwent the source–sink transition. Asterisks indicate differences among treatments ($*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$). Letters represent significant differences found using the minimum adequate model ($n = 5$). For abbreviations, see Figure 1.

is consistent with the hypothesis that total leaf N content and TSP (RuBisCO content) are controlled by different mechanisms, as has been shown for *Oryza sativa* (rice; Ishimaru *et al.*, 2001; Makino *et al.*, 2000).

The TSP pools differed between the lines in all three leaf types (ANOVA; oRL, $F_{1,27} = 8.70$, $P = 0.007$; yRL, $F_{1,27} = 12.95$, $P = 0.001$; S1, $F_{1,27} = 44.77$, $P = 3.5*10^{-07}$; Figure 3). The TSP pools of ir*LOX3* and ir*MYB8* in the yRL were reduced by about 50% after OS elicitation, whereas WT TSP pools were reduced by 91%. Both transgenic lines had constitutively larger TSP pools in the S1 leaf than the WT, and although S1 TSP decreased after elicitation in ir*LOX3* plants, TSP pools of both transgenic lines were still 2.5–3.0 times larger than those of WT after elicitation (Figure 3), suggesting that the biosynthesis of N-containing metabolites affects protein pool sizes, and that inducible defenses also have constitutive costs.

The recently developed method for the absolute quantification of single proteins allowed us to quantitatively compare the investment in defense metabolites with that in growth-related compounds, specifically, the photosynthetic protein RuBisCO, with similar accuracy (Ullmann-Zeunert *et al.*, 2012). Being the most abundant soluble protein in plants, the total level of RuBisCO (sum of LSU and SSU) reflected the TSP pattern in the different leaves, independent of genotypes (ANCOVA; oRL, $P < 0.0001$; yRL, $P < 0.0001$; S1, $P = 0.42$; Figures 3 and S3a). Overall, the data revealed a decrease in pool sizes of both RuBisCO subunits after OS elicitation (Figure S3a), which coincided with an increase in N-containing defense metabolites (Figure S3b), but the effects differed among lines, and for most traits measured ir*LOX3* showed an intermediate phenotype between WT and ir*MYB8*. The two transgenic lines with either reduced (ir*LOX3*) or undetectable levels of CP and DCS (ir*MYB8*; Figure S3b) showed a smaller decrease of RuBisCO LSU and SSU than the WT in the elicited yRL (47–59% in ir*MYB8*/ir*LOX3* compared with 92–95% in WT; Figure S3a). RuBisCO LSU and SSU levels were unaltered after elicitation in the systemic S1 leaf of ir*MYB8*, but strongly declined in WT and ir*LOX3*. The nicotine pool sizes showed similar induction patterns for all lines, except in the yRL, where the OS-elicited nicotine levels were higher in the WT than in the transgenic lines. These data suggest that the growth–defense trade-offs at the leaf scale are probably influenced by the capacity to biosynthesize and accumulate phenolamides, and that this also affects growth investments in the systemic S1 leaf.

As all transgenic lines used in this study accumulated similar levels of nicotine, it is unclear whether the biosynthesis of this alkaloid might affect N allocation to proteins

(Figure S3b). To answer this question rigorously, experiments with transgenic lines completely with no flux of N into nicotine biosynthesis are needed. In the nicotine-silenced transgenic lines we have produced in our laboratory by silencing putrescine N-methyl transferase, nicotine biosynthesis is silenced, but the elicited flux of N into other alkaloids (anatabine) is not (Steppuhn *et al.*, 2004).

Similarly, the induction of proteinase inhibitors could have additional influence on N allocation; however, preliminary experiments with virus-induced empty vector and *MYB8*-silenced plants showed a similar trypsin proteinase activity in both plants after elicitation (H. Kaur, personal communication), indicating that the synthesis of proteinase inhibitors does not seem to play a key role in the reallocation of N from primary to secondary metabolism.

A comparison of the two locally elicited leaves revealed differences in their defense and growth pool sizes: whereas the oRL accumulated the largest defense metabolite pools, with only slight reductions in TSP and both RuBisCO subunits after elicitation, the yRL had the strongest reductions in protein pools, with less pronounced increases in N-containing defense metabolite levels than the oRL (Figures 3 and S3). The optimal defense theory predicts that the allocation of defense metabolites is directly proportional to the fitness value of different plant parts (McKey, 1974, 1979; Rhoades, 1979), and many studies have demonstrated that younger leaves of *N. attenuata*, presumed to have a higher fitness value than older leaves, contain higher defense metabolite levels (Zavala *et al.*, 2004a; Kaur *et al.*, 2010; Onkokesung *et al.*, 2012). These results appear to contradict our findings, because the oRL contained higher metabolite levels than the yRL; however, the previous studies compared concentrations of metabolites in elicited rosette leaves at different stages of plant development, whereas here we analyzed metabolite pool sizes of two elicited rosette leaves, of different maturity, harvested simultaneously from the same plant. As the plants were just beginning stalk elongation at the time of OS elicitation, both oRL and yRL are likely to be important tissues for later plant growth and reproduction. Thus the larger defense metabolite pools of the mature oRL – which was a source leaf at the time of the first elicitation – may result from its larger nutrient pools, which are probably important for regrowth capacity. Meanwhile, the smaller pools of TSP and RuBisCO in elicited yRL – which was in the transition stage from sink to source during the first W + OS treatment – may reflect a lower N allocation to proteins in developing leaves, which could enhance their defense status by reducing the food quality for herbivores. This is in agreement with the model from Orians *et al.* (2011), assuming that the mature source leaf allocates resources not only to defense and growth, but also to storage, thus making it relatively more valuable for the whole plant, and therefore better protected. Regardless of their ultimate explanations, these data demonstrate that growth–defense trade-offs are dependent on leaf development.

Many previous studies have demonstrated that inducible defenses are costly, often leading to a decrease in reproductive performance (Heil and Baldwin, 2002): e.g. growth–defense trade-offs at the leaf scale affect the N allocation to capsules in *Nicotiana sylvestris* (Ohnmeiss and Baldwin, 2000). However, here, neither the time of flowering and seed ripening, nor the number of mature capsules, the mass of the first mature seed capsule, nor the total N content of the first seed capsule were significantly different from controls after repeated simulated herbivory (Figure S4). This lack of observed fitness effects could result from species-specific differences or differences in the experimental design. In our experiment, OS elicitation may have been too early to affect seed set (the first capsules were harvested on average 18 days after the last elicitation), or the W + OS treatment was too weak to elicit changes in allocation to seeds, compared with the relatively stronger MeJA elicitation used in other experiments (Voelckel *et al.*, 2001). In nature, wild tobacco faces strong intraspecific competition because of its mass-germination behavior, and strong alterations in N allocation to reproductive units in glasshouse-cultivated tobacco were only found when MeJA-elicited plants competed with control plants for the same limited resources (Van Dam and Baldwin, 2001). Thus, the costs and benefits of N allocation for a plant after herbivore attack may only become obvious if neighboring plants competing for the same limited resources are present. Additional experiments with plants grown in competition and exposed to simulated and natural herbivory are necessary to further explore the impact of growth–defense trade-offs within the leaf on plant fitness.

MYB8 indirectly affects nitrogen investment into proteins

The pool sizes of proteins and defense metabolites of the two transgenic lines suggest an influence of N-containing metabolite biosynthesis on the observed growth–defense trade-offs, but did not allow for a direct comparison of the levels of N demanded for metabolite biosynthesis, and the decreased N partitioned into TSP and RuBisCO after herbivory. By calculating the N investment into growth and defense per mg of fresh tissue mass after elicitation, we were able to further explore the role of phenolamide biosynthesis on N reallocation. We combined this approach with ^{15}N pulse labeling to follow the investment of a defined N pool into both plant functions.

For all lines, and in locally treated leaves, elicitation decreased the N investment into rest TSP and RuBisCO per mg fresh mass, compared with controls. Particularly in yRLs, the decrease in N investment into TSP (rest TSP and RuBisCO) was much more pronounced in the WT (89%)

compared with 28% in ir*MYB8* and 47% in ir*LOX3* (Figure 4a). Ir*LOX3* plants, for all parameters measured here, showed similar but less pronounced N-allocation patterns after elicitation as WT. These patterns are consistent with the correlation analysis of all measured N pools (Figure 4a, heat maps). Correlating all genotype/treatment groups with each other revealed that OS-elicited WT plants did not correlate with the other genotype/treatment groups in all three leaf types. Only OS-elicited ir*LOX3* oRL and S1 leaves showed a weak correlation with WT-OS. In contrast, ir*MYB8*-OS did not correlate with any other genotype by treatment group.

Interestingly, the observed N-investment pattern is congruent with previous results on the patterns of *MYB8* transcript accumulation in *N. attenuata* as*LOX3* plants (which are comparable with ir*LOX3*; Allmann *et al.*, 2010; Halitschke *et al.*, 2004). After elicitation, as*LOX3* leaves have four times lower *MYB8* transcript levels, whereas ir*MYB8* have 10 times lower levels than WT leaves (Kaur *et al.*, 2010; Onkokesung *et al.*, 2012). Furthermore, MYB8

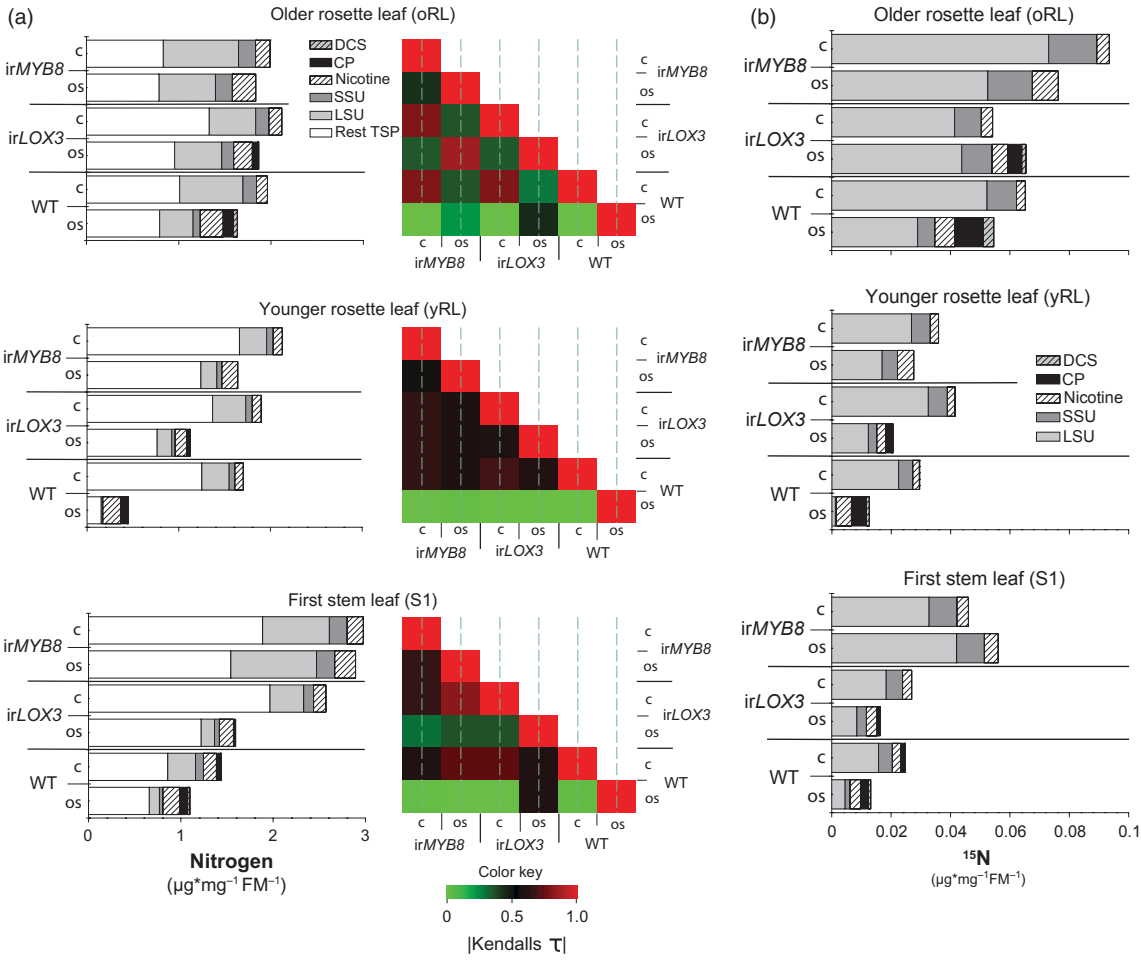


Figure 4. The increased nitrogen (N) investment in nicotine, caffeoyl-putrescine (CP) and dicaffeoyl-spermidine (DCS) is accompanied by a decreased N investment in protein.

(a) Investment of N in residual total soluble protein (TSP) [TSP – (SSU + LSU)], RuBisCO large (LSU) and small (SSU) subunits, nicotine, CP and DCS in older rosette leaves (oRL), younger rosette leaves (yRL) and first stem leaves (S1) was calculated by multiplying the proportion of N in each compound with the concentration of the compound for each leaf. The level of TSP was quantified by the Bradford assay, RuBisCO LSU and SSU were determined by LC-MS^E, and the defense metabolites were determined by UPLC-UV-ToF-MS. Plants were elicited as described in Figure 2 and leaves were harvested as described in Figure 3 (n = 5). FM, fresh mass; for other abbreviations, see Figure 1. Heat maps represent Kendall's τ coefficient for pairwise correlation of N investment in all of the above compounds among all genotype/elicitation groups.

(b) Investment of ^{15}N in RuBisCO LSU and SSU and defense metabolites was calculated as ^{15}N-incorporation multiplied by the N investment. Plants were pulse-labeled with $K^{15}NO_3$ 3 days before the first treatment. ^{15}N-incorporation was determined based on the MS spectra with the Excel spreadsheet ProSIPQuant (Taubert *et al.*, 2011).

functions downstream of JA signaling, and OS-elicited JA levels are not altered in ir*MYB8* plants (Kaur *et al.*, 2010), whereas they are significantly reduced in *LOX3*-silenced lines (to about one-third of that in the WT, but roughly six to seven times higher than in untreated controls; Allmann *et al.*, 2010). Thus, the observed phenotypes of the two transgenic lines are consistent with their respective MYB8 transcript levels, but not their JA levels; the *MYB8* expression after elicitation in the three lines used in this study is inversely proportional to the N investments into soluble proteins. Based on these results we conclude that the observed changes in N allocation after simulated herbivory only indirectly depend on JA signaling, and are probably caused by differences in *MYB8* expression or the MYB8-regulated synthesis of phenolamides. MYB8 could regulate defense induction by playing a role in N assimilation and allocation. In other plants and algae, members of the R2R3-MYB transcription factor family, to which NaMYB8 belongs, have been shown to be crucial for increases in the abundance of transcripts of N assimilation genes (Miyake *et al.*, 2003; Imamura *et al.*, 2009). To further elucidate the putative role of MYB8 in N reallocation, more detailed expression and enzyme activity studies targeting N metabolism at later time points after herbivory are necessary.

Based on our data we cannot differentiate whether MYB8 itself or the synthesis of phenolamides, in particular CP and DCS, mediate the changes in N investment into growth and defense. Silencing *MYB8* also silences genes further downstream of the transcription factor, and in addition to CP and DCS, the synthesis of at least 29 different coumaroyl-, caffeoyl- and feruloyl-containing metabolites (Onkokesung *et al.*, 2012). It is difficult to pinpoint the effects of single compounds in the complex biosynthetic network of a leaf, but applying phenolamides in different concentrations to control and elicited leaves of ir*MYB8* plants, and evaluating their effects on protein (RuBisCO) levels, or using plants silenced in genes affecting phenolamide biosynthesis downstream of MYB8, can help to evaluate if either MYB8 alone or MYB8 indirectly through phenolamide biosynthesis mediates the changes in N investment into proteins.

A comparison of the total N investment with the ^{15}N investment per mg fresh mass revealed a similar pattern, with increased ^{15}N in defense compounds and decreased ^{15}N in both RuBisCO subunits after elicitation. One major difference was that WT and ir*LOX3* plants allocated proportionally more ^{15}N than total N into CP and DCS, and less into nicotine, after elicitation, whereas the ^{15}N investment into the RuBisCO subunits was proportionally similar to the total N investment in both control and elicited leaves (Figure 4b; for a clearer comparison of N and ^{15}N investment, see Figure S5). Larger investments of recently assimilated ^{15}N into CP and DCS, compared with nicotine, makes ecological sense, because the OS used was from *M. sexta* larvae, a tobacco specialist, which is nicotine-tolerant but negatively affected by phenolamides (Kaur *et al.*, 2010).

A comparison of the decrease in total N investment into RuBisCO and TSP after OS elicitation with the N requirements of nicotine and phenolamide biosynthesis (Figures 4a and S6, showing a time-course analysis) suggested that RuBisCO metabolism could be a source of reallocated N to defense metabolite biosynthesis. Based on concentrations in the yRL, about 54% of N from RuBisCO or 13% of N from TSP could have been invested into phenolamides and nicotine (Figure S5). This comparison does not take into account the N requirements of biosynthetic enzymes or other N-containing inducible defense compounds, such as proteinase inhibitors (Zavala *et al.*, 2004b). Hence, the N demands for defense metabolite biosynthesis are likely to be underestimated; however, considering the dramatic decline in TSP it is likely that more N is released from the turnover of primary metabolism than N invested into defense metabolites.

Nitrogen invested into phenolamides does not originate from RuBisCO after herbivory

To further elucidate the N flux into defense metabolites and to investigate whether RuBisCO N is used as a source of N for CP and DCS biosynthesis after OS elicitation, the ^{15}N-incorporation (atomic percentage, At%) into N-containing metabolites and RuBisCO was determined in a time-course experiment (for details, see Figure 1b). This approach allows us to follow the N flux of a known quantity of ^{15}N, independently of within-leaf N pool sizes. The experiment was carried out with the yRL, because this leaf showed the greatest differences in N investment after elicitation (Figure 4a,b). It is important to note that during the experimental period, the ^{15}N-incorporation of the whole leaf was constant in all three lines, independent of elicitation (Figure S7), indicating that N is mainly redistributed within the leaves and that there is no increased net N influx into the leaf after elicitation.

As the ^{15}N-incorporation into RuBisCO LSUs and SSUs was similar, we only report on the incorporation into LSUs. Incorporation into RuBisCO increased at a constant rate until it reached a maximum of about 8 At% between 4 and 7 days after the first OS elicitation in all three lines, independent of elicitation (Figure 5). In contrast, ^{15}N was rapidly incorporated into CP and DCS in OS-elicited leaves until these compounds attained a maximum of about 10–12 At%, 4 days after the first elicitation in WT and ir*LOX3* plants. Had RuBisCO degradation provided the precursors for PA biosynthesis, it should have a similar or higher ^{15}N-incorporation as phenolamides, because the precursor pools will have similar or higher labeled isotope incorporation rates as their derived compounds. The large differences in ^{15}N-incorporation between CP and DCS and

3.2 L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, K. Groten. *The Plant Journal*, 75(3):417-429, July 2013.

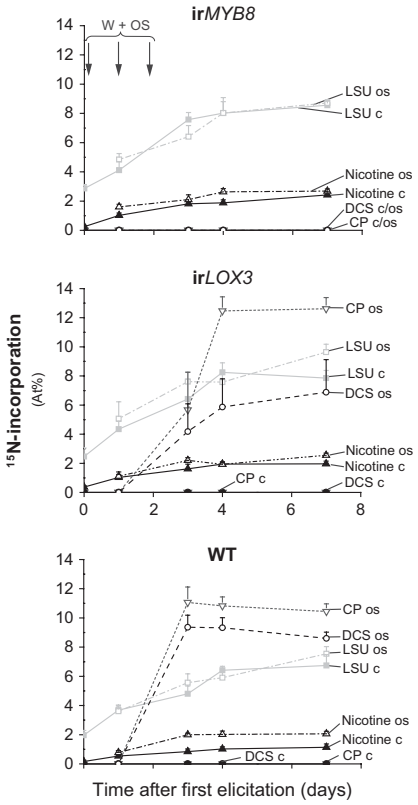


Figure 5. The dynamics of ^{15}N-incorporation into nicotine, caffeoyl-putrescine (CP), dicaffeoyl-spermidine (DCS) and RuBisCO large subunit (LSU) demonstrates that recently assimilated N, not N derived from LSU metabolism, is rapidly invested into CP and DCS biosynthesis after elicitation. Three days before the first W + OS treatment plants were pulse-labeled with $K^{15}NO_3$ (see Figure 1a). The yRL at the time of labeling was harvested at the time points indicated. ^{15}N-incorporation (n = 5) of RuBisCO LSU, nicotine, CP and DCS was determined as described for Figure 4. For abbreviations see Figure 1.

LSU make it unlikely that N derived from RuBisCO was used for CP and DCS biosynthesis. This result challenges the common conception that N released from products of primary metabolism (proteins) is a direct source for the production of defense metabolites (Herms and Mattson, 1992; Schwachtje *et al.*, 2006). In contrast, the data indicate that recently assimilated N is channeled into defense metabolite synthesis (Figure 4b). We hypothesize that N released from TSP turnover is mainly reinvested into other compounds, enabling the plant to react in different ways upon attack. Thus, plants may reduce the nutritive value of the tissue by reducing the level of TSP, and at the same time investing N not only in defense metabolites, but also in other N-containing compounds that are less digestible for the herbivore or more easily reallocated.

The incorporation of ^{15}N into nicotine only increased slightly after elicitation, and reached a maximum of around 2 At% in all three lines (Figure 5), although roots had a labeling of about 8 At%, similar to leaves (Figure S6). These findings differ from previous results showing the rapid incorporation of recently assimilated ^{15}N into nicotine after elicitation, but those results were obtained from plants that were starved of N for 24 h before application of the ^{15}N pulse, and ^{15}N was applied at the same time as MeJA to the roots (Baldwin *et al.*, 1994, 1998; Lynds and Baldwin, 1998). Elicitation of roots and shoots is known to differentially affect the accumulation of defense metabolites (van Dam and Oomen, 2008). Furthermore, MeJA is a stronger elicitor than OS elicitation (Voelckel *et al.*, 2001), and N-starved plants are known to transport N preferentially to the strongest sink (Ohtake *et al.*, 2001). These differences in experimental design probably led to different source–sink relationships within the plant, resulting in different patterns of ^{15}N investments.

Nicotine is a constitutively synthesized pool in the roots of *N. attenuata* that is transported to the shoot, but not metabolized, and contains 5–8% of the total N in the plant (Baldwin and Hamilton, 2000). It is possible that the newly synthesized nicotine might be diluted by the large pool of previously synthesized unlabeled nicotine, resulting in a low ^{15}N-incorporation. Alternatively, it may be derived from previously synthesized (and therefore unlabeled) precursors.

In summary, the ^{15}N-incoporation illustrates the flux of a defined ^{15}N pulse, independent of pool size, and indicates that N invested into CP and DCS is unlikely to be derived from RuBisCO, but is allocated directly to defense processes after assimilation instead of growth processes.

CONCLUSION

In this study, we quantified simulated herbivory-induced growth–defense trade-offs in a unified currency by measuring the investments of the limited resource of N into RuBisCO as proxy for growth and into small defense-related compounds (nicotine and phenolamides). In *N. attenuata*, OS elicitation reconfigures N allocation on multiple scales. Figure 6 summarizes the relative changes in the different N pool sizes after repeated simulated herbivory in the yRL. At the whole-plant scale, OS elicitation induced a weak N reallocation from the shoot to the root, thus presumably giving attacked plants a higher tolerance against herbivores by reducing the chances that valuable resources are removed by herbivores, and by increasing regrowth capacity after attack. At the within-leaf scale, changes between different N pools are much more dramatic. Taking the N level of RuBisCO after elicitation as a reference (x), RuBisCO-N declined 21x, and TSP declined from 73x to 9x, whereas N investment into defense metabolites increased, but to a far lesser extent (an increase from 5.5x to 11.5x for nicotine, and of 5x for CP and DCS).

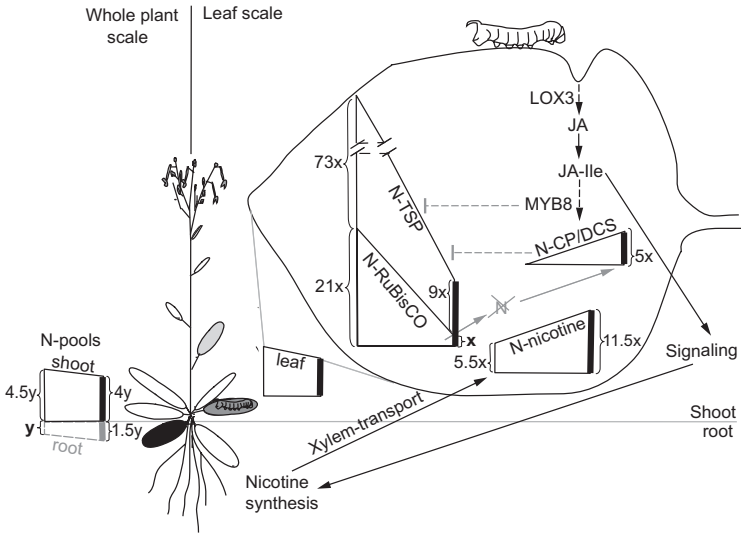


Figure 6. Herbivory-induced trade-offs of nitrogen (N) investment into growth and defense are mediated by MYB8. N investment in defense causes a reallocation of N from the shoot to the root. We suggest that the transcription factor MYB8, probably via the synthesis of phenolamides (caffeoyl-putrescine, CP; dicaffeoyl-spermidine, DCS), is involved in the reallocation of N within the local leaf. N invested in phenolamides, and in the root-synthesized alkaloid nicotine, increases after herbivory, whereas the N-investment in total soluble protein (TSP) and RuBisCO strongly decreases, but it is unlikely that N invested into phenolamides originates from RuBisCO metabolism. The height of the left and right side of the quadrangles represent relative changes in N pool sizes for each compound of C- and OS-elicited plants, respectively, using the N level of RuBisCO after elicitation as a reference (*x*). All N pools within the depicted leaf show the ratios of measured values per mg fresh mass. Shoot, root and whole-leaf N pools depicted outside the plant represent ratios of N determined per mg dry mass. For abbreviations, see Figure 1.

The transcription factor Na*MYB8*, possibly by regulating the production of metabolically dynamic phenolamides, CP and DCS, indirectly mediates the reconfiguration of N allocation after elicitation. The comparison of two elicited rosette leaves indicated that the extent of reconfiguration and the total concentration of defense metabolites produced depends on the developmental stage of the leaf, and on source–sink relationships.

Flux studies with ^{15}N strongly indicated that the N for PA biosynthesis comes from recently assimilated N rather than RuBisCO turnover. These results suggest that the drastic reallocation of resources and the shut-down of investments into growth within the leaf are not primarily driven by the direct costs for defense metabolite biosynthesis, but rather that N release from primary metabolism may enable the plants to react to attack in multiple ways. It remains to be elucidated if and how these allocation costs are translated into ecological costs. This question can only be answered if plants are grown in competition under different levels of herbivore attack.

Future experiments will seek to validate these results under more natural settings by comparing the results shown here for simulated herbivory, using OS of a specialist folivore, with damage by the natural herbivore community. An additional focus will be on tracing N investments into further metabolites and non-soluble proteins, and following the flux of N at the whole-plant level in more detail.

EXPERIMENTAL PROCEDURES

Plant germination and growth conditions

Seeds of the 31st generation of an inbred WT line of *N. attenuata* Torr. ex. Watts (Solanaceae) and two stably transformed lines, ir*MYB8* with reduced expression of the transcription factor Na*MYB8* (A-08-810, Kaur *et al.*, 2010), and ir*LOX3* silenced in lipoxygenase 3 (Na*LOX3*, A-03-562-2, Allmann *et al.*, 2010), were sterilized and germinated according to Kruegel *et al.* (2002) and cultivated in 1-L pots. For details on cultivation and fertilization, see the Appendix S1. The transgenic lines were homozygous, near-isogenic to the WT and representative of several independent transformation events.

Plant treatment

Pre-experiment to determine elicitation time points. Seven days after transfer to 1-L single pots, rosette-stage plants were pulse-labeled with 5.1 mg ^{15}N in 50 ml of a 0.694 g l^{-1} solution of $K^{15}NO_3$ (modified from Van Dam and Baldwin, 2001), and the oldest sink leaf (hereafter yRL) and the youngest source leaf (hereafter oRL; Pluskota *et al.*, 2007) were labeled for later sampling. The leaves and roots were harvested at 0, 4 and 12 h, and at 4, 7 and 10 days after the ^{15}N pulse. Roots were washed to remove excess soil and all samples were dried for 48 h at 60°C. Between 3 and 10 days after the pulse, the leaves and roots had a constant ^{15}N concentration (Figure 1b), indicating that an equilibrium had been reached. This time period was chosen for further experiments (Figure 1b, indicated by the grey arrows), as a stable ^{15}N-incorporation facilitates the analysis of proportional allocation to single compounds. The N pulse did not have any obvious effects on plant growth.

Pulse-labeling experiments. Three days after the ^{15}N pulse the oldest sink, youngest source and transition leaf at the time point of labeling were wounded with a pattern wheel and treated with *M. sexta* OS (10 μl per leaf per day, 1:5 diluted) on three consecutive days (Ullmann-Zeunert *et al.*, 2012). Unelicited plants were used as controls. For the whole-shoot N analysis, the aboveground biomass of control and elicited plants was harvested 4 days after the first elicitation, and dried as above.

For the N-partitioning analysis, both the locally elicited yRL and oRL were harvested 4 days after the first elicitation and flash-frozen in liquid N_2. After stalk elongation, the first stem leaf (S1) was harvested when it reached the source–sink transition stage.

3.2 L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, K. Groten. *The Plant Journal*, 75(3):417-429, July 2013.

For all three leaves, only the right leaf blade was harvested to standardize sampling and minimize changes in source–sink relationships arising from repeated sampling. The harvest time points of the S1 leaf differed depending on plant development. The first mature seed capsules were harvested at the day of opening: seeds were counted, weighed and analyzed for N content. For the kinetic analysis, plants received a ^{15}N pulse and were elicited as described above, and the locally elicited yRL was harvested at 0, 1, 3, 4 and 7 days after the first elicitation (Figure 1b). The sample size for all analyses was five.

Protein extraction and quantification

The TSP and RuBisCO LSUs and SSUs were extracted and quantified by Bradford assay and LC-MS^E, respectively, as described by Ullmann-Zeunert *et al.* (2012). The ^{15}N-incorporation of RuBisCO was determined with the Excel spreadsheet ProSIPQuant (Taubert *et al.*, 2011).

Metabolite extraction and quantification

Small metabolites were extracted as in Gaquerel *et al.* (2010) and analyzed by UPLC/UV/ToF-MS, using a Dionex RSLC system with a diode array detector (Dionex, http://www.dionex.com) and a Micro-ToF Mass Spectrometer (Bruker, http://www.bruker.com). Further details on instrument parameters and quantification are described in Appendix S1. Average mass spectra were extracted for ^{15}N-incorporations using the Excel spreadsheet ProSIPQuant (Taubert *et al.*, 2011), modified for small metabolites based on compound sum formulae.

Isotope ratio mass spectrometry analysis (IRMS)

The IRMS sample preparation, analysis and following calculations of total N content (% dry mass) and ^{15}N-incorporation were carried out as described in Meldau *et al.* (2012).

Statistical analysis

The R environment was used for statistical analysis (Team, 2009). For ANOVA and ANCOVA analyses, if the assumption of homoscedasticity of variances was violated or the residuals did not follow a normal distribution, response variables were transformed prior to the analyses using Box–Cox transformation (see Appendix S2). The Box–Cox lambda was estimated using Venables' and Ripley's MASS library for R. All ANOVA models were simplified to the minimum adequate model using Aikaike's information criterion (Ronchetti, 1985). For the correlation analysis (Figure 4a, heat maps) the data were imported into the environment and vectors containing the following variables were generated: N-rest protein $\mu g\ mg^{-1}$, N-RuBisCO LSU $\mu g\ mg^{-1}$, N-RuBisCO SSU $\mu g\ mg^{-1}$, N-nicotine $\mu g\ mg^{-1}$, N-CP $\mu g\ mg^{-1}$ and N-DCS $\mu g\ mg^{-1}$. These vectors were pairwise correlated, calculating Kendall's τ coefficient (Kendall, 1938). In contrast to Pearson's correlation coefficient, Kendall's τ is more robust and not sensitive to the data distribution.

ACKNOWLEDGEMENTS

The authors thank Franziska Hufsky for bioinformatics help with RuBisCO quantification and Dr Matthias Schöttner for technical support with metabolite measurements. This research was supported by the Max Planck Society, M.A.S. was supported by a grant of the International Max Planck Research School and I.T.B. was supported by an advanced ERC grant, ClockworkGreen (293926).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. The dry mass and the absolute amount of nitrogen (N) of the shoot are not influenced by genotype.

Figure S2. Average leaf size produced by transgenic (ir*LOX3*, ir*MYB8*) and WT plants with and without the W + OS treatment.

Figure S3. Silencing of *LOX3* and *MYB8* alters the absolute pools of RuBisCO (a) and N-containing small metabolites (b) in leaves.

Figure S4. Reproductive timing and output produced by transgenic (ir*LOX3*, ir*MYB8*) and WT plants with and without the W + OS treatment.

Figure S5. Increased N investment into nicotine, CP and DCS is accompanied by a decreased N investment into RuBisCO.

Figure S6. The decrease of N investment into protein pools is greater than the amount of N required for the biosynthesis of the N-containing defense metabolites.

Figure S7. ^{15}N-incorporation in the yRL is not influenced by treatment or genotype.

Appendix S1. Supplemental procedures.

Appendix S2. Supplemental statistical information.

REFERENCES

Allen, D.K., Laclair, R.W., Ohlrogge, J.B. and Shachar-Hill, Y. (2012) Isotope labelling of Rubisco subunits provides in vivo information on subcellular biosynthesis and exchange of amino acids between compartments. *Plant Cell Environ.* **35**, 1232–1244.

Allmann, S., Halitschke, R., Schuurink, R.C. and Baldwin, I.T. (2010) Oxylipin channelling in *Nicotiana attenuata*: lipoxygenase 2 supplies substrates for green leaf volatile production. *Plant Cell Environ.* **33**, 2028–2040.

Baldwin, I.T. (1999) Inducible nicotine production in native *Nicotiana* as an example of adaptive phenotypic plasticity. *J. Chem. Ecol.* **25**, 3–30.

Baldwin, I.T. and Hamilton, W. (2000) Jasmonate-induced responses of *Nicotiana sylvestris* results in fitness costs due to impaired competitive ability for nitrogen. *J. Chem. Ecol.* **26**, 915–952.

Baldwin, I.T. and Ohnmeiss, T.E. (1994) Coordination of photosynthetic and alkaloidal responses to damage in uninducible and inducible *Nicotiana sylvestris*. *Ecology*, **75**, 1003–1014.

Baldwin, I.T., Karb, M.J. and Ohnmeiss, T.E. (1994) Allocation of ^{15}N from nitrate to nicotine - production and turnover of a damage-induced mobile defense. *Ecology*, **75**, 1703–1713.

Baldwin, I.T., Gorham, D., Schmelz, E.A., Lewandowski, C.A. and Lynds, G.Y. (1998) Allocation of nitrogen to an inducible defense and seed production in *Nicotiana attenuata*. *Oecologia*, **115**, 541–552.

Bazzaz, F.A., Chiariello, N.R., Coley, P.D. and Pitelka, L.F. (1987) Allocating rescources to reproducation and defense. *Bioscience*, **37**, 58–67.

Chapin, F.S., Schulze, E.D. and Mooney, H.A. (1990) The ecology and economics of storage in plants. *Annu. Rev. Ecol. Syst.* **21**, 423–447.

van Dam, N.M. and Oomen, M.W.A.T. (2008) Root and shoot jasmonic acid applications differentially affect leaf chemistry and herbivore growth. *Plant Signal. Behav.* **3**, 91–98.

Diezel, C., von Dahl, C.C., Gaquerel, E. and Baldwin, I.T. (2009) Different lepidopteran elicitors account for cross-talk in herbivory-induced phytohormone signaling. *Plant Physiol.* **150**, 1576–1586.

Ellis, R.J. (1979) Most abundant protein in the world. *Trends Biochem. Sci.* **4**, 241–244.

Frost, C.J. and Hunter, M.D. (2008) Herbivore-induced shifts in carbon and nitrogen allocation in red oak seedlings. *New Phytol.* **178**, 835–845.

Gaquerel, E., Heiling, S., Schoettner, M., Zurek, G. and Baldwin, I.T. (2010) Development and validation of a liquid chromatography-electrospray ionization-time-of-flight mass spectrometry method for Induced changes in *Nicotiana attenuata* leaves during simulated herbivory. *J. Agric. Food Chem.* **58**, 9418–9427.

Giri, A.P., Wuensche, H., Mitra, S., Zavala, J.A., Muck, A., Svatos, A. and Baldwin, I.T. (2006) Molecular interactions between the specialist herbi-

3.2 L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, K. Groten. *The Plant Journal*, 75(3):417-429, July 2013.

vore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. VII. Changes in the plant's proteome. *Plant Physiol.* **142**, 1621–1641.

Gomez, S., Ferrieri, R.A., Schueller, M. and Orians, C.M. (2010) Methyl jasmonate elicits rapid changes in carbon and nitrogen dynamics in tomato. *New Phytol.* **188**, 835–844.

Gomez, S., Steinbrenner, A.D., Osorio, S., Schueller, M., Ferrieri, R.A., Fernie, A.R. and Orians, C.M. (2012) From shoots to roots: transport and metabolic changes in tomato after simulated feeding by a specialist lepidopteran. *Entomol. Exp. Appl.* **144**, 101–111.

Halitschke, R., Gase, K., Hui, D.Q., Schmidt, D.D. and Baldwin, I.T. (2003) Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. VI. Microarray analysis reveals that most herbivore-specific transcriptional changes are mediated by fatty acid-amino acid conjugates. *Plant Physiol.* **131**, 1894–1902.

Halitschke, R., Ziegler, J., Keinanen, M. and Baldwin, I.T. (2004) Silencing of hydroperoxide lyase and allene oxide synthase reveals substrate and defense signaling crosstalk in *Nicotiana attenuata*. *Plant J.* **40**, 35–46.

Heil, M. and Baldwin, I.T. (2002) Fitness costs of induced resistance: emerging experimental support for a slippery concept. *Trends Plant Sci.* **7**, 61–67.

Herms, D.A. and Mattson, W.J. (1992) The dilemma of plants - to grow or defend. *Q. Rev. Biol.* **67**, 283–335.

Hibi, N., Higashiguchi, S., Hashimoto, T. and Yamada, Y. (1994) Gene-expression in tobacco low-nicotine mutants. *Plant Cell*, **6**, 723–735.

Imai, K., Suzuki, Y., Mae, T. and Makino, A. (2008) Changes in the synthesis of rubisco in rice leaves in relation to senescence and N influx. *Ann. Bot.* **101**, 135–144.

Imamura, S., Kanesaki, Y., Ohnuma, M., Inouye, T., Sekine, Y., Fujiwara, T., Kuroiwa, T. and Tanaka, K. (2009) R2R3-type MYB transcription factor, CmMYB1, is a central nitrogen assimilation regulator in *Cyanidioschyzon merolae*. *Proc. Natl Acad. Sci. USA*, **106**, 12548–12553.

Ishimaru, K., Kobayashi, N., Ono, K., Yano, M. and Ohsugi, R. (2001) Are contents of Rubisco, soluble protein and nitrogen in flag leaves of rice controlled by the same genetics? *J. Exp. Bot.* **52**, 1827–1833.

Karban, R. and Baldwin, I.T. (1997) *Induced Responses to Herbivory*. Chicago: University of Chicago Press.

Kaur, H., Heinzel, N., Schoettner, M., Baldwin, I.T. and Galis, I. (2010) R2R3-NaMYB8 regulates the accumulation of phenylpropanoid-polyamine conjugates, which are essential for local and systemic defense against insect herbivores in *Nicotiana attenuata*. *Plant Physiol.* **152**, 1731–1747.

Kendall, M. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–89.

Kessler, A., Halitschke, R. and Baldwin, I.T. (2004) Silencing the jasmonate cascade: Induced plant defenses and insect populations. *Science*, **305**, 665–668.

Kruegel, T., Lim, M., Gase, K., Halitschke, R. and Baldwin, I.T. (2002) Agrobacterium-mediated transformation of *Nicotiana attenuata*, a model ecological expression system. *Chemoecology*, **12**, 177–183.

Lou, Y.G. and Baldwin, I.T. (2004) Nitrogen supply influences herbivore-induced direct and indirect defenses and transcriptional responses to *Nicotiana attenuata*. *Plant Physiol.* **135**, 496–506.

Lynds, G.Y. and Baldwin, I.T. (1998) Fire, nitrogen, and defensive plasticity in *Nicotiana attenuata*. *Oecologia*, **115**, 531–540.

Makino, A., Mae, T. and Ohira, K. (1984) Relation between nitrogen and ribulose-1,5-bisphosphate carboxylase in rice leaves from emergence through senescence. *Plant Cell Physiol.* **25**, 429–437.

Makino, A., Harada, M., Kaneko, K., Mae, T., Shimada, T. and Yamamoto, N. (2000) Whole-plant growth and N allocation in transgenic rice plants with decreased content of ribulose-1,5-bisphosphate carboxylase under different CO_2 partial pressures. *Aust. J. Plant Physiol.* **27**, 1–12.

Matt, P., Krapp, A., Haake, V., Mock, H.P. and Stitt, M. (2002) Decreased Rubisco activity leads to dramatic changes of nitrate metabolism, amino acid metabolism and the levels of phenylpropanoids and nicotine in tobacco antisense RBCS transformants. *Plant J.* **30**, 663–677.

McCloud, E.S. and Baldwin, I.T. (1997) Herbivory and caterpillar regurgitants amplify the wound-induced increases in jasmonic acid but not nicotine in *Nicotiana sylvestris*. *Planta*, **203**, 430–435.

McKey, D. (1974) Adaptive patterns in alkaloid physiology. *Am. Nat.* **108**, 305–320.

McKey, D. (1979) Distribution of secondary compounds within plants. In *Herbivores: Their Interaction with Secondary Plant Metabolites* (Rosenthal, G.A. and Janzen, D.H., eds). New York: Academic Press, pp. 1–55.

Meldau, S., Ullmann-Zeunert, L., Govind, G., Bartram, S. and Baldwin, I.T. (2012) Basal and herbivory-induced defense trade-offs are mediated by mitogen-activated protein kinases, jasmonic acid and salicylic acid in the native tobacco, *Nicotiana attenuata*. *BMC Plant Biol.* **12**, 213.

Millard, P. (1988) The accumulation and storage of nitrogen by herbaceous plants. *Plant Cell Environ.* **11**, 1–8.

Miyake, K., Ito, T., Senda, M., Ishikawa, R., Harada, T., Niizeki, M. and Akada, S. (2003) Isolation of a subfamily of genes for R2R3-MYB transcription factors showing up-regulated expression under nitrogen nutrient-limited conditions. *Plant Mol. Biol.* **53**, 237–245.

Mole, S. (1994) Trade-offs and constraints in plant-herbivore defense theory - a life-history perspective. *Oikos*, **71**, 3–12.

Ohnmeiss, T.E. and Baldwin, I.T. (2000) Optimal Defense theory predicts the ontogeny of an induced nicotine defense. *Ecology*, **81**, 1765–1783.

Ohtake, N., Sato, T., Fujikake, H. *et al.* (2001) Rapid N transport to pods and seeds in N-deficient soybean plants. *J. Exp. Bot.* **52**, 277–283.

Onkokesung, N., Galis, I., von Dahl, C.C., Matsuoka, K., Saluz, H.-P. and Baldwin, I.T. (2010) Jasmonic acid and ethylene modulate local responses to wounding and simulated herbivory in *Nicotiana attenuata* leaves. *Plant Physiol.* **153**, 785–798.

Onkokesung, N., Gaquerel, E., Kotkar, H., Kaur, H., Baldwin, I.T. and Galis, I. (2012) MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A:polyamine transferases in *Nicotiana attenuata*. *Plant Physiol.* **158**, 389–407.

Orians, C.M., Thorn, A. and Gomez, S. (2011) Herbivore-induced resource sequestration in plants: why bother? *Oecologia*, **167**, 1–9.

Pluskota, W.E., Qu, N., Maitrejean, M., Boland, W. and Baldwin, I.T. (2007) Jasmonates and its mimics differentially elicit systemic defence responses in *Nicotiana attenuata*. *J. Exp. Bot.* **58**, 4071–4082.

Preston, C.A. and Baldwin, I.T. (1999) Positive and negative signals regulate germination in the post-fire annual, *Nicotiana attenuata*. *Ecology*, **80**, 481–494.

Rhoades, D.F. (1979) Evolution of plant chemical defense against herbivores. In *Herbivores: Their Interaction with Secondary Plant Metabolites* (Rosenthal, G.A. and Janzen, D.H., eds). New York: Academic Press, pp. 1–55.

Ronchetti, E. (1985) Robust model selection in regression. *Stat. Probab. Lett.* **3**, 21–23.

Schittko, U., Hermsmeier, D. and Baldwin, I.T. (2001) Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. II. Accumulation of plant mRNAs in response to insect-derived cues. *Plant Physiol.* **125**, 701–710.

Schwachtje, J., Minchin, P.E.H., Jahnke, S., van Dongen, J.T., Schittko, U. and Baldwin, I.T. (2006) SNF1-related kinases allow plants to tolerate herbivory by allocating carbon to roots. *Proc. Natl Acad. Sci. USA*, **103**, 12935–12940.

Simon, J., Gleadow, R.M. and Woodrow, I.E. (2010) Allocation of nitrogen to chemical defence and plant functional traits is constrained by soil N. *Tree Physiol.* **30**, 1111–1117.

Skibbe, M., Qu, N., Galis, I. and Baldwin, I.T. (2008) Induced plant defenses in the natural environment: nicotiana attenuata WRKY3 and WRKY6 coordinate responses to herbivory. *Plant Cell*, **20**, 1984–2000.

Stamp, N. (2003) Out of the quagmire of plant defense hypotheses. *Q. Rev. Biol.* **78**, 23–55.

Steinbrenner, A.D., Gomez, S., Osorio, S., Fernie, A.R. and Orians, C.M. (2011) Herbivore-induced changes in tomato (*Solanum lycopersicum*) primary metabolism: a whole plant perspective. *J. Chem. Ecol.* **37**, 1294–1303.

Steppuhn, A., Gase, K., Krock, B., Halitschke, R. and Baldwin, I.T. (2004) Nicotine's defensive function in nature. *PLoS Biol.* **2**, 1074–1080.

Stitt, M. and Krapp, A. (1999) The interaction between elevated carbon dioxide and nitrogen nutrition: the physiological and molecular background. *Plant Cell Environ.* **22**, 583–621.

Stitt, M. and Schulze, D. (1994) Does Rubisco control the rate of photosynthesis and plant-growth - an exercise in molecular ecophysiology. *Plant Cell Environ.* **17**, 465–487.

Stork, W., Diezel, C., Halitschke, R., Galis, I. and Baldwin, I.T. (2009) An ecological analysis of the herbivory-elicited JA burst and its metabolism:

3.2 L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, K. Groten. *The Plant Journal*, 75(3):417-429, July 2013.

plant memory processes and predictions of the moving target model. *PLoS ONE*, **4**, e4697.

Takano, A., Kakehi, J.I. and Takahashi, T. (2012) Thermospermine is not a minor polyamine in the plant kingdom. *Plant Cell Physiol.* **53**, 606–616.

Taubert, M., Jehmlich, N., Vogt, C., Richnow, H.H., Schmidt, F., von Bergen, M. and Seifert, J. (2011) Time resolved protein-based stable isotope probing (Protein-SIP) analysis allows quantification of induced proteins in substrate shift experiments. *Proteomics*, **11**, 2265–2274.

Team, R.D.C. (2009) *R: A Language and Environment for Statistical Computing*.Vienna: R Foundation for Statistical Computing http://www.r-project.org.

Trumble, J.T., Kolodnyhirsch, D.M. and Ting, I.P. (1993) Plant compensation for arthropod herbivory. *Annu. Rev. Entomol.* **38**, 93–119.

Ullmann-Zeunert, L., Muck, A., Wielsch, N., Hufsky, F., Stanton, M.A., Bartram, S., Böcker, S., Baldwin, I.T., Groten, K. and Svatos, A. (2012) Determination of ^{15}N-Incorporation into plant proteins and their absolute quantitation: A new tool to study nitrogen flux dynamics and protein pool sizes elicited by plant–herbivore interactions. *J. Proteome Res.* **11**, 4947.

Van Dam, N.M. and Baldwin, I.T. (2001) Competition mediates costs of jasmonate-induced defences, nitrogen acquisition and transgenerational plasticity in *Nicotiana attenuata*. *Funct. Ecol.* **15**, 406–415.

Van Dam, N.M., Hermenau, U. and Baldwin, I.T. (2001) Instar-specific sensitivity of specialist Manduca sexta larvae to induced defences in their host plant Nicotiana attenuata. *Ecol. Entomol.* **26**, 578–586.

Voelckel, C. and Baldwin, I.T. (2004a) Generalist and specialist lepidopteran larvae elicit different transcriptional responses in *Nicotiana attenuata*, which correlate with larval FAC profiles. *Ecol. Lett.* **7**, 770–775.

Voelckel, C. and Baldwin, I.T. (2004b) Herbivore-induced plant vaccination. Part II. Array-studies reveal the transience of herbivore-specific transcriptional imprints and a distinct imprint from stress combinations. *Plant J.* **38**, 650–663.

Voelckel, C., Krugel, T., Gase, K., Heidrich, N., van Dam, N.M., Winz, R. and Baldwin, I.T. (2001) Anti-sense expression of putrescine N-methyltransferase confirms defensive role of nicotine in *Nicotiana sylvestris* against *Manduca sexta*. *Chemoecology*, **11**, 121–126.

Woldemariam, M.G., Baldwin, I.T. and Galis, I. (2011) Transcriptional regulation of plant inducible defenses against herbivores: a mini-review. *J. Plant Interact.* **6**, 113–119.

Zangerl, A.R., Hamilton, J.G., Miller, T.J., Crofts, A.R., Oxborough, K., Berenbaum, M.R. and de Lucia, E.H. (2002) Impact of folivory on photosynthesis is greater than the sum of its holes. *Proc. Natl Acad. Sci. USA*, **99**, 1088–1091.

Zavala, J.A., Patankar, A.G., Gase, K. and Baldwin, I.T. (2004a) Constitutive and inducible trypsin proteinase inhibitor production incurs large fitness costs in *Nicotiana attenuata*. *Proc. Natl Acad. Sci. USA*, **101**, 1607–1612.

Zavala, J.A., Patankar, A.G., Gase, K., Hui, D.Q. and Baldwin, I.T. (2004b) Manipulation of endogenous trypsin proteinase inhibitor production in *Nicotiana attenuata* demonstrates their function as antiherbivore defenses. *Plant Physiol.* **134**, 1181–1190.

Kapitel 4

Diskussion

Die vorliegende Doktorarbeit enthält vier Veröffentlichungen, die Daten aus biologischer und medizinischer Forschung hinsichtlich ihrer Varianzen untersuchen. Kapitel 2 zeigt zwei Arbeiten zum Umgang mit in Microarray-Experimenten immanenten Messfehlern. Zwei Beispiele biologischer Varianz, davon eines wieder in einem Microarray-Experiment, werden in Kapitel 3 gezeigt.

4.1 Zur technischen Varianz

In der Genexpressionsanalyse mit Microarrays ist der Messfehler traditionell hoch (Heber and Sick [2006], Modlich and Munnes [2007], Workman et al. [2002]). In der Literatur wird für diesen Messfehler eine Vielzahl von Gründen angegeben. Einige Arbeiten legen nahe, dass das Phänomen der Kreuzhybridisierung, das heißt des unspezifischen Bindens von RNA-Fragmenten, die Messungen maßgeblich verfälscht (Cambon et al. [2007], Chen et al. [2007], Wu et al. [2005]). Andere Autoren behaupten hingegen, der Einfluss wäre vernachlässigbar (Elbez et al. [2006]). Insofern war die Frage, inwieweit Kreuzhybridisierungen die Qualität der Messergebnisse beeinflussen, bisher offen. Diese Arbeit zeigt nun, dass der Einfluss von Kreuzhybridisierungen auf den Messfehler ganz erheblich ist.

In einer ersten Untersuchung, die im Kapitel 2.1 dargestellt ist, der Arbeit

Creation and comparison of different chip definition files for Affymetrix microarrays (Hummert et al. [2011]), wird zunächst ein neues Chip Definition File (CDF) erstellt, das solche Probes ausschließt, die möglicherweise kreuzhybridisieren. Solche mehrdeutigen Probes sind auf dem Chip materiell vorhanden, werden bei der Auswertung aber nicht mehr berücksichtigt. Hierdurch können falsch positive Ergebnisse vermieden werden, also solche bei denen 'falsche' RNA-Fragmente an einer Sonde hybridisieren, deren 'richtige' RNA-Fragmente aber eigentlich nicht oder viel weniger vorhanden sind. Falsch negative Ergebnisse können so jedoch nicht vermieden werden. Dies sind durch den sogenannten Stealing-Effekt (siehe Seite 10) verursachte Einflüsse. Es fehlen also RNA-Fragmente, die an einer 'falschen' Sonde binden und dann einen zu niedrigen Wert an der eigentlich 'richtigen' Sonde verursachen. Da die 'falschen' Sonden materiell noch vorhanden sind, hybridisieren an ihnen auch RNA-Fragmente.

Beim Affymetrix Genechip HG-U133 Plus 2.0 werden beispielsweise durch das neue CDF, also das Weglassen in der Auswertung von Sonden, die nicht eindeutig hybridisieren oder das angegebene Gen gar nicht repräsentieren, nur noch 37 % der Probes verwendet. Etwa 63 % der Probes sind also mehrdeutige Probes. Vergleiche zeigen, dass dieses CDF eine mindestens ebenso gute Auswertung liefert wie mittels des bisherigen CDFs, obwohl deutlich weniger Sonden pro Probeset zur Verfügung stehen. Dieses Ergebnis bestätigt die Hypothese, dass Kreuzhybridisierungen tatsächlich eine bedeutende Quelle für die technische Varianz sind.

Das neue CDF liefert eine kostenneutrale Lösung für das Problem der Kreuzhybridisierungen. Es müssen keine neuen Arrays gespottet werden, sondern es muss lediglich das neue CDF in der Auswertung verwendet werden. Da aber die mehrdeutigen Probes immer noch auf dem Chip vorhanden sind, und der beschriebene Stealing-Effekt rechnerisch nicht korrigiert werden kann, bleibt als einzige mögliche Lösung die Erstellung eines neuen Microarrays, und zwar ganz ohne mehrdeutige Probes.

Dieser weitergehende Ansatz wird in der im Kapitel 2.2 angeführten Ar-

beit *Optimization of a microarray probe design focusing on the minimization of cross-hybridization* (Horn et al. [2011]) mit dem Entwurf eines neuen Arrays mit besonderer Beachtung auf der Vermeidung von Kreuzhybridisierungen umgesetzt. Für den Organismus *Aspergillus nidulans* gab es vor der Arbeit ein bestehendes Microarray-Design. Nun sollte eine neue Generation von Microarrays für den Organismus gespottet werden. In dem Entwurf wird konsequent darauf geachtet, Kreuzhybridisierungen zu vermeiden. Mit beiden Designs wird ein Experiment unter gleichen Bedingungen durchgeführt. Es zeigt sich, dass die mittlere Varianz der technischen Replikate zwischen 14 % und 24 % abnimmt. Hiermit wird überzeugend gezeigt, dass der Effekt der Kreuzhybridisierungen einen erheblichen Einfluss auf die Messgüte von Microarray-Experimenten hat.

Inzwischen wird zur qualitativen und quantitativen Messung der Genexpression regelmäßig Next-Generation RNA-Sequenzierung verwendet. RNA-Seq erlaubt eine größere Auflösung und damit auch Einblicke in nur schwach exprimierte Transkripte. Der Hauptvorteil ist aber, dass mittels RNA-Seq eine *de-novo*-Transkriptomanalyse von solchen Organismen, deren Genomsequenz zuvor nicht sequenziert wurde, möglich ist (Paszkiewicz and Studholme [2010]). Dies erlaubt eine Abwendung von den Modellorganismen und der Forscher kann sich direkt dem *organism of interest* zuwenden.

Die Spearman-Korrelation (siehe Seite 23) zwischen mit RNA-Seq und Microarrays gemessenen Genexpressionsdaten liegt in einer Studie (Marioni et al. [2008]) bei 0,73. Das macht die beiden Technologien vergleichbar. Die Korrelation zwischen RNA-Seq-Daten und mit dem Goldstandard qRT-PCR gemessenen Werten ist höher als die zwischen Microarray- und qRT-PCR-Daten, was bedeutet, dass RNA-Seq-Daten genauer sind.

Nichtsdestotrotz sind Microarrays weiterhin eine wichtige Technologie zur Bestimmung der Genexpression. Sie sind in der Anschaffung günstiger. Im Jahr 2011 waren die Kosten pro Messung bei Microarrays um das 10-fache geringer als bei RNA-Seq (Malone and Oliver [2011]). Zudem erlauben Microarrays eine schnellere Messung und das Verfahren ist in der Durchführung einfacher

(Mockler et al. [2005]). Der höhere Preis pro Replikat bei RNA-Seq führt in der Praxis auch zu dem Effekt, dass, um Kosten zu sparen, weniger technische sowie biologische Replikate gemessen werden. Dadurch wird die bessere Genauigkeit des Verfahrens „statistisch aufgefressen"(Auer and Doerge [2010]).

Desweiteren ist es nicht so, dass das Problem der Kreuzhybridisierung mit der Einführung von Sequenzern der Vergangenheit angehört. Da auch beim RNA-Seq der Digestionsschritt, in dem die extrahierte RNA in kleinere Fragmente zerlegt wird, stattfindet, ergibt sich bei Genen mit gleichen Abschnitten wieder das Problem der Zuordnung, analog zum Problem der Kreuzhybridisierung bei Microarrays (Mardis [2008]). Die Ableselänge beträgt beispielsweise bei dem verbreiteten Illumina Sequencer zwischen 32 und 40 bp, was mit den 25meren bei Affymetrix Arrays durchaus vergleichbar ist (Fox et al. [2009]). Eine in *Nature Methods* veröffentlichte Studie aus dem Jahr 2008 behauptet, die durchschnittliche Ableselänge läge zwischen 25 und 35 bp und 15 bis 20 % der gelesenen Sequenzen seien nicht eindeutig zuordenbar (Wold and Myers [2008]). Beim RNA-Seq wird dieser Effekt mit *Cross-Alignment* analog zu *Cross-Hybridization* bezeichnet. Neuere Studien ergeben, dass sich die Verfahren eher gegenseitig ergänzen, als dass das eine das andere verdrängen wird (Valdes et al. [2013], Malone and Oliver [2011], Liu et al. [2007a]).

Die Berücksichtigung von Kreuzhybridisierungen kann optimalerweise nur beim Design der Arrays stattfinden. Es ist allerdings so, dass beim Entwurf neuer Arrays mehrere Vorgaben konkurrieren. Insbesondere konkurriert das Ziel möglichst viele oder idealerweise alle Gene eines Organismus abzubilden, mit dem Ziel, keine Kreuzhybridisierungen zuzulassen. In der Regel wird die Vermeidung von Kreuyhybridisierungen ersterem Ziel untergeordnet. Vor allem kommerzielle Arrays werden häufig als *Full Genome Arrays* angeboten. Auch weitere Vorgaben, wie GC-Gehalt oder Schmelztemperatur, können gegen ein Oligomer sprechen, das Kreuzhybridisierungen vermeidet.

Bei kommerziellen Full Genome Arrays ist das Problem eklatant. Beispielsweise sind von den untersuchten Affymetrix Arrays sämtliche, in unterschied-

lich großem Ausmaß, von dem Problem betroffen. Beim Affymetrix GeneChip U133A keuzhybridisieren 45 % der Probes, beim U133B 2 %, beim U133 2.0 Plus 28 % und beim Mouse Genome 430 2.0 Array 5 % der Sonden.

In der Auswertekette von Microarray-Experimenten ist die Behandlung von Kreuzhybridisierungen im Preprocessing angesiedelt. Häufig wird auf diesen Schritt allerdings völlig verzichtet. Sogar die von Affymetrix vorgegebene Unterteilung der Probes in die Gruppen mit den auf Seite 8 aufgeführten Suffixen _s_at, _x_at, _i_at oder _a_at wird häufig ignoriert (Dixon et al. [2007], Rohle et al. [2013], Sethu et al. [2006]).

Dabei ist die Auswirkung von Kreuzhybridisierungen auf das Ergebnis oft ganz erheblich. In der im Kapitel 2.1 dargestellten Studie *Creation and comparison of different chip definition files for Affymetrix microarrays* (Hummert et al. [2011]) verbessert sich die Korrelation zwischen mit Microarrays und qRT-PCR gemessenen Expressionsdaten von 0,61 auf 0,72 um über 18 %. In der im Kapitel 2.2 angeführten Arbeit *Optimization of a microarray probe design focusing on the minimization of cross-hybridization* (Horn et al. [2011]) wird gezeigt, dass die mittlere Varianz der technischen Replikate zwischen 14 % und 24 % abnimmt, wenn im Arraydesign konsequent auf die Vermeidung von Kreuzhybridisierungen geachtet wird.

Solche Abweichungen sind zum Beispiel beim Reverse Engineering genregulatorischer Netzwerke sehr bedeutend. Bei einer Studie, in der genregulatorische Netzwerke mit FastNCA aus künstlichen Microarray-Daten rekonstruiert werden, können bei einem Messfehler von 10 % nur noch ein Zehntel der Netzwerkknoten korrekt rekonstruiert werden (Chang et al. [2008]). In einer Arbeit, die genregulatorische Netzwerke mit Hilfe von Singular Value Decomposition (SVD) basierten Methoden rekonstruiert, wird auch auf die Notwendigkeit eines kleinen Fehlers hingewiesen (Guthke et al. [2005]). Generell ist es so, dass die Qualität eines rekonstruierten Netzwerks von der Qualität der zugrundeliegenden Daten direkt abhängt (Marseguerra et al. [2005]).

Auch bei den dem Reverse Engineering genregulatorischer Netzwerke vorge-

lagerten Schritten wirkt sich der Messfehler negativ aus. Beispielsweise werden der eigentlichen Netzwerkrekonstruktion regelmäßig Clusteranalysen vorangestellt. Obwohl unüberwachte Clusterverfahren als extrem robust gegenüber Rauschen gelten, sind bei Microarray-Daten die Abweichungen teilweise so hoch, dass es zu Fehlern beim Clustern kommt (Liu et al. [2007b], Liu and Rattray [2010]).

Zusammenfassend kann gesagt werden: Mit Daten geringer Qualität lassen sich keine guten Ergebnisse erzielen. Zumindest ist der Aufwand, fehlerbehaftete Daten im Nachhinein zu korrigieren, in der Regel höher als der, während der Datenerfassung Fehler zu vermeiden. Im englischen Sprachraum hat sich hierfür das geflügeltes Wort: „*Garbage in, Garbage out*“ etabliert (Leming et al. [2003], Bininda-Emonds et al. [2004]).

Konsequenz dieser Arbeit soll zum einen sein, die Qualität der Daten gerade bei Hochdurchsatzexperimenten immer im Auge zu haben. Dabei sollte für jeden Schritt überprüft werden, ob die Daten weiter verrauscht werden. Da sich Messfehler durch Fehlerfortpflanzung im Verlauf der Auswertekette zum Ende hin immer stärker auswirken können, ist es von Bedeutung, gerade beim Preprocessing, auf eine strikteste Qualitätskontrolle zu achten (Jochum et al. [1981], Taylor [1997]). Für Microarray-Experimente heißt dies insbesondere, möglichen Kreuzhybridisierungen ein besonderes Augenmerk zu schenken. Wenn immer möglich sollten diese bereits beim Chipdesign konsequent vermieden werden. Wird auf fertige, kommerzielle Arrays zurückgegriffen, sollten Kreuzhybridisierungen durch Auswahl geeigneter CDFs herausgerechnet werden. Kann der anfängliche Messfehler allerdings abgeschätzt werden, ist es leicht möglich die Abschätzung mitzurechnen und die Fehlerfortpflanzung so nachzuvollziehen (Quackenbush [2002]).

Die von Ronald Aylmer Fisher in seinem berühmt gewordenen Buch „The Design of Experiments“ aus dem Jahr 1935 festgehaltene Grundregel:

> ≪Replication, randomization, and blocking are essential components of any well planned and properly analyzed design. ≫ (Fisher [1935])

gilt uneingeschränkt auch für Microarray- oder RNA-Seq-Experimente (Auer and Doerge [2010]).

4.2 Zur biologischen Varianz

Bei der Auswertung biologischer Experimente spielt die Varianz häufig eine untergeordnete Rolle. In der Regel interessiert sich der Experimentator für den Mittelwertvergleich verschiedener Gruppen. Hierzu werden oft sogenannte Fold-Changes berechnet. Da biologische Daten häufig normalverteilt sind, kommt auch regelmäßig der Student-t-Test zum Einsatz. Tatsächlich geht die Varianz der Gruppen auch in die Berechnung des t-Tests ein, dies ist für den Experimentator mit seiner Auswertesoftware aber häufig nicht ersichtlich.

Auch Best-Practice-Manuals für die Auswertung von Microarray-Experimenten empfehlen obengenannte Vorgehensweise: Preprocessing, Fold-Change, t-Test und dann weitergehende Auswertungen (Choe et al. [2005]). Für mehr als zwei Transkripte kommt dann der multiple t-Test (*pairwise t-test*) mit verschiedenen Korrekturtermen, wie der Bonferroni-Korrektur (Abdi [2007]), Holm-Korrektur (Holm [1979]), Benjamini-Hochberg-Korrektur (Benjamini and Hochberg [1995]) oder Benjamini-Yekutieli-Korektur (Benjamini and Yekutieli [2001]) zum Einsatz.

Im Begriff des ANOVA (Analysis of Variance) (Chambers et al. [1992]) ist zwar der Begriff der Varianz im Titel prominent enthalten, tatsächlich handelt es sich aber ebenfalls um einen statistischen Test zum Erwartungswertvergleich mehrerer Gruppen. Die einfaktorielle ANOVA ist eine Verallgemeinerung des t-Tests für mehr als zwei Gruppen. Für genau zwei abhängige Variablen ist sie äquivalent mit dem t-Test.

Ein eigentlicher Test auf Varianzhomogenität, wie der Levene-Test, Bartlett-Test oder auch der in der Einleitung beschriebene Brown-Forsythe-Test (siehe Seite 18), wird dann zumeist nur verwendet, um die Vorausetzungen für die ANOVA zu prüfen. Diese sind nämlich Varianzhomogenität und Normalverteilung der Zufallsgrößen.

In der Auswertung biologischer Experimente wird Varianz innerhalb der Gruppen zumeist als Ungenauigkeit gewertet. Statt mit Individuen umzugehen, wünschen sich viele Biologen exakt reproduzierbare Ergebnisse und möglichst konforme Messdaten, wie sie in anderen Naturwissenschaften vorkommen. Zum Beispiel verlangt das Robert Koch-Institut von seinen Forschern Anstrengungen,

> ≪... damit ungleiche Einflüsse einzelner Individuen vermieden und Transparenz und Reproduzierbarkeit gewährleistet werden können.≫ (Krause [2008])

Die im Kapitel 3.1 dargestellte Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) im Kapitel „Analyse der biologischen Varianz“ zeigt jedoch, dass tatsächlich auch unterschiedliche Varianzen innerhalb biologischer Gruppen existieren, deren Auswertung lohnt. Varianz innerhalb der Gruppe ist nicht automatisch auf Messfehler zurückzuführen. Es kann hilfreich sein, die Gründe für diese Varianz zu untersuchen.

Für Krebserkrankungen, insbesondere Epitheltumore, ist bekannt, dass die Proteinexpression von Tumorzellen eine geringere Varianz als die von gesunden Zellen aufweist (Müller et al. [2006]).

Im Kapitel 3.1 wird die Frage: „Was ist Krankheit?“ aufgeworfen. Im Fall von rheumatoider Arthritis gibt es viele Studien, die Ursachen auf der Ebene der Transkripte suchen (Hoffmann et al. [2006], Pohlers et al. [2007], Szekanecz et al. [1995]). Alle diese Arbeiten werten die Transkriptomdaten aber nach

eben genanntem Schema aus, also Berechnung des Fold-Changes und Student-t-Test, eventuell gefolgt von weiteren Auswerteschritten.

Diese Arbeiten suchen also Unterschiede in den Erwartungswerten der Transkription. Vereinfacht gesagt wird Krankheit dann als ein „Zuviel von A“ oder ein „Zuwenig von B“ aufgefasst. Mit dem hier verwendeten Ansatz des Varianzvergleichs ließe sich Krankheit dann vereinfacht als ein zu starkes Schwanken von biologischen Größen interpretieren. Der Patient hat heute zu viel A, morgen aber zu wenig.

Wird ein Datensatz jedoch ausschließlich nach dem Standardschema anhand von Erwartungswertvergleichen untersucht, kann eine solche Störung nicht detektiert werden.

Zur Auswertung von Varianzunterschieden zwischen verschiedenen Gruppen definiert der Autor der vorliegenden Arbeit den Varianzfold (siehe Seite 19). Damit werden solche Transkripte identifiziert, für die die Varianzen stark differieren. Durch das Abbilden dieser Gene auf KEGG-Pathways (Ogata et al. [1999]) wird weiterhin gezeigt, dass diese Gene tatsächlich mit rheumatoider Arthritis im Zusammenhang stehen. Damit wird eine neue Auswertemethode für Transkriptomdaten bei einem Vergleich „krank vs. gesund“ dargestellt.

Die Kovarianz und Korrelation sind eng mit der Varianz verwandt. Wie im Kapitel 1.7 beschrieben, lässt sich der Korrelationskoeffizient nicht nur als Zusammenhangsmaß, sondern auch als Ähnlichkeitsmaß oder, nach einer geeigneten Transformation, als Abstandsmaß interpretieren. Selbstverständlich lassen sich biologische Datensätze auch mit Hilfe dieser Maße auswerten.

In der im Kapitel 3.2 angeführten Arbeit *Quantification of growth-defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover* (Ullmann-Zeunert et al. [2013]) wird eine Ähnlichkeitsanalyse an biologischen Daten durchgeführt. Statt Transkriptomdaten werden hier aber Metabolomdaten, also die Konzentrationen verschiedener Metabolite im Organismus, ausgewertet.

Im Detail wird die Stickstoffverteilung innerhalb einer Pflanze untersucht. Die Verteilung des Stickstoffs auf verschiedene Verbindungen, wie Nikotin, Caffeoylputrescin oder Dicaffeoylspermidin, ergibt ein bestimmtes Muster für jedes Individuum. Für die Mutanten und den Wildtyp werden unter jeweils zwei Bedingungen (Kontrolle und unter Einfluss von herbivorem Speichelsekret) die Stickstoffmuster aufgenommen. Nun wird nach Ähnlichkeiten beziehungsweise Unterschieden zwischen den Stickstoffmustern in verschiedenen Pflanzenteile gesucht.

Für diese Aufgabe wird der Korrelationskoeffizient als Abstand zweier Vektoren interpretiert. In der Arbeit wird paarweise Kendalls τ (siehe Seite 23) für alle Paare berechnet und in einer Heatmap dargestellt. So kann sehr schnell erkannt werden, welche Mutanten respektive Bedingungen sich gleichmäßig (konkordant) und welche sich verschieden (diskordant) verhalten. Durch diese Analysen werden Hypothesen über den funktionalen Pathway in der induzierten Abwehr gewonnen.

In diesem Fall ist eine einfache paarweise Berechnung der Kendall-Korrelation als Ähnlichkeitsmaß, aufwändigen Clusterverfahren, wie sie häufig bei ähnlich gelagerten Fragestellungen angewandt werden (Catchpole et al. [2005], Li et al. [2009], Guthke et al. [2007]), vorzuziehen. Tatsächlich wurden in der Vorbereitung der Arbeit verschiedene Clusterverfahren, wie hierarchisches Clustern (Johnson [1967]) oder der k-Median-Algorithmus (MacQueen [1967]), verwendet. Die Ergebnisse bestätigen die Ergebnisse der Korrelationsanalyse, sind aber aufgrund der komplexeren Struktur der Verfahren bei so kleinen Datensätzen viel anfälliger für Artefakte (Yeung et al. [2001]).

Es ist auch denkbar, die Ähnlichkeitsanalyse mit Korrelationskoeffizienten deutlich zu erweitern. Analog zum Varianzfold (siehe Seite 19), der vom Autor dieser Arbeit in der im Kapitel 3.1 angeführten Veröffentlichung *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008])

definiert wird, ließe sich ein Korrelationsfold:

$$Corfold((x,y),(z,m)) = \begin{cases} \rho_{x,y} \geq \rho_{z,m} : \rho_{x,y}/\rho_{z,m} \\ \rho_{x,y} < \rho_{z,m} : \text{-}(\rho_{z,m}/\rho_{x,y}) \end{cases} \tag{4.1}$$

definieren, um Unterschiede zwischen den Korrelationskoeffizienten zweier Gruppenpaare (x, y) und (z, m) quantifizierbar zu machen. In der Praxis wäre häufig der Fall $x = z$ von Interesse, um die Korrelationsunterschiede zwischen einer Referenzgruppe und zunächst einer und dann einer anderen Gruppe zu quantifizieren. So könnten Korrelationsunterschiede deutlicher dargestellt werden und mit statistischen Tests auf ihre Signifikanz überprüft werden.

4.3 Fazit und Ausblick

Die in der Einleitung formulierten Fragestellungen werden in der vorliegenden Arbeit vollständig beantwortet.

Im Kapitel 2.1, in der Arbeit *Creation and comparison of different chip definition files for Affymetrix microarrays* (Hummert et al. [2011]), wird gezeigt, dass die kreuzhybridisierenden Sonden auf den Chips ohne Wert für die Auswertung sind. Werden diese einfach nicht ausgewertet, verschlechtert sich das Ergebnis nicht, obwohl weniger Sonden pro Probeset zur Verfügung stehen. Gleichzeitig wird mit den neuen Chip Definition Files (CDFs) ein Werkzeug zur Verfügung gestellt, um dieses Problem bei der Auswertung zu berücksichtigen.

Als logische Schlussfolgerung dieser Arbeit wird nachfolgend, wie im Kapitel 2.2 *Optimization of a microarray probe design focusing on the minimization of cross-hybridization* (Horn et al. [2011]) dargestellt, ein kreuzhybridisierungsfreies Microarray geschaffen. Der Vergleich dieses neuen Arrays mit einem schon vorhandenem Microarray, für denselben Organismus aber mit möglichen Kreuzhybridisierungen, zeigt, dass Kreuzhybridisierungen die Ergebnisse von Microarray-Experimenten tatsächlich erheblich beeinflussen. Die Varianz der

technischen Replikate nimmt beim kreuzhybridisierungsfreien Array gegenüber dem vorhandenen Array mit möglichen Kreuzhybridisierungen signifikant ab.

In der im Kapitel 3.1 angeführten Arbeit *Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane* (Huber et al. [2008]) wird ein Microarray-Experiment mit Hilfe des neu definierten Varianzfolds und des Brown-Forsythe-Tests (Siehe Seiten 18 und 19) ausgewertet. Es zeigt sich, dass sich zwei Erkrankungen an Varianzunterschieden in der Genexpression differenzieren lassen. Durch Abbildung auf KEGG-Pathways zeigt sich, dass es sich nicht um falsch positive Artefakte handelt, sondern tatsächlich biologisch relevante Transkripte gefunden werden. Bei einer auf Mittelwertvergleichen basierenden Auswertung werden diese Gene teilweise nicht detektiert.

Die Ergebnisse der Auswertung eines Datensatzes mit Hilfe von Kendalls τ als Ähnlichkeitsmaß werden in der Arbeit im Kapitel 3.2, *Quantification of growth-defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover* (Ullmann-Zeunert et al. [2013]), dargestellt. Dabei werden Hypothesen über den funktionalen Pathway in der induzierten Abwehr von wilden Tabakpflanzen (*Nicotiana attenuata*) gewonnen.

In zukünftigen Arbeiten wäre es wünschenswert, weitere Auswertungen mit von der Varianz abgeleiteten Messgrößen durchzuführen. Denkbar sind dabei nicht nur der auf Seite 84 vorgeschlagene Korrelationsfold, sondern auch das Clustern von Varianzen oder die Verwendung alternativer Ähnlichkeits- beziehungsweise Abstandsmaße als Abstandsfunktionen in Clusterverfahren. Der Autor der vorliegenden Arbeit hat sich bereits in der Arbeit *Gene acquisition, duplication and metabolic specification: the evolution of fungal methylisocitrate lyases* (Müller et al. [2011]) mit alternativen Clusterverfahren beschäftigt. In dieser Arbeit werden die Sequenzabstände zwischen den Arten mithilfe der LG-Matrix (Le and Gascuel [2008]) berechnet. Eine Weiterentwicklung der Varianzvergleiche für biologische Daten wird sicher zu neuen Auswertestrategien

und neuen Erkenntnissen in der Bioinformatik führen.

Neben diesen Publikationen hat sich der Autor der vorliegenden Dissertation mit anderen Möglichkeiten der Modellierung biologischer Daten beschäftigt. In der Arbeit *Game theoretical modelling of survival strategies of Candida albicans inside macrophages* (Hummert et al. [2010]) wird die Interaktion von Sporen des humanpathogenen Pilzes *Candida albicans* mit dem menschlichen Immunsystem modelliert.

Darüberhinaus wären weitere Arbeiten zur technischen Varianz bei neuen Verfahren der Genexpressionsmessung, im Besonderen der RNA-Seq-Technik, wünschenswert. Das beim RNA-Seq auftretende *Cross-Alignment* ist dem Effekt der Kreuzhybridisierungen sehr ähnlich (Mardis [2008], Pirovano [2010]). Obwohl der Effekt experimentell nachgewiesen wurde (Valdes et al. [2013]), gibt es bislang nur wenige technische Verfahren mit dem Problem umzugehen. Zu nennen ist hier HAXAT (Homopolymer Aware Cross Alignment Tool), das einen Ansatz bietet (Lysholm [2012]). An dieser Stelle könnte gepüft werden, ob sich aus den hier vorgestellten Techniken für Microarrays neue Methoden für RNA-Seq ableiten lassen.

Mit Hilfe neu entwickelter Techniken wie Taqman-Arrays (Life Technologies [2011]) oder Fluidigm-Chips kann für eine Auswahl von Genen kostengünstig die Genexpression bestimmt werden. Im Gegensatz zu Microarrays und RNA-Seq wird hier aber nicht das komplette Transkriptom gemessen, sondern eine kleinere Anzahl von Transkripten. Gebräuchliche Taqman-Arrays messen beispielsweise 96 Transkripte gleichzeitig.

Eine Reihe von in neuerer Zeit veröffentlichten Algorithmen befassen sich mit der technischen Intergruppenvarianz (den sogenannten Batch-Effekten) (Johnson et al. [2007], Kupfer et al. [2012]). Ziel dieser Algorithmen ist es, zu verhindern, dass ungewollte Gruppen durch Messfehler oder Unterschiede bei der Datenerhebung entstehen. Auch wenn es mit diesen Techniken nicht möglich ist, systematische Fehler, die der Technik immanent sind, zu beheben, handelt es sich hierbei um ein wichtiges Forschungsfeld, das Bezüge zur hier

dargestellten Problematik aufweist.

Trotz neuer Entwicklungen existieren heute im Jahr 2014 nur zwei Technologien, die in der Lage sind das ganze Transkriptom in einem Experiment zu messen. Dies sind Microarrays und RNA-Seq. Die RNA-Seq-Technologie ist fortschrittlicher und wird von vielen Autoren als die Zukunft der Transkriptom-Analyse angesehen. Jay Shendure titelte in seinem in *Nature Methods* erschienenem Aufsatz bereits: ≪The beginning of the end for microarrays?≫ (Shendure [2008]).

Die beim NCBI (National Center for Biotechnology Information) angesiedelte GEO-Datenbank (Gene Expression Omnibus) speichert sowohl von Autoren eingereichte Studienbeschreibungen als auch betreute Datensätze mit Genexpressionsdaten (Barrett et al. [2007]). Es ist möglich, die GEO-Datenbank nach dem Datum der Eintragung (Submission Date) zu durchsuchen und nach Array-Daten respektive RNA-Seq-Daten zu filtern. Tabelle 4.1 zeigt die Anzahl der Datensätze, die seit dem Jahr 2008, als Shendure das Ende der Microarrays vorhersagte, bis heute eingetragen wurden.

Jahr der Eintragung	Anzahl Microarray-Einträge	Anzahl RNA-Seq-Einträge
2008	2797	17
2009	3969	34
2010	4043	136
2011	4771	316
2012	5266	565
2013	5578	1039
2014 (bis 27. März)	1290	301

Tabelle 4.1: Eintragungen zu Microarray-Experimenten beziehungsweise RNA-Seq-Experimenten in der GEO-Datenbank nach Jahr.

Die Zahlen in der Tabelle geben nicht die tatsächliche Anzahl an hybridisierten Microarrays wieder, sondern die Anzahl der eingetragenen Experimente. Die GEO-Datenbank unterscheidet zwischen „Series (GSE)“ und „Datasets (GDS)“. Die Series enthalten die Zusammenfassungen von Experimenten. Die Datasets sind die von GEO aufbereiteten Daten zu den Series, nicht alle eingetragenen Daten werden jedoch zu einem Dataset aufbereitet. Somit ist die

Anzahl der Series aussagekräftiger.[1]

Natürlich sind nicht alle durchgeführten Genexpressionsexperimente in der GEO-Datenbank gespeichert und es gibt auch andere Datenbanken, die Genexpressionsdaten speichern. Dennoch ist leicht zu sehen, dass die Verwendung von Microarrays immer noch zunimmt und die Technologie nicht vor ihrem Ende zu stehen scheint.

Microarrays sind in der Anschaffung günstiger. Sie erlauben eine deutlich schnellere Messung und sind in der Durchführung einfacher (Mockler et al. [2005]). RNA-Sequencing ist genauer, erlaubt *de-novo*-Transkriptomanalyse und weitergehende Untersuchungen wie SNP-Analysen oder Messungen von microRNA (Priebe [2012]).

Viele Autoren sind zu dem Schluss gelangt, dass sich die Verfahren eher gegenseitig ergänzen, als dass das eine das andere verdrängen wird (Valdes et al. [2013], Malone and Oliver [2011], Liu et al. [2007a]).

Bei beiden Technologien muss ein besonderes Augenmerk auf die Datenqualität gelegt werden. Dies gilt sowohl für die Planung und Durchführung von Genexpressionsmessungen als auch für das Preprocessing und die weitere Auswertung der Daten. In jedem Fall ist auf höchste Datenqualität zu achten.

[1] Der Suchstring, mit dem die Zahlen in der Tabelle abgefragt wurden, lautet: `"gse"[Filter] AND "Expression profiling by array"[Filter] AND ("JAHR/01/01"[PDAT] : "JAHR/12/31"[PDAT])` respektive `"gse"[Filter] AND ("JAHR/01/01"[PDAT] : "JAHR/12/31"[PDAT]) AND "Expression profiling by high throughput sequencing"[Filter]`

Literaturverzeichnis

H. Abdi. Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 103–107. SAGE Publications, Thousand Oaks, CA, 2007.

A. M. Abeles and M. H. Pillinger. The role of the synovial fibroblast in rheumatoid arthritis: cartilage destruction and the regulation of matrix metalloproteinases. *Bulletin of the NYU Hospital for Joint Diseases*, 64(1-2):20–24, 2006.

Affymetrix Inc. *Statistical algorithms description document. Whitepaper. Part No. 701137 Rev. 3.* Santa Clara, 2002. `http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`.

Affymetrix Inc. *GeneChip® Customexpress™ array design guide. Part no. 700506 Rev. 4.* Santa Clara, 2003a. `http://www.affymetrix.com/support/technical/other/custom_design_manual.pdf`.

Affymetrix Inc. *Array design and performance of the GeneChip® mouse expression set 430. Technical note. Part No. 701405 Rev. 1.* Santa Clara, 2003b. `http://www.affymetrix.com/support/technical/technotes/mouse430_technote.pdf`.

Affymetrix Inc. *Array design for the GeneChip® Human genome U133 Set. Technical note. Part No. 701133 Rev. 2.* Santa Clara, 2007. `http://www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf`.

Affymetrix Inc. *GeneChip® HT PM Array Plate System for Human, Mouse, and Rat. Data Sheet. Part No. 702733 Rev. 3.* Santa Clara,

2009. `http://media.affymetrix.com/support/technical/datasheets/ht_pm_array_plates_system.pdf`.

D. K. Allen, R. W. Laclair, J. B. Ohlrogge, and Y. Shachar-Hill. Isotope labelling of Rubisco subunits provides in vivo information on subcellular biosynthesis and exchange of amino acids between compartments. *Plant, Cell & Environment*, 35 (7):1232–1244, July 2012.

S. Allmann, R. Halitschke, R. C. Schuurink, and I. T. Baldwin. Oxylipin channelling in Nicotiana attenuata: lipoxygenase 2 supplies substrates for green leaf volatile production. *Plant, Cell & Environment*, 33(12):2028–2040, December 2010.

R. Altman, E. Asch, D. Bloch, G. Bole, D. Borenstein, K. Brandt, W. Christy, T. D. Cooke, R. Greenwald, M. Hochberg, D. Howell, D. Kaplan, W. Koopman, S. Longley III, H. Mankin, D. J. McShane, T. Medsger Jr., R. Meenan, W. Mikkelsen, R. Moskowitz, W. Murphy, B. Rothschild, M. Segal, L. Sokoloff, and F. Wolfe. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. *Arthritis & Rheumatism*, 29 (8):1039–1049, August 1986.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

S. Amu, K. Strömberg, M. Bokarewa, A. Tarkowski, and M. Brisslert. CD25-expressing B-lymphocytes in rheumatic diseases. *Scandinavian Journal of Immunology*, 65(2):182–191, February 2007.

E. Andreakos, S. Sacre, B. M. Foxwell, and M. Feldmann. The toll-like receptor-nuclear factor kappaB pathway in rheumatoid arthritis. *Frontiers in Bioscience*, 10:2478–2488, 2005.

J. H. Anolik, R. Ravikumar, J. Barnard, T. Owen, A. Almudevar, E. C. Milner, C. H. Miller, P. O. Dutcher, J. A. Hadley, and I. Sanz. Cutting edge: anti-tumor necrosis factor therapy in rheumatoid arthritis inhibits memory B lymphocytes

via effects on lymphoid germinal centers and follicular dendritic cell networks. *Journal of Immunology*, 180(2):688–692, January 2008.

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, January 2001.

F. C. Arnett, S. M. Edworthy, D. A. Bloch, D. J. McShane, J. F. Fries, N. S. Cooper, L. A. Healey, S. R. Kaplan, M. H. Liang, H. S. Luthra, T. A. Medsger Jr., D. M. Mitchell, D. H. Neustadt, R. S. Pinals, J. G. Schaller, J. T. Sharp, R. L. Wilder, and G. G. Hunder. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis & Rheumatism*, 31(3): 315–324, March 1988.

P. L. Auer and R. W. Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185:405–416, 2010.

I. T. Baldwin. Inducible nicotine production in native nicotiana as an example of adaptive phenotypic plasticity. *Journal of Chemical Ecology*, 25(1):3–30, January 1999.

I. T. Baldwin and W. Hamilton. Jasmonate-induced responses of Nicotiana sylvestris results in fitness costs due to impaired competitive ability for nitrogen. *Journal of Chemical Ecology*, 26(4):915–952, April 2000.

I. T. Baldwin and T. E. Ohnmeiss. Coordination of photosynthetic and alkaloidal responses to damage in uninducible and inducible Nicotiana sylvestris. *Ecology*, 75(4):1003–1014, June 1994.

I. T. Baldwin, M. J. Karb, and T. E. Ohnmeiss. Allocation of ^{15}N from nitrate to nicotine – production and turnover of a damage-induced mobile defense. *Ecology*, 75(6):1703–1713, September 1994.

I. T. Baldwin, D. Gorham, E. A. Schmelz, C. A. Lewandowski, and G. Y. Lynds. Allocation of nitrogen to an inducible defense and seed production in Nicotiana attenuata. *Oecologia*, 115(4):541–552, July 1998.

M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Research*, 33(18):5914–5923, October 2005.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Research*, 35(Database issue):D760–D765, January 2007.

M. S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London Series A*, 160:268–282, 1937.

F. M. Batliwalla, E. C. Baechler, X. Xiao, W. Li, S. Balasubramanian, H. Khalili, A. Damle, W. A. Ortmann, A. Perrone, A. B. Kantor, P. S. Gulko, M. Kern, R. Furie, T. W. Behrens, and P. K. Gregersen. Peripheral blood gene expression profiling in rheumatoid arthritis. *Genes & Immunity*, 6(5):388–397, August 2005.

F. A. Bazzaz, N. R. Chiariello, P. D. Coley, and L. F. Pitelka. Allocating rescources to reproducation and defense. *Bioscience*, 37(1):58–67, January 1987.

S. Beer, M. Oleszewski, P. Gutwein, C. Geiger, and P. Altevogt. Metalloproteinase-mediated release of the ectodomain of L1 adhesion molecule. *Journal of Cell Science*, 112:2667–2675, August 1999.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testingy. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, March 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, April 2001.

P. Bertone, V. Trifonov, J. S. Rozowsky, F. Schubert, O. Emanuelsson, J. Karro, M. Y. Kao, M. Snyder, and M. Gerstein. Design optimization methods for genomic DNA tiling arrays. *Genome Research*, 16(2):271–281, February 2006.

F. Bidard, S. Imbeaud, N. Reymond, O. Lespinet, P. Silar, C. Clave, H. Delacroix, V. Berteaux-Lecellier, and R. Debuchy. A general framework for optimization of

probes for gene expression microarray and its application to the fungus Podospora anserina. *BMC Research Notes*, 3:171, 2010.

H. Binder and S. Preibisch. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophysical Journal*, 89(1):337–352, July 2005.

O. R. P. Bininda-Emonds, K. E. Jones, S. A. Price, M. Cardillo, R. Grenyer, and A. Purvis. Garbage in, garbage out. In O. R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees*, volume 4 of *Computational Biology*, pages 267–280. Springer, 2004.

Y. Bira, K. Tani, Y. Nishioka, J. Miyata, K. Sato, A. Hayashi, Y. Nakaya, and S. Sone. Transforming growth factor beta stimulates rheumatoid synovial fibroblasts via the type II receptor. *Modern Rheumatology*, 15(2):108–113, 2005.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(3):185–193, January 2003.

P. C. Boutros. Systematic evaluation of the microarray analysis pipeline. In G. Sherlock, editor, *Proceedings of the First 11th MGED Meeting: 1-4 September 2008; Riva del Garda*, pages 16–27. MGED, 2008.

Z. Bozdech, J. Zhu, M. P. Joachimiak, F. E. Cohen, B. Pulliam, and J. L. DeRisi. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biology*, 4(2):R9, 2003.

A. A. Brakhage and V. Schroeckh. Fungal secondary metabolites – strategies to activate silent gene clusters. *Fungal Genetics and Biology*, 48(1):15–22, January 2011.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME) –

toward standards for microarray data. *Nature Genetics*, 29(4):365–371, December 2001.

A. Brazma, K. Ikeo, and Y. Tateno. Standardization of microarray data. *Tanpakushitsu Kakusan Koso*, 48(3):280–285, March 2003. (in japanisch).

M. B. Brown and A. B. Forsythe. Robust tests for equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, April 1974.

A. M. Byrne, D. J. Bouchier-Hayes, and J. H. Harmey. Angiogenic and cell survival functions of vascular endothelial growth factor (VEGF). *Journal of Cellular and Molecular Medicine*, 9(4):777–794, 2005.

A. C. Cambon, A. Khalyfa, N. G. F. Cooper, and C. M. Thompson. Analysis of probe level patterns in Affymetrix microarray data. *BMC Bioinformatics*, 8:146, May 2007.

R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology*, 24(9): 1115–1122, September 2006.

D. A. Casciano and J. Woodcock. Empowering microarrays in the regulatory setting. *Nature Biotechnology*, 24(9):1103, September 2006.

P. Casel, F. Moreews, S. Lagarrigue, and C. Klopp. sigReannot: an oligo-set reannotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters. *BMC Proceedings*, 3 Suppl 4:S3, 2009.

G. S. Catchpole, M. Beckmann, D. P. Enot, M. Mondhe, B. Zywicki, J. Taylor, N. Hardy, A. Smith, R. D. King, D. B. Kell, O. Fiehn, and J. Draper. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14458–14462, 2005.

M. Centola, M. B. Frank, A. I. Bolstad, P. Alex, A. Szanto, M. Zeher, T. O. Hjelmervik, R. Jonsson, B. Nakken, G. Szegedi, and P. Szodoray. Genome-scale assessment of molecular pathology in systemic autoimmune diseases using microarray technology: a potential breakthrough diagnostic and individualized therapy-design tool. *Scandinavian Journal of Immunology*, 64(3):236–242, September 2006.

V. Chalifa-Caspi, I. Yanai, R. Ophir, N. Rosen, M. Shmoish, H. Benjamin-Rodrig, M. Shklar, T. I. Stein, O. Shmueli, M. Safran, and D. Lancet. GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. *Bioinformatics*, 20(9):1457–1458, June 2004.

J. M. Chambers, A. E. Freeny, and R. M. Heiberger. Analysis of variance; designed experiments. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, pages 145–195. Wadsworth & Brooks/Coles, Pacific Grove, 1992.

C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung. Fast network component analysis (fastNCA) for gene regulatory network reconstruction from microarray data. *Bioinformatics*, 24(11):1349–1358, 2008.

F. S. Chapin, E. D. Schulze, and H. A. Mooney. The ecology and economics of storage in plants. *Annual Review of Ecology and Systematics*, 21:423–447, 1990.

Y. A. Chen, C.-C. Chou, X. Lu, E. H. Slate, K. Peck, W. Xu, E. O. Voit, and J. S. Almeida. A multivariate prediction model for microarray cross-hybridization. *BMC Bioinformatics*, 7:101, March 2006.

Z. Chen, M. McGee, Q. Liu, and R. H. Scheuermann. A distribution free summarization method for Affymetrix GeneChip® arrays. *Bioinformatics*, 23(3):321–327, February 2007.

C. Cheng and L. M. Li. Systematic identification of cell cycle regulated transcription factors from microarray time series. *BMC Genomics*, 9:116, March 2008.

S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16, January 2005.

H. H. Chou. Shared probe design and existing microarray reanalysis using PICKY. *BMC Bioinformatics*, 11:196, 2010.

I. Chowers, D. Liu, R. H. Farkas, T. L. Gunatilaka, A. S. Hackam, S. L. Bernstein, P. A. Campochiaro, G. Parmigiani, and D. J. Zack. Gene expression variation in the adult human retina. *Human Molecular Genetics*, 12(22):2881–2893, November 2003.

G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32 Suppl:490–495, December 2002.

W. G. Cochran. The distribution of the lagest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, 11(1):47–52, 1941.

X. Cui and A. E. Loraine. Consistency analysis of redundant probe sets on Affymetrix three-prime expression arrays and applications to differential mRNA processing. *PLoS One*, 4(1):e4229, January 2009.

M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175, November 2005.

D. D. Dalma Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada. The Affymetrix GeneChip platform: an overview. *Methods in Enzymology*, 410:3–28, 2006.

J. M. Dayer. Interleukin 1 or tumor necrosis factor-alpha: which is the real target in rheumatoid arthritis? *Journal of Rheumatology*, 65 Suppl:10–15, September 2002.

F. Dieterle and E. Marrer. New technologies around biomarkers and their interplay with drug development. *Analytical and Bioanalytical Chemistry*, 390(1):141–154, January 2008.

C. Diezel, C. C. von Dahl, E. Gaquerel, and I. T. Baldwin. Different lepidopteran elicitors account for cross-talk in herbivory-induced phytohormone signaling. *Plant Physiology*, 150(3):1576–1586, July 2009.

A. L. Dixon, L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, G. M. Lathrop, G. R. Abecasis, and W. O. C. Cookson. A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–1207, September 2007.

R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30 (1):207–210, January 2002.

B. J. Edwards, C. Haynes, M. A. Levenstien, S. J. Finch, and D. Gordon. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genetics*, 6:18, 2005.

L. Egghe and L. Leydesdorff. The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, 60(5):1027–1036, May 2009.

M. Eisenstein. Microarrays: quality control. *Nature*, 442:1067–1070, August 2006.

Y. Elbez, S. Farkash-Amar, and I. Simon. An analysis of intra array repeats: the good, the bad and the non informative. *BMC Genomics*, 7:136, June 2006.

R. J. Ellis. Most abundant protein in the world. *Trends in Biochemical Sciences*, 4 (11):241–244, November 1979.

European Bioinformatics Institute. HUGO Gene Nomenclature Committee. `http://www.genenames.org`, 2008. zuletzt abgerufen am 15.07.2008.

P. G. Febbo. cDNA microarrays. In J. M. Walker and R. Rapley, editors, *Medical Biomethods Handbook*, pages 255–271. Humana Press, New York, 2005.

F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. A. Danieli, and S. Bicciato. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*, 8:446, November 2007.

G. S. Firestein. Evolving concepts of rheumatoid arthritis. *Nature*, 423(6937):356–361, May 2003.

G. S. Firestein. Immunologic mechanisms in the pathogenesis of rheumatoid arthritis. *Journal of Clinical Rheumatology*, 11(3 Suppl):39–44, June 2005.

G. S. Firestein and A. M. Manning. Signal transduction and transcription factors in rheumatic disease. *Arthritis & Rheumatism*, 42(4):609–621, April 1999.

G. S. Firestein, J. M. Alvaro-Gracia, and R. Maki. Quantitative analysis of cytokine gene expression in rheumatoid arthritis. *Journal of Immunology*, 144(9):3347–3353, May 1990.

R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, January 1919.

R. A. Fisher. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85:597–612, 1922.

R. A. Fisher. *The Design of Experiments*. Oliver & Boyd, Edinburgh, 2 edition, 1935.

R. A. Fisher. *Statistical Methods and Scientific Inference*. Hafner Press, New York, 1973.

O. Førre, K. Waalen, J. B. Natvig, and J. Kjeldsen-Kragh. Evidence for activation of rheumatoid synovial T lymphocytes – development of rheumatoid T cell clones. *Scandinavian Journal of Rheumatology*, 76 Suppl:153–160, 1988.

S. Fox, S. Filichkin, and T. C. Mockler. Applications of ultra-high-throughput sequencing. *Methods in Molecular Biology*, 553:79–108, 2009.

C. J. Frost and M. D. Hunter. Herbivore-induced shifts in carbon and nitrogen allocation in red oak seedlings. *New Phytologist*, 178(4):835–845, 2008.

J. E. Galagan, S. E. Calvo, C. Cuomo, L. J. Ma, J. R. Wortman, S. Batzoglou, S. I. Lee, M. Baştürkmen, C. C. Spevak, J. Clutterbuck, V. Kapitonov, J. Jurka, C. Scazzocchio, M. Farman, J. Butler, S. Purcell, S. Harris, G. H. Braus, O. Draht, S. Busch, C. D'Enfert, C. Bouchier, G. H. Goldman, D. Bell-Pedersen,

S. Griffiths-Jones, J. H. Doonan, J. Yu, K. Vienken, A. Pain, M. Freitag, E. U. Selker, D. B. Archer, M. A. Penalva, B. R. Oakley, M. Momany, T. Tanaka, T. Kumagai, K. Asai, M. Machida, W. C. Nierman, D. W. Denning, M. Caddick, M. Hynes, M. Paoletti, R. Fischer, B. Miller, P. Dyer, M. S. Sachs, S. A. Osmani, and B. W. Birren. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature*, 438(7071):1105–1115, December 2005.

E. Gaquerel, S. Heiling, M. Schöttner, G. Zurek, and I. T. Baldwin. Development and validation of a liquid chromatography-electrospray ionization-time-of-flight mass spectrometry method for induced changes in Nicotiana attenuata leaves during simulated herbivory. *Journal of Agricultural and Food Chemistry*, 58(17): 9418–9427, September 2010.

U. Gaur and B. B. Aggarwal. Regulation of proliferation, survival and apoptosis by members of the TNF superfamily. *Biochemical Pharmacology*, 66(8):1403–1408, October 2003.

A. P. Giri, H. Wünsche, S. Mitra, J. A. Zavala, A. Muck, A. Svatoš, and I. T. Baldwin. Molecular interactions between the specialist herbivore Manduca sexta (Lepidoptera, Sphingidae) and its natural host Nicotiana attenuata. VII. Changes in the plant's proteome. *Plant Physiology*, 142(4):1621–1641, December 2006.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

S. Gómez, R. A. Ferrieri, M. Schueller, and C. M. Orians. Methyl jasmonate elicits rapid changes in carbon and nitrogen dynamics in tomato. *New Phytologist*, 188 (3):835–844, November 2010.

S. Gómez, A. D. Steinbrenner, S. Osorio, M. Schueller, R. A. Ferrieri, A. R. Fernie, and C. M. Orians. From shoots to roots: transport and metabolic changes in tomato after simulated feeding by a specialist lepidopteran. *Entomologia Experimentalis et Applicata*, 144(1):101–111, July 2012.

M. Grabe. *Measurement Uncertainties in Science and Technology.* Springer, Berlin, 2005.

S. Gräf, F. G. Nielsen, S. Kurtz, M. A. Huynen, E. Birney, H. Stunnenberg, and P. Flicek. Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, 23(13):195–204, July 2007.

W. Grassi, R. De Angelis, G. Lamanna, and C. Cervini. The clinical features of rheumatoid arthritis. *European Journal of Radiology*, 27(Suppl 1):18–24, May 1998.

Q.-M. Guo. DNA microarray and cancer. *Current Opinion in Oncology*, 15(1):36–43, January 2003.

R. Guthke, U. Möller, M. Hoffmann, and F. Thies. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8):1626–1634, 2005.

R. Guthke, O. Kniemeyer, D. Albrecht, A. A. Brakhage, and U. Möller. Discovery of gene regulatory networks in Aspergillus fumigatus. In K. Tuyls, R. Westra, Y. Saeys, and A. Nowé, editors, *Knowledge Discovery and Emergent Complexity in Bioinformatics*, volume 4366 of *Lecture Notes in Computer Science*, pages 22–41. Springer, Berlin, 2007.

C. S. Haas, M. A. Amin, J. H. Ruth, B. L. Allen, S. Ahmed, A. Pakozdi, J. M. Woods, S. Shahrara, and A. E. Koch. In vivo inhibition of angiogenesis by interleukin-13 gene therapy in a rat model of rheumatoid arthritis. *Arthritis & Rheumatism*, 56 (8):2535–2548, August 2007.

H. Hache. *Computational Analysis of Gene Regulatory Networks.* PhD thesis, Humboldt-Universität zu Berlin, June 2009.

G. Hahn, B. Stuhlmüller, N. Hain, J. R. Kalden, K. Pfizenmaier, and G. R. Burmester. Modulation of monocyte activation in patients with rheumatoid arthritis by leukapheresis therapy. *Journal of Clinical Investigation*, 91(3):862–870, March 1993.

R. Halitschke, K. Gase, D. Hui, D. D. Schmidt, and I. T. Baldwin. Molecular interactions between the specialist herbivore Manduca sexta (Lepidoptera, Sphingidae) and its natural host Nicotiana attenuata. VI. Microarray analysis reveals that most herbivore-specific transcriptional changes are mediated by fatty acid-amino acid conjugates. *Plant Physiology*, 131(4):1894–1902, April 2003.

R. Halitschke, J. Ziegler, M. Keinänen, and I. T. Baldwin. Silencing of hydroperoxide lyase and allene oxide synthase reveals substrate and defense signaling crosstalk in Nicotiana attenuata. *The Plant Journal*, 40(1):35–46, October 2004.

A. Halperin, A. Buhot, and E. B. Zhulina. On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. *Journal of Physics: Condensed Matter*, 18(18):S463–S490, April 2006.

D. R. Hammaker, D. L. Boyle, M. Chabaud-Riou, and G. S. Firestein. Regulation of c-Jun N-terminal kinase by MEKK-2 and mitogen-activated protein kinase kinase kinases in rheumatoid arthritis. *Journal of Immunology*, 172(3):1612–1618, February 2004.

D. R. Hammaker, D. L. Boyle, T. Inoue, and G. S. Firestein. Regulation of the JNK pathway by TGF-beta activated kinase 1 in rheumatoid arthritis synoviocytes. *Arthritis Research & Therapy*, 9(3):R57, 2007.

S. W. Han, G. W. Kim, J. S. Seo, S. J. Kim, K. H. Sa, J. Y. Park, J. Lee, S. Y. Kim, J. J. Goronzy, C. M. Weyand, and Y. M. Kang. VEGF gene polymorphisms and susceptibility to rheumatoid arthritis. *Rheumatology*, 43(9):1173–1177, September 2004.

Z. Han, D. L. Boyle, A. M. Manning, and G. S. Firestein. AP-1 and NF-kappaB regulation in rheumatoid arthritis and murine collagen-induced arthritis. *Autoimmunity*, 28(4):197–208, 1998.

Z. Han, D. L. Boyle, K. R. Aupperle, B. Bennett, A. M. Manning, and G. S. Firestein. Jun N-terminal kinase in rheumatoid arthritis. *Journal of Pharmacology and Experimental Therapeutics*, 291(1):124–130, October 1999.

L. Hanriot, C. Keime, N. Gay, C. Faure, C. Dossat, P. Wincker, C. Scoté-Blachon, C. Peyron, and O. Gandrillon. A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. *BMC Genomics*, 9:418, September 2008.

J. Harbig, R. Sprinkle, and S. A. Enkemann. A sequence-based identification of the genes detected by probesets on Affymetrix U133 plus 2.0 array. *Nucleic Acids Research*, 33(3):e31, January 2005.

G. Hardiman. Microarrays technologies 2006: an overview. *Pharmacogenomics*, 7 (8):1153–1158, December 2006.

A. P. Harrison, C. E. Johnston, and C. A. Orengo. Establishing a major cause of discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinformatics*, 8: 195, June 2007.

N. Haslam, N. Whiteford, G. Weber, J. Essex, A. Prügel-Bennett, and C. Neylon. Estimating the likelihood of cross-hybridisation in DNA hybridisation. In C. Falter, B. Regierer, J. Selbig, M. Vingron, and D. Walther, editors, *German Conference on Bioinformatics GCB 2007 Short Papers and Abstracts*, page 43, September 2007.

T. Häupl, M. Yahyawi, C. Lübke, J. Ringe, T. Rohrlach, G. R. Burmester, M. Sittinger, and C. Kaps. Gene expression profiling of rheumatoid arthritis synovial cells treated with antirheumatic drugs. *Journal of Biomolecular Screening*, 12(3): 328–340, April 2007.

A. J. Hayes. Angioneogenesis in rheumatoid arthritis. *Lancet*, 354(9176):423–424, July 1999.

Z. He, L. Wu, X. Li, M. W. Fields, and J. Zhou. Empirical establishment of oligonucleotide probe design criteria. *Applied and Environmental Microbiology*, 71(7): 3753–3760, July 2005.

S. Heber and B. Sick. Quality assessment of Affymetrix GeneChip data. *OMICS A Journal of Integrative Biology*, 10(3):358–368, Fall 2006.

M. Heil and I. T. Baldwin. Fitness costs of induced resistance: emerging experimental support for a slippery concept. *Trends in Plant Science*, 7(2):61–67, February 2002.

D. A. Herms and W. J. Mattson. The dilemma of plants – to grow or defend. *The Quarterly Review of Biology*, 67(3):283–335, September 1992.

N. Hibi, S. Higashiguchi, T. Hashimoto, and Y. Yamada. Gene expression in tobacco low-nicotine mutants. *The Plant Cell*, 6(5):723–735, May 1994.

S. Hirohata and J. Sakakibara. Angioneogenesis as a possible elusive triggering factor in rheumatoid arthritis. *Lancet*, 353(9161):1331, April 1999.

M. Hoffmann, D. Pohlers, D. Koczan, H.-J. Thiesen, S. Wölfl, and R. W. Kinne. Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics*, 7: 369, 2006.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, September 1979.

F. Horn, H.-W. Nützmann, V. Schroeck, R. Guthke, and C. Hummert. Optimization of a microarray probe design focusing on the minimization of cross-hybridization. In H. R. Arabnia and Q.-N. Tran, editors, *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11), Vol. I*, pages 3–9. CSREA Press, July 2011.

L. C. Huber, O. Distler, I. Tarner, R. E. Gay, S. Gay, and T. Pap. Synovial fibroblasts: key players in rheumatoid arthritis. *Rheumatology*, 45(6):669–675, June 2006.

R. Huber, E. Kunisch, B. Glück, R. Egerer, S. Sickinger, and R. W. Kinne. Comparison of conventional and real-time RT-PCR for the quantitation of jun protooncogene mRNA and analysis of junB mRNA expression in synovial membranes and isolated synovial fibroblasts from rheumatoid arthritis patients. *Zeitschrift für Rheumatologie*, 62(4):378–389, August 2003.

R. Huber, C. Hummert, U. Gausmann, D. Pohler, D. Koczan, R. Guthke, and R. W. Kinne. Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. *Arthritis Research & Therapy*, 10(4):R98, August 2008.

T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephaniants, G. Cavet, W. L. Walker, A. West, E. Coffey, D. D. Shoemaker, R. Stoughton, A. P. Blanchard, S. H. Friend, and P. S. Linsley. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, 19(4):342–347, April 2001.

C. Hummert, F. Mech, F. Horn, M. Weber, S. Drynda, U. Gausmann, and R. Guthke. Creation and comparison of different chip definition files for Affymetrix microarrays. In H. R. Arabnia and Q.-N. Tran, editors, *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCOMP'11), Vol. I*, pages 16–22. CSREA Press, July 2011.

S. Hummert, C. Hummert, A. Schröter, B. Hube, and S. Schuster. Game theoretical modelling of survival strategies of Candida albicans inside macrophages. *Journal of Theoretical Biology*, 264:312–318, 2010.

K. Imai, Y. Suzuki, T. Mae, and A. Makino. Changes in the synthesis of Rubisco in rice leaves in relation to senescence and N influx. *Annals of Botany*, 101(1): 135–144, January 2008.

S. Imamura, Y. Kanesaki, M. Ohnuma, T. Inouye, Y. Sekine, T. Fujiwara, T. Kuroiwa, and K. Tanaka. R2R3-type MYB transcription factor, CmMYB1, is a central nitrogen assimilation regulator in Cyanidioschyzon merolae. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12548–12553, July 2009.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003.

R. A. Irizarry, Z. Wu, and H. A. Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7):789–794, July 2006.

A.-K. Järvinen, S. Hautaniemi, H. Edgren, P. Auvinen, J. Saarela, O.-P. Kallioniemo, and O. Monni. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168, June 2004.

C. Jochum, P. Jochum, and B. R. Kowalski. Error propagation and optimal performance in multicomponent analysis. *Analytical Chemistry*, 53(1):85–92, January 1981.

S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241–254, 1967.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

L. Jourdren, A. Duclos, C. Brion, T. Portnoy, H. Mathis, A. Margeot, and S. Le Crom. Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments. *Nucleic Acids Research*, 38(10):e117, June 2010.

M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, 28(22):4552–4557, November 2000.

R. Karban and I. T. Baldwin. *Induced Responses to Herbivory*. University of Chicago Press, Chicago, 1997.

E. Karouzakis, M. Neidhart, R. E. Gay, and S. Gay. Molecular and cellular basis of rheumatoid joint destruction. *Immunology Letters*, 106(1):8–13, July 2006.

H. Kaur, N. Heinzel, M. Schöttner, I. T. Baldwin, and I. Gális. R2R3-NaMYB8 regulates the accumulation of phenylpropanoid-polyamine conjugates, which are essential for local and systemic defense against insect herbivores in Nicotiana attenuata. *Plant Physiology*, 152(3):1731–1747, March 2010.

A. Kawakami, S. Urayama, S. Yamasaki, A. Hida, T. Miyashita, M. Kamachi, K. Nakashima, F. Tanaka, H. Ida, Y. Kawabe, T. Aoyagi, I. Furuichi, K. Migita, T. Origuchi, and K. Eguchi. Anti-apoptogenic function of TGFbeta1 for human synovial cells: TGFbeta1 protects cultured synovial cells from mitochondrial perturbation induced by several apoptogenic stimuli. *Annals of the Rheumatic Diseases*, 63(1): 95–97, January 2004.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, June 1938.

A. Kessler, R. Halitschke, and I. T. Baldwin. Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science*, 305(5684):665–668, July 2004.

E. Keystone. B cell targeted therapies. *Arthritis Research & Therapy*, 7 Suppl 3: S13–S18, 2005.

S. Y. Kim, S. W. Han, G. W. Kim, J. M. Lee, and Y. M. Kang. TGF-beta1 polymorphism determines the progression of joint damage in rheumatoid arthritis. *Scandinavian Journal of Rheumatology*, 33(6):389–394, 2004.

R. W. Kinne, E. Palombo-Kinne, and F. Emmrich. Activation of synovial fibroblasts in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 54(6):501–504, June 1995.

R. W. Kinne, T. Liehr, V. Beensen, E. Kunisch, T. Zimmermann, H. Holland, R. Pfeiffer, H. D. Stahl, W. Lungershausen, G. Hein, A. Roth, F. Emmrich, U. Claussen, and U. G. Froster. Mosaic chromosomal aberrations in synovial fibroblasts of patients with rheumatoid arthritis, osteoarthritis, and other inflammatory joint diseases. *Arthritis Research*, 3(5):319–330, 2001.

R. W. Kinne, E. Kunisch, V. Beensen, T. Zimmermann, F. Emmrich, P. Petrow, W. Lungershausen, G. Hein, R. K. Braun, M. Förster, C. Krögel, R. Winter, E. Liesaus, R. A. Fuhrmann, A. Roth, U. Claussen, and T. Liehr. Synovial fibroblasts and synovial macrophages from patients with rheumatoid arthritis and

other inflammatory joint diseases show chromosomal aberrations. *Genes, Chromosomes and Cancer*, 38(1):53–67, September 2003.

H. Kiyosawa, N. Mise, S. Iwase, Y. Hayashizaki, and K. Abe. Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Research*, 15(4):463–474, April 2005.

E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, Weinheim, 2005.

D. Koczan, S. Drynda, M. Hecker, R. Guthke, J. Kekow, and H.-J. Thiessen. Molecular discrimination of responders and nonresponders to anti-TNFalpha in rheumatoid arthritis therapy by Etanercept. *Arthritis Research & Therapy*, 10:R50, May 2008.

R. T. Koehler and N. Peyret. Effects of DNA secondary structure on oligonucleotide probe binding efficiency. *Computational Biology and Chemistry*, 29(6):393–397, December 2005.

G. Kollias, E. Douni, G. Kassiotis, and D. Kontoyiannis. The function of tumour necrosis factor and receptors in models of multi-organ inflammation, rheumatoid arthritis, multiple sclerosis and inflammatory bowel disease. *Annals of the Rheumatic Diseases*, 58 Suppl 1:I32–39, November 1999.

G. Krause. Zur Priorisierung von Infektionskrankheiten im öGD. *Epidemiologisches Bulletin*, (40):343–347, October 2008.

T. Krügel, M. Lim, K. Gase, R. Halitschke, and I. T. Baldwin. Agrobacterium-mediated transformation of Nicotiana attenuata, a model ecological expression system. *Chemoecology*, 12(4):177–183, November 2002.

W. P. Kuo, T.-K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–12, March 2002.

P. Kupfer, R. Guthke, D. Pohlers, R. Huber, D. Koczan, and R. W. Kinne. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC Medical Genomics*, 5:23, June 2012.

U. Kurzik-Dumke, C. Schick, R. Rzepka, and I. Melchers. Overexpression of human homologs of the bacterial DnaJ chaperone in the synovial tissue of patients with rheumatoid arthritis. *Arthritis & Rheumatism*, 42(2):210–220, February 1999.

S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320, 2008.

N. Le Novère. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*, 17(12):1226–1227, December 2001.

I. Lee, A. A. Dombkowski, and B. D. Athey. Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Research*, 32(2):681–690, 2004.

D. L. Leiske, A. Karimpour-Fard, P. S. Hume, B. D. Fairbanks, and R. T. Gill. A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray. *BMC Genomics*, 7:72, 2006.

J. Leming, L. Ellington, and K. Porter-Magee. *Where Did Social Studies Go Wrong?*, chapter 6: Garbage In, Garbage Out, pages 111–123. Thomas B. Fordham Foundation, Washington, D.C., January 2003.

S. Lemoine, F. Combes, and S. Le Crom. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Research*, 37(6): 1726–1739, April 2009.

N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

H. Levene. Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, editors, *Contributions to Probability and*

Statistics: Essays in Honor of Harold Hotelling, pages 278–292. Stanford University Press, Palo Alto, CA, 1960.

H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, May 2010.

X. Li, X. Lu, J. Tian, P. Gao, H. Kong, and G. Xu. Application of fuzzy c-means clustering in data analysis of metabolomics. *Analytical Chemistry*, 81(11):4468–4475, 2009.

B. Liagre, P. Vergne-Salle, D. Y. Leger, and J. L. Beneytout. Inhibition of human rheumatoid arthritis synovial cell survival by hecogenin and tigogenin is associated with increased apoptosis, p38 mitogen-activated protein kinase activity and upregulation of cyclooxygenase-2. *International Journal of Molecular Medicine*, 20(4):451–460, October 2007.

Life Technologies. *TaqMan® Array Plates. User Guide. Part No. 4391016 Rev. F.* Carlsbad, 2011. `http://tools.lifetechnologies.com/content/sfs/manuals/cms_053406.pdf`.

F. Liu, T.-K. Jenssen, J. Trimarchi, C. Punzo, C. L. Cepko, L. Ohno-Machado, E. Hovig, and W. P. Kuo. Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC Genomics*, 8: 153, June 2007a.

H.-C. Liu, C.-Y. Chen, Y.-T. Liu, C.-B. Chu, D.-C. Liang, L.-Y. Shih, and C.-J. Lin. Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods. *Journal of Biomedical Informatics*, 41(4):570–579, August 2008.

X. Liu and M. Rattray. Including probe-level measurement error in robust mixture clustering of replicated microarray gene expression. *Statistical Applications in Genetics and Molecular Biology*, 9(1), December 2010.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22 (17):2107–2113, September 2006.

X. Liu, K. K. Lin, B. Andersen, and M. Rattray. Including probe-level uncertainty in model-based gene expression clustering. *BMC Bioinformatics*, 8:89, March 2007b.

Y. Lou and I. T. Baldwin. Nitrogen supply influences herbivore-induced direct and indirect defenses and transcriptional responses in Nicotiana attenuata. *Plant Physiology*, 135(1):496–506, May 2004.

T. Love and A. Carriquiry. Repeated measurements on distinct scales with censoring – a bayesian approach applied to microarray analysis of maize. *Journal of the American Statistical Association*, 104(486):524–540, June 2009.

J. Lu, J. C. Lee, M. L. Salit, and M. C. Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, March 2007.

G. Y. Lynds and I. T. Baldwin. Fire, nitrogen, and defensive plasticity in Nicotiana attenuata. *Oecologia*, 115(4):531–540, July 1998.

F. Lysholm. Highly improved homopolymer aware nucleotide-protein alignments with 454 data. *BMC Bioinformatics*, 13:230, September 2012.

J. E. Mabey, M. J. Anderson, P. F. Giles, C. J. Miller, T. K. Attwood, N. W. Paton, E. Bornberg-Bauer, G. D. Robson, S. G. Oliver, and D. W. Denning. CADRE: the Central Aspergillus Data REpository. *Nucleic Acids Research*, 32(Database issue):D401–D405, January 2004.

D. W. MacGlashan. Relationship between spleen tyrosine kinase and phosphatidylinositol 5' phosphatase expression and secretion from human basophils in the general population. *Journal of Allergy and Clinical Immunology*, 119(3):626–633, March 2007.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, 1967. University of California Press.

A. Makino, T. Mae, and K. Ohira. Relation between nitrogen and ribulose-1,5-bisphosphate carboxylase in rice leaves from emergence through senescence. *Plant and Cell Physiology*, 25(3):429–437, 1984.

A. Makino, M. Harada, K. Kaneko, T. Mae, T. Shimada, and Y. Naoki. Whole-plant growth and N allocation in transgenic rice plants with decreased content of ribulose-1,5-bisphosphate carboxylase under different CO_2 partial pressures. *Australian Journal of Plant Physiology*, 27(1):1–12, 2000.

C. J. Malemud. Growth hormone, VEGF and FGF: involvement in rheumatoid arthritis. *Clinica Chimica Acta*, 375(1-2):10–19, January 2007.

J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9:34, May 2011.

E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, February 2008.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

M. Marseguerra, E. Zio, L. Podofillini, and D. W. Coit. Optimal design of reliable network systems in presence of uncertainty. *IEEE Transactions on Reliability*, 54 (2):243–253, June 2005.

A. Martínez, A. Valdivia, D. Pascual-Salcedo, A. Balsa, B. Fernández-Gutiérrez, E. G. de la Concha, and E. Urcelay. Role of SLC22A4, SLC22A5, and RUNX1 genes in rheumatoid arthritis. *Journal of Rheumatology*, 33(5):842–846, May 2006.

N. Maruotti, F. P. Cantatore, E. Crivellato, A. Vacca, and D. Ribatti. Angiogenesis in rheumatoid arthritis. *Histology and Histopathology*, 21(5):557–566, May 2006.

P. Matt, A. Krapp, V. Haake, H. P. Mock, and M. Stitt. Decreased Rubisco activity leads to dramatic changes of nitrate metabolism, amino acid metabolism and the levels of phenylpropanoids and nicotine in tobacco antisense RBCS transformants. *The Plant Journal*, 30(6):663–677, June 2002.

S. S. McCachren. Expression of metalloproteinases and metalloproteinase inhibitor in human arthritic synovium. *Arthritis & Rheumatism*, 34(9):1085–1093, September 1991.

E. S. McCloud and I. T. Baldwin. Herbivory and caterpillar regurgitants amplify the wound-induced increases in jasmonic acid but not nicotine in Nicotiana sylvestris. *Planta*, 203(4):403–435, November 1997.

S. McGinnis and T. L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32:W20–W25, July 2004.

D. McKey. Adaptive patterns in alkaloid physiology. *The American Naturalist*, 108 (961):305–320, June 1974.

D. McKey. The distribution of secondary compounds within plants. In G. A. Rosenthal and D. H. Janzen, editors, *Herbivores: Their Interaction with Secondary Plant Metabolites*, pages 55–133. Academic Press, New York, 1979.

S. Meldau, L. Ullmann-Zeunert, G. Govind, S. Bartram, and I. T. Baldwin. Basal and herbivory-induced defense trade-offs are mediated by mitogen-activated protein kinases, jasmonic acid and salicylic acid in the native tobacco, Nicotiana attenuata. *BMC Plant Biology*, 12:213, November 2012.

J. Middleton, L. Americh, R. Gayon, D. Julien, L. Aguilar, F. Amalric, and J. P. Girard. Endothelial cell phenotypes in the rheumatoid synovium: activated, angiogenic, apoptotic and leaky. *Arthritis Research & Therapy*, 6(2):60–72, 2004.

J. Mieczkowski, M. E. Tyburczy, M. Dabrowski, and P. Pokarowski. Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. *BMC Bioinformatics*, 11:104, 2010.

P. Millard. The accumulation and storage of nitrogen by herbaceous plants. *Plant, Cell & Environment*, 11(1):1–8, January 1988.

K. Miyake, T. Ito, M. Senda, R. Ishikawa, T. Harada, M. Niizeki, and S. Akada. Isolation of a subfamily of genes for R2R3-MYB transcription factors showing up-

regulated expression under nitrogen nutrient-limited conditions. *Plant Molecular Biology*, 53(1-2):237–245, September 2003.

T. C. Mockler, S. Chan, A. Sundaresan, H. Chen, S. E. Jacobsen, and J. R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85(1): 1–15, May 2005.

O. Modlich and M. Munnes. Statistical framework for gene expression data analysis. In M. J. Korenberg, editor, *Microarray Data Analysis. Methods and Applications*, volume 377 of *Methods in Molecular Biology*, chapter 6, pages 111–130. Humana Press, Totowa, May 2007.

S. Mole. Trade-offs and constraints in plant-herbivore defense theory – a life-history perspective. *Oikos*, 71(1):3–12, October 1994.

J. S. Morey, J. C. Ryan, and F. M. Van Dolah. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biological Procedures Online*, 8(1):175–193, December 2006.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5:621–628, May 2008.

S. Müller, C. B. Fleck, D. Wilson, C. Hummert, B. Hube, and M. Brock. Gene acquisition, duplication and metabolic specification: the evolution of fungal methylisocitrate lyases. *Environmental Microbiology*, 13(6):1534–1548, 2011.

S. Müller, C. Baldin, M. Groth, R. Guthke, O. Kniemeyer, A. A. Brakhage, and V. Valiante. Comparison of transcriptome technologies in the pathogenic fungus Aspergillus fumigatus reveals novel insights into the genome and MpkA dependent gene expression. *BMC Genomics*, 13:519, October 2012.

U. Müller, G. Ernst, C. Melle, R. Guthke, and F. von Eggeling. Convergence of the proteomic pattern in cancer. *Bioinformatics*, 22(11):1293–1296, March 2006.

Y. Nakamura, M. Nawata, and S. Wakitani. Expression profiles and functional

analyses of Wnt-related genes in human joint disorders. *American Journal of Pathology*, 167(1):97–105, July 2005.

National Center for Biotechnology Information. Gene Expression Omnibus. `http://www.ncbi.nlm.nih.gov/geo`, 2008. zuletzt abgerufen am 15.07.2008.

P. B. Neerincx, H. Rauwerda, H. Nie, M. A. Groenen, T. M. Breit, and J. A. Leunissen. OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity. *BMC Proceedings*, 3 Suppl 4:S4, 2009.

D. H. Nguyen and P. D'haeseleer. Deciphering principles of transcription regulation in eukaryotic genomes. *Molecular Systems Biology*, 2(1), April 2006.

D. S. Nuyten and M. J. van de Vijver. Using microarray analysis as a prognostic and predictive tool in ocology: focus on breast cancer and normal tissue toxity. *Seminars in Radiation Oncology*, 18(2):105–114, April 2008.

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, January 1999.

T. E. Ohnmeiss and I. T. Baldwin. Optimal defense theory predicts the ontogeny of an induced nicotine defense. *Ecology*, 81(7):1765–1783, July 2000.

N. Ohtake, T. Sato, H. Fujikake, K. Sueyoshi, T. Ohyama, N. S. Ishioka, S. Watanabe, A. Osa, T. Sekine, S. Matsuhashi, T. Ito, C. Mizuniwa, T. Kume, S. Hashimoto, H. Uchida, and A. Tsuji. Rapid N transport to pods and seeds in N-deficient soybean plants. *Journal of Experimental Botany*, 52(355):277–283, February 2001.

N. Onkokesung, I. Gális, C. C. von Dahl, K. Matsuoka, H.-P. Saluz, and I. T. Baldwin. Jasmonic acid and ethylene modulate local responses to wounding and simulated herbivory in Nicotiana attenuata leaves. *Plant Physiology*, 153(2):785–798, June 2010.

N. Onkokesung, E. Gaquerel, H. Kotkar, H. Kaur, I. T. Baldwin, and I. Galis. MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-

coenzyme A: polyamine transferases in Nicotiana attenuata. *Plant Physiology*, 158 (1):389–407, January 2012.

N. Ono, S. Suzuki, C. Furusawa, T. Agata, A. Kashiwagi, H. Shimizu, and T. Yomo. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics*, 24(10):1278–1285, May 2008.

C. M. Orians, A. Thorn, and S. Gómez. Herbivore-induced resource sequestration in plants: why bother? *Oecologia*, 167(1):1–9, September 2011.

C. Ospelt and S. Gay. Antirheumatic drugs and gene signatures. *Current Opinion in Investigational Drugs*, 8(5):385–389, May 2007.

C. Ospelt, M. Neidhart, R. E. Gay, and S. Gay. Synovial activation in rheumatoid arthritis. *Frontiers in Bioscience*, 9:2323–2334, September 2004.

A. Paoloni-Giacobino, R. Grimble, and C. Pichard. Genetics and nutrition. *Clinical Nutrition*, 22(5):429–435, October 2003.

A. Pardo and M. Selman. MMP-1: the elder of the family. *International Journal of Biochemistry & Cell Biology*, 37(2):283–288, February 2005.

H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Zheng-Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–D872, January 2009.

K. Paszkiewicz and D. J. Studholme. De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11(5):457–472, 2010.

K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, June 1895.

W. R. Pearson. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, 132:185–219, 2000.

S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller. The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics*, 8:273, July 2007.

N. Philippe, A. Boureux, L. Bréhélin, J. Tarhio, T. Commes, and É. Rivals. Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Research*, 37(15):e104, June 2009.

M. Pierer, J. Rethage, R. Seibl, R. Lauener, F. Brentano, U. Wagner, H. Hantzschel, B. A. Michel, R. E. Gay, S. Gay, and D. Kyburz. Chemokine secretion of rheumatoid arthritis synovial fibroblasts stimulated by Toll-like receptor 2 ligands. *Journal of Immunology*, 172(2):1256–1265, January 2004.

W. A. Pirovano. *Comparing building blocks of life: sequence alignment and evaluation of predicted structural and functional features.* PhD thesis, Vrije Universiteit Amsterdam, January 2010.

W. E. Pluskota, N. Qu, M. Maitrejean, W. Boland, and I. T. Baldwin. Jasmonates and its mimics differentially elicit systemic defence responses in Nicotiana attenuata. *Journal of Experimental Botany*, 58(15-16):4071–4082, 2007.

D. Pohlers, A. Beyer, D. Koczan, T. Wilhelm, H. J. Thiesen, and R. W. Kinne. Constitutive upregulation of the transforming growth factor-beta pathway in rheumatoid arthritis synovial fibroblasts. *Arthritis Research & Therapy*, 9(3):R59, 2007.

V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms for Molecular Biology*, 6:25, 2011.

J. U. Pontius, L. Wagner, and G. D. Schuler. UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. National Center for Biotechnology Information, Bethesda, 2003.

A. E. Postlethwaite, M. A. Holness, H. Katai, and R. Raghow. Human fibroblasts synthesize elevated levels of extracellular matrix proteins in response to interleukin 4. *Journal of Clinical Investigation*, 90(4):1479–1485, October 1992.

C. A. Preston and I. T. Baldwin. Positive and negative signals regulate germination in the post-fire annual, Nicotiana attenuata. *Ecology*, 80(2):481–494, March 1999.

S. Priebe. *Transcriptome analysis of multiple species using novel sequencing technologies*. PhD thesis, Friedrich-Schiller-Universität Jena, 2012.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33:D501–D504, January 2005.

E. Purdom and S. P. Holmes. Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), July 2005.

J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32 Suppl:496–501, Dezember 2002.

R Foundation for Statistical Computing. R: a language and environment for statistical computing. `http://www.r-project.org`, 2013. zuletzt abgerufen am 10.01.2013.

F. Rannou, M. François, M. T. Corvol, and F. Berenbaum. Cartilage breakdown in rheumatoid arthritis. *Joint Bone Spine*, 73(1):29–36, January 2006.

S. K. Rhee, X. Liu, L. Wu, S. C. Chong, X. Wan, and J. Zhou. Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Applied and Environmental Microbiology*, 70(7):4303–4317, July 2004.

D. F. Rhoades. Evolution of plant chemical defense against herbivores. In G. A. Rosenthal and D. H. Janzen, editors, *Herbivores: Their Interaction with Secondary Plant Metabolites*, pages 4–55. Academic Pres, New York, 1979.

S. Rimour, D. Hill, C. Militon, and P. Peyret. GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, 21(7):1094–1103, April 2005.

D. Rohle, J. Popovici-Muller, N. Palaskas, S. Turcan, C. Grommes, C. Campos, J. Tsoi, O. Clark, B. Oldrini, E. Komisopoulou, K. Kunii, A. Pedraza, S. Schalm, L. Silverman, A. Miller, F. Wang, H. Yang, Y. Chen, A. Kernytsky, M. K. Rosenblum, W. Liu, S. A. Biller, S. M. Su, C. W. Brennan, T. A. Chan, T. G. Graeber, K. E. Yen, and I. K. Mellinghoff. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science*, 340(6132):626–630, May 2013.

E. Ronchetti. Robust model selection in regression. *Statistics & Probability Letters*, 3(1):21–23, February 1985.

L. Sachs. *Angewandte Statistik*. Springer, Berlin, 2004.

G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence. Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21(19):3748–3754, October 2005.

J. SantaLucia and D. Hicks. The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33:415–440, 2004.

M. A. Sartor, M. Medvedovic, and B. J. Aronow. Normalization: correcting for technical variance in oder to study biological variation. In E. M. Blalock, editor, *A Beginner's Guide to Microarrays*, pages 151–178. Kluwer Academic Publishers, Norwell, 2003.

U. Schittko, D. Hermsmeier, and I. T. Baldwin. Molecular interactions between the specialist herbivore Manduca sexta (Lepidoptera, Sphingidae) and its natural host Nicotiana attenuata. II. Accumulation of plant mRNAs in response to insect-derived cues. *Plant Physiology*, 125(2):701–710, February 2001.

V. Schroeckh, K. Scherlach, H.-W. Nützmann, E. Shelest, W. Schmidt-Heck, J. Schümann, K. Martin, C. Hertweck, and A. A. Brakhage. Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans. *Proceedings of the National Acadamy of Science of the United States of America*, 106(34):14558–14563, August 2009.

J. Schwachtje, P. E. Minchin, S. Jahnke, J. T. van Dongen, U. Schittko, and I. T. Baldwin. SNF1-related kinases allow plants to tolerate herbivory by allocating

carbon to roots. *Proceedings of the National Academy of Sciences of the United States of America*, 103(34):12935–12940, August 2006.

J. Seo and E. P. Hoffman. Probe set algorithms: is there a rational best bet? *BMC Bioinformatics*, 7:395, August 2006.

P. Sethu, L. L. Moldawer, M. N. Mindrinos, P. O. Scumpia, C. L. Tannahill, J. Wilhelmy, P. A. Efron, B. H. Brownstein, R. G. Tompkins, and M. Toner. Microfluidic isolation of leukocytes from whole blood for phenotype and gene expression analysis. *Analytical Chemistry*, 78(15):54535461, June 2006.

K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K. R. Cho, T. J. Giordano, S. B. Gruber, E. R. Fearon, J. M. G. Taylor, and S. Hanash. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics*, 6:26, 2005.

J. Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7): 585–587, July 2008.

A. Shervington, R. Patel, C. Lu, N. Cruickshanks, R. Lea, G. Roberts, T. Dawson, and L. Shervington. Telomerase subunits expression variation between biopsy samples and cell lines derived from malignant glioma. *Brain Research*, 1134(1): 45–52, February 2007.

L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Scherf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. Bergstrom Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle,

S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker Jr. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, September 2006.

H. J. Shin, H. Y. Park, S. J. Jeong, H. W. Park, Y. K. Kim, S. H. Cho, Y. Y. Kim, M. L. Cho, H. Y. Kim, K. U. Min, and C. W. Lee. STAT4 expression in human T cells is regulated by DNA methylation but not by promoter polymorphism. *Journal of Immunology*, 175(11):7143–7150, December 2005.

J. Simon, R. M. Gleadow, and I. E. Woodrow. Allocation of nitrogen to chemical defence and plant functional traits is constrained by soil N. *Tree Physiology*, 30 (9):1111–1117, September 2010.

M. Skibbe, N. Qu, I. Galis, and I. T. Baldwin. Induced plant defenses in the natural environment: Nicotiana attenuata WRKY3 and WRKY6 coordinate responses to herbivory. *The Plant Cell*, 20(7):1984–2000, July 2008.

J. Smolen and D. Aletaha. The burden of rheumatoid arthritis and access to treatment: a medical overview. *European Journal of Health Economics*, 8 Suppl 2: 39–47, January 2008.

G. W. Snedecor. *Calculation and Interpretation of Analysis of Variance and Covariance*. Collegiate Press, Ames, 1934.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, January 1904.

M. A. Stalteri and A. P. Harrison. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8:13, January 2007.

N. Stamp. Out of the quagmire of plant defense hypotheses. *The Quarterly Review of Biology*, 78(1):23–55, March 2003.

K. A. Stangier and H. Hegele. New sequencers and opportunities in genome and transcriptome research. Technical report, GATC Biotech, Konstanz, March 2011. Version III 2011 3.

A. D. Steinbrenner, S. Gómez, S. Osorio, A. R. Fernie, and C. M. Orians. Herbivore-induced changes in tomato (Solanum lycopersicum) primary metabolism: a whole plant perspective. *Journal of Chemical Ecology*, 37(12):1294–1303, December 2011.

A. Steppuhn, K. Gase, B. Krock, R. Halitschke, and I. T. Baldwin. Nicotine's defensive function in nature. *PLoS Biology*, 2(8):E217, August 2004.

M. Stitt and A. Krapp. The interaction between elevated carbon dioxide and nitrogen nutrition: the physiological and molecular background. *Plant, Cell & Environment*, 22(6):583–621, March 1999.

M. Stitt and D. Schulze. Does Rubisco control the rate of photosynthesis and plant growth? An exercise in molecular ecophysiology. *Plant, Cell & Environment*, 17 (5):465–487, May 1994.

J. D. Storey, J. Madeoy, J. L. Strout, M. Wurfel, J. Ronald, and J. M. Akey. Gene-expression variation within and among human populations. *American Journal of Human Genetics*, 80(3):502–509, March 2007.

W. Stork, C. Diezel, R. Halitschke, I. Gális, and I. T. Baldwin. An ecological analysis of the herbivory-elicited JA burst and its metabolism: plant memory processes and predictions of the moving target model. *PLoS ONE*, 4(3):e4697, 2009.

M. Sultan, I. Piccini, D. Balzereit, R. Herwig, N. G. Saran, H. Lehrach, R. H.

Reeves, and M. L. Yaspo. Gene expression variation in Down's syndrome mice allows prioritization of candidate genes. *Genome Biology*, 8(5):R91, 2007.

S. E. Sweeney and G. S. Firestein. Rheumatoid arthritis: regulation of synovial inflammation. *International Journal of Biochemistry & Cell Biology*, 36(3):372–378, March 2004.

Z. Szekanecz and A. E. Koch. Update on synovitis. *Current Rheumatology Reports*, 3(1):53–63, February 2001.

Z. Szekanecz, G. K. Haines, L. A. Harlow, M. R. Shah, T. W. Fong, R. Fu, S. J. Lin, G. Rayan, and A. E. Koch. Increased synovial expression of transforming growth factor (TGF)-beta receptor endoglin and TGF-beta 1 in rheumatoid arthritis: possible interactions in the pathogenesis of the disease. *Clinical Immunology and Immunopathology*, 76(2):187–194, August 1995.

P. Szodoray, P. Alex, M. B. Frank, M. Turner, S. Turner, N. Knowlton, C. Cadwell, I. Dozmorov, Y. Tang, P. C. Wilson, R. Jonsson, and M. Centola. A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis. *Rheumatology*, 45(12):1466–1476, December 2006.

P. P. Tak. IFN-beta in rheumatoid arthritis. *Frontiers in Bioscience*, 9:3242–3247, September 2004.

T. Takai. Fc receptors and their role in immune regulation and autoimmunity. *Journal of Clinical Immunology*, 25(1):1–18, January 2005.

A. Takano, J. I. Kakehi, and T. Takahashi. Thermospermine is not a minor polyamine in the plant kingdom. *Plant and Cell Physiology*, 53(4):608–616, February 2012.

P. K. Tan, T. J. Downey, E. L. Spitznagel Jr, P. Xu, D. Fu, D. S. Dimitrov, R. A. Lempicki, B. M. Raaka, and M. C. Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19): 5676–5684, October 2003.

E. Taranto and M. Leech. Expression and function of cell cycle proteins in rheumatoid arthritis synovial tissue. *Histology and Histopathology*, 21(2):205–211, February 2006.

M. Taubert, N. Jehmlich, C. Vogt, H. H. Richnow, F. Schmidt, M. von Bergen, and J. Seifert. Time resolved protein-based stable isotope probing (Protein-SIP) analysis allows quantification of induced proteins in substrate shift experiments. *Proteomics*, 11(11):2265–2274, June 2011.

J. R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, 2 edition, 1997.

J. T. Trumble, D. M. Kolodnyhirsch, and I. P. Ting. Plant compensation for arthropod herbivory. *Annual Review of Entomology*, 38:93–119, 1993.

T. Tucker, M. Marra, and J. M. Friedman. Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2): 142–154, 2009.

H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino. Universality and flexibility in gene expression from bacteria to human. *The Proceedings of the National Academy of Sciences (US)*, 101(11):3765–3769, March 2004.

L. Ullmann-Zeunert, A. Muck, N. Wielsch, F. Hufsky, M. A. Stanton, S. Bartram, S. Böcker, I. T. Baldwin, K. Groten, and A. Svatoš. Determination of ^{15}N-incorporation into plant proteins and their absolute quantitation: a new tool to study nitrogen flux dynamics and protein pool sizes elicited by plant-herbivore interactions. *Journal of Proteome Research*, 11(10):4947–4960, October 2012.

L. Ullmann-Zeunert, M. A. Stanton, N. Wielsch, S. Bartram, C. Hummert, A. Svatoš, I. T. Baldwin, and K. Groten. Quantification of growth-defense trade-offs in a common currency: nitrogen required for phenolamide biosynthesis is not derived from ribulose-1,5-bisphosphate carboxylase/oxygenase turnover. *The Plant Journal*, 75(3):417–429, 2013.

C. Valdes, P. Seo, N. Tsinoremas, and J. Clarke. Characteristics of cross-hybridization and cross-alignment of expression in pseudo-xenograft samples by RNA-Seq and microarrays. *Journal of Clinical Bioinformatics*, 3(8), April 2013.

N. M. van Dam and I. T. Baldwin. Competition mediates costs of jasmonate-induced defences, nitrogen acquisition and transgenerational plasticity in Nicotiana attenuata. *Functional Ecology*, 15(3):578–586, June 2001.

N. M. van Dam and M. W. Oomen. Root and shoot jasmonic acid applications differentially affect leaf chemistry and herbivore growth. *Plant Signaling and Behavior*, 3(2):91–98, February 2008.

N. M. van Dam, U. Hermenau, and I. T. Baldwin. Instar-specific sensitivity of specialist Manduca sexta larvae to induced defences in their host plant Nicotiana attenuata. *Ecological Entomology*, 26(6):578–586, December 2001.

T. C. van der Pouw Kraan, F. A. van Gaalen, P. V. Kasperkovitz, N. L. Verbeet, T. J. Smeets, M. C. Kraan, M. Fero, P. P. Tak, T. W. Huizinga, E. Pieterman, F. C. Breedveld, A. A. Alizadeh, and C. L. Verweij. Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. *Arthritis & Rheumatism*, 48(8):2132–2145, August 2003.

K. Vermeulen, Z. N. Berneman, and D. R. Van Bockstaele. Cell cycle and apoptosis. *Cell Proliferation*, 36(3):165–175, June 2003.

C. Voelckel and I. T. Baldwin. Generalist and specialist lepidopteran larvae elicit different transcriptional responses in Nicotiana attenuata, which correlate with larval FAC profiles. *Ecology Letters*, 7(9):770–775, September 2004a.

C. Voelckel and I. T. Baldwin. Herbivore-induced plant vaccination. Part II. Array-studies reveal the transience of herbivore-specific transcriptional imprints and a distinct imprint from stress combinations. *The Plant Journal*, 38(4):650–663, May 2004b.

C. Voelckel, T. Krügel, K. Gase, N. Heidrich, N. M. van Dam, R. Winz, and I. T. Baldwin. Anti-sense expression of putrescine N-methyltransferase confirms defen-

sive role of nicotine in Nicotiana sylvestris against Manduca sexta. *Chemoecology*, 11(3):121–126, September 2001.

W. Vongsangnak and J. Nielsen. Bioinformatics and systems biology of Aspergillus. In M. Machida and K. Gomi, editors, *Aspergillus: Molecular Biology and Genomics*, pages 61–84. Caister Academic Press, Norfolk, 2010.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):5763, January 2009.

T. Werner. Bioinformatics applications for pathway analysis of microarray data. *Current Opinion in Biotechnology*, 19(1):50–54, February 2008.

C. M. Weyand, T. M. Seyler, and J. J. Goronzy. B cells in rheumatoid synovitis. *Arthritis Research & Therapy*, 7 Suppl 3:9–12, 2005.

J. Wishart. Statistics in agricultural research. *Journal of the Royal Statistical Society*, 1 Suppl:26–62, 1934.

B. Wold and R. M. Myers. Sequence census methods for functional genomics. *Nature Methods*, 5(1):19–21, January 2008.

M. G. Woldemariam, I. T. Baldwin, and I. Galis. Transcriptional regulation of plant inducible defenses against herbivores: a mini-review. *Journal of Plant Interactions*, 6(2-3):113–119, March 2011.

Y. Woo, J. Affourtit, S. Daigle, A. Viale, K. Johnson, J. K. Naggert, and G. A. Churchill. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *Journal of Biomolecular Techniques*, 3: 579–588, December 2004.

C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielser, H.-H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3(9), August 2002.

C. Wu, R. Carta, and L. Zhang. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Research*, 33(9):e84, May 2005.

C. Wu, H. Zhao, R. Baggerly, Keith, R. Carta, and L. Zhang. Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics*, 23(19):2566–2572, 2007.

J. J. Wu, M. W. Lark, L. E. Chun, and D. R. Eyre. Sites of stromelysin cleavage in collagen types II, IX, X, and XI of cartilage. *Journal of Biological Chemistry*, 266(9):5625–5628, March 1991.

I. V. Yang. Use of external controls in microarray experiments. *Methods in Enzymology*, 411:50–63, 2006.

Y. H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 32 Suppl:490–495, August 2002.

C. L. Yauk, M. L. Berndt, A. Williams, and G. R. Douglas. Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*, 32(15):e124, 2004.

R. Yelin, D. Dahary, R. Sorek, E. Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, and G. Rotman. Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, 21(4):379–386, April 2003.

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17 (10):977–987, 2001.

C. Yoo and G. F. Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine*, 31(2):169–182, 2004.

D. A. Young, M. Hegen, H. L. Ma, M. J. Whitters, L. M. Albert, L. Lowe, M. Senices, P. W. Wu, B. Sibley, Y. Leathurby, T. P. Brown, C. Nickerson-Nutter, J. C. Keith, and M. Collins. Blockade of the interleukin-21/interleukin-21 receptor pathway ameliorates disease in animal models of rheumatoid arthritis. *Arthritis & Rheumatism*, 56(4):1152–1163, April 2007.

A. R. Zangerl, J. G. Hamilton, T. J. Miller, A. R. Crofts, K. Oxborough, M. R. Berenbaum, and E. H. de Lucia. Impact of folivory on photosynthesis is greater than the sum of its holes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):1088–1091, January 2002.

J. A. Zavala, A. G. Patankar, K. Gase, and I. T. Baldwin. Constitutive and inducible trypsin proteinase inhibitor production incurs large fitness costs in Nicotiana attenuata. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1607–1612, February 2004a.

J. A. Zavala, A. G. Patankar, K. Gase, D. Hui, and I. T. Baldwin. Manipulation of endogenous trypsin proteinase inhibitor production in Nicotiana attenuata demonstrates their function as antiherbivore defenses. *Plant Physiology*, 134(3): 1181–1190, March 2004b.

L. Zhang, M. F. Miles, and K. D. Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, 21(7):818–821, July 2003.

Z. Zhang, C. Gorman, J. M. Clark, and A. P. Cope. Rheumatoid arthritis: a disease of chronic, low-amplitude signals transduced through T cell antigen receptors? *Wiener Medizinische Wochenschrift*, 156(1-2):2–10, January 2006.

X. Zhou, L. Ren, Q. Meng, Y. Li, Y. Yu, and J. Yu. The next-generation sequencing technology and application. *Protein & Cell*, 1(6):520–536, 2010.

Sachregister

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe. Mir ist die geltende Promotionsordnung bekannt und ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des oben genannten akademischen Grades an einer anderen Hochschule beantragt.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich der Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena unter der Leitung von Prof. Dr. Stefan Schuster und die Arbeitsgruppe Bioinformatik/Systembiologie am Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie - Hans-Knöll-Institut unter der Leitung von Prof. Dr. Reinhard Guthke unterstützt.

Jena, den 24. November 2014

.....

(Christian Hummert)