Strategien für die Expressionsanalyse in funktionellen Gengruppen

Manuela Benita Hummel

München 2009

Dissertation

Aus dem Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. rer. nat. Ulrich Mansmann

Strategien für die Expressionsanalyse in funktionellen Gengruppen

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Manuela Benita Hummel

aus Aschaffenburg

2009

Dissertation

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. Ulrich Mansmann
Mitberichterstatter:	Prof. Dr. Peter B. Becker Prof. Dr. Peter Lohse
Dekan:	Prof. Dr. med. Dr. n.c. M. Reiser, FACR, FRCR
Tag der mündlichen Prüfung:	20.02.2009

Zusammenfassung

Die Analyse von Genexpressionsdaten, die durch die *Microarray*-Technologie bereit gestellt werden, ist in den letzten Jahren zu einem interessanten Forschungsfeld der Statistik geworden. Die ersten Verfahren auf diesem Gebiet zielen darauf ab, differentiell exprimierte Gene aus der riesigen Menge aller Gene eines Microarrays heraus zu filtern. Das Resultat einer solchen genweisen Analyse ist eine Liste interessanter Gene. Derartige Listen einzeln ausgewählter Gene sind allerdings schwer in einen biologischen Kontext zu bringen. Überdies hängen sie stark von der verwendeten Analysemethode und vom jeweiligen Datensatz ab, so daß Genlisten verschiedener Arbeitsgruppen meist eine relativ schlechte Übereinstimmung aufweisen.

Eine Alternative beziehungsweise Weiterführung der genweisen Herangehensweise bietet die Analyse funktioneller Gengruppen. Diese beinhalten biologisches Vorwissen über das Zusammenspiel von Genen. Somit sind relevante Gengruppen sinnvoller interpretierbar als einzelne relevante Gene. Es werden verschiedene Verfahren für die Untersuchung funktioneller Gengruppen hinsichtlich differentieller Expression vorgestellt und auf methodischer Ebene sowie anhand von realen Datenbeispielen und Simulationsstudien verglichen. Von speziellem Interesse ist hier die Familie von Gengruppen, die durch die *Gene Ontology* definiert wird. Die hierarchische Struktur dieser Ontologien bedeutet eine zusätzliche Herausforderung für die Analyse, insbesondere für die Adjustierung für multiples Testen.

Ein globaler Test auf differentielle Expression in Gengruppen ist das *GlobalAncova* Verfahren, welches im Rahmen dieser Arbeit weiter entwickelt und als *R* Paket bereit gestellt wurde. Die Signifikanz von Gengruppen kann dabei durch ein Permutationsmodell sowie über die asymptotische Verteilung der Teststatistik bewertet werden. Wir legen die theoretischen Grundlagen und Aspekte der Programmierung des Verfahrens dar. GlobalAncova eignet sich für die Analyse komplexer Fragestellungen. Hierzu werden einige ausführliche Auswertungen präsentiert, die im Rahmen von Kooperationen mit Medizinern und Biologen durchgeführt wurden.

Abstract

The analysis of gene expression data that is provided by *microarray* technology has evolved to an interesting field of statistical research. Primary analysis approaches aim to select the differentially expressed genes out of the huge amount of all genes on a microarray. The result of such a gene-wise analysis is a list of interesting genes. However, lists of separately selected genes are difficult to interpret in a biological context. Moreover, they depend heavily on the applied analysis method and the respective data set. Therefore gene lists detected by different research groups in most cases do not show much overlap.

An alternative or an additional step to the gene-wise approach is given by the analysis of functional groups of genes. Gene sets encode the current biological knowledge about the relations and interactions between genes. For this reason relevant gene sets can be interpreted in a more meaningful way than single relevant genes. In this work we describe several methods for the exploration of functional gene sets and compare them on a theoretical basis and by means of real and simulated data. The family of gene groups given by the *Gene Ontology* is of particular interest. The hierarchical structure of these ontologies implies a special challenge for the analysis and in particular for the adjustment for multiple testing.

A global test for differential expression in gene sets is the *GlobalAncova* method that was extended and advanced within this work. The significance of gene sets is assessed by a permutation approach or alternatively by the asymptotic distribution of the test statistic. We describe the theoretical background and details of the implementation of the corresponding R package. GlobalAncova can be used for the analysis of complex research questions. For demonstration of its potentiality some real data analyses, that resulted from cooperations with physicians and biologists, are presented in detail.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Expressionsanalyse in funktionellen Gengruppen

Moderne biotechnologische Verfahren ermöglichen die simultane Messung tausender Merkmale eines Individuums, wie zum Beispiel bestimmter Abschnitte des Erbguts oder verschiedener Proteine. Diese Arbeit befaßt sich mit der Analyse von Genexpressionsdaten, die mit Hilfe von *Microarrays* gewonnen werden. Die gewaltigen Datenmengen bedeuten eine große Herausforderung für Informatik und Statistik.

Die erste und derzeit immer noch übliche Herangehensweise an Genexpressionsdaten besteht in einer Gen-für-Gen Analyse. Das heißt die Gene werden einzeln betrachtet und ihre Expressionsniveaus hinsichtlich bestimmter Fragestellungen untersucht. Allerdings geht man davon aus, daß die meisten biologischen Phänomene und Krankheiten nicht durch einzelne Gene sondern durch das komplexe Zusammenspiel mehrerer genetischer Faktoren beeinflußt werden. Deshalb bietet alternativ oder zusätzlich zu einer genweisen Auswertung die Analyse *funktioneller Gengruppen* die Möglichkeit, die biologischen Hintergründe der untersuchten Krankheit besser zu verstehen. Häufig werden dazu Sammlungen von Gengruppen herangezogen, die das aktuelle Wissen über das komplexe Zusammenspiel der Gene repräsentieren. Das Konzept der Gengruppe kann aber noch allgemeiner aufgefaßt werden. Ein Beispiel für eine Gengruppenanalyse findet sich bei Lamb u. a. (2003). Die Autoren untersuchen die Gruppe von Genen, die durch Cyclin D1 reguliert werden, und zeigen, daß die globale Expression dieser Gruppe deutlich erhöht ist gegenüber einer zufällig zusammen gestellten Gruppe gleicher Größe.

Die wohl häufigste Form der Gengruppenanalyse beschäftigt sich mit *differentieller Expression*. Dabei ist von Interesse, welche funktionellen Gengruppen eine Rolle spielen für die Unterschiede in der Expression zwischen Patienten mit verschiedenen klinischen Eigenschaften. Für die Analyse solcher Fragestellungen gibt es bereits eine Vielzahl von Methoden. Derzeit mangelt es teilweise noch an Gegenüberstellungen der einzelnen Verfahren und an Hinweisen, welche Methode in der Praxis für welche Situation die passende

ist. Deshalb soll diese Arbeit einen Überblick und systematischen Vergleich der gängigsten Verfahren aufzeigen und somit eine Art „Orientierungshilfe“ bei der Wahl einer geeigneten Analysemethode bieten. Darüber hinaus wird der Gengruppentest *GlobalAncova* (Mansmann und Meister, 2005; Hummel u. a., 2008a) weiter entwickelt und seine Nützlichkeit anhand von ausführlichen praktischen Auswertungsbeispielen dargestellt.

Ein weiteres Beispiel, bei dem Gruppenanalyse benötigt wird, ist die Validierung und biologische Ergründung von *Gensignaturen*. Prognostische Signaturen werden durch explorative Methoden oder Klassifizierungsalgorithmen erstellt. Zumeist endet derzeit eine Analyse an diesem Punkt. Ebenso wichtig wie das Erstellen einer Signatur ist aber auch ihre Validierung. Ihre Nützlichkeit für die Prognose neuer Patienten kann durch geeignete Gengruppentests überprüft werden. Weiterhin sollte versucht werden, die zugrunde liegenden biologischen Mechanismen einer Signatur zu verstehen. Zu diesem Thema findet sich noch nicht viel in der aktuellen Literatur. In der vorliegenden Arbeit wird gezeigt, wie mehr über die Biologie einer Signatur zu erfahren ist, beispielsweise indem Zusammenhänge zwischen der Signatur und bekannten *pathways* untersucht werden.

1.2 Aufbau der Arbeit

Abschnitt 2 gibt eine Einführung in die Technologie und gängige Auswertung von Microarrays. Es werden einige Verfahren für die genweise Analyse differentieller Expression kurz genannt. Da eines der Hauptprobleme bei der Auswertung von Microarray Daten die enorme Anzahl von Tests darstellt, wird an dieser Stelle eine Einführung in Konzepte des multiplen Testens gegeben. Auch auf weitere generelle Schwierigkeiten der Microarray Technologie und der üblichen genweisen Analysestrategie wird eingegangen.

In Kapitel 3 wird die Analyse von Gengruppen als sinnvolle Alternative oder Ergänzung zur genweisen Auswertung motiviert. Es wird vorgestellt, welche Arten von Gengruppen für eine Analyse denkbar sein können. Der Hauptteil dieses Abschnittes besteht aus der Beschreibung bestehender Methoden für die Expressionsanalyse in Gengruppen. Die Verfahren werden eingeteilt gemäß der zwei Grundkonzepte *Gene Set Enrichment* und *holistische Strategie*.

Das im Rahmen dieser Arbeit weiter entwickelte Verfahren *GlobalAncova* wird in Kapitel 4 ausführlich dargestellt. Nach den Grundlagen des Tests werden die zwei Möglichkeiten für die Bestimmung von p-Werten gezeigt. Ebenso gibt es eine kurze Übersicht über das *R* Programmpaket. Dabei werden Details zur Programmierung und graphische Instrumente besprochen.

Die Gengruppen, die durch die *Gene Ontology* (Ashburner u. a., 2000) definiert werden, sind in der vorliegenden Arbeit von speziellem Interesse. Die Analyse dieser Gengruppen wird in Kapitel 5 genauer betrachtet. Die besondere Struktur der Gene Ontology erfordert

geeignete Analyseverfahren. Einige Konzepte hierfür gibt es bereits.

In Kapitel 6 werden die verschiedenen vorgestellten Verfahren miteinander verglichen. Zunächst werden noch einmal die beiden Konzepte des Gene Set Enrichment und der holistischen Strategie gegenüber gestellt. Desweiteren werden Zusammenhänge und Unterschiede zwischen einzelnen Verfahren genauer betrachtet. Eine Simulationsstudie soll dabei helfen, ein Gefühl dafür zu entwickeln, in welcher Situation welche Methode am geeignetsten erscheint.

Abschnitt 7 liefert schließlich einige praktische Beispiele für die Analyse funktioneller Gengruppen. Ein Schwerpunkt liegt darin, verschiedene Anwendungsmöglichkeiten für den globalen Test GlobalAncova zu zeigen. Dementsprechend geht es in den Beispielen nicht um die Analyse einfacher differentieller Expression zwischen zwei klinischen Gruppen, sondern um komplexere Fragestellungen.

Die Auswertungen werden alle mit dem unter http://www.r-project.org frei zugänglichen Statistikprogramm *R* (R Development Core Team, 2006) durchgeführt. Auf der Basis dieses Programmes wurde der *Bioconductor* (http://www.bioconductor.org, Gentleman u. a., 2004), eine Sammlung von Paketen speziell für die Analyse genomischer Daten, entwickelt. Einen guten Überblick über die Auswertung mit Hilfe des Bioconductors liefert das Buch von Gentleman u. a. (2005).

Kapitel 2

Microarray Analyse

2.1 Biologische Grundlagen

Die Grundlage der molekularen Biologie ist die Übersetzung des Erbgutes, der *DNA*, in Proteine. Zunächst werden dabei Abschnitte (*Gene*) der zweisträngigen DNA in einsträngige *RNA* umgeschrieben (Transkription). Diese RNA Stücke bilden dann den „Bauplan“ für die entsprechenden Proteine (Translation) (Campbell, 2000). Je nach Typ oder Zustand einer Zelle werden unterschiedliche Proteine benötigt und demnach unterschiedliche Gene kopiert und übersetzt. Mißt man die Mengen der RNA Abschnitte, die sich in einer Zelle befinden, gewinnt man einen Eindruck über die Aktivität der entsprechenden Gene. Dadurch lassen sich Rückschlüsse auf die Mengen der zugehörigen Proteine und damit letztendlich auf die Biologie der betrachteten Zellen ziehen. Zwar wäre es erstrebenswert direkt die letztlich interessierenden Proteine zu quantifizieren, dies ist derzeit allerdings viel aufwendiger als die Messung von RNA Mengen.

Microarrays bieten die Möglichkeit, die Aktivität zehntausender Gene gleichzeitig zu messen. Die folgende Einführung ist angelehnt an Ewens und Grant (2005). Die Technologie beruht auf dem Prinzip der spezifischen Bindung einsträngiger Erbgutabschnitte. Die Erbgutinformation ist in Sequenzen aus vier verschiedenen Bausteinen, den *Nukleotiden*, kodiert. Jeweils zwei Nukleotide passen nach dem Schlüssel-Schloß-Prinzip zueinander. Darauf beruht die Bindung der beiden komplementären Stränge der DNA. Die Microarrays sind mit bekannten einsträngigen DNA Stücken besetzt. Wird nun aus ebenfalls einsträngigen Sequenzen bestehendes Probenmaterial aufgetragen, so binden diese Sequenzen an die passenden Stellen auf dem array. Zuvor wird das Probenmaterial mit fluoreszierenden Molekülen markiert, so daß die von einem Scanner abgelesenen Leuchtintensitäten der einzelnen Punkte ein Maß für die entsprechenden RNA Mengen darstellen.

2.2 Technologien und Schwierigkeiten

Es gibt verschiedene Microarray Technologien. Für die Genexpressionsanalyse sind die zwei wohl gängigsten *spotted* (bzw. *cDNA* bzw. *two-colour*) arrays und *Oligonukleotid* (bzw. *probeset* bzw. *one-channel*) arrays der Firma Affymetrix.

Two-colour arrays
Bei two-colour arrays wird Probenmaterial von je zwei verschiedenen Beobachtungseinheiten mit roten und grünen Farbstoffen (*dyes*) markiert und zusammen auf ein Glas- oder Nylon array aufgetragen. Das array ist bedruckt mit tausenden von *spots*, wobei jeder spot einsträngige DNA Fragmente eines Gens enthält. Die Länge der Fragmente variiert zwischen ca. 70 und mehreren tausend Nukleotiden. An die spots bindet das einsträngige Probenmaterial spezifisch, das heißt es binden an einen spot (im Idealfall) nur Proben mit der entsprechenden komplementären Nukleotidsequenz. Je mehr Proben an einem spot binden, desto stärker ist die Leuchtintensität. Über die Verhältnisse der roten und grünen Intensitäten der einzelnen Gene lassen sich Expressionsunterschiede zwischen den Beobachtungseinheiten feststellen.

Affymetrix arrays
Affymetrix arrays arbeiten mit nur einer Farbe. Das heißt, daß ein array die Genexpression nur einer klinischen Entität darstellt. Ein Vergleich verschiedener klinischer Gruppen muß folglich über den Vergleich von mindestens zwei arrays geschehen. Die arrays sind mit Erbgutabschnitten, den *probes*, von je 25 Nukleotiden Länge besetzt. Mehrere probes setzen sich jeweils zu *probesets* zusammen, welche letztlich den Genen entsprechen sollen. Als Maß für die Expression eines Gens wird also ein aus den Intensitäten der einzelnen probes zusammengesetzter Wert benötigt. Zusätzlich gibt es zu jedem probe, das einem real vorkommenden Erbgutabschnitt entspricht (*perfect match*), ein *mismatch* probe, bei dem ein Nukleotid in der Mitte der Sequenz durch ein falsches ausgetauscht ist. Affymetrix verfolgte dabei die Idee, das Ausmaß an unspezifischen Bindungen quantifizieren zu können. Es zeigt sich allerdings, daß die Berücksichtigung der mismatches eher zu noch größerer Variabilität in den Daten führt (Irizarry u. a., 2003b). Deshalb gehen viele Normalisierungsverfahren (siehe unten) rein von den perfect match probes aus. In der vorliegenden Arbeit werden ausschließlich Datensätze von Affymetrix-arrays betrachtet.

Die komplizierte und vielschrittige Microarray-Technologie ist anfällig für vielerlei zufällige und systematische Fehler, nämlich für

- Array spezifische Effekte: Da die vielen technischen Arbeitsschritte nie exakt wiederholt werden können, sind keine zwei arrays gleich. Es liegt eine zufällige Variabilität zwischen den arrays vor.

- Genspezifische Effekte: Die Hybridisierung des Probenmaterials mit den arrays erfolgt unterschiedlich gut für verschiedene Gene.

- Hintergrundrauschen und Artefakte: Es wird immer eine gewisse Hintergrundintensität auf den arrays beobachtet. Zudem können die arrays durch Staub und Kratzer etc. verunreinigt sein.

- Effekte durch die Aufbereitung: Variabilität kann entstehen durch verschiedene Bearbeiter der arrays, unterschiedliche Herstellungszeiten etc.

Am Anfang jeder Microarray Analyse muß deshalb eine Qualitätskontrolle durchgeführt werden, um zu prüfen ob grobe Verunreinigungen oder andere Mängel vorliegen. Beispielsweise kann man sich *images* der gemessenen Leuchtintensitäten ansehen oder Grafiken, die das Maß der RNA Degradierung veranschaulichen. Mit boxplots werden die arrays untereinander verglichen. Auch der Zusammenhang zwischen Stärke und Variabilität der Intensitäten innerhalb eines arrays sollte mit sogenannten MA plots überprüft werden.

Auch wenn nur arrays guter Qualität analysiert werden, beobachtet man wegen der oben genannten Gründe unerwünschte experimentelle Variabilität. Interessiert ist man nur an Expressionsunterschieden zwischen Genen und zwischen arrays von tatsächlich biologischer Herkunft. Vor der eigentlichen statistischen Analyse müssen die Expressionsdaten deshalb geeignet bereinigt werden. Für ein solche *Normalisierung* der Daten gibt es verschiedene Verfahren. Für Affymetrix arrays werden häufig *RMA* (Irizarry u. a., 2003a), *GCRMA* (Wu u. a., 2004) und *vsn* (Huber u. a., 2002) verwendet. RMA (robust multi-array analysis) setzt sich aus drei Schritten zusammen: 1) Korrektur des Hintergrundrauschens über ein globales Modell der Intensitätenverteilung je array, 2) Quantilnormalisierung, durch die die Expressionsverteilungen der arrays aneinander angeglichen werden, 3) Zusammenfassen der Proben zu probesets über Medianglättung. Bei GCRMA wird eine Hintergrundkorrektur verwendet, bei der die Anfälligkeiten verschiedener Probensequenzen zu unspezifischen Bindungen mit berücksichtigt werden. Quantilnormalisierung und Zusammenfassen der Proben erfolgt wie bei RMA. Die Grundlage zu vsn bildet die Beobachtung, daß die Expressionsvariabilität von der Stärke der Intensität abhängt. Diese Abhängigkeit wird durch eine varianzstabilisierende Transformation ausgeglichen. Vor der Transformation werden die Intensitäten hinsichtlich Hintergrundrauschen kalibriert. Ein array übergreifendes Modell sorgt für Vergleichbarkeit der Expressionsverteilungen. Das Zusammefassen der Proben kann zum Beispiel wie bei RMA über Medianglättung erfolgen.

In dieser Arbeit werden wir meist von geeignet normalisierten Daten ausgehen. Der Ausgangspunkt einer jeden Auswertung stellt somit die normalisierte *Expressionsmatrix* X dar. Im Bereich der Microarray Analyse hat es sich (entgegen der üblichen Konvention in der Statistik) eingebürgert, daß die Zeilen von X den m Genen entsprechen und die Spalten den n Beobachtungen. Im Rest dieses einleitenden Kapitels wird kurz eine gängige Herangehensweise an Expressionsdaten vorgestellt, bei der die Gene einzeln nacheinander prozessiert werden. Der Schwerpunkt liegt in den folgenden Abschnitten dagegen auf der Analyse funktioneller Gengruppen.

2.3 Genweise Analyse differentieller Expression

Die Verwendung von Microarrays ist vor allem in Situationen angebracht, in denen noch nicht genau bekannt ist, welche Gene für die untersuchte klinische Fragestellung von Bedeutung sind. Sie dienen in diesem Fall also als screening Werkzeug. Die einfachste Art einer möglichen Fragestellung betrachtet zwei veschiedene biologische Konditionen, zum Beispiel eine bestimmte Krankheit und gesunde Kontrollen oder zwei Subtypen eines Karzinoms. Mit Hilfe der Microarrays sollen diejenigen Gene gefunden werden, deren Expressionsmuster sich zwischen den beiden Gruppen unterscheiden. Solche Gene nennt man *differentiell exprimiert*. Um diese Gene zu detektieren werden derzeit in der Regel einfache statistische Tests einzeln auf jedes Gen angewandt. Zwar ist bekannt, daß Genexpression ein koordiniertes System darstellt und die Gene also nicht unabhängig voneinander reguliert werden. Dennoch beschränken sich viele Ansätze aufgrund der sehr hohen Dimensionalität der Daten und fehlendem Wissen über die genauen Zusammenhänge zwischen den Genen auf eine genweise, univariate Analyse. Einige der gängigsten Verfahren werden im folgenden Unterkapitel kurz dargestellt. Dieser und der folgende Abschnitt sind größtenteils angelehnt an Scholtens und von Heydebreck (2005).

2.3.1 Einige Verfahren

Das einfachste Maß für differentielle Expression sind Verhältnisse, beziehungsweise auf logarithmischer Skala Differenzen, der Expressionsmittelwerte (*fold changes*) zwischen den klinischen Gruppen. Bei sehr kleinen Fallzahlen, wie sie bei Microarray Experimenten aufgrund der hohen Kosten häufig vorliegen, ist der fold change wohl die einzige Möglichkeit zur Bewertung differentieller Expression. Er hat den entscheidenden Nachteil, daß die experimentelle Variabilität nicht berücksichtigt werden kann. Aus diesem Grund werden wenn möglich statistische Tests bevorzugt. Für nicht-parametrische Tests wie den Mann-Whitney-Test müssen weniger Voraussetzungen erfüllt sein als für parametrische. Allerdings haben erstere gerade bei kleinen Fallzahlen eine geringere Macht. Deshalb werden meist parametrische Tests wie der zwei-Stichproben t-Test oder die Varianzanalyse verwendet. Betrachtet man normalisierte Daten auf einer logarithmischen Skala, die zudem hinreichend varianzstabilisiert sind, so sind die vorausgesetzten Verteilungsannahmen weitgehend erfüllt. Dennoch sind die resultierenden p-Werte mit Vorsicht zu genießen, nicht zuletzt auch wegen der unabhängigen Behandlung der Gene und der enormen multiplen Testsituation. Diese Probleme werden ausführlicher im nächsten Abschnitt behandelt.

Die kleinen Fallzahlen in Microarray Experimenten erschweren die Schätzung der genweisen Varianzen, die zum Beispiel in der t-Teststatistik verwendet werden. Einige Verfahren, unter anderen *Significance Analysis for Microarrays (SAM)* von Tusher u. a. (2001) und *limma* von Smyth (2005), verbessern die Varianzschätzungen, indem sie Information aus der Streuung aller Gene mit einbeziehen. In den beiden genannten Methoden wird jeweils die Varianz s_j^2 von Gen j durch eine globale Varianzschätzung auf der Basis aller Gene s_0^2 adjustiert, wodurch die resultierenden t-Statistiken stabiler werden.

In SAM wird ein konstanter Wert s_0 zu den genweisen Standardabweichungen s_j addiert. Die Konstante wird so gewählt, daß der Variationskoeffizient der angepaßten t-Statistik minimiert wird. Im Programmpaket SAM werden die unter der Nullhypothese erwarteten Statistiken durch einen Permutationsansatz simuliert. Die beobachteten Werte werden dann gegen die erwarteten aufgetragen, so daß differentielle Gene als die stark von der Hauptdiagonalen abweichenden erkennbar werden. Durch den Permutationsansatz wird außerdem die *false discovery rate* geschätzt, die dem Anteil der fälschlicherweise detektierten Gene entspricht.

Bei limma wird als Varianzschätzung ein gewichtetes Mittel aus s_j^2 und s_0^2 verwendet. Dabei ist s_0^2 der *a priori* Mittelwert der wahren Varianz σ_j^2 in einem empirischen Bayes-Ansatz. Der Name „limma“ leitet sich von „linearen Modellen“ ab. Dementsprechend ist dieses Analysewerkzeug nicht auf den Vergleich von zwei Gruppen beschränkt, sondern man kann damit vielfältige Studiendesigns untersuchen, insbesondere faktorielle Designs, Zeitreihen etc.

2.3.2 Korrektur für multiples Testen

Eine besondere Herausforderung bei der Analyse von Microarray Daten liegt in der Adjustierung für multiples Testen. Die „Datenflut“ der neuen Technologien hat in den letzten Jahren auf diesem Gebiet der Statistik viele Neuentwicklungen bewirkt. Jeder einzelne Test hat eine gewisse Fehlerwahrscheinlichkeit, das heißt daß bei tausenden von Tests eine nicht zu vernachlässigende Anzahl von falschen Ergebnissen zu erwarten ist. In solchen multiplen Testsituationen muß demnach eine Korrektur durchgeführt werden. Einen guten Überblick über Fehlerkonzepte und Adjustierungsmethoden beim multiplen Testen geben Dudoit u. a. (2003).

Allgemein gibt es vier mögliche Ausgänge eines statistischen Tests, je nachdem ob die zugrunde liegende Nullhypothese wahr oder falsch ist und ob der Test die Hypothese ablehnt oder nicht. Tabelle 2.1 zeigt für ein multiples Testproblem mit m getesteten Hypothesen die Anzahlen der jeweiligen Testausgänge. Die Anzahlen wahrer und falscher Nullhypothesen m_0 und m_1 sind unbekannte Parameter. Die Anzahl abgelehnter Hypothesen R ist eine beobachtbare, S, T, U und V dagegen sind unbeobachtbare Zufallsvariablen. Dabei

	nicht abgelehnt	**abgelehnt**	
H_0 **wahr**	U	V	m_0
H_0 **falsch**	T	S	m_1
	$m - R$	R	m

Tabelle 2.1: Anzahlen möglicher Testausgänge beim multiplen Testen.

ist V die Anzahl *falsch positiver* Ergebnisse, beziehungsweise *Fehler 1. Art* oder α*-Fehler*, das heißt die Anzahl der fälschlicherweise abgelehnten wahren Hypothesen. Ebenso kommt es vor, daß Hypothesen nicht abgelehnt werden, obwohl tatsächlich die Alternative gilt. T ist die Anzahl solcher *falsch negativer* Ergebnisse, beziehungsweise *Fehler 2. Art*. Da man nicht beide Fehler zugleich minimieren kann, wählt man in der Regel einen maximalen Wert für den Fehler 1. Art (α- bzw. *Signifikanzniveau*) und sucht Tests, die zusätzlich einen möglichst kleinen Fehler 2. Art, das heißt eine möglichst große *Macht* besitzen. Für einen einzelnen Test muß gelten $P(H_0 \text{ ablehnen } | H_0 \text{ wahr}) \leq \alpha$. Für das multiple Testproblem müssen geeignete Fehlerraten definiert werden, sowie mächtige Testprozeduren, die diese Fehlerraten kontrollieren und dabei die gemeinsame Verteilung der Teststatistiken berücksichtigen. Die wohl gängigsten *Typ I Fehlerraten* sind

- die *family-wise error rate (FWER)*: Wahrscheinlichkeit mindestens eines Fehlers 1. Art $FWER = P(V \geq 1)$
- die *false discovery rate (FDR)*: erwarteter Anteil an Fehlern 1. Art unter den abgelehnten Hypothesen $FDR = E(V/R)$ wenn $R > 0$ und $FDR = 0$ wenn $R = 0$

Eine multiple Testprozedur kontrolliert eine bestimmte Typ I Fehlerrate, wenn die Fehlerrate kleiner oder gleich dem Signifikanzniveau α ist, wenn also zum Beispiel gilt $FWER \leq \alpha$.

Das einfachste Adjustierungsverfahren, das die family-wise error rate kontrolliert, ist die Bonferroni-Korrektur. Das Signifikanzniveau α eines einzelnen Tests wird durch α/m ersetzt. Das heißt es werden nur Hypothesen abgelehnt, deren p-Werte kleiner oder gleich α/m sind. Alternativ kann man die Korrektur auch mit Hilfe *adjustierter p-Werte* $\tilde{p}_j$ beschreiben. Ein adjustierter p-Wert für eine Hypothese H_j entspricht dem α Niveau der gesamten Testprozedur, zu dem H_j, bei gegebenen beobachteten Teststatistiken aller beteiligten Hypothesen, gerade noch abgelehnt werden kann. Will man beispielsweise die FWER kontrollieren, so gilt $\tilde{p}_j = \inf\{\alpha \in [0,1] : H_j \text{ wird abgelehnt, wobei } FWER = \alpha\}$. Die Hypothese H_j wird dann abgelehnt, wenn gilt $\tilde{p}_j \leq \alpha$, wenn also der adjustierte p-Wert der individuellen Hypothese kleiner oder gleich dem nominalen α level der gesamten Testprozedur ist. Der Vorteil bei der Verwendung adjustierter p-Wert ist (wie im Falle des üblichen p-Wertes bei einem einzelnen Test), daß das Signifikanzniveau nicht im Voraus festgelegt werden muß. Bei der Bonferroni-Korrektur sind die adjustierten p-Werte definiert durch $\tilde{p}_j = \min(m \cdot p_j, 1)$. Die Methode von Holm (1979) zur Kontrolle der FWER ist weniger konservativ. Seien $p_{(1)} \leq \ldots \leq p_{(m)}$ die geordneten rohen p-Werte. Ist der kleinste p-Wert $p_{(1)} \leq \alpha/m$, so wird H_1 abgelehnt und es wird überprüft, ob der nächste p-Wert $p_{(2)} \leq \alpha/(m-1)$ ist. Die Prozedur läuft weiter, das heißt $p_{(j)}$ wird verglichen mit $\alpha/(m-j+1)$, so lange bis die erste Hypothese nicht mehr abgelehnt werden kann. Ab diesem Punkt wird keine weitere Hypothese mehr abgelehnt. Die entsprechenden adjustierten p-Werte lauten

$$\tilde{p}_{(j)} = \max_{k=1,\ldots,j} \left\{ \min\left((m-k+1)p_{(k)}, 1 \right) \right\}.$$

Es gibt noch viele weitere Verfahren zur Kontrolle der FWER, zum Beispiel permutationsbasierte Methoden wie die von Westfall und Young (1993).

Die false discovery rate ist weniger restriktiv als die family-wise error rate, das heißt FDR $\leq$ FWER. Im Falle, daß alle Nullhypothesen wahr sind, sind die beiden Fehlerraten identisch. Bei der FDR werden im Gegensatz zur FWER einige Fehler 1. Art toleriert, solange deren Anzahl klein ist im Vergleich zur Anzahl der abgelehnten Hypothesen. Das Konzept der FDR wurde von Benjamini und Hochberg (1995) entwickelt. Sie schlagen folgende Prozedur vor. Gibt es ein $j^* = \max\{j : p_{(j)} \leq (j/m)\alpha\}$, so können die Hypothesen $H_{(1)}, \ldots, H_{(j^*)}$ abgelehnt werden. Alternativ können auch im Falle der FDR-Kontrolle adjustierte p-Werte definiert werden. Für die Methode von Benjamini und Hochberg (1995) lauten sie

$$\tilde{p}_{(j)} = \min_{k=j,\ldots,m} \left\{ \min \left(\frac{m}{k} p_{(k)}, 1 \right) \right\}.$$

Durch diese Prozedur wird die FDR unter Annahme bestimmter Abhängigkeitsstrukturen kontrolliert. Will man FDR-Kontrolle für beliebige Abhängigkeitsstrukturen gewährleisten, bietet sich das etwas konservativere Verfahren von Benjamini und Yekutieli (2001) an. Die adjustierten p-Werte sind hier definiert als

$$\tilde{p}_{(j)} = \min_{k=j,\ldots,m} \left\{ \min \left(\frac{m \sum_{j=1}^{m} 1/j}{k} p_{(k)}, 1 \right) \right\}.$$

Der „Korrekturfaktor“ m/k aus der Benjamini & Hochberg Adjustierung wird hier also noch um $\sum 1/j$ erhöht. Es gibt zahlreiche Abwandlungen der FDR und weitere Adjustierungsverfahren, viele davon permutationsbasiert.

2.3.3 Probleme der genweisen Analyse

Die beschriebenen Ansätze zur Detektion differentiell exprimierter Gene durch genweise Tests werfen einige Probleme auf. Das entscheidenste ist wohl das der Multiplizität. Die Adjustierung hinsichtlich der FWER erweist sich im Rahmen der Microarray Analyse meist als zu konservativ, da die sehr große Anzahl an Genen zu einer massiven Korrektur führt, die oft nur sehr wenige oder gar keine signifikanten Tests bestehen läßt. Deshalb wird üblicherweise das Fehlerkonzept der FDR bevorzugt. Oft wird auch durch unspezifisches Filtern die Anzahl an Tests im Vorhinein reduziert, um das Ausmaß der Adjustierung einzuschränken. Es werden zum Beispiel nur Gene, die in einem gewissen Prozentsatz der arrays einen bestimmten Expressionswert überschreiten, oder solche, die eine gewisse Variabilität aufweisen, in die Analyse aufgenommen. Dennoch bleibt die Anzahl der Tests meist sehr groß. Außerdem will gut überlegt sein, ob das Filtern tatsächlich unspezifisch ist oder ob die nachfolgenden Tests bereits dadurch beeinflußt werden.

Ein weiteres Problem bildet die Tatsache, daß die Gene unabhängig voneinander betrachtet werden. Expressionsmuster entstehen aber bekanntermaßen durch komplexe Regulationssysteme. Es bestehen also starke Korrelationen zwischen den Genen, die möglichst berücksichtigt werden sollten, um eine biologisch sinnvolle Analyse und Interpretation zu

ermöglichen.

Aufgrund der vielfältigen, in Abschnitt 2.2 erwähnten Quellen ungewollter Variabilität, der enormen Anzahl an Tests und der oft kleinen Fallzahlen sind die aus einer genweisen Analyse resultierende Genlisten in den meisten Fällen eher „instabil“. Verschiedene Arbeitsgruppen publizieren völlig verschiedene Listen interessanter Gene, selbst wenn die Studienpopulationen ähnlich sind. Auch verschiedene Normalisierungs- oder Auswertungsmethoden, zum Beispiel fold changes und t-Tests, liefern unterschiedliche rankings der Gene.

Ein letzter Schwachpunkt der genweisen Analyse ist die schwere Interpretierbarkeit der Ergebnisse. Es bedeutet für Mediziner und Biologen einen erheblichen Aufwand, aus einer langen Liste von Genen Rückschlüsse über Wirkungsweise und Zusammenhänge zu ziehen. Die Analyse sollte also nicht an diesem Punkt enden. Vielmehr sollten die biologischen Hintergründe der Genlisten ergründet werden, indem die Zugehörigkeit der relevanten Gene zu *funktionellen Gengruppen* untersucht wird. Alternativ kann gleich auf eine genweise Auswertung verzichtet und stattdessen globale Strategien für die Gruppenanalyse zum Einsatz gebracht werden.

Kapitel 3

Analyse von Gengruppen

3.1 Motivation

Wie im vorigen Abschnitt erwähnt liefert das übliche Vorgehen einer genweisen Analyse eher instabile Genlisten hinsichtlich verschiedener Datensätze und Verfahren. Sucht man dagegen statt nach einzelnen interessanten Genen nach relevanten funktionellen Gengruppen besteht die Hoffnung auf robustere Ergebnisse. Manoli u. a. (2006) bestätigen die bessere Vergleichbarkeit von Datensätzen bei Verwendung von Gengruppen- statt Einzelgenuntersuchungen. Vergleicht man verschiedene Datensätze, um dieser Behauptung näher nachzugehen, muß natürlich genau geprüft werden, ob das Patientengut beziehungsweise Probenmaterial tatsächlich aus ähnlichen Bedingungen stammt. Außerdem sind große Sammlungen von Datensätzen derzeit noch nicht vorhanden oder nur schwer zugänglich. Deshalb soll die bessere Stabilität von Gengruppen gegenüber Listen „unzusammenhängender“ Gene hier anhand einer kleinen Studie gezeigt werden, bei der Unterschiede in den Ergebnissen nur durch verschiedene Normalisierungsverfahren herrühren können. Dazu werden vier Datensätze mit RMA und mit vsn normalisiert. Danach wird mit den je zwei entstandenen Expressionsdatensätzen zunächst eine genweise Analyse durchgeführt. Anschließend wird eine Sammlung von *pathways* mit globalen Tests hinsichtlich differentieller Expression untersucht. Es werden die folgenden Datensätze betrachtet:

- Zeitliche Entwicklung der Expression in **Prion-infizierten Mäusehirnen**: In der Studie von Xiang u. a. (2007) wurde versucht, Unterschiede in zeitlichen Expressionsmustern zwischen mit Prionen infizierten Mäusen und Kontrolltieren festzustellen. Zu jedem von drei Zeitpunkten und in jeder der beiden Gruppen gibt es drei Versuchstiere. Mit den verwendeten Microarrays können 22690 probesets gemessen werden. Für die genweise Analyse werden lineare Modelle verwendet, wobei die probesets mit signifikanter Interaktion zwischen Zeit und Gruppe selektiert werden sollen. Einen entsprechenden globalen Test für Gengruppen bietet *GlobalAncova*, siehe Kapitel 4. Insgesamt 185 pathways werden mit diesem Verfahren auf differentielle zeitliche Trends untersucht. Dies sind alle KEGG pathways (http://www.genome.jp/kegg/, siehe Abschnitt 3.2), denen mindestens ein Gen aus dem Experiment zugeordnet

werden kann.

- Vergleich der Expressionsprofile in **Lymphknoten und Mandeln**: Um das Verhalten von Metastasezellen in Lymphknoten zu untersuchen, wurden RNA Expressionsprofile in Lymphknoten- und zum Vergleich in Mandelzellen mit Hilfe von Microarrays erfasst (Martens u. a., 2006). Der Datensatz besteht aus jeweils 10 Beobachtungen aus Lymphknoten und Mandeln und 22283 probesets. Mit Zweistichproben-t-Tests wird auf differentielle Expression zwischen Lymphknoten und Mandeln getestet. Wiederum wird GlobalAncova für die Analyse von 191 erhältlichen KEGG pathways verwendet.

- Vergleich zwischen **AML-Patienten mit oder ohne Mutation des NPM-Gens**: Aus einem recht umfangreichen Datensatz (Metzeler u. a., 2008) werden für dieses Beispiel 25 Patienten zufällig ausgewählt. Davon tragen 14 eine Mutation im NPM-Gen. 22283 probesets werden mit t-Tests auf differentielle Expression hinsichtlich des Mutationsstatus' untersucht, 183 pathways werden mit GlobalAncova entsprechend getestet.

- Vegleich zwischen **therapie-assoziierter und de-novo AML**: Die Fragestellung dieser unveröffentlichten Untersuchung beschäftigt sich damit, ob es möglich ist, anhand des Genexpressionsprofils zu erkennen, ob die Leukämie de novo oder aufgrund einer Therapie (Strahlen oder Chemotherapie) entstanden ist. Aus dem genetisch heterogenen Patientengut werden Paare mit ähnlichen Karyotypen gebildet. Die Unterschiede zwischen de novo und therapie-assoziierter AML werden mit 10 solchen Paaren für 54675 probesets mit Hilfe von gepaarten t-Tests anlaysiert. Wie in den anderen Beispielen werden 183 pathways mit GlobalAncova getestet. Die Paarung kann dabei mit Hilfe einer zusätzlichen Kovariablen berücksichtigt werden.

Die Ergebnisse der Analysen sind jeweils geordnete Listen von Genen beziehungsweise Gengruppen. Die folgenden Vergleiche zwischen genweiser und Gruppenanalyse sollen einen rein deskriptiven Charakter haben und sind nicht weiter in Richtung statistischer Aussagekraft vertieft. Die Güte der Übereinstimmung der Ergebnisse für die beiden Normalisierungen kann man wie in Abbildung 3.1 darstellen. Für die Anzahlen der „besten“ Gene (Gengruppen), die aus der Analyse nach RMA Normalisierung resultieren (x-Achse), zeigt die y-Achse die entsprechenden Anzahlen der Überschneidungen mit der Liste der besten Gene (Gengruppen) nach vsn Normalisierung. Je näher die Kurven an der Hauptdiagonalen verlaufen, desto besser ist also die Übereinstimmung der beiden Analysen. Es zeigt sich, daß die Linien der Gruppenanalyse fast durchweg oberhalb derer der genweisen Analyse liegen. Auch die Kurven der genweisen Auswertung enden natürlich letztlich wieder an der Hauptdiagonalen, hier werden die Linien aber nur bis zu den ersten 1000 Genen gezeichnet. Aufgrund der enorm unterschiedlichen Anzahlen von Tests ist diese Darstellung allerdings nicht geeignet, die beiden Konzepte zu vergleichen – die Kurve der Gengruppenanalyse *muß* bereits nach knapp 200 Tests wieder mit der Hauptdiagonalen zusammenfallen, so

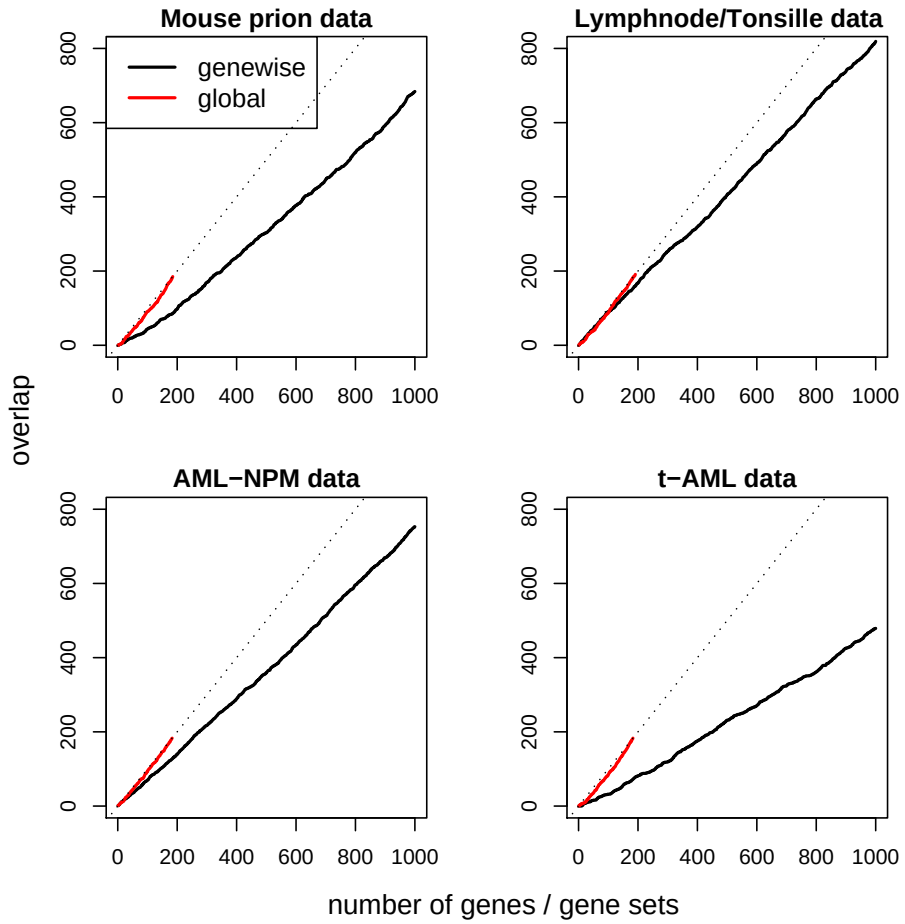


Abbildung 3.1: Übereinstimmung zwischen Ergebnissen aus genweiser (schwarz) und Gengruppenanalyse (rot) bei unterschiedlicher Normalisierung von vier Beispieldatensätzen.

daß sie fast zwangsweise über der genweisen Kurve verläuft.

Gerechter wird der Vergleich, wenn die Kurve der Gruppenanalyse künstlich auf die Länge der genweisen Analyse ausgeweitet wird. Bei m Genen und G pathways enthält ein pathway durchschnittlich m/G Gene. Jede Gengruppe wird $[m/G]$ mal wiederholt, so daß die Liste die gleiche Länge wie die der Einzelgene bekommt. Zeichnet man für diese Listen die Überschneidungen ergibt sich Abbildung 3.2. Auch in dieser Darstellung erscheint die Gruppenanalyse zumindest in drei von vier Datenbeispielen der genweisen Auswertung gleichwertig oder sogar überlegen.

Da die kompletten Kurven an der Hauptdiagonalen starten und enden, egal wie viele Tests betrachtet werden, könnte man auch die Fläche unter der Kurve als Maß für die Güte der Übereinstimmung zwischen den Analysen nach RMA und vsn Normalisierung betrachten. Im optimalen Fall beträgt die Fläche die Hälfte des Quadrats mit Fläche m^2 (bzw. G^2). Im

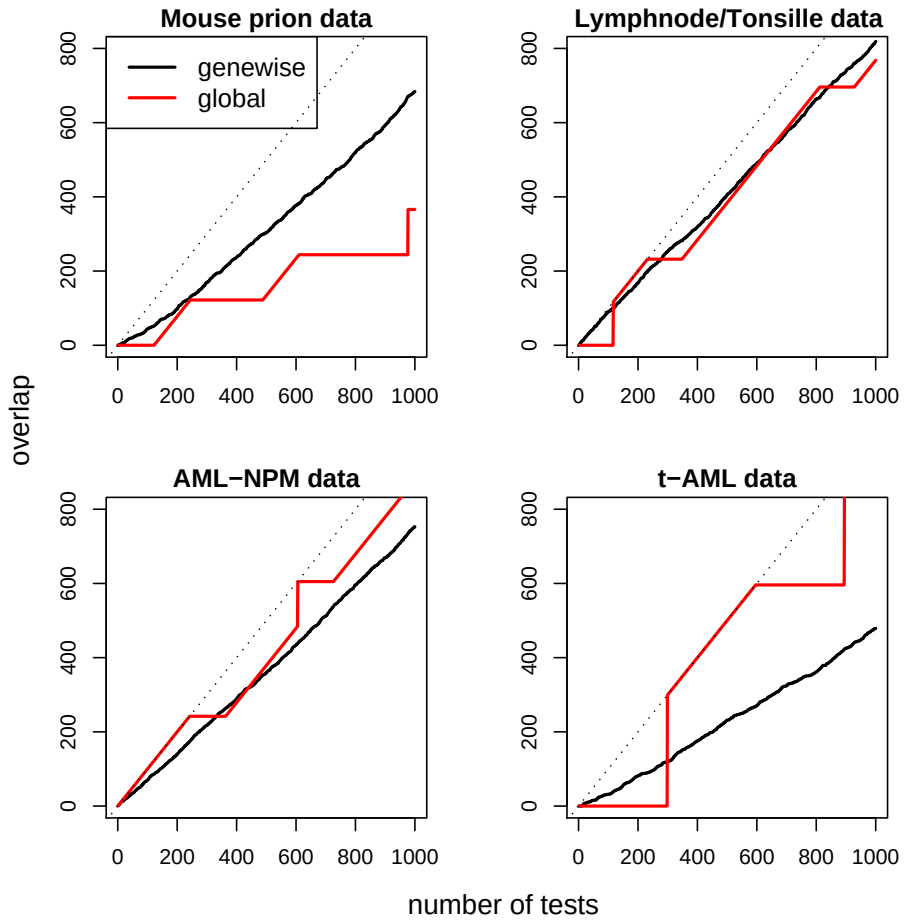


Abbildung 3.2: Übereinstimmung zwischen Ergebnissen aus genweiser (schwarz) und Gengruppenanalyse (rot) bei unterschiedlicher Normalisierung von vier Beispieldatensätzen. Die Kurve der Gruppenanalyse wurde an die Länge der genweisen Kurve angeglichen.

schlechtesten Fall hat die Liste der einen Auswertung die genau gegensätzliche Reihenfolge der anderen. Die Kurve verläuft dann bis zur Hälfte der Tests bei 0 und steigt danach mit Steigung 2 an. Die Fläche unter der Kurve beträgt in diesem Fall ein Viertel der des gesamten Quadrats. Es läßt sich dementsprechend ein Maß mit Werten zwischen 0.25 und 0.5 berechnen. Je näher der Wert bei 0.5 liegt, desto stärker ist die Überschneidung zwischen den beiden Auswertungen. Tabelle 3.1 zeigt diese Maß für die genweise und die Gruppenanalyse für die vier betrachteten Datensätze. Die Gruppenanalyse (mit der tatsächlichen Kurve aus Abbildung 3.1) weist in jedem Fall einen höheren Flächenanteil unterhalb der beschriebenen Kurven auf. Wie gesagt ist dies eine rein deskriptive Feststellung. Für eine statistische Aussage bräuchte man wie üblich Konfidenzintervalle oder ähnliches, um einen tatsächlichen Unterschied belegen zu können.

Datensatz	genweise Analyse	Gruppenanalyse
Mouse prion data	0.441	0.448
Lymphnode/Tonsille data	0.432	0.463
AML-NPM data	0.447	0.472
t-AML data	0.411	0.439

Tabelle 3.1: Anteile der Flächen unter den Kurven aus Abbildung 3.1 an den gesamten Flächen der entsprechenden Quadrate.

Neben der wohl besseren Robustheit der Gruppenanalyse bietet sie weitere praktische Vorteile gegenüber der genweisen Auswertung. Es wurde bereits die schlechte Interpretierbarkeit von Genlisten angesprochen. Vordefinierte Gruppen von Genen beinhalten per Definition mehr Information, da sie aus über Jahrzehnte hinweg gesammeltem biologischen Wissen über Zusammenhänge zwischen Genen resultieren. Die Aussage über die Relevanz einer bestimmten Gengruppe ist somit biologisch gehaltvoller und leichter interpretierbar als über die Relevanz eines einzelnen Gens.

Für die Methodik bietet die Analyse von Gengruppen im Vergleich zu Genen den Vorteil, daß in der Regel viel weniger Tests durchgeführt werden müssen und somit die Korrektur für multiples Testen nicht so gravierend ausfällt. In manchen Studien werden aufgrund der Adjustierung keine signifikant differentiell exprimierten Gene nachgewiesen. Die Gruppenanalyse bietet dennoch die Möglichkeit, gezielt die biologischen Hintergründe eines Unterschieds im Phänotypen zu ergründen. Beispiele hierfür liefern Grond-Ginsbach u. a. (2008) und Barry u. a. (2005). Allgemein können auch für die Adjustierung bei Gengruppen die in Abschnitt 2.3.2 beschriebenen Fehlerkonzepte und Adjustierungsverfahren angewendet werden.

3.2 Arten von Gengruppen

Zunächst stellt sich die Frage, welche Gengruppen betrachtet beziehungsweise wie sie definiert werden können. In vielen Fällen wird man auf 'vorgegebene' Gruppen zurückgreifen, die das Resultat langjähriger Forschung sind und somit das aktuelle biologische Wissen über Zusammenhänge zwischen Genen repräsentieren.

Gene, die über Regulationsmechanismen miteinander interagieren, werden zu *pathways* zusammengefaßt. Allgemein ist ein pathway eine Reihe von molekularen Interaktionen und Reaktionen, die in Form eines Netzwerks dargestellt werden können. Es gibt bereits große Sammlungen von pathways, beispielsweise die KEGG Datenbank (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/, Kanehisa u. a., 2006).

Eine weitere umfangreiche Zusammenstellung von Gengruppen bietet die *Gene Ontology*, die von einem Konsortium überprüft und weiter entwickelt wird (Ashburner u. a., 2000,

http://www.geneontology.org/). Sie stellt eine hierarchische Sammlung biologischer Begriffe zur Beschreibung von Genen und Genprodukten aus den drei Bereichen *Biologischer Prozeß*, *Molekulare Funkion* und *Zelluläre Komponente* dar. Jedem Begriff werden entsprechende Gene zugeordnet, so daß sich wiederum Gengruppen ergeben. Die spezielle Struktur der Gene Ontology (GO) sowie Probleme und Möglichkeiten bei der Analyse von GO-Gruppen werden in Kapitel 5 behandelt.

Neben funktionellen Gengruppen ist auch eine Gruppierung hinsichtlich der Lokalisierung der Gene im Erbgut denkbar. Zum Beispiel können alle Gene eines bestimmten Abschnitts auf einem Chromosom von Interesse sein.

Neben solchen vorgegebenen Gengruppen können durch Literaturrecherche oder explorative Analysen neue Sammlungen definiert werden. Eine interessante Anwendungsmöglichkeit der Gruppenanalyse ist beispielsweise die Validierung bereits veröffentlichter *Gensignaturen*. Anstatt die Auswertung einer Forschungsgruppe an neuen Daten reproduzieren zu wollen, und dabei sicherlich nicht die identische Gensignatur wieder zu finden, kann mit einem geeigneten Gruppentest direkt der Nutzen der veröffentlichten Genliste überprüft werden.

3.3 Gene Set Enrichment und holistische Verfahren

Es gibt mittlerweile eine Fülle von Methoden für die Analyse von Gengruppen. Grundsätzlich liegen diesen Verfahren zwei verschiedene Strategien zugrunde. In wohl den meisten Fällen geht der Gruppenanalyse eine Suche nach einzelnen interessanten Genen voraus. Erst in einem zweiten Schritt wird versucht, die biologischen Zusammenhänge der selektierten Gene zu ergründen, indem diese mit funktionellen Gengruppen in Verbindung gebracht werden. Und zwar stellt man die Frage, ob eine Gengruppe mit interessanten Genen *angereichert* (*enriched*) ist. Methoden, die diese Strategie verfolgen, nennen wir demnach Zwei-Schritt- oder *Gene Set Enrichment Verfahren*. Die zweite Strategie geht im Gegensatz dazu direkt von den Gengruppen aus, ohne vorangehende genweise Analyse. Die globalen Expressionsprofile innerhalb der Gruppen werden untersucht. Es handelt sich also um einen (einschrittigen) *holistischen Ansatz*.

Die beiden Strategien unterscheiden sich auch in der Behandlung der Gene, die nicht zu der jeweiligen betrachteten Gruppe G gehören. Bei den Gene Set Enrichment Verfahren wird über die Signifikanz einer Anreicherung von G mit interessanten Genen dadurch entschieden, daß G in Bezug gesetzt wird zu allen übrigen Genen $\bar{G}$ (z.B. zu allen Genen, die mit dem entsprechenden microarray gemessen werden). Die Nullhypothese lautet

H_0: Die Gene in G sind höchstens so oft differentiell exprimiert wie Gene in $\bar{G}$.

Bei holistischen Verfahren wird in der Regel kein Vergleich mit einer Gen-Referenzpopulation gezogen. Man betrachtet die Gengruppen per se. Hier lautet die Nullhypothese

H_0: Keine Gene in G sind differentiell exprimiert.

Goeman und Bühlmann (2007) sprechen deshalb von einer Einteilung in *kompetitive* (*competitive*) einerseits und *in sich geschlossene* (*self-contained*) Verfahren andererseits.

Die Verschiedenartigkeit der beiden Ansätze zeigt sich auch in den zugrunde liegenden Verteilungen unter der Nullhypothese. Bei den meisten Gene Set Enrichment Methoden wird geprüft, ob die betrachtete Gengruppe extrem ist im Vergleich zu zufällig zusammen gestellten Genmengen. Das H_0-Modell beruht also auf *Genrandomisierung*. Bei holistischen Verfahren dagegen wird die Außergewöhnlichkeit des Gengruppen-Expressionsprofils hinsichtlich einer zufälligen Auswahl von Beobachtungen (arrays, Patienten), also einer Stichprobe im klassischen Sinne, getestet. Man würde die entsprechende Nullverteilung durch *Permutation der Beobachtungen* simulieren. Es gibt jedoch auch holistische Methoden, die mit Genrandomisierung arbeiten können, und umgekehrt Gene Set Enrichment Ansätze, bei denen der klassische Permutationsansatz möglich ist.

3.4 Überblick über derzeitige Methoden

Die beiden Strategien Gene Set Enrichment und holistischer Ansatz werden nochmals ausführlich in Kapitel 6.1.1 verglichen. In diesem Abschnitt werden die gängigsten Verfahren zur Gengruppenanalyse im Einzelnen kurz beschrieben. Tabelle 3.2 zeigt eine Übersicht über die aufgeführten Methoden.

3.4.1 Gene Set Enrichment Verfahren

Gene Set Enrichment Verfahren fragen nach der *Anreicherung* einer Gengruppe G mit interessanten, zum Beispiel differentiell exprimierten, Genen. In einem ersten Schritt erfolgt eine genweise Analyse, durch die die Gene in eine Ordnung gebracht werden können, zum Beispiel gemäß t-Teststatistiken oder p-Werten. Der zweite Schritt besteht aus einer Analyse hinsichtlich der Überrepräsentation der interessanten Gene in den Gengruppen. Dabei gibt es zwei Strategien. Bei der ersten wird eine Liste $\mathcal{L}$ interessanter Gene definiert. Es wird angenommen, daß ein Zusammenhang zwischen $\mathcal{L}$ und der vordefinierten Gengruppe G besteht, wenn besonders viele Gene aus $\mathcal{L}$ in G enthalten sind (oder äquivalent dazu viele Gene aus G der Liste $\mathcal{L}$ angehören). Bei der zweiten Strategie wird keine Einteilung in interessante und nicht interessante Gene vorgenommen. Stattdessen werden die Ränge der Gene analysiert. Haben viele Gene aus G hohe Ränge, so wird die Gengruppe als angereichert angesehen.

Tests für Kontingenztafeln

Viele Gene Set Enrichment Methoden basieren darauf, eine Liste $\mathcal{L}$ mit L interessanten Gene mit Gengruppen G in Verbindung zu bringen. Zunächst muß also die Liste $\mathcal{L}$ definiert werden. Wenn differentielle Expression von Interesse ist, kann dies mit Verfahren wie

Methode	Autoren	Hypothese	Permutationen
Tests für Kontingenztafeln	z.B. Beissbarth und Speed (2004) Falcon und Gentleman (2007)	kompetitiv	Gene
GSEA	Mootha u. a. (2003) Subramanian u. a. (2005)	kompetitiv	(Gene) / Beob.
t-Test für genweise Statistiken	Tian u. a. (2005) Alexa und Rahnenführer (2007)	kompetitiv	Gene
Category	Gentleman und Falcon (2007)	komp. / in sich geschl.	Gene / Beob.
Mittelwert genweiser Statistiken	Tian u. a. (2005)	in sich geschlossen	Beobachtungen
SAM-GS	Dinu u. a. (2007)	in sich geschlossen	Beobachtungen
Restandardisierung	Efron und Tibshirani (2007)	komp. + in sich geschl.	Gene + Beob.
Hotelling's T^2	Kong u. a. (2006) Song und Black (2007)	in sich geschlossen	Beobachtungen
Pathway activity levels	Tomfohr u. a. (2005)	in sich geschlossen	Beobachtungen
globaltest	Goeman u. a. (2004)	in sich geschlossen	Beobachtungen
GlobalAncova	Mansmann und Meister (2005) Hummel u. a. (2008a)	in sich geschlossen	Beobachtungen

Tabelle 3.2: Übersicht über Verfahren für die Gengruppenanalyse.

in Abschnitt 2.3.1 genannt geschehen, zum Beispiel mit t-Tests. Die Gene werden gemäß der resultierenden genweisen Statistiken oder p-Werte geordnet. Allgemein ist die im Folgenden beschriebene Analyse nicht auf differentielle Expression beschränkt. Man kann die Gene je nach Fragestellung beispielsweise auch nach ihrer mittleren Expression über alle arrays sortieren. Nun muß ein Schwellenwert für die Einteilung in „interessant“ und „nicht interessant“ gewählt werden. Zu diesem Zweck werden die genweisen Tests üblicherweise zunächst für multiples Testen korrigiert. Die Liste $\mathcal{L}$ besteht dann aus allen Genen mit adjustierten p-Werten $p_j^{adj} < \alpha$. Oder aber man wählt ad hoc zum Beispiel die „besten 100“ Gene.

Von den insgesamt m Genen im Experiment erweisen sich also L als interessant. Wir betrachten eine Gengruppe G mit m_G Genen, von denen x aus der Liste der L interessanten Gene stammen. Es stellt sich nun die Frage, ob x eine extrem große Anzahl an interessanten (z.B. differentiell exprimierten) Genen ist in Hinblick auf die Anzahlen L und m der gesamten Population. Die Expression der Gengruppe wird verglichen mit der Expression aller übrigen Gene. Demnach liegt hier die kompetitive Gruppenteststrategie vor. Die Situation entspricht einem Urnenmodell: die Urne enthält m Kugeln, L davon sind „rot“. Es werden m_G Kugeln zufällig gezogen. Wie hoch ist die Wahrscheinlichkeit, dabei x rote Kugeln zu

ziehen? Diese Wahrscheinlichkeit ist gegeben durch die hypergeometrische Verteilung

$$P(X = x|m, L, m_G) = \frac{\left(\begin{array}{c} L \\ x \end{array}\right)\left(\begin{array}{c} m-L \\ m_G - x \end{array}\right)}{\left(\begin{array}{c} m \\ m_G \end{array}\right)}.$$

Der p-Wert bezüglich der Anreicherung entspricht der Wahrscheinlichkeit x oder noch mehr rote Kugeln, also interessante Gene in der Gengruppe zu beobachten

$$p = P(X \geq x|m, L, m_G) = 1 - P(X \leq x-1|m, L, m_G) = 1 - \sum_{\nu=1}^{x-1} \frac{\left(\begin{array}{c} L \\ \nu \end{array}\right)\left(\begin{array}{c} m-L \\ m_G - \nu \end{array}\right)}{\left(\begin{array}{c} m \\ m_G \end{array}\right)}. \tag{3.1}$$

Falcon und Gentleman (2007) verwenden diese Gene Set Enrichment p-Werte.

Äquivalent zu obigen Berechnungen ist die Verwendung des exakten Tests von Fisher. Er begründet sich in der Darstellung des Urnenmodells als Kontingenztafel wie in Tabelle 3.3. Mit dem Fisher-Test wird die Assoziation zwischen den beiden Eigenschaften 'Gen ist interessant' und 'Gen gehört der Gengruppe an' getestet. Der Fisher-Test wird häufig für

	$\in$ Gengruppe	$\notin$ Gengruppe	
interessant	x	$L - x$	L
nicht interessant	$m_G - x$	$(m - m_G) - (L - x)$	$m - L$
	m_G	$m - m_G$	m

Tabelle 3.3: Kontingenztafel für den Zusammenhang zwischen differentieller Expression und Zugehörigkeit zur Gengruppe

die Gene Set Enrichment Analyse verwendet (Zeeberg u. a., 2003; Al-Shahrour u. a., 2004; Draghici u. a., 2003).

Für große m kann die hypergeometrische auch durch die Binomialverteilung $B(m_G, L/m)$ approximiert werden (Draghici u. a., 2003). In dieser Formulierung wird überprüft, ob die Anteile an differentiellen Genen in der Gesamtpopulation und in der Gengruppe übereinstimmen $H_0 : \pi_1 = \pi_2$ mit $\hat{\pi}_1 = \frac{L}{m}$ und $\hat{\pi}_2 = \frac{x}{m_G}$. Alternativ wird auch der χ^2-Test für den Vergleich der beiden Anteile verwendet (Draghici u. a., 2003; Beissbarth und Speed, 2004; Zhong u. a., 2004b). Weiterhin kann die Binomialverteilung durch die Normalverteilung approximiert werden. Rivals u. a. (2007) vergleichen die verschiedenen Varianten des Gruppentests und stellen dar, daß allen genannten Verfahren die hypergeometrische

als tatsächliche Nullverteilung zugrunde liegt. Deshalb werden diese Methoden im Folgenden allgemein unter 'Tests basierend auf 2×2 Kontingenztafeln bzw. dem Urnenmodell', 'Fisher-Test ähnliche Verfahren' oder 'Tests basierend auf der hypergeometrischen Verteilung' zusammen gefaßt. Stets liegt das Urnenmodell zugrunde, dessen Verteilung unter der Nullhypothese auf dem zufälligen Ziehen von Genen, also Genrandomisierung, beruht.

Ein kritischer Punkt bei dem beschriebenen Vorgehen ist stets die Definition der Liste $\mathcal{L}$ interessanter Gene. Im Falle differentieller Expression ist eine solche Einteilung in differentielle und nicht differentielle Gene sehr künstlich. Tatsächlich beobachtet man über alle Gene hinweg meist ein Kontinuum an differentieller Expression. Demnach gibt es keine natürliche Gruppierung – der Schwellenwert, der entscheidet ob ein Gen zu $\mathcal{L}$ gehört oder nicht, wird mehr oder weniger willkürlich gewählt. Eine alternative Strategie ohne Einteilung der Gene liegt in der Analyse der Genränge.

Kolmogorov-Smirnov-Rang-Statistik *(GSEA)*

Die sogenannte *Gene Set Enrichment Analysis (GSEA)* wurde erstmal von Lamb u. a. (2003) angewendet und von Mootha u. a. (2003), Barry u. a. (2005) und Subramanian u. a. (2005) weiter ausgeführt. Wie bei den Verfahren im vorigen Abschnitt wird für jedes Gen zunächst eine Statistik $x_j, j = 1, \ldots, m$ ermittelt, zum Beispiel die enstprechende (gegebenenfalls absolute) t-Statistik. Diese Statistik legt eine Ordnung der Gene fest $x_{(1)} \geq \ldots \geq x_{(m)}$. Wieder ist man nicht auf Statistiken für differentielle Expression beschränkt. Bei Lamb u. a. (2003) werden die Gene beispielsweise hinsichtlich ihrer Ähnlichkeit zum Gen Cyclin D1 geordnet. Hier gehen wir aber zumeist von der Analyse differentieller Expression aus. Im Gegensatz zur Fisher-Test Strategie muß nun kein Schwellenwert bestimmt werden, um eine Liste differentieller Gene zu definieren. Stattdessen wird die Verteilung der *Genränge* innerhalb der interessierenden Gengruppe G untersucht. Die Gengruppe ist angereichert, wenn viele Gene aus G hohe Ränge aufweisen, also in der geordneten Liste aller Gene eher „früh auftreten“. Es werden folglich die Rängeverteilungen von G und $\bar{G}$ verglichen. Demnach liegt auch hier die kompetitive Teststrategie vor. Der Vergleich von Rängeverteilungen kann über eine Kolmogorov-Smirnov ähnliche Teststatistik geschehen. Jedem Gen in der geordneten Liste wird ein score

$$s_{(j)} = \begin{cases} m_{\bar{G}}, & \text{Gen } j \in G \\ -m_G, & \text{Gen } j \in \bar{G} \end{cases}$$

zugewiesen, wobei m_G und $m_{\bar{G}}$ den Anzahlen der Gene in G beziehungsweise $\bar{G}$ entsprechen. Das Maximum der kumulativen Summe $M = max\{S_1, \ldots, S_m\}$ mit $S_j = \sum_{\nu=1}^{j} s_{(\nu)}$ gibt Aufschluß darüber wie gut sich die Gengruppe G von den übrigen Genen „abhebt“. Abbildung 3.3 zeigt zwei Beispiele der kumulativen Summe $S = (S_1, \ldots, S_m)$. In der linken Grafik sind die Gene aus G zufällig über die geordnete Liste aller Gene verteilt. Dies bewirkt ein eher „willkürliches“ auf und ab der kumulativen Summe. Das Maximum M ist dementsprechend nicht besonders hoch. Anders sieht es im rechten Bild aus. Hier haben

mehrere Gene aus G recht hohe Ränge in der geordneten Liste, das heißt sie liegen in der Grafik eher links. Diese Separierung der Gruppe G von den übrigen Genen wird in dem recht großen Wert von M deutlich.

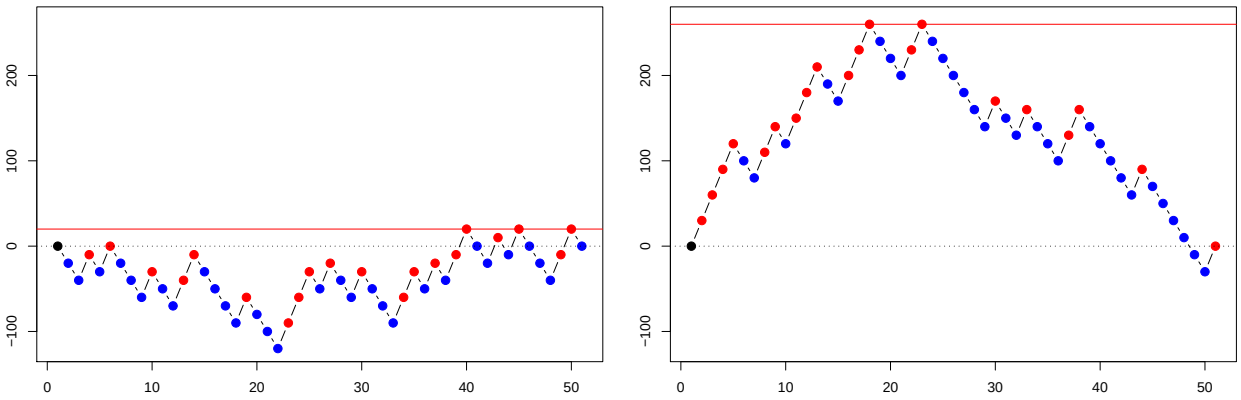


Abbildung 3.3: Kumulative Summe der scores $s_{(j)}$ in der geordneten Genliste. Die Farben geben die Positionen von Genen aus G (rot) oder $\bar{G}$ (blau) an. Die rote Linie zeigt das Maximum der kumulativen Summe. Links: Gene in G sind nicht extrem im Vergleich zu den übrigen. Rechts: Einige Gene in G haben hohe Ränge. Dies bewirkt ein großes Maximum der kumulativen Summe.

Bisher sind wir davon ausgegangen, daß nur Gene mit hohen Rängen interessant sind. Genauso sind auch Situationen denkbar, in denen eine Anreicherung der Gruppe mit Genen sowohl sehr hoher als auch sehr niedriger Ränge von Interesse sein kann. Betrachtet man zum Beipspiel t-Statistiken (nicht absolute) hätte man auf der einen Seite der geordneten Liste die hochexprimierten und auf der anderen Seite die herunter regulierten Gene. Die GSEA-Statistik kann leicht auf diese Situation angepaßt werden, indem nicht das Maximum der kumulativen Summe betrachtet wird sondern $M = \max\{\max(S), |\min(S)|\}$. Im Folgenden gehen wir meist wieder nur von dem einseitigen Fall aus.

Um zu entscheiden, welche Werte von M tatsächlich extrem sind, werden in der Regel Permutationstests durchgeführt. Ob G mehr angereichert ist mit hochrangigen Genen als eine zufällig zusammengestellte Gengruppe, würde man anhand von mehrfachen Genrandomisierungen beurteilen, bei denen jeweils die neu berechnete Statistik M_b mit dem Wert M der tatsächlichen Gengruppe verglichen wird. Das folgende Schema zeigt das Vorgehen zur Berechnung empirischer p-Werte durch Permutation der Gene.

Empirische GSEA p-Werte durch Genrandomisierung

Für die b-te Permutation, $b = 1, \ldots, B$:

1. Permutiere die scores $s_{(j)}$ in der geordneten Liste und erhalte $s^b_{(j)}$, $j = 1, \ldots, m$. Dies entspricht einer Permutation der Gene.
2. Berechne die kumulative Summe $S^b_j = \sum_{\nu=1}^{j} s^b_{(\nu)}$, $j = 1, \ldots, m$, und deren Maximum $M_b = max\{S^b_1, \ldots, S^b_m\}$.

Ein empirischer p-Wert ist gegeben durch $p = \frac{\sum_{b=1}^{B} I(M_b \geq M)}{B}$,
wobei $I(\cdot)$ die Indikatorfunktion ist.

Diesem Permutationsansatz entspricht der Kolmogorov-Smirnov Test zum Vergleich der Rängeverteilungen von G und $\bar{G}$. Die Nullverteilung des Tests entspricht der einer Genrandomisierung. Bei der Permutation der Gene werden allerdings Korrelationen nicht berücksichtigt, was gerade im Falle von Gruppen interagierender Gene ein erhebliches Problem darstellt. Die Genrandomisierung als H_0-Modell hat noch weitere problematische Aspekte, wie in Abschnitt 6.1 erläutert wird.

Aus klassischer statistischer Sicht würde man daher eher zur Permutation der Beobachtungen für die Simulation der H_0-Verteilung tendieren. Die Nullhypothese lautet in diesem Fall, daß das Expressionsprofil der Gengruppe nicht extremer ist als das der übrigen Gene. Das folgende Schema zeigt das entsprechende Vorgehen.

Empirische GSEA p-Werte durch Permutation der Beobachtungen

Für die b-te Permutation, $b = 1, \ldots, B$:

1. Permutiere die Spalten der Expressionsmatrix X und erhalte X_b. Dies entspricht einer Permutation der Beobachtungen (arrays).
2. Berechne für X_b die genweisen Statistiken x^b_j. Dadurch ergibt sich eine neue Ordnung der Gene $x^b_{(1)}, \ldots, x^b_{(m)}$.
3. Berechne für diese neue Ordnung $s^b_{(j)}$, S_b und M_b.

Ein empirischer p-Wert ist gegeben durch $p = \frac{\sum_{b=1}^{B} I(M_b \geq M)}{B}$,
wobei $I(\cdot)$ die Indikatorfunktion ist.

Die GSEA Analyse ist eine kompetitive Teststrategie, da ja stets die interessierende Gruppe G mit allen übrigen Genen verglichen wird. Für die Nullverteilung kann sowohl ein Genran-

domisierungsmodell als auch ein Modell basierend auf der Permutation der Beobachtungen heran gezogen werden. Meist wird aus oben erwähnten Gründen letzteres gewählt.

t-Test für Verteilungen der genweisen Statistiken

Ein Alternative zur Kolmogorov-Smirnov Statistik wird in Tian u. a. (2005) und Alexa und Rahnenführer (2007) vorgeschlagen. Es werden die Verteilungen der genweisen Statistiken, zum Beispiel der absoluten t-Statistiken, innerhalb der interessierenden Gruppe G und in der Gruppe aller übrigen Gene $\bar{G}$ betrachtet. Abbildung 3.4 ist angelehnt an das Beispiel aus Abbildung 3.3 und zeigt Kerndichteschätzer der entsprechenden Verteilungen. Ist die Gengruppe nicht angereichert mit differentiell exprimierten Genen, so wird man kaum einen Unterschied in den Verteilungen von G und $\bar{G}$ erkennen, so wie es in der linken Grafik der Fall ist. Auf der rechten Seite dagegen liegt die Verteilung der genweisen Statistiken aus G zu einem großen Teil rechts von der Referenzverteilung. In G sind in diesem Fall viele differentiell exprimierte Gene mit folglich hohen absoluten t-Statistiken. Als Gruppenstatistik liegt bei dieser Betrachtung der klassische Zwei-Stichproben t-Test nahe. Entsprechende t-Gruppenstatistiken und p-Werte sind in Abbildung 3.4 eingefügt.

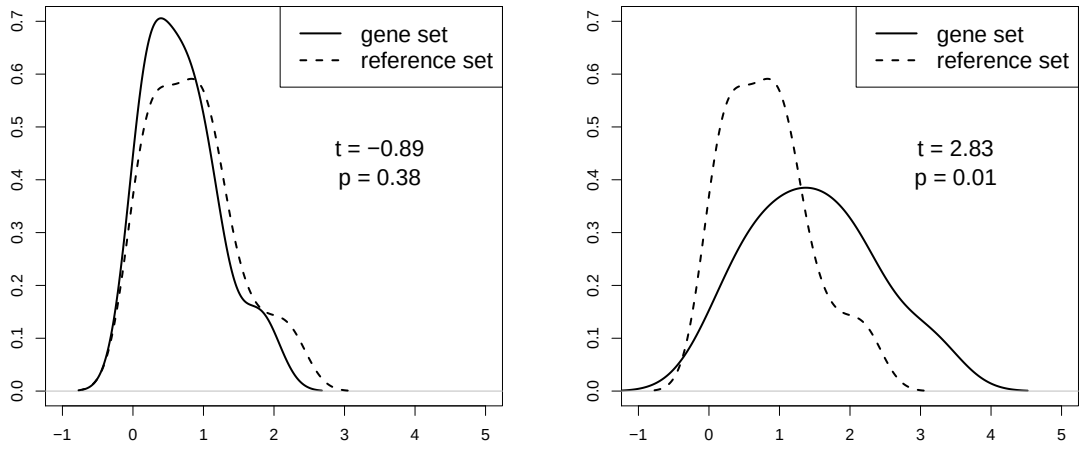


Abbildung 3.4: Kerndichteschätzer für die genweisen Statistiken in G (durchgezogene Linie) und $\bar{G}$ (gestrichelte Linie). Links: Gene in G sind nicht extrem im Vergleich zu den übrigen, die Verteilungen der genweisen Statistiken sind ähnlich. Rechts: Einige Gene in G haben hohe genweise Statistiken, deshalb unterscheiden sich die Verteilungen deutlich. Die jeweiligen t-Statistiken und p-Werte für den Vergleich der Verteilungen sind gegeben.

Neben den theoretischen sind auch permutationsbasierte p-Werte denkbar. Wiederum sind sowohl Permutationen der Beobachtungen als auch Permutationen der Gene möglich. Die theoretische Nullverteilung des beschriebenen Tests entspricht der Genrandomisierung, bei der die Zugehörigkeit der genweisen Statistiken zu G bzw. $\bar{G}$ permutiert wird. Bei einer

Permutation der Beobachtungen ergeben sich jeweils andere genweise Statistiken und somit auch andere Verteilungen dieser Statistiken. Wie bei der GSEA Analyse liegt also ein kompetitives Verfahren vor, dessen zugrunde liegendes Randomisierungsmodell unterschiedlich gewählt werden kann.

3.4.2 Holistische Verfahren

Im Gegensatz zu Gene Set Enrichment Verfahren basieren holistische Ansätze nicht auf einer vorangehenden Analyse der einzelnen Gene. Das primäre Ziel besteht nicht darin, interessante Gene aus der großen Masse aller auf einem Microarray gemessenen heraus zu filtern. Statt dessen geht man direkt von vordefinierten Gengruppen aus und untersucht diese hinsichtlich differentieller Expression.

Der *Category* Ansatz

Das Verfahren *Category* von Gentleman und Falcon (2007) beginnt mit der Berechnung genweiser Statistiken für differentielle Expression, zum Beispiel zwei-Stichproben t-Statistiken. Von den Autoren wird die Methode als Erweiterung des Gene Set Enrichment vorgestellt. Dennoch zählen wir es eher zu den holistischen Ansätzen, da hier nicht in einem ersten Schritt die einzelnen Gene im Vordergrund stehen, sondern „sofort“ aus den genweisen Statistiken durch geeignete Funktionen globale Maße für differentielle Expression in Gengruppen, den *Kategorien*, gewonnen werden.

Die Category Analyse ist im Grunde ein allgemeines Prinzip, eine Menge K vordefinierter Gengruppen mit der Expression von m Genen zu verknüpfen und so Gengruppen mit interessanten globalen Expressionsprofilen identifizieren zu können. Die Gengruppen können beliebige Kategorien sein, die weder disjunkt sein noch insgesamt alle Gene enthalten müssen. Die Verknüpfung der beiden Mengen kann durch einen Graphen dargestellt werden, dessen Knoten die Gene und Gruppen sind. Befindet sich ein Gen in einer Gruppe, so gibt es eine Kante zwischen den beiden entsprechenden Knoten. Die gleiche Information kann auch in einer $K \times m$ Inzidenzmatrix A kodiert werden. Ein Element $a_{k,j}$ der Matrix hat den Wert 1, wenn sich Gen j in Kategorie k befindet und 0 andererseits. Demnach entsprechen die Zeilensummen von A den Anzahlen der Gene m_k in jeder Gruppe. Die Spaltensummen geben an, in wie vielen Kategorien sich ein Gen befindet.

Für das Erstellen eines globalen Maßes für differentielle Expression für die einzelnen Kategorien wird zunächst der m-Vektor von univariaten Teststatistiken x benötigt. Gentleman und Falcon (2007) schlagen gewöhnliche t-Teststatistiken vor. Generell sind aber beliebige andere Statistiken, je nach Zielsetzung, denkbar. Die genweisen Statistiken werden über eine Funktion des Vektors x und der Inzidenzmatrix A zu einem k-Vektor von Gruppenstatistiken zusammengefaßt $z = f(A, x)$. Die Funktion soll so gewählt werden, daß interessante Kategorien durch extreme Werte z_k auffallen. Der Vorschlag der Autoren lautet eine gewichtete Summe der x_j aus der jeweiligen Gruppe $z = Ax/\sqrt{(m_1, \ldots, m_K)}$.

Enthält eine Kategorie G_k viele Gene, die *gleichgerichtet* differentiell exprimiert sind, so wird das entsprechende z_k einen großen positiven beziehungsweise kleinen negativen Wert annehmen. Sind zum Beispiel alle Gene der Kategorie in einer der beiden Phänotypgruppen höher exprimiert als in der anderen, sind alle t-Statistiken positiv. Dadurch wird auch die (gewichtete) Summe z_k relativ groß sein. Selbst wenn die Expressionsunterschiede nur gering sind, also keines der Gene für sich alleine betrachtet als differentiell gelten kann, kann die Gruppenstatistik dennoch extrem sein. Dies ist ein bedeutender Unterschied zu den zuvor vorgestellten Gene Set Enrichment Verfahren, bei denen solch schwache aber gleichgerichtete differentielle Expression innerhalb einer Gengruppe nicht detektiert werden kann. Andererseits könnten sich bei dieser globalen Statistik Effekte von stark hoch und stark herunter geregelten Genen gegenseitig bei der Summierung „auslöschen“. Wenn Gengruppen, die solche Gene enthalten, als interessant angesehen werden, sollten andere genweise Statistiken verwendet werden, zum Beispiel absolute t-Statistiken.

Die Verwendung gewöhnlicher t-Statistiken, wie von den Autoren vorgeschlagen, hat den Vorteil schöner Verteilungseigenschaften. Da die genweisen Statistiken x in diesem Fall approximativ normalverteilt sind, gilt dies auch für die standardisierte Summe $z \sim N(0,1)$. Diese Annahme ist allerdings nur zulässig, falls die t-Statistiken unabhängig sind, was gerade im Fall von Expressionsdaten natürlich sehr unrealistisch ist. Die Verteilungsapproximation wird verwendet, um zum Beispiel mit Hilfe von Quantil-Quantil-plots (qq-plots) heraus stechende Gruppen visuell zu detektieren. Die Signifikanz einer Gengruppe dagegen wird in der Regel über einen Permutationsansatz bestimmt.

Wie zuvor beim Kolmogorov-Smirnov Ansatz könnte man die Gene oder aber auch die Beobachtungen permutieren. Bei der Genrandomisierung lautet das H_0-Modell, daß das Expressionsprofil der betrachteten Gengruppe nicht extremer ist als das einer Gruppe aus zufällig zusammengestellten Genen. Das Schema zur Berechnung empirischer p-Werte durch Permutation der Gene ist in der folgenden Box dargestellt.

Empirische Category p-Werte durch Genrandomisierung

Für die b-te Permutation, $b = 1, \ldots, B$:

1. Permutiere die Spalten der Matrix A und erhalte A_b. Dies entspricht einer Permutation der Gene.
2. Berechne die globalen Statistiken $z_b = z_1^b, \ldots, z_K^b = f(A_b, x)$ für alle Gengruppen.

Empirische p-Werte sind gegeben durch $p_k = \frac{\sum_{b=1}^{B} I(z_k^b \geq z_k)}{B}$,
wobei $I(\cdot)$ die Indikatorfunktion ist.

Die Permutation der Gene hat unter anderem den Nachteil, daß dabei implizit von Un-

abhängigkeit zwischen Genen ausgegangen wird.

Deshalb wird eher die Permutation der Beobachtungen für die Simulation der H_0-Verteilung empfohlen. Das folgende Schema zeigt das entsprechende Vorgehen für die Category Analyse.

Empirische Category p-Werte durch Permutation der Beobachtungen

Für die b-te Permutation, $b = 1, \ldots, B$:

1. Permutiere die Spalten der Expressionsmatrix X und erhalte X_b. Dies entspricht einer Permutation der Beobachtungen.
2. Berechne für X_b die genweisen Statistiken x_b.
3. Berechne die globalen Statistiken $z_b = z_1^b, \ldots, z_K^b = f(A, x_b)$ für alle Gengruppen

Empirische p-Werte sind gegeben durch $p_k = \frac{\sum_{b=1}^{B} I(z_k^b \geq z_k)}{B}$,
wobei $I(\cdot)$ die Indikatorfunktion ist.

Die Category Analyse bietet die Möglichkeit von Vergleichen sowohl innerhalb als auch zwischen Gengruppen. Man kann fragen, ob eine bestimmte Kategorie eine bezüglich der Permutationsverteilung extreme Statistik hat. Ebenso kann man die Gruppen untereinander vergleichen und dadurch besonders heraus stechende identifizieren. Die Methode kann also sowohl als kompetitive als auch als in sich geschlossene Teststrategie verwendet werden. Ebenso kann die H_0-Verteilung auf Genrandomsierung oder auf Permutation der Beobachtungen basieren. Demnach ist die Einordnung von Category in entweder Gene Set Enrichment oder holistische Verfahren, wie zuvor auch schon erwähnt, nicht eindeutig.

Neben der vorgestellten Variante von Category schlagen Jiang und Gentleman (2007) zahlreiche Alternativen und Erweiterungsmöglichkeiten vor. Das Verfahren kann über die Wahl der Inzidenzmatrix A, der genweisen Statistiken x und der Funktion f für die Zusammenfassung der genweisen zu Gruppenstatistiken flexibel gestaltet werden. Die Einträge von A können mit Vorzeichen und Gewichten versehen werden, beispielsweise um Hoch- und Herunter Regulierung oder die Wahrscheinlichkeit für differentielle Expression widerzugeben. Die genweisen Statistiken kann man zum Beispiel über lineare Modelle oder Bayesianische Ansätze definieren. Für die globalen Statistiken stehen neben der standardisierten Summe zahlreiche andere Methoden, wie der Median- oder der Vorzeichentest, zur Verfügung.

Auch Tian u. a. (2005) diskutieren sowohl die kompetitive als auch die in sich geschlossene Hypothese. Zum Testen der ersteren schlagen sie, wie im letzten Abschnitt von 3.4.1 bereits erwähnt, einen t-Test zum Vergleich der Mittelwerte der genweisen Statistiken in

der interessierenden Gengruppe einerseits und in der Menge aller übrigen Genen andererseits vor. Um dagegen die in sich geschlossene Hypothese zu überprüfen, bei der nur die Gene innerhalb der interessierenden Gruppe eine Rolle spielen, verwenden sie den Mittelwert der genweisen Statistiken und bewerten dessen Signifikanz über Permutationen der Phänotypzugehörigkeiten. Dieser Ansatz paßt demnach genau in das Schema von Category.

Das Verfahren SAM-GS von Dinu u. a. (2007) basiert ebenfalls auf einer einfachen Zusammenfassung von genweisen scores zu Gruppenstatistiken. Hierfür werden die in Abschnitt 2.3.1 kurz beschriebenen SAM Statistiken verwendet. Die SAM Statistiken der Gene innerhalb einer interessierenden Gengruppe werden quadriert und aufsummiert. Somit erhält man wieder ein globales Maß für differentielle Expression in der Gengruppe.

Das *R* Paket `Category` bietet einige Funktionen für die im Prinzip einfach zu verwirklichende Analyse. Es beinhaltet keine komplette Prozedur – schließlich versteht sich Category auch eher als allgemeines Prinzip statt als abgegrenztes Verfahren.

Der Restandardisierungsansatz

Eine ähnliche globale Statistik für differentielle Expression in Gengruppen wie die der Category Analyse schlagen Efron und Tibshirani (2007) vor. Es handelt sich ebenfalls um die Zusammenfassung von genweisen Statistiken x_G einer Gruppe G mit m_G Genen, zum Beispiel t-Statistiken, zu einer Gruppenstatistik z_G. Man könnte diese Methode also auch als eine spezielle Form der Category Analyse bezeichnen. Für die *maxmean* Statistik von Efron und Tibshirani (2007) werden die Mittelwerte der positiven und der negativen Elemente von x_G berechnet

$$\bar{x}_G^{(+)} = \frac{1}{m_G} \sum_{Genj \in G} x_j^{(+)} \quad \text{und} \quad \bar{x}_G^{(-)} = \frac{1}{m_G} \sum_{Genj \in G} x_j^{(-)}$$

mit

$$x_j^{(+)} = \max\{x_j, 0\} \quad \text{und} \quad x_j^{(-)} = -\min\{x_j, 0\}.$$

Die *maxmean* Statistik ist das Maximum dieser beiden Mittelwerte

$$z_G^{maxmean} = \max\{\bar{x}_G^{(+)}, \bar{x}_G^{(-)}\}.$$

Die Gruppenstatistik $z_G^{maxmean}$ ist so gestaltet, daß Gengruppen mit entweder ungewöhnlich großen positiven oder negativen genweisen Statistiken oder ungewöhnlichen Statistiken in „beide Richtungen“ detektiert werden. Da für die Berechnung von $\bar{x}_G^{(+)}$ und $\bar{x}_G^{(-)}$ durch die gesamte Anzahl der Gene m_G in der Gruppe dividiert wird und diese Größen nicht etwa die Mittelwerte der (nur) positiven und negativen Werte x_j sind, ist $z_G^{maxmean}$ robust gegenüber wenigen Genen mit stark positiven oder negativen Werten x_j. Man stelle sich zum Beispiel eine Gengruppe mit 100 Genen vor, von denen 99 eine t-Statistik von -0.5 haben

und eines eine t-Statistik von 10. Die Mittelwerte der jeweils positiven und negativen Statistiken sind folglich -0.5 und 10 und somit ergäbe sich eine Gruppenstatistik von 10. Nach obiger Definition dagegen ist $\bar{x}_G^{(+)} = 10/100 = 0.1$ und $\bar{x}_G^{(-)} = -99(-0.5)/100 = 0.495$ und demnach $z^{maxmean} = 0.495$. Hier dominieren also die deutlich in der Überzahl vorhandenen negativen Elemente von x_G.

Interessanter als die Einführung einer neuen Gruppenstatistik ist bei Efron und Tibshirani (2007) die Kombination aus Genrandomisierung und Permutation von Beobachtungen zur Feststellung der Signifikanz einer Gengruppe. Nachteilig bei der Genrandomisierung ist (unter anderem) wie bereits erwähnt die Annahme der Unabhängigkeit zwischen den Genen. Aber auch die Permutation der Beobachtungen als Grundlage der Nullverteilung hat ihre Grenzen: die gesamte *Verteilung der Gruppenstatistiken* wird nicht berücksichtigt. Aus der Sicht der in sich geschlossenen Teststrategie ist dies kein Nachteil. Es wird jede Gruppe einzeln für sich betrachtet und ihre Signifikanz mit Hilfe von Permutationen der Beobachtungen bestimmt. Geht man nun einmal davon aus, daß alle Gene eines Experiments differentiell exprimiert sind, so bekäme man für obiges Permutationsschema auch für *alle* Gengruppen ein signifikantes Resultat. Das kann ein interessantes Gesamtergebnis sein. Man kann aber ebenso gut auf dem Standpunkt stehen, daß in einem solchen Fall *keine* Gruppe signifikant sein sollte, da keine im Vergleich zu allen anderen besonders heraus sticht.

Abbildung 3.5 verdeutlicht die Problematik. Es wurden hierfür Expressionswerte für 1000 Gene und 50 Beobachtungen als standardnormalverteilte Zufallszahlen simuliert. Zur Demonstration sollen alle Gene differentiell exprimiert sein. Zu diesem Zweck wird zu den Expressionswerten der ersten zehn Beobachtungen jeweils 0.5 hinzu addiert. Als genweiser score für differentielle Expression dienen absolute t-Statistiken. Die Gene werden in 50 Gengruppen á 20 Gene aufgeteilt, und die Mittelwerte der genweisen scores je Gruppe als Gruppenstatistiken ermittelt. Die durchgezogene Linie in Abbildung 3.5 zeigt das Histogramm dieser Gruppenstatistiken. Die Verteilung liegt relativ weit rechts, da aufgrund der erzeugten differentiellen Expression in allen Gengruppen Gene mit großen t-Statistiken auftreten. Permutiert man nun die Beobachtungen, so simuliert man die Nullverteilung, unter der keine differentielle Expression vorliegt. Die Verteilung der Permutations-t-Statistiken und demnach auch die Verteilung der entsprechenden Gruppenstatistiken (graues Histogramm) liegen deshalb deutlich weiter links als die der beobachteten Werte. Gemäß der Permutationsverteilung würde man für fast alle Gengruppen die Nullhypothese ablehnen. Da keine Gruppe extremer ist als die anderen könnte man hier gegen die Permutation der Beobachtungen zur Simulation der Nullverteilung beziehungsweise gegen die Nullhypothese 'die jeweilige Gengruppe enthält keine differentiell exprimierten Gene' argumentieren. Stattdessen würde man eher formulieren 'die jeweilige Gengruppe enthält keine stärker differentiell exprimierten Gene als andere Gruppen'.

Zum Prüfen der letzteren Hypothese schlagen Efron und Tibshirani (2007) ein Kombination aus beiden Permutationsstrategien vor. Dadurch ergibt sich gleichermaßen eine Mi-

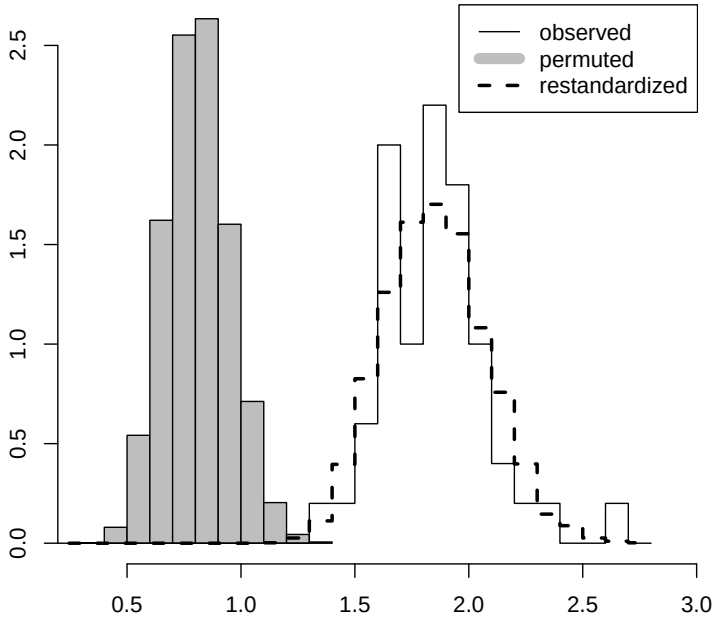


Abbildung 3.5: Verteilung von 50 tatsächlichen Gruppenstatistiken $z_G = \sum_G x_j/m_G$ (durchgezogene Linie) sowie Verteilungen der Permutationsstatistiken z_G^* (graues Histogramm) und der restandardisierten Statistiken z_G^{**} (gestrichelte Linie).

schung aus der kompetitiven und der in sich geschlossenen Teststrategie. Zunächst werden auf Permutationen der Beobachtungen basierende Gruppenstatistiken z^* berechnet. Diese werden gemäß Mittelwert μ^* und Standardabweichung σ^* der Permutationsverteilung standardisiert. Zusätzlich wird die Statistik nun noch *restandardisiert* gemäß Mittelwert $\mu^\dagger$ und Standardabweichung $\sigma^\dagger$ der Verteilung, die für die Gruppenstatistiken aus einer Randomisierung der Gene resultiert

$$z_G^{**} = \mu^\dagger + \frac{\sigma^\dagger}{\sigma^*}(z_G^* - \mu^*).$$

In unserem Beispiel beschreibt die durch Restandardisierung simulierte Nullverteilung (gestrichelte Linie in Abbildung 3.5) gut die Verteilung der tatsächlichen Statistiken. Nur Gengruppen, die sich deutlich von allen übrigen abheben, sind gemäß dieser Nullverteilung signifikant.

Wählt man einfache Gruppenstatistiken z_G, so lassen sich die für die Restandardisierung benötigten Parameter relativ einfach berechnen. Zum Beispiel im Falle des Mittelwertes $z_G = \sum x_j/m_G$ erhält man $\mu^\dagger$ und $\sigma^\dagger$ als Mittelwert und Standardabweichung aller genweisen scores x_j, sowie μ^* und σ^* als Mittelwert und Standardabweichung über alle Gene

und Permutationen x_j^*. Bei komplexeren Gruppenstatistiken wird dagegen eine genestete Simulation benötigt. Restandardisierte p-Werte ergeben sich bei B Permutationen durch

$$p_G = \frac{\sum_{b=1}^{B} I(z_G^{**} \geq z_G)}{B}.$$

Die beiden Strategien zur Permutation der Gene einerseits und der Beobachtungen andererseits beantworten unterschiedliche Fragestellungen und es liegen somit verschiedene Nullhypothesen zugrunde. Demnach stellt sich die Frage, wie die Mischung der beiden Konzepte im Restandardisierungsansatz tatsächlich sinnvoll interpretiert werden kann. Weiterhin erwähnen die Autoren, daß die Methode hauptsächlich bei Unabhängigkeit zwischen den Genen von Vorteil ist. Diese Annahme ist, wie bereits öfters erwähnt, in den wenigsten Fällen gerechtfertigt.

Hotelling's T^2 Test

Kong u. a. (2006) und Song und Black (2007) stellen eine klar in sich geschlossene Teststrategie vor, da ihre Gruppenstatistik nur anhand der Gene innerhalb der Gruppe ermittelt wird, und zwar wird eine globale *Hotelling's* T^2 Statistik berechnet. Im Gegensatz zum Category oder Restandardisierungsansatz werden hier nicht lediglich genweise Statistiken geeignet zu Gruppenstatistiken zusammen gefaßt. Der große Vorteil dieser multivariaten Methode ist, daß die Korrelationsstruktur der Gene mit berücksichtigt werden kann. Sei X die $m_G \times n$ Expressionsmatrix (Zeilen = Gene, Spalten = Beobachtungen) einer Gengruppe G mit m_G Genen, deren Expression für n Beobachtungen gemessen wurde. Wir betrachten einen Zwei-Gruppenvergleich, wobei n_1 Beobachtungen zur einen und n_2 Beobachtungen zur anderen Phänotypgruppe gehören. Mit $X_{p,i}$ wird der m_G-dimensionale Expressionsvektor der i-ten Beobachtung in Gruppe p ($p = 1, 2$) bezeichnet. Hotelling's T^2 Statistik lautet

$$T^2 = \frac{n_1 n_2}{n}(\bar{X}_1 - \bar{X}_2)^t S^{-1}(\bar{X}_1 - \bar{X}_2),$$

wobei $\bar{X}_p = 1/n_p \sum_{i=1}^{n_p} X_{p,i}$ der Mittelwertsvektor in Gruppe p ist. S bezeichnet die gepoolte Kovarianzmatrix $S = ((n_1 - 1)S_1 + (n_2 - 1)S_2)/(n - 2)$, wobei $S_p = 1/(n_p - 1) \sum_{i=1}^{n_p}(X_{p,i} - \bar{X}_p)(X_{p,i} - \bar{X}_p)^t$. Unter der Nullhypothese $H_0 : \bar{X}_1 = \bar{X}_2$ und der Annahme gleicher Kovarianzen ist die Statistik F-verteilt. Da im Allgemeinen nicht von gleichen Varianzen ausgegangen werden kann, wird die Signifikanz des Tests über einen Permutationsansatz bewertet.

Die Hotelling's T^2 Statistik kann nur direkt angewendet werden, wenn die Anzahl der Gene in der betrachteten Gengruppe kleiner ist als die Stichprobengröße ($m_G < n - 1$). Wenn $m_G \geq n - 1$ ist eine Modifikation notwendig, da die Kovarianzmatrix S in diesem Fall singulär ist. Kong u. a. (2006) und Song und Black (2007) diagonalisieren S, indem sie die Daten durch *Hauptkomponentenanalyse (principal components analysis)* auf einen orthogonalen Unterraum projizieren. Nach dieser Transformation sind die Koordinaten unkorreliert und jede Hauptkomponente hat Varianz eins. Mit der Hauptkomponentenanalyse

wird die neue Datenmatrix X' berechnet

$$X' = D^{-\frac{1}{2}} U^t X,$$

wobei die diagonale Matrix D und die orthogonale Matrix U durch die Zerlegung der Kovarianzmatrix $S = UDU^t$ bestimmt werden.

Trotz der Dimensionsreduktion durch die Hauptkomponentenanalyse kann der Test von sehr großen Gengruppen problematisch sein. Der Hotelling's T^2 Test ist außerdem auf die Analyse differentieller Expression, also den Vergleich von zwei klinischen Gruppen, beschränkt. Die Autoren schlagen für den Fall komplexerer Fragestellungen als mögliche Erweiterung multivariate Varianzanalyse (MANOVA) vor.

Gruppentest basierend auf *pathway activity levels*

Eine weitere Möglichkeit der Gengruppenanalyse bietet das Verfahren von Tomfohr u. a. (2005). Es basiert darauf, die „Aktivität“ einer gegebenen Gengruppe innerhalb der verschiedenen Beobachtungen mit Hilfe von Singulärwertzerlegung zu bestimmen. Diese *activity levels* bilden dann die Grundlage für die Analyse der interessierenden phänotypischen Struktur der Patienten. Beispielsweise wird ein Zwei-Gruppenvergleich über einen gewöhnlichen t-Test hinsichtlich der activity levels durchgeführt.

Zunächst wird die $m \times n$ Expressionsmatrix X standardisiert, so daß die Expressionswerte Mittelwert 0 und Varianz 1 über alle Beobachtungen haben. Wie im Verfahren des vorigen Abschnitts wird X für jede Gengruppe G auf die m_G zugehörigen Gene, also auf die entsprechenden Zeilen, eingeschränkt. Für die reduzierte $m_G \times n$ Matrix X_G wird eine Singulärwertzerlegung

$$X_G = WDC$$

berechnet. Dabei sind die Spalten von W die Eigenvektoren von X_G und D ist eine Diagonalmatrix mit den entsprechenden Eigenwerten. Für die Definition der activity levels wird lediglich derjenige Eigenvektor $w = (w_1, \ldots, w_{m_G})$, genannt *Metagen*, in Betracht gezogen, der mit dem größten Eigenwert λ assoziiert ist. Die Motivation hierfür ist, daß dieser erste Eigenvektor den Hauptteil der Variabilität in den Daten erklärt. Die activity levels entsprechen der zugehörigen Zeile von C. Sie können auch geschrieben werden als gewichtete Summe der standardisierten Expressionswerte der einzelnen Gene in G, wobei die Gewichte durch das erste Metagen w gegeben sind. Als activity level c_i für die i-te Beobachtung ergibt sich

$$c_i = \frac{1}{\lambda} \sum_{j=1}^{m_G} w_j x_{ji}.$$

Statt der $m_G \times n$ Matrix X_G von Expressionswerten liegt nun der n-Vektor c von activity levels vor. Dieser Vektor wird für die Analyse der klinischen Fragestellungen verwendet. Für die Frage nach differentieller Expression bietet sich der übliche t-Test an. Es können auch komplexere Modelle, beispielsweise mit Hilfe von Varianz- oder Korrelationsanalysen

betrachtet werden. Die Autoren weisen zudem auf die Ausweitungsmöglichkeit der Methode auf andere Bereiche wie Genregulation, Clusteranalyse und Klassifikation hin.

Der *globaltest*

Das Verfahren *globaltest* von Goeman u. a. (2004) (und Goeman u. a., 2006) ist ein globaler Test für die Assoziation zwischen der Expression einer Gruppe von Genen und einer klinischen Variablen. Dabei wird stets nur das globale Expressionsprofil innerhalb der interessierenden Gengruppe betrachtet, alle übrigen Gene des Experiments spielen dafür keine Rolle. Es handelt sich also um eine klar in sich geschlossene Teststrategie. Wie andere holistische Methoden zielt er nicht nur darauf ab, Gruppen mit sehr extremen Genen zu identifizieren, sondern es können auch Gruppen mit vielen mäßig differentiell exprimierten Genen detektiert werden. Geht man von spezifisch interagierenden Genen innerhalb der funktionellen Gruppen aus, so sind gerade solche Genmengen besonders interessant, die als Ganzes betrachtet Unterschiede im Phänotypen erklären können, auch wenn die einzelnen Gene nicht herausragend differentiell sind. Ein signifikantes Ergebnis des globaltest kann so interpretiert werden, daß die Gene in der betrachteten Gruppe im Mittel mit der klinischen Variablen in Beziehung stehen. Diese Beziehung kann sowohl positiv als auch negativ sein (Hoch- oder Herunterregulierung).

Das Modell des globaltest beschreibt die *Prädiktion* der klinischen Variablen Y anhand der Genexpressionsdaten X. Die entsprechende Nullhypothese lautet demnach

$$H_0 : P(Y|X, C) = P(Y|C),$$

wobei C weitere Kovariablen sein können, die für die Vorhersage von Y von Bedeutung sind. Für den Zusammenhang zwischen X, C und Y wird ein generalisiertes lineares Modell verwendet

$$E(y_i|\beta) = h^{-1}\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right).$$

Dabei ist h eine *link Funktion* (z.B. die logit Funktion) und β_j ist der Regressionskoeffizient für Gen $j, j = 1, \ldots, m$. Es wird angenommen, daß die Regressionsparameter eine gemeinsame Verteilung mit Erwartungswert 0 und Varianz τ^2 besitzen. Demnach ist die Nullhypothese $H_0 : \beta_1 = \ldots = \beta_m = 0$ äquivalent zu $H_0 : \tau^2 = 0$. Die Hypothese wird mittels eines *score Tests* überprüft. Die Teststatistik Q kann sowohl als Summe über die Gene als auch als Summe über die Beobachtungen dargestellt werden. In der ersten Notation kann Q als Mittelwert über die entsprechenden genweisen Statistiken Q_i betrachtet werden

$$Q = \frac{1}{m}\sum_{j=1}^{m} Q_j = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{\mu_2}(X_j^t(Y-\mu))^2,$$

wobei μ und μ_2 der Erwartungswert und des zweite zentrale Moment von Y unter H_0 sind. Die zweite Schreibweise

$$Q = \frac{1}{\mu_2}\sum_{i=1}^{n}\sum_{l=1}^{n} R_{il}(Y_i-\mu)(Y_l-\mu)$$

zeigt, daß globaltest interpretiert werden kann als Überprüfung, ob Subjekte (Patienten) mit ähnlichen Genexpressionsprofilen auch ähnliche klinische Parameter aufweisen. Wenn nämlich die Kovarianz R der Expression zwischen den Beobachtungen mit der Kovarianz $(Y - \mu)(Y - \mu)^t$ des klinischen responses korreliert, so nimmt die Statistik Q einen hohen Wert an. Der globaltest ist nicht auf den Vergleich zweier klinischer Gruppen beschränkt. Ebenso kann die Zielvariable multikategorial oder stetig sein oder Überlebenszeiten darstellen.

Das R Paket `globaltest` beinhaltet drei Möglichkeiten zur Bewertung der Signifikanz der globalen Statistik Q. Die volle asymptotische Verteilung der Teststatistik steht für die Berechnung von p-Werten zur Verfügung. Stattdessen kann man auch eine Gamma-Approximation dieser Verteilung sowie einen Permutationsansatz wählen. Dem Permutationstest liegt die Randomisierung der Beobachtungen zugrunde. Laut einer Simulationsstudie von Liu u. a. (2007) liefert die asymptotische Verteilung zu konservative p-Werte, mit der Gamma-Approximation wird dagegen das α-Niveau nicht immer eingehalten. Sie schlagen deshalb die Verwendung des Permutationsansatzes vor.

Kapitel 4

GlobalAncova

Neben dem globaltest von Goeman u. a. (2004) ist *GlobalAncova* von Mansmann und Meister (2005) ein weiterer globaler Test auf differentielle Expression in funktionellen Gengruppen. Das Verfahren wurde im Rahmen dieser Arbeit weiter entwickelt (Hummel u. a., 2008a) und als *R* Paket im *Bioconductor* (http://www.bioconductor.org) öffentlich zugänglich gemacht. Die Beschreibung des *R* Paketes in Form einer *Vignette* findet sich im Anhang A.

4.1 Grundlagen

4.1.1 Globaler F-Test für differentielle Expression

Das Ziel von GlobalAncova ist ganz allgemein, Zusammenhänge zwischen der Expression X einer Gruppe von Genen mit einer interessierenden Variablen Y zu quantifizieren. Zunächst betrachten wir wie in den vorigen Kapiteln den einfachen Fall der differentiellen Expression zwischen zwei Gruppen von Individuen, zum Beispiel Patienten mit zwei verschiedenen Unterarten eines bestimmten Karzinoms. Das heißt Y ist eine binäre Variable. Wie beim globaltest (Goeman u. a., 2004) lautet auch bei GlobalAncova die Nullhypothese 'kein Gen in der Gengruppe ist differentiell exprimiert'. Allerdings betrachten wir nicht die Genexpression als Prädiktor für die klinische Variable, sondern fragen umgekehrt danach wie Y die Genexpression X beeinflußt. Der Ausgangspunkt von GlobalAncova ist der Vergleich der Expressionsmittelwerte in den verschiedenen klinischen Gruppen. Dies führt zu einem Varianzanalyse-Ansatz. Da auch für Kovariablen adjustiert werden kann, erklärt sich die Bezeichnung 'ANCOVA', und da es sich um einen globalen Test für mehrere Gene gleichzeitig handelt ergibt sich der Name 'GlobalAncova'.

Dem globalen Test liegen genweise lineare Modelle für jedes Gen $j, j = 1, \ldots, m$ zugrunde, die den systematischen Teil μ_j der Expressionswerte $x_j = (x_{1j}, \ldots, x_{nj})^t$ und das zufällige 'Rauschen' ε_j quantifizieren

$$x_j = \mu_j + \varepsilon_j = \beta_{0j} + \beta_{1j} Y + \varepsilon_j.$$

Der genspezifische Expressionsmittelwert wird durch β_{0j} beschrieben und β_{1j} ist der Effekt der klinischen Variablen auf die Expression. Zusätzlich zur interessierenden Variablen Y können Kovariablen wie Alter, Geschlecht etc. in das Modell eingeführt werden. Sämtliche Variablen werden wie üblich zu einer $n \times (p+1)$ Designmatrix D zusammen gefaßt (p: Anzahl Variablen), wobei $d_{i\nu}$ die Ausprägung der ν-ten Variablen von Beobachtung i kodiert. Allgemein läßt sich das genweise lineare Modell demnach schreiben als

$$x_j = D\beta_j + \varepsilon_j = \begin{pmatrix} 1 & d_{11} & \dots & d_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & d_{n1} & \dots & d_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_{0j} \\ \vdots \\ \beta_{pj} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1j} \\ \vdots \\ \varepsilon_{nj} \end{pmatrix}.$$

Für die Fehlerkomponente ε_j wird ein Mittelwert von 0 und eine diagonale Kovarianzmatrix $Cov(\varepsilon_j) = \sigma_j^2 \cdot I_n$ angenommen, wobei I_n die n-dimensionale Einheitsmatrix ist.

Generell setzt sich die Designmatrix D zusammen aus einem Teil interessierender Variablen D_1 und einem Teil mit Intercept und zusätzlichen Kovariablen D_0. Im Fall des Zweigruppenvergleichs und ohne weitere Kovariablen entspricht D_1 der Gruppenvariablen $D_1 = Y$ und D_0 dem Intercept $D_0 = \mathbf{1}_n$. Um die Relevanz der interessierenden Variablen für die beobachtete Expression zu überprüfen, wird dieses *volle Modell* (VM) mit einem *reduzierten Modell* (RM) verglichen, welches nur die übrigen Variablen D_0 enthält. Wir betrachten also die genweisen Modelle

$$\begin{aligned} &\text{volles Modell:} && E(x_j) = (D_0, D_1) \begin{pmatrix} \beta_{j,0} \\ \beta_{j,1} \end{pmatrix} = D_{VM}\beta_{j,VM} \\ &\text{reduziertes Modell:} && E(x_j) = D_0\beta_{j,0} = D_{RM}\beta_{j,RM}. \end{aligned}$$

Der Vergleich der beiden Modelle erfolgt über das Prinzip der Extra-Residuenquadratsumme (Draper und Smith, 1998). Die Residuenquadratsumme quantifiziert wie gut ein zugrunde liegendes Modell die Daten anpaßt. Sind die interessierenden Variablen für die Expression von Bedeutung, so wird das volle Modell deutlich besser zu den Daten passen als das reduzierte und somit wird sich ein Unterschied in den entsprechenden Residuenquadratsummen zeigen. Die Residuen $\hat{\varepsilon}_{j,VM}$ des vollen Modells erhält man durch

$$\hat{\varepsilon}_{j,VM} = x_j - \hat{x}_j = x_j - D_{VM}\hat{\beta}_{j,VM} = x_j - D_{VM}(D_{VM}^t D_{VM})^{-1} D_{VM}^t x_j = (I_n - H_{VM})x_j,$$

wobei H_{VM} die übliche *hat-Matrix* der linearen Modelle ist. Die Residuenquadratsumme $RSS_{j,VM}$ (*residual sum of squares*) lautet

$$RSS_{j,VM} = \hat{\varepsilon}_{j,VM}^t \hat{\varepsilon}_{j,VM} = x_j^t (I_n - H_{VM}) x_j. \tag{4.1}$$

Residuen und Residuenquadratsumme des reduzierten Modells berechnen sich analog. Die Extra-Residuenquadratsumme ist definiert als die Differenz der Residuenquadratsummen aus reduziertem und vollem Modell

$$RSS_{j,extra} = RSS_{j,RM} - RSS_{j,VM}. \tag{4.2}$$

Die genweisen Informationen werden in GlobalAncova zu einem globalen linearen Modell für den gesamten $m \cdot n$-Expressionsvektor $\tilde{X}$ zusammen gefaßt

$$\tilde{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} D & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & D \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \tilde{D}\tilde{\beta} + \tilde{\varepsilon}. \tag{4.3}$$

Die Designmatrix $\tilde{D}$ ist blockdiagonal und hat Dimension $(mn) \times (p+1)m$. Der $(p+1)m$-Vektor $\tilde{\beta}$ beinhaltet die genspezifischen Regressionsparameter. Die Fehlerkomponente $\tilde{\varepsilon}$ hat Mittelwert 0 und eine positiv definite Kovarianzmatrix $Cov(\tilde{\varepsilon}) = \tilde{\Sigma}$. Wie zuvor am genweisen Modell veranschaulicht werden ein volles und ein reduziertes Modell auf diese Weise beschrieben und entsprechende Residuen berechnet. Die Residuenquadratsummen berechnen sich wiederum durch $RSS_{VM} = \hat{\tilde{\varepsilon}}^t_{VM}\hat{\tilde{\varepsilon}}_{VM}$, beziehungsweise einfacher ausgedrückt als Summe der genweisen Residuenquadratsummen $RSS_{VM} = \sum_{j=1}^{m} RSS_{j,VM}$ (RSS_{RM} analog).

Das Prinzip der Extra-Residuenquadratsummen erlaubt die Konstruktion einer multivariaten GlobalAncova Teststatistik

$$F = \frac{RSS_{RM} - RSS_{VM}}{RSS_{VM}} \cdot \frac{df_{VM}}{df_{extra}},$$

wobei $df_{VM} = (n - p_{VM})m$ mit $p_{VM} = p + 1$ und $df_{extra} = (n - p_{RM})m - (n - p_{VM})m = (p_{VM} - p_{RM})m$ mit p_{RM} = Anzahl Parameter im reduzierten Modell, also Anzahl Spalten von D_0. Unter der Annahme unabhängiger, homoskedastischer Gene $\tilde{\varepsilon} \sim N(0, \sigma^2 I_{mn})$ ist die Statistik F-verteilt mit Freiheitsgraden df_{extra} und df_{VM} unter der Nullhypothese 'die interessierenden Variablen haben keinen Einfluß auf die globale Expression'. Da bei Genexpressionsdaten diese Annahmen üblicherweise nicht erfüllt sind, werden GlobalAncova p-Werte anhand eines Permutationsansatzes oder durch Approximation der Verteilung berechnet. Darauf wird in Abschnitt 4.2 näher eingegangen.

4.1.2 Analyse komplexer linearer Modelle

Durch die Verwendung linearer Modelle stellt GlobalAncova ein sehr flexibles Werkzeug für die Analyse komplexer Fragestellungen dar. Es müssen lediglich geeignete Designmatrizen $D_{VM} = (D_0, D_1)$ und $D_{RM} = D_0$ entworfen werden. Es kann wie bereits erwähnt stets für wichtige Kovariablen adjustiert werden. Hierfür erweitert man die Matrix D_0 um die entsprechenden Faktoren. Man ist nicht auf die Analyse differentieller Expression zwischen zwei klinischen Gruppen beschränkt, das heißt D_1 muß keine binäre Variable sein, sondern kann beliebig viele Gruppen kodieren. Desweiteren kann D_1 eine lineare Variable darstellen, wie zum Beispiel die Dosis eines verabreichten Medikaments. Es können auch Interaktionen zwischen mehreren Variablen getestet werden. Dies ermöglicht beispielsweise die Analyse von Unterschieden in zeitlichen Expressionsverläufen zwischen verschiedenen klinischen

Gruppen. Eine weitere interessante Anwendung ist die Untersuchung von Co-Expression. Hierbei wird der Zusammenhang der Expression eines oder mehrerer Gene, zum Beispiel aus einer Gensignatur, mit der Expression einer funktionellen Gengruppe betrachtet. Viele weitere Modell-Szenarien sind denkbar. Einige davon sind in Tabelle 4.1 dargestellt.

Design	Volles Modell (D_0, D_1)	Reduziertes Modell D_0
Klinische Gruppen	$\sim group + cov$	$\sim cov$
Dosiseffekt	$\sim dose + cov$	$\sim cov$
Gruppen-Dosis-Interaktion	$\sim group * dose + cov$	$\sim group + dose + cov$
Zeitliche Trends in Gruppen	$\sim group * time + cov$	$\sim group + time + cov$
Gen-Gen-Interaktion (Co-Expression)	$\sim gene + cov$	$\sim cov$
Differentielle Co-Expression	$\sim group * gene + cov$	$\sim group + gene + cov$
Metaanalyse	$\sim group * dataset$	$\sim dataset$

Tabelle 4.1: Modell-Szenarien für GlobalAncova. Für die Definition der vollen und reduzierten Modelle wird die in R *übliche Formelschreibweise verwendet. Zusätzliche Kovariablen sind mit 'cov' bezeichnet.*

Aus diesen Beispielen wird klar, daß die Bezeichnung 'globaler Test auf differentielle Expression in Gengruppen' für GlobalAncova eigentlich nicht ausreichend ist. Vielmehr kann der Effekt einer beliebigen Struktur innerhalb der Beobachtungen auf die globale Expression einer Gengruppe untersucht werden. In Kapitel 7 werden einige der hier aufgeführten Modellierungsansätze an praktischen Beispielen ausführlich dargestellt.

4.2 Bestimmung der Signifikanz

Eine F-Verteilung der GlobalAncova Statistik kann nur angenommen werden, wenn man von Unabhängigkeit und gleichen Varianzen zwischen den Expressionen der einzelnen Gene ausgeht. Da diese Annahmen unrealistisch sind, muß die Signifikanz einer Gengruppe auf andere Weise bewertet werden. Dies kann durch einen Permutationsansatz geschehen. Desweiteren kann die Verteilung des Zählers der F-Statistik geeignet approximiert werden.

4.2.1 Empirische Signifikanz durch Permutationstest

Der Permutationsansatz basiert auf Randomisierung der Beobachtungen. Es werden die Zeilen der Designmatrix permutiert und zwar nur der Spalten, die den zu testenden Variablen entsprechen, also die Zeilen von D_1. Die Zeilen von D_0 bleiben dagegen unverändert. Dadurch wird die Kovariablenstruktur der Beobachtungen beibehalten. Da $D_0 = D_{RM}$ ändert sich das reduzierte Modell bei der Permutation nicht. Demnach müssen für die Berechnung der F-Statistiken nur die Residuenquadratsummen des vollen Modells in jedem

Permutationsschritt neu ermittelt werden. Empirische p-Werte sind gegeben durch den Anteil an Permutations-F-Statistiken, die größer sind als der tatsächlich beobachtete Wert. Der GlobalAncova Permutationstest ist in der folgenden Box noch einmal schematisch dargestellt.

Empirische GlobalAncova p-Werte

Für die b-te Permutation, $b = 1, \ldots, B$:

1. Permutiere die Zeilen der Matrix D_1 und erhalte D_1^b und damit $D_{VM}^b = (D_1^b, D_0)$. Dies entspricht einer Permutation der Beobachtungen.
2. Berechne daraus die Residuenquadratsumme des vollen Modells RSS_{VM}^b und damit die F-Statistik $F_b = \frac{RSS_{RM} - RSS_{VM}^b}{RSS_{VM}^b} \cdot \frac{df_{VM}}{df_{extra}}$

Empirische p-Werte sind gegeben durch $p = \frac{\sum_{b=1}^{B} I(F_b > F)}{B}$,
wobei $I(\cdot)$ die Indikatorfunktion ist.

Simulationen zeigen, daß unter der Nullhypothese GlobalAncova bei Verwendung der empirischen p-Werte das α-Niveau in der Regel gut einhält. Üblicherweise werden 10,000 Permutationen durchgeführt, um verläßliche p-Werte zu erlangen. Dies bedeutet, besonders für große Gengruppen, einen erheblichen Zeitaufwand. Stets wird zunächst die Anzahl aller möglichen Permutationen ermittelt. Bei sehr kleinen Stichproben kann diese Anzahl kleiner sein als die vorgegebene Anzahl an Permutationen. In diesem Fall werden alle möglichen Permutationen durchgeführt.

4.2.2 Asymptotische Nullverteilung

Wir gehen von dem mn-Expressionsvektor $\tilde{X}$ aus (4.3) mit Kovarianz $\tilde{\Sigma}$ aus. Der Zähler der GlobalAncova Statistik lautet

$$RSS_{RM} - RSS_{VM} = \tilde{X}^t(I_{nm} - \tilde{H}_{RM})\tilde{X} - \tilde{X}^t(I_{nm} - \tilde{H}_{VM})\tilde{X} = \tilde{X}^t(\tilde{H}_{VM} - \tilde{H}_{RM})\tilde{X}.$$

Dies ist eine quadratische Form, deren Verteilungsfunktion sich laut Robbins und Pitman (1949) und Kotz u. a. (1967) schreiben läßt als

$$F(\alpha, y) = P\left(\sum_{\nu=1}^{mn} \alpha_\nu Z_\nu^2 \leq y\right),$$

wobei $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{mn}$ die Eigenwerte von $\tilde{\Sigma}(\tilde{H}_{VM} - \tilde{H}_{RM})$ sind und die Z_ν's unabhängig standardnormalverteilte Zufallsgrößen. Die Verteilung kann approximiert werden durch eine Reihe von χ^2-Verteilungen

$$F_{mn}(\alpha, y) = \sum_{k=1}^{\infty} c_k \chi^2_{mn+2k}(y/\beta) \approx \sum_{k=1}^{mn} c_k \chi^2_{mn+2k}(y/\beta).$$

Die Koeffizienten werden in Kotz u. a. (1967) gegeben als

$$\begin{aligned} \beta &= \min_{\nu=1,\ldots,mn}\{\alpha_\nu\} = \alpha_{mn} \\ c_0 &= \prod_{\nu=1}^{mn}(\beta/\alpha_\nu)^{\frac{1}{2}} \\ c_k &= \tfrac{1}{k}\sum_{r=0}^{k-1} d_{k-r}c_r \\ d_k &= \tfrac{1}{2}\sum_{\nu=1}^{mn}(1-\beta\alpha_\nu^{-1})^k. \end{aligned}$$

Die Werte $\alpha_1, \ldots, \alpha_{mn}$ berechnen sich schließlich durch

$$\alpha_1, \ldots, \alpha_{mn} = \{\rho_i \cdot \lambda_j; i = 1, \ldots, n; j = 1, \ldots, m\},$$

wobei $\rho_1, \ldots, \rho_n$ die Eigenwerte der hat-Matrixdifferenz $(H_{VM} - H_{RM})$ und $\lambda_1, \ldots, \lambda_m$ die Eigenwerte der $m \times m$ Expressionskovarianzmatrix Σ sind.

Die Schwierigkeit bei der Approximation ist die Schätzung der Expressionskovarianzmatrix Σ, da m sehr groß sein kann und insbesondere $m > n$. In diesem Fall hat die empirische Kovarianzmatrix keinen vollen Rang. Ledoit und Wolf (2004) schlagen eine *shrinkage* Schätzung

$$\Sigma_\varphi = \varphi \cdot T + (1 - \varphi) \cdot U$$

vor mit Schrumpffaktor φ, shrinkage Ziel T und unrestringiertem Schätzer U. Wenn nur wenige Gene korrelieren, kann für T eine Diagonalmatrix mit ungleichen Varianzen gewählt werden. Für eine solches shrinkage Ziel lautet das optimale φ

$$\varphi^* = \sum_{i \neq j} Var(s_{ij}) / \sum_{i \neq j} s_{ij},$$

wobei s_{ij} ein unverzerrter Schätzer für die Kovarianz zwischen Genen i und j ist. Die Berechnung der shrinkage Schätzung erfolgt mit Hilfe der Funktion `cov.shrink` aus dem *R* Paket `corpcor` (Schäfer u. a., 2006). Bei sehr großen Gengruppen ist die Schätzung allerdings sehr langsam, beziehungsweise gar nicht mehr möglich.

In Simulationen zeigt sich, daß unter Verwendung der approximativen p-Werte das α-Niveau oft deutlich überschritten wird, der Test also antikonservativ ist. Abbildung 4.1 zeigt die logarithmierten asymptotischen und permutationsbasierten p-Werte für zwei Simulationsszenarien im Vergleich. Es wurden für jeweils 100 Datensätze mit 200 Genen und 40 Beobachtungen $N(0, 1)$-verteilte Expressionswerte simuliert. Mit GlobalAncova wurde auf differentielle Expression zwischen den ersten und den letzten 20 Beobachtungen getestet. Es wurde aber keine differentielle Expression vorgegeben, das heißt es wurde die Nullhypothese simuliert. In der linken Grafik gibt es keine Abhängigkeiten zwischen Genen. Für das Szenario der rechten Grafik wurde eine gleichmäßige Korrelation von $\rho = 0.2$ zwischen je zwei Genen eingeführt. Die vertikalen und horizontalen Linien geben das 5%-Niveau an.

Gemäß der Permutations-p-Werte werden in beiden Fällen vier von 100 Tests abgelehnt, mit den asymptotischen p-Werten erhält man dagegen acht und sieben falsch positive Ergebnisse und damit eine Überschreitung des α-Niveaus. Dieses antikonservative Verhalten wird auch durch die Simulationsstudie von Liu u. a. (2007) bestätigt. Der Grund hierfür liegt wahrscheinlich in der nicht optimalen Wahl des shrinkage Ziels T in der Approximation. Die Funktion `cov.shrink` müßte für die Verwendung in GlobalAncova entsprechend angepaßt werden. Beispielsweise könnte die Annahme von *compound symmetry*, die kurze Blöcke interagierender Gene darstellt, besser geeignet sein als die Diagonalmatrix.

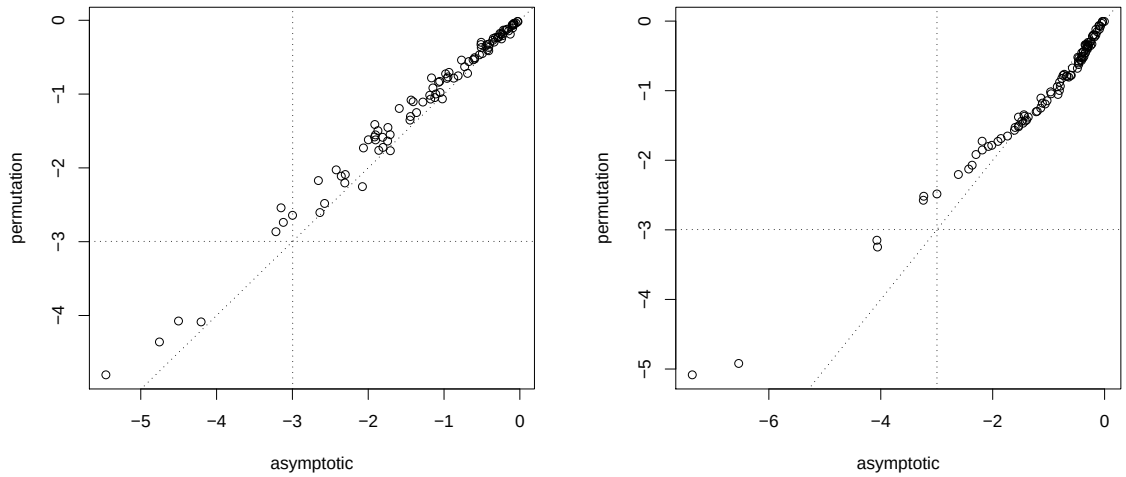


Abbildung 4.1: Vergleich der logarithmierten asymptotischen (x-Achse) und permutationsbasierten (y-Achse) p-Werte von GlobalAncova. Berechnet für 100 simulierte Datensätze á 200 Genen und 40 Beobachtungen; ohne differentielle Expression. Links: ohne Abhängigkeiten zwischen den Genen, rechts: mit gleicher Korrelation $\rho = 0.2$ zwischen je zwei Genen. Vertikale und horizontale Linien markieren das 5%-Niveau.

4.3 Programmierung und graphische Darstellung

4.3.1 Das *R* Paket `GlobalAncova`

Das Programm GlobalAncova ist in einem *R* Paket zusammen gefaßt, das beim *Bioconductor* unter http://www.bioconductor.org frei zugänglich ist. Das Paket beinhaltet den programmierten code, ein ausführliches Manual (*Vignette* 'GlobalAncova.pdf', siehe Anhang A), html-Hilfeseiten für jede Benutzerfunktion und kleine Beispieldatensätze für die mögliche direkte Anwendung der Funktionen in den Hilfeseiten und der Vignette. Der Programm code besteht aus folgenden Komponenten

- Kernfunktion `GlobalAncova()`, sowie nicht für den Benutzer bestimmte Hilfsfunktionen, zum Beispiel für die Berechnung der permutationsbasierten und der asymptotischen p-Werte
- Funktionen `Plot.genes()` und `Plot.subjects()` für diagnostische Plots, siehe Abschnitt 4.3.3
- Funktion `GlobalAncova.closed()` für die Adjustierung beim Testen einiger (weniger) Gengruppen mit Hilfe der *geschlossenen Testprozedur* (Marcus u. a., 1976)
- Funktion `GAGO()` zum Finden signifikanter Subgraphen innerhalb der *Gene Ontology* mit Hilfe der *focus level* Methode (Goeman und Mansmann, 2008), siehe Abschnitt 5.2
- Tests und Graphikfunktionen von sequentiellen und Typ III Zerlegungen (Searle, 1971, implementiert von Ramona Scheufele); diese Funktionen sind in einer eigenen Vignette 'GlobalAncovaDecomp.pdf' beschrieben

Ein Aufruf der GlobalAncova Hauptfunktion könnte beispielsweise so aussehen

```
GlobalAncova(xx = exprs(eset), formula.full = ~ group + sex,
             formula.red = ~ sex, model.dat = pData(eset), method = "both",
             perm=1000, test.genes = 1:100)
```

Dabei ist `eset` ein Objekt der Klasse `"ExpressionSet"`, die in *R* für die Speicherung von Expressionsdaten üblich ist. Die zwei wichtigsten Komponenten, nämlich die Expressionsmatrix und die klinischen Informationen zu den Beobachtungen, erhält man über die Funktionen `exprs()` und `pData()`, wie es in obigem Beispiel ausgeführt ist. Die zu vergleichenden vollen und reduzierten Modelle werden über die Parameter `formula.full` und `formula.red` definiert. In diesem Fall wollen wir die Expressionsunterschiede zwischen verschiedenen Phänotypgruppen (`group`) testen und dabei für das Geschlecht (`sex`) adjustieren. Die Option `method` gibt an, ob permutationsbasierte oder asymptotische oder, wie hier, beide p-Werte berechnet werden sollen. Mit `perm` kann die Anzahl der Permutationen manipuliert werden (Standardeinstellung `perm = 10000`). Es kann mit `GlobalAncova()` eine einzelne Gengruppe beziehungsweise die ganze Expressionsmatrix getestet werden. Für die Definition einer Gengruppe bietet sich der Parameter `test.genes` an, der die Gennamen oder -indizes angibt, wie hier zum Beispiel die der ersten 100 Gene. Die Funktion gibt die Namen der verwendeten und getesteten Variablen zurück, sowie eine übliche ANOVA Tabelle mit Residuenquadratsummen, Freiheitsgraden und mittleren Residuenquadratsummen des Zählers (`Effect`) und Nenners (`Error`) der GlobalAncova Statistik. Ein weiterer Teil des outputs beinhaltet die F-Statistik und die p-Werte.

```
$effect
[1] "group"
```

```
$ANOVA
              SSQ   DF        MS
Effect   93.79648  100 0.9379648
Error  1727.58686 1700 1.0162276

$test.result
              [,1]
F.value  0.9229869
p.perm   0.6820000
p.approx 0.6098411

$terms
[1] "(Intercept)" "group"       "sex"
```

Mit der Option `test.genes` können auch mehrere Gengruppen gleichzeitig getestet werden, indem man eine Liste von entsprechendene Gennamen beziehungsweise -indizes übergibt. Der output ist in diesem Fall etwas knapper, zum Beispiel

```
     genes   F.value p.perm  p.approx
[1,]   100 0.9229869  0.682 0.6098411
[2,]    50 1.1509893  0.218 0.1680226
[3,]   200 0.9956069  0.523 0.3971742
```

4.3.2 Programmierung des Permutationstests

Um beim Permutationstest Rechenzeit einzusparen, wird beim Testen mehrerer Gengruppen $G_1, \ldots, G_K$ nicht für jede Gengruppe ein eigener Permutationstest durchgeführt. Stattdessen werden für jede Permutation der Beobachtungen gleich die entsprechenden Statistiken aller Gengruppen $F^b_{G_1}, \ldots, F^b_{G_K}$, $b = 1, \ldots, B$, berechnet. Die Anzahl an benötigten Permutationen läßt sich dadurch von $B \cdot K$ auf B reduzieren.

Beim Testen funktioneller Gengruppen gibt es zudem häufig Überschneidungen zwischen den Gruppen, da Gene oft mehreren Funktionen oder Prozessen zugeordnet werden können. Wir nutzen diesen Umstand aus, um den Permutationstest noch effizienter zu gestalten. Anstatt je Permutation für jede Gengruppe einzeln die entsprechende globale Statistik $F^b_{G_k}, k = 1, \ldots, K$, zu berechnen, werden für die Vereinigung der Gene aus allen Gruppen die genweisen Residuenquadratsummen ermittelt. Für Gengruppen $G_1, \ldots, G_K$ mit Genindexmengen $\mathcal{G}_1, \ldots, \mathcal{G}_K$ betrachten wir also die Vereinigung $\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}_k$ und bestimmen B Permutationen der genweisen Nenner- und Zählerstatistiken $RSS^b_{j,VM}$ und $RSS^b_{j,extra} = RSS_{j,RM} - RSS^b_{j,VM}$, $j = 1, \ldots, |\mathcal{G}|$. Die Permutations-F-statistiken für die einzelnen Gengruppen erhält man dann einfach durch Summation der entsprechenden gen-

weisen Zähler- und Nennerstatistiken und anschließende Division

$$F^b_{G_k} = \frac{\sum\limits_{j \in \mathcal{G}_k} RSS^b_{j,extra}}{\sum\limits_{j \in \mathcal{G}_k} RSS^b_{j,VM}} \cdot df, \tag{4.4}$$

wobei *df* vereinfacht den Quotienten der entsprechenden Freiheitsgrade darstellt. Je mehr Überschneidungen zwischen den Gengruppen bestehen, desto mehr Rechenzeit kann durch diese Art der Berechnung eingespart werden.

Der Permutationstest wurde von Sven Knüppel in der Programmiersprache *C* implementiert. Diese ist, im Gegensatz zu *R*, keine Interpreter Sprache. Dadurch können insbesondere Schleifenkonstruktionen, wie sie für den Permutationstest benötigt werden, schneller abgearbeitet werden. Durch die Implementation in *C* verringert sich die Rechenzeit um etwa die Hälfte. Beim Testen einzelner Gengruppen ist bei sehr vielen Genen oder sehr großen Fallzahlen der Vorteil von *C* gegenüber *R* nicht so deutlich. Beim Testen vieler Gengruppen dagegen beobachtet man wiederum eine gute Zeitersparnis.

4.3.3 Diagnostische Graphiken

Wie zuvor beschrieben basiert GlobalAncova auf der Analyse der Extra-Residuenquadratsummen. Daher kann die Zerlegung der totalen Quadratsumme in Bezug auf die Modellkomponenten auf zweierlei Weise betrachtet werden: sowohl als Summe der genweisen Beiträge als auch als Summe der Beiträge der einzelnen Beobachtungen. Diese genweise und beobachtungsweise Sicht wird in den diagnostischen Graphen *gene plot* und *subject plot* visualisiert.

Der *gene plot*

Der gene plot zeigt den Einfluß der einzelnen Gene auf die globale Teststatistik. Je Gen wird ein Balken gezeichnet. Die Länge $b_j, j = 1, \ldots, m$ des Balkens entspricht der Extra-Residuenquadratsumme für das entsprechende Gen

$$b_j = RSS_{j,extra} = RSS_{j,RM} - RSS_{j,VM} = \sum_{i=1}^{n} (\hat{\varepsilon}^2_{ij,RM} - \hat{\varepsilon}^2_{ij,VM}).$$

Die Balken haben immer positive Länge, da ein Reduktion der Residuenquadratsumme durch das volle Modell im Vergleich zum reduzierten immer möglich ist. Die genweise Residuenquadratsumme im vollen Modell (*mean squared error*, MSE) ist als Referenzlinie zugefügt. Sie entspricht der Höhe des Balkens unter der Nullhypothese, die besagt, daß kein Zusammenhang zwischen der Expression des Gens und den interessierenden Modellvariablen besteht. Dadurch gewinnt man einen optischen Eindruck der genweisen F-Statistiken, wobei die Balkenhöhe dem Zähler und die Höhe der Referenzlinie dem Nenner entsprechen. Das Verhältnis der beiden Werte ist ein Maß für die Assoziation zwischen dem entsprechenden Gen und den untersuchten Parametern. Durch den gene plot können diejenigen Gene

detektiert werden, die für ein signifikantes Ergebnis des globalen Tests „verantwortlich“ sind. Zudem läßt sich beurteilen, ob ein kleiner p-Wert duch einige wenige stark differentielle Gene verursacht wurde, oder ob die meisten Gene moderate Expressionsunterschiede aufweisen.

Abbildung 4.2 zeigt zwei gene plots für simulierte Daten. Für die linke Graphik wurden für 30 Gene und 20 Beobachtungen Expressionswerte als standardnormalverteilte Zufallszahlen erzeugt. Es liegt also keine differentielle Expression vor. Im gene plot liegen die meisten Balken unterhalb oder in der Nähe der MSE Linie. Einige Gene haben zufällig einen relativ differentiellen Charakter – deren Balken ragen über die Referenzlinie hinaus. Dies reicht aber bei weitem nicht für ein signifikantes Ergebnis der ganzen Gengruppe (GlobalAncova p-Werte ≈ 0.5). Für die rechte Graphik sollten die ersten fünf Gene differentiell sein. Hierfür wurde zu ihren Expressionswerten bei den ersten 10 Beobachtungen der Wert 1 addiert. Bei vier der fünf Gene erreicht man tatsächlich die gewünschten Unterschiede in den Expressionsmittelwerten. (Beim vierten Gen lagen zuvor in der ersten Phänotypgruppe zufällig deutlich geringere Werte vor als in der zweiten. Durch die Erhöhung der Werte in der ersten Gruppe wurde demnach kein Gruppeneffekt erzeugt, sondern er hat sich im Gegenteil sogar aufgehoben.) Für die vier Gene zeigen sich im gene plot entsprechend große Balken. Diese wenigen Gene sind extrem genug, um ein signifikantes globales Ergebnis zu erzielen (GlobalAncova p-Werte < 0.01).

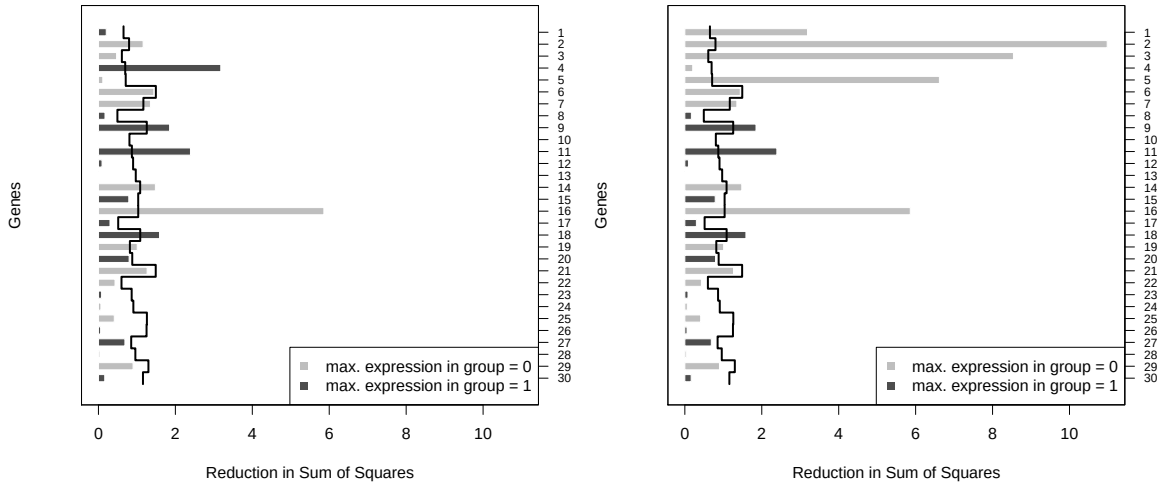


Abbildung 4.2: Gene plot für simulierte Daten (30 Gene, 20 Beobachtungen). Links: ohne differentielle Expression, rechts: die ersten 5 Gene wurden als differentiell simuliert.

Gene plots werden mit der Funktion `Plot.genes()` erzeugt. Die Funktion nimmt die gleichen Argumente für die Modelldefinition entgegen wie die Hauptfunktion `GlobalAncova()`,

sowie weitere Parameter zur Manipulation der graphischen Darstellung. Unter anderem können die Balken entsprechend der Phänotypgruppen gefärbt werden, in denen sie die höchste mittlere Expression haben (vergleiche Abbildung 4.2). Die Variable zur Bestimmung der Färbung kann vom Benutzer frei gewählt werden. Im Fall der Analyse differentieller Expression zwischen zwei Gruppen wird dies wohl meist die binäre Gruppenvariable sein, wie auch im obigen Beispiel. Desweiteren ist es möglich, die einzelnen Balkenlängen ausgeben zu lassen.

Der *subject plot*

Der subject plot ist ebenso ein Balkendiagramm, das Information über die Extra-Residuenquadratsumme pro Beobachtung gibt. Die Länge des Balkens b_i für Beobachtung i, $i = 1, \ldots, n$, entspricht der Summe der genweisen Beiträge

$$b_i = RSS_{i,extra} = \sum_{j=1}^{m} (\hat{\varepsilon}^2_{ij,RM} - \hat{\varepsilon}^2_{ij,VM}).$$

Ein langer Balken bedeutet eine gute Anpassung der jeweiligen Beobachtung durch das volle Modell. Negative Balken können hier auftreten, wenn Beobachtungen nicht in das Expressionsprofil ihrer klinischen Gruppe „passen“, das heißt wenn ihre Expressionsprofile nicht gut durch das Modell beschrieben werden. Demnach gehen kleine GlobalAncova p-Werte in der Regel mit vielen positiven Balken einher. Gibt es trotz eines signifikanten Ergebnisses große negative Balken, so können die entsprechenden Beobachtungen durch den subject plot als „Ausreißer“ identifiziert werden.

Abbildung 4.3 zeigt subject plots für die selben simulierten Daten, die im vorigen Abschnitt für die gene plots verwendet wurden. Die rechte Graphik, mit differentieller Expression, zeigt haupsächlich positive Balken, während das Bild auf der linken Seite nicht so klar ist. Wir erkennen außerdem, daß in der linken Graphik die erste Phänotypgruppe (in diesem Fall zufällig) homogenere Expressionsprofile aufweist als die zweite.

Für die Erstellung von subject plots wird die Funktion `Plot.subjects()` verwendet. Die Parameter stimmen weitgehend mit denen des gene plots überein. Eine zusätzliche Option bietet die Möglichkeit, die Beobachtungen hinsichtlich ihrer Zugehörigkeiten zu den verglichenen klinischen Gruppen zu sortieren.

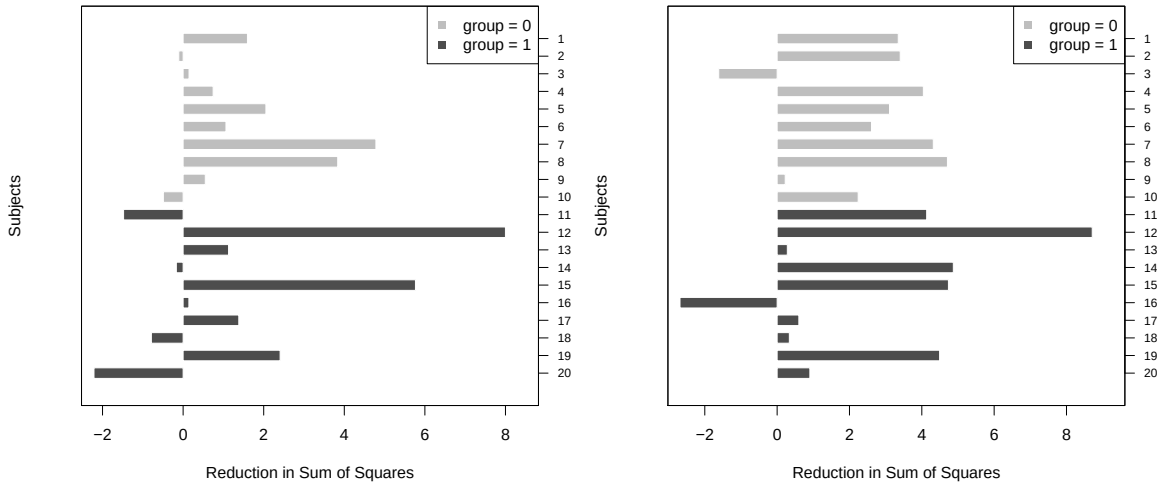


Abbildung 4.3: Subject plot für simulierte Daten (30 Gene, 20 Beobachtungen, vergleiche Abbildung 4.2). Links: ohne differentielle Expression, rechts: mit 5 differentiell exprimierten Genen.

Kapitel 5

Gene Ontology Analyse

Die im vorigen Kapitel vorgestellten Verfahren können auf beliebige Gruppierungen von Genen angewendet werden. Immer beliebter wird dabei derzeit die Analyse von *Gene Ontology* (GO) Kategorien (Ashburner u. a., 2000). Sie beschreibt für verschiedenste Organismen bekannte Funktionen und Zugehörigkeiten von Genen oder Genprodukten. Die Zuordnung (*Annotation*) von Genen zu Gene Ontology Kategorien wird laufend weiterentwickelt und von einem speziellen Konsortium überwacht, um konsistente Beschreibungen und Bezeichnungen zu gewährleisten. Die Gene Ontology hatte ihren Ursprung 1998 in der Zusammenarbeit der Genomdatenbanken dreier Modellorganismen. Heute beinhaltet und verknüpft sie sehr viele Datenbanken der wichtigsten pflanzlichen, tierischen und mikrobiellen Genome. Unter http://www.geneontology.org/ kann man die Datenbanken abfragen. Beispielsweise kann man nach allen Genprodukten suchen, die im Mausgenom etwas mit Signaltransduktion zu tun haben. Ebenso können für ein Gen alle Zugehörigkeiten zu Gene Ontology Begriffen, oder anders herum für einen Begriff alle annotierten Gene gefunden werden. Die Gene Ontology ist somit für Biologen ein wichtiges Hilfsmittel für die Recherche. Daneben wird sie immer häufiger dazu verwendet, die biologischen Hintergründe der Ergebnisse von Microarray Experimenten zu entschlüsseln. Hierfür werden die statistischen Methoden der Gengruppenanalyse benötigt.

5.1 Struktur der Gene Ontology

Die Gene Ontology besteht aus den drei Ontologien *Biologischer Prozeß (BP)*, *Molekulare Funkion (MF)* und *Zelluläre Komponente (CC)*. Zelluläre Komponenten sind Substrukturen innerhalb der Zelle wie zum Beispiel Proteinkomplexe. Eine molekulare Funktion beschreibt Aktivitäten auf molekularer Ebene wie Stoffwechselvorgänge oder Bindung und Transport von Molekülen. Ein biologischer Prozeß ist eine Reihe von Ereignissen, die durch das Zusammenspiel mehrerer molekularer Funktionen hervorgerufen werden. Dabei ist ein biologischer Prozeß nicht das gleiche wie ein *pathway*, da für einen pathway auch Interaktionen und Abhängigkeiten zwischen den einzelnen Komponenten genau beschrieben sein müßten.

Jede Ontologie stellt ein geregeltes Vokabular von biologischen Begriffen (Termen) dar. Jeder Term ist durch eine Identifikationsnummer und einen Namen, zum Beispiel 'GO:0003750' und 'cell cycle regulator', eindeutig gekennzeichnet. Die einzelnen Begriffe stehen hierarchisch miteinander in Beziehung. Es gibt die zwei Beziehungen *is-a* und *part-of*. Die Beziehung 'A is-a B' bedeutet, daß 'A' eine Subklasse von 'B' ist. Beispielsweise stehen 'nucleid acid binding' und der generelle Begriff 'binding' so in Verbindung. Eine Beziehung 'A part-of B' besagt, daß der speziellere Term 'A' ein Teil von 'B' ist, zum Beispiel ist 'nucleus' ein Teil von 'cell'. In beiden Fällen ist der spezielle Term 'A' jeweils in dem allgemeineren Term 'B' logisch enthalten. Man nennt 'A' deshalb auch ein „Kind“ von 'B'.

Diese Struktur läßt sich am besten mit Hilfe eines gerichteten azyklischen Graphen (*directed acyclic graph, DAG*) darstellen. Abbildung 5.1 zeigt einen kleinen GO Teilgraphen. Der oberste Knoten ist die Gene Ontology selbst und wird meist nicht mit abgebildet. Die eigentliche „Wurzel“ des Graphen entspricht einer der drei Ontologien, in diesem Fall 'Molekulare Funktion'. Hier werden die zwei molekularen Funktionen 'binding' und 'transcription regulator activity' gezeigt, deren „Nachfahren“ weitere Spezialisierungen dieser Begriffe sind. Wie man sieht, kann ein Knoten mehrere Kinder haben und ebenso auch mehrere Eltern. Weiter fällt auf, daß die Knoten nicht leicht in „Generationen“ eingeteilt werden können. So kommt zum Beispiel 'transcription factor activity' im linken Pfad des Graphen in der zweiten Generation nach der Wurzel, im rechten Pfad dagegen erst in der vierten. Es liegt also keine einfache Verzweigungstruktur vor.

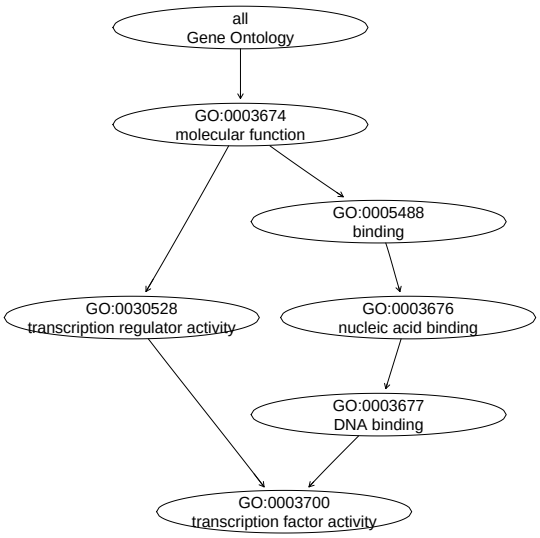


Abbildung 5.1: Ausschnitt der Ontologie 'Molekulare Funktion', dargestellt als gerichteter azyklischer Graph.

Jeder GO Term definiert eine Gengruppe, nämlich die Gruppe von Genen, die mit dem entsprechenden Begriff bekanntermaßen assoziiert sind. „Bekanntermaßen“ bezieht sich dabei immer auf den aktuellen wissenschaftlichen Kenntnisstand, der sich laufend fort entwickelt. Demnach werden die GO Gruppen mit der Zeit größer und es kommen außerdem ständig neue GO Begriffe hinzu. Derzeit enthalten die drei Ontologien folgende Anzahlen an Termen

Biologischer Prozess (BP):	13,860
Molekulare Funktion (MF):	7,825
Zelluläre Komponente (CC):	1,993

Ein Gen oder Genprodukt kann gleichzeitig in jeder der drei Ontologien aufgeführt sein. Für die Annotation von Genen zu den GO Termen gibt es verschiedene Möglichkeiten wie Literaturrecherche oder experimentelle Analyse. Die Herkunft jeder Annotation wird durch eine von 14 *evidence codes* (Tabelle 5.1) dokumentiert, so daß bei einer GO Analyse beispielsweise weniger verläßliche Quellen ausgeschlossen werden können. Bei der Verwendung von Affymetrix Microarrays ist zu beachten, daß oftmals mehrere probesets dem selben Gen entsprechen. Man muß sich also überlegen, ob jedes probeset einzeln annotiert werden soll oder ob jedes tatsächliche Gen nur einmal vorkommen darf. Hier gehen die Meinungen auseinander, und für letztere Variante ist nicht klar, nach welchen Kriterien dasjenige probeset auszuwählen ist, das das Gen repräsentieren soll.

Ein besonders wichtiger Aspekt der Gene Ontology ist die „Vererbungsregel“ (*true path rule*): ist ein Gen einem bestimmten Begriff zugeordnet, so gehört es auch allen Vorfahren dieses Begriffes an. So hat zum Beipsiel ein Gen, das mit 'nucleid acid binding' zu tun hat, auch automatisch mit dem allgemeineren Begriff 'binding' zu tun. Demnach sind GO Gruppen immer Teilmengen ihrer Eltern. Die Wurzel des GO Graphen enthält alle Gene, das heißt alle Gene eines Experiments, die mit der entsprechenden Ontologie in Verbindung gebracht werden können. Die Aufteilung aller Gene auf den GO Graphen ist nicht disjunkt, da Gene mit ganz verschiedenen Begriffen assoziiert sein können. Dadurch gibt es auch zwischen „nicht verwandten“ Termen Überschneidungen. Desweiteren ist die Aufteilung nicht komplett, das heißt nicht jedes Gen erscheint in einem Endknoten des Graphen. Alles in allem ergibt sich für die Gene Ontology eine recht komplexe Struktur, die eine große Herausforderung für eine adäquate Analyse darstellt. Insbesondere ist es derzeit nicht klar, wie eine passende Adjustierung für multiples Testen im Falle der Gene Ontology auszusehen hat.

Für die Verwendung der Gene Ontology in *R* gibt es beim *Bioconductor* das Paket `GO`, das vier mal im Jahr aktualisiert wird. Die Zuordnung von Affymetrix probesets zu GO Termen ist in den speziellen Annotationspaketen für die vielen verschiedenen arrays (verschiedene Organismen, verschiedene Teile des Genoms) enhalten.

Evidenz-code	Quelle der Annotation
IMP	gefolgert aus mutantem Phänotypen
IGC	gefolgert aus genomischem Kontext
IGI	gefolgert aus genetischer Interaktion
IPI	gefolgert aus physikalischer Interaktion
ISS	gefolgert aus Sequenzen- oder Strukturähnlichkeit
IDA	gefolgert aus direkter Untersuchung
IEP	gefolgert aus Expressionsprofil
IEA	gefolgert aus elektronischer Annotation
TAS	nachvollziehbare Aussage aus Veröffentlichung
NAS	nicht nachvollziehbare Aussage aus Veröffentlichung
RCA	gefolgert aus geprüfter rechnerischer Analyse
IC	gefolgert durch Kurator
ND	keine biologischen Daten verfügbar
NR	nicht beschrieben

Tabelle 5.1: Evidence codes der Annotation von Genen und Genprodukten zu Gene Ontology Kategorien.

5.2 Spezielle Verfahren für die Gene Ontology

In den meisten derzeit verwendeten Verfahren für die Suche nach relevanten GO Begriffen (zum Beispiel Al-Shahrour u. a., 2004; Beissbarth und Speed, 2004; Doniger u. a., 2003; Draghici u. a., 2003; Falcon und Gentleman, 2007; Zeeberg u. a., 2003; Zhong u. a., 2004a) wird auf das Urnenmodell zurück gegriffen, das die Anreicherung von GO Gruppen mit interessanten (meist differentiell exprimierten Genen) aufzeigen soll. Es werden also Gene Set Enrichment Methoden wie in Kapitel 3.4.1 beschrieben verwendet. Einen Überblick und Vergleiche der verschiedenen Verfahren bieten Khatri und Draghici (2005) und Rivals u. a. (2007). Bei all diesen Methoden wird allerdings nicht die spezielle Struktur der Gene Ontology berücksichtigt. Eine Adjustierung für multiples Testen wird in manchen Fällen vorgeschlagen. Dabei werden aber üblicherweise ebenfalls die komplexen Abhängigkeiten zwischen den GO Gruppen ignoriert.

Im folgenden Abschnitt 5.2.1 werden Verfahren vorgestellt, die ebenfalls auf Gene Set Enrichment basieren und die auf die GO Struktur eingehen. Es besteht auch das Interesse an speziellen GO Verfahren, die alternativ zum Gene Set Enrichment mit holistischen Methoden arbeiten. Erst seit kurzem werden solche Verfahren entwickelt. Ansätze hierfür werden in Abschnitt 5.2.2 beschrieben.

5.2.1 Verfahren basierend auf Gene Set Enrichment

Bei dem ersten hier vorgestellten Verfahren wird die Verteilung unter der Nullhypothese dahingehend verändert, daß die Abhängigkeiten zwischen Gengruppen, die aufgrund von Überlappungen der zugehörigen Genmengen vorliegen, mit berücksichtigt werden.

Die beiden weiteren Methoden begründen sich in der Beobachtung, daß bei einer GO Analyse mit üblichem Gene Set Enrichment die gefundenen „signifikanten“ GO Begriffe oftmals „enge Verwandte“ sind, das heißt Knoten im GO Graphen, die in direkter Verbindung stehen. Wegen der Vererbungsregel gibt es sehr starke Überlappungen zwischen aufeinander folgenden GO Knoten und demnach ist dieses Ergebnis nicht verwunderlich. Für die Interpretation kann es von Interesse sein, innerhalb solcher „signifikanten Familien“ die tatsächlich relevanten Begriffe heraus zu filtern, also solche, die für die Signifikanz der übrigen Knoten „verantwortlich“ sind. Die beiden Verfahren benutzen hierfür recht unterschiedliche Ansätze.

Hypergeometrische Verteilung mit Abhängigkeiten

Die Methode von Gold u. a. (2007) basiert auf der üblichen Idee des Gene Set Enrichment. Allerdings werden die Gengruppen nicht einzeln betrachtet, sondern alle möglichen Paare von Gruppen, deren Überlappung mit in die Verteilung unter der Nullhypothese mit eingeht. Dadurch entfällt die implizite Annahme der Unabhängigkeit zwischen den Gruppen. Bezeichnen m_G und $m_{G'}$ die Genanzahlen zweier GO Gruppen G und G'. Von insgesamt L interessanten Genen befinden sich L_G und $L_{G'}$ Gene *ausschließlich* in G beziehungsweise G' und $L_{G\cap G'}$ Gene sowohl in G als auch in G'. Man ist nun interessiert an der Verteilung von $X_G = L_G + L_{G\cap G'}$ und $X_{G'} = L_{G'} + L_{G\cap G'}$. Die gemeinsame Verteilung von $(X_G, X_{G'})$ lautet

$$P(X_G = x_G, X_{G'} = x_{G'}) = \sum_{L_{G\cap G'}=0}^{min(x_G,x_{G'})} P(L_G = x_G - L_{G\cap G'}, L_{G'} = x_{G'} - L_{G\cap G'}, L_{G\cap G'}), \quad (5.1)$$

wobei

$$P(L_G, L_{G'}, L_{G\cap G'}|L, m_G, m_{G'}, m_{G\cap G'}, m) =$$

$$= \frac{\begin{pmatrix} m_G \\ L_G \end{pmatrix}\begin{pmatrix} m_{G'} \\ L_{G'} \end{pmatrix}\begin{pmatrix} m_{G\cap G'} \\ L_{G\cap G'} \end{pmatrix}\begin{pmatrix} m - (m_G + m_{G'} + m_{G\cap G'}) \\ L - (L_G + L_{G'} + L_{G\cap G'}) \end{pmatrix}}{\begin{pmatrix} m \\ L \end{pmatrix}}.$$

Die gemeinsame Verteilung in (5.1) und deren Erwartungswert, Varianz und Kovarianz erhält man durch Ausintegrieren von $L_{G\cap G'}$.

Exakte p-Werte für die Verteilung (5.1) zu berechnen ist mühsam. Deshalb werden empirische p-Werte über Genrandomisierung ermittelt. Alternativ kann die gemeinsame Verteilung (5.1) als multivariate Normalverteilung approximiert werden. Für diese Approximation müssen allerdings große Anzahlen m_G vorausgesetzt werden. Die Autoren empfehlen ihre Verwendung nur für Gruppen mit mindestens 300 Genen. Im Falle der Gene Ontology könnte man also die meisten spezifischen Terme gar nicht analysieren.

Mit Hilfe von Simulationen und Auswertungen an realen Daten argumentieren Gold u. a. (2007), daß es nur kleine Unterschiede zwischen den Ergebnissen mit der klassischen Variante des Fisher-Tests und mit der neuen Methode gibt und somit die Berücksichtigung der Abhängigkeiten in der Nullverteilung nicht von zentraler Wichtigkeit ist. Hier ist allerdings anzumerken, daß die Autoren scheinbar für jede GO Gruppe nur die Gene als zugehörig auffassen, die direkt dem entsprechenden Begriff zugeordnet wurden. Die Vererbung der Gene an alle Elternknoten wird nicht berücksichtigt. Das heißt es liegt hier gar nicht die komplette Hierarchie der Gene Ontology vor. Dementsprechend berichten die Autoren über gar nicht allzu viele und große Überlappungen zwischen den GO Gruppen und somit auch nicht so bedeutende Abhängigkeiten.

Für eine Adjustierung für multiples Testen werden übliche Verfahren zur FWER oder FDR Kontrolle vorgeschlagen.

Dekorrelierung der Gene Ontology

Der Ansatz von Alexa u. a. (2006) zielt darauf ab, die Abhängigkeiten zwischen verwandten signifikanten GO Begriffen zu reduzieren. Dadurch werden Familien von signifikanten Knoten „aufgebrochen“ und es können zusätzlich noch weitere interessante Regionen im GO Graphen detektiert werden. Die Idee besteht darin, die Signifikanz eines Knotens auf der Basis der Signifikanz seiner Kinder zu berechnen. Ist ein spezifischer Begriff G stark angereichert mit differentiell exprimierten Genen, so wird dessen Elternknoten mit hoher Wahrscheinlichkeit ebenfalls ein signifikantes Ergebnis erzielen, weil er alle Gene aus G enthält. Aus der Sicht von Alexa u. a. (2006) ist die Signifikanz des Elternknotens lediglich bedingt durch die Signifikanz des Kindes und sollte neu bewertet werden.

In dem ersten Algorithmus *elim* geschieht dies dadurch, daß die Gene eines signifikanten Knoten in all seinen Vorfahren eliminiert werden. Zunächst wird ein Gene Set Enrichment Test (üblicherweise der Fisher-Test) für alle Endknoten der GO berechnet. Die Gene signifikanter Knoten werden aus den entsprechenden Vorfahren entfernt. Als signifikant gelten dabei in der Regel Knoten, deren Bonferroni adjustierte p-Werte $< \alpha$ sind. Der Fisher-Test wird dann für die Elterngeneration durchgeführt, nur auf Basis der dort verbleibenden Gene. Ist ein Elternknoten signifikant, so ist er es nun nicht mehr aufgrund seiner signifikanten Kinder, sondern wegen der Anreicherung seiner eigenen spezifisch annotierten Gene mit differentiellen Genen. Der Algorithmus wird iterativ fortgesetzt bis alle Knoten prozessiert sind. Abbildung 5.2 skizziert die Idee des elim Algorithmus'.

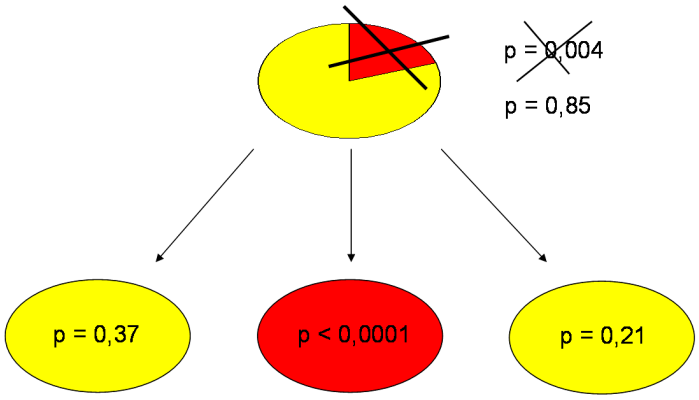


Abbildung 5.2: Skizze des elim Algorithmus': Von drei Endknoten hat der rot gekennzeichnete einen signifikanten Fisher-Test p-Wert. Die diesem Knoten zugehörigen Gene werden im Elternknoten eliminiert. Auf der Basis der verbleibenden Gene ist der Elternknoten nicht signifikant, während er nach der klassischen Methode unter Berücksichtigung aller Gene signifikant wäre (durchgestrichener p-Wert).

Die Autoren präsentieren einen weiteren Algorithmus *weight*, bei dem die Gene nicht tatsächlich eliminiert werden. Stattdessen werden den Genen eines Knotens G Gewichte zugeteilt, die von den p-Werten der verwandten Knoten abhängen. Hat G einen kleineren Fisher-Test p-Wert als all seine Kinder, das heißt G repräsentiert am besten die differentiell exprimierten Gene, so sollten die Kinderknoten nicht als signifikant detektiert werden. Deshalb werden die Gene in den Kindern heruntergewichtet. Gibt es dagegen zumindest ein Kind, das einen kleineren p-Wert hat als G, so erhalten die Gene dieses Kindes in G und allen weiteren Vorfahren kleine Gewichte. Nach jeder neuen Gewichtung werden die Knoten wieder mit dem Fisher-Test bewertet. Dieser wird hinsichtlich der Gewichtungen entsprechend angepaßt: an die Stelle der Anzahlen in der Kontingenztafel für die Eigenschaften 'Gen ist differentiell exprimiert' und 'Gen gehört zu Knoten G' (vergleiche Tabelle 3.3) treten die (aufgerundeten) Summen der entsprechenden Gengewichte. Kurz gesagt hebt der weight Algorithmus die Knoten hervor, die mehr mit differentiellen Genen angereichert sind als all ihre direkten Vor- und Nachfahren.

Wie der Ansatz mit dem klassischen Fisher-Test sind auch die Algorithmen elim und weight nicht an eine spezielle Methode für die genweise Analyse gebunden. Es genügt, eine Liste „interessanter“ Gene zu definieren. Der einfachere elim Algorithmus kann außerdem mit beliebigen anderen Gruppenstatistiken außer dem Fisher-Test angewendet werden. Der Kern dieser Methode ist lediglich die Anpassung der Genmengen innerhalb der GO Knoten. Wie die Knoten anschließend bewertet werden, kann im Prinzip frei gewählt werden. Man kann die Kolmogorov-Smirnov Statistik oder den t-Test für die Genrängeverteilung (siehe Kapitel 3.4.1) verwenden. Ebenso ist auch eine Kombination zwischen dem elim

Algorithmus und globalen Tests denkbar. Dieser Ansatz läßt sich demnach als allgemeines Prinzip für die GO Analyse mit Berücksichtigung der Graphtopologie betrachten, anstatt als ein abgeschlossenes, spezialisiertes Testverfahren. Eine ähnliche Verallgemeinerung des weight Algorithmus' ist nicht so einfach, da die Gewichtung bei der erneuten Bewertung der Knoten mit berücksichtigt werden muß.

Die Verfahren von Alexa u. a. (2006) bieten für die Praxis hilfreiche Instrumente, um interessante GO Begriffe zu detektieren. Sie sind heuristischer Natur und die zugrunde liegenden Nullhypothesen sind nicht einfach offensichtlich. Es handelt sich auch nicht um eine Adjustierung für multiples Testen. Die klassischen Gruppenteststrategien werden lediglich hinsichtlich der Abhängigkeitsstruktur der GO angepaßt.

Der *parent-child* Ansatz

Auch Grossmann u. a. (2007) beschäftigen sich mit der Problematik, daß die stark überlappenden GO Knoten nicht getrennt voneinander betrachtet werden sollten, so wie es aber in der üblichen Gene Set Enrichment Strategie gehandhabt wird. Wie im vorigen Kapitel beschrieben werden bei unabhängiger Prozessierung der GO Gruppen aufgrund der Vererbungsregel oftmals eng verwandte Knoten detektiert. Ist ein GO Knoten mit differentiellen Genen angereichert, so steigt auch für dessen Kinder die Chance eines signifikanten Testergebnisses. Aus der Sicht der Autoren ist die Signifikanz der Kinder demnach nicht immer gerechtfertigt, und es sollte hinsichtlich der Eltern-Kind-Beziehungen eine Korrektur vorgenommen werden.

Der *parent-child* Ansatz arbeitet mit dem üblichen Urnenmodell (vergleiche Kapitel 3.4.1). Allerdings wird die Anreicherung eines Knotens G hier in Bezug auf die Anreicherung seiner Eltern $pa(G)$ bewertet. Seien m_G und $m_{pa(G)}$ die Anzahlen der zu G beziehungsweise der zu den Elternknoten von G gehörenden Gene. Hat ein Knoten mehrere Eltern, so werden zwei Definitionen für $m_{pa(G)}$ vorgeschlagen:

(i) Es wird die Vereinigung $\bigcup_{G' \in pa(G)} G'$ der Genmengen aller Elternknoten als Referenzpopulation betrachtet. Demnach ist $m_{pa(G)}$ die Anzahl der Gene, die zu *mindestens einem* Elternknoten von G gehören.

(ii) Es wird der Durchschnitt $\bigcap_{G' \in pa(G)} G'$ der Genmengen aller Elternknoten als Referenzpopulation betrachtet. Demnach ist $m_{pa(G)}$ die Anzahl der Gene, die zu *allen* Elternknoten von G gehören.

Von den insgesamt L differentiell exprimierten Genen im Experiment bezeichnen L_G und $L_{pa(G)}$ die entsprechenden Anzahlen der differentiellen Gene, die in G beziehungsweise $pa(G)$ gefunden werden. Der p-Wert des klassischen hypergeometrischen Tests wird be-

rechnet durch

$$p_G = 1 - \sum_{\nu=1}^{L_G-1} \frac{\begin{pmatrix} L \\ \nu \end{pmatrix} \begin{pmatrix} m-L \\ m_G-\nu \end{pmatrix}}{\begin{pmatrix} m \\ m_G \end{pmatrix}},$$

vergleiche (3.1). Für den parent-child Ansatz werden die Anzahl m der Gene insgesamt und die Anzahl L der interessanten Gene ersetzt durch $m_{pa(G)}$ und $L_{pa(G)}$. Der p-Wert lautet demnach

$$p_G = 1 - \sum_{\nu=1}^{L_G-1} \frac{\begin{pmatrix} L_{pa(G)} \\ \nu \end{pmatrix} \begin{pmatrix} m_{pa(G)}-L_{pa(G)} \\ m_G-\nu \end{pmatrix}}{\begin{pmatrix} m_{pa(G)} \\ m_G \end{pmatrix}}.$$

Spezifische GO Begriffe werden nach diesem Ansatz nur noch für signifikant befunden, wenn sie im Vergleich zu ihren allgemeineren Vorfahren deutlich mit interessanten Genen angereichert sind. Das Resultat ist, daß durch die Methode weniger spezifische Begriffe detektiert werden. Der Schwerpunkt liegt eher auf den allgemeineren Termen. Die Autoren erläutern, daß solche speziellen Begriffe, die nach der klassischen Methode für interessant befunden werden, nicht aber im parent-child Ansatz, nicht als irrelevant betrachtet werden, aber daß in den Daten nicht genügend Information vorhanden ist, um deren Relevanz zu begründen. Auf der anderen Seite seien spezifische Begriffe, die laut parent-child Ansatz signifikant sind, besonders wichtig.

Wie die Verfahren von Alexa u. a. (2006) ist auch die parent-child Methode durch heuristische Überlegungen motiviert. Ebenso handelt es sich nicht um eine Adjustierung für multiples Testen. Grossmann u. a. (2007) schlagen zusätzlich die Adjustierung mit einer der üblichen Methoden vor.

5.2.2 Verfahren basierend auf globalen Tests

Im Vergleich zur Gene Set Enrichment Strategie sind holistische Verfahren wie in den Kapiteln 3.4.2 und 4 beschrieben in der Theorie besser zu begründen. Deshalb ist es wünschenswert, GO Analysemethoden zu entwickeln, die mit solchen Verfahren arbeiten. Der Schwerpunkt liegt hier auf den globalen Tests GlobalAncova (Mansmann und Meister, 2005; Hummel u. a., 2008a) und globaltest (Goeman u. a., 2004). Ebenso bedarf es einer theoretisch fundierten Adjustierung für multiples Testen, welche die hierarchische Struktur der Gene Ontology berücksichtigt.
Die folgenden Abschnitte zeigen hierfür geeignete Ansätze.

Mit globalen Tests wird für jede GO Gengruppe G die entsprechende Nullhypothese H_G getestet, daß sie keine differentiell exprimierten Gene aufweist. Es wird von den klassischen

logischen Zusammenhängen zwischen Hypothesen ausgegangen

$$G \subseteq G' \Rightarrow \text{wenn } H_{G'} \text{ wahr, dann auch } H_G \text{ wahr.} \tag{5.2}$$

Umgekehrt kann $H_{G'}$ abgelehnt werden, wenn H_G falsch ist. Somit spiegelt sich die hierarchische Struktur der Gene Ontology in der logischen Struktur der Hypothesen wider. Dies ist bei den im vorigen Abschnitt gezeigten Verfahren nicht der Fall. Auf diesen Unterschied wird in Kapitel 6.1 nochmals eingegangen.

Die *focus level* Methode

Die *focus level* Methode von Goeman und Mansmann (2008) verwendet globale Tests zur Untersuchung der GO Terme hinsichtlich differentieller Expression und korrigiert für multiples Testen mit Hilfe einer Kombination aus der Adjustierung nach Holm (1979) und der geschlossenen Testprozedur von Marcus u. a. (1976). Es wird dabei die family-wise error rate (FWER) auf dem gesamten GO Graphen kontrolliert. Die Korrektur von Holm ist sehr schnell und einfach durchzuführen, sie ist allerdings sehr konservativ. Die geschlossene Testprozedur ist effizient für korrelierte Tests, die im Falle der Gene Ontology vorliegen. Für die gesamte GO ist das geschlossene Testen aber rechnerisch kaum realisierbar. Das Verfahren von Goeman und Mansmann (2008) bietet einen „Mittelweg“ zwischen den beiden Methoden, wobei vom Anwender ein *focus level*, das heißt ein Grad an Spezifität von meistem Interesse irgendwo in der Mitte des GO Graphen, vorgegeben wird.

Nutzt man die logischen Beziehungen (5.2) zwischen GO Hypothesen, so könnte man sich auf das Testen der Endknoten im Graphen beschränken. Nach entsprechender Adjustierung können alle Vorfahren der verbleibenden signifikanten Endknoten ebenfalls als signifikant deklariert werden (*upward propagation*). Für die Adjustierung nach Holm (1979) sortiert man die (rohen) p-Werte der K' Endknoten $p_{(1)} \leq \ldots \leq p_{(K')}$. Der kleinste p-Wert wird mit α/K' verglichen. Ist er kleiner, so wird als nächstes überprüft ob $p_{(2)} \leq \alpha/(K'-1)$ gilt. Die Prozedur wird weitergeführt bis k^* gefunden wird mit $p_{(k^*)} > \alpha/(K'-k^*+1)$. Die Hypothesen $H_{(1)}, \ldots, H_{(k^*-1)}$ werden abgelehnt, alle übrigen nicht. Da der GO Graph weit ausfächert und es sehr viele Endknoten gibt, ist diese Korrektur eher streng. Schwächere aber konsistente Effekte in weniger speziellen Knoten in höheren Ebenen des Graphen werden leicht übersehen.

Die zweite im focus level Ansatz verwendete Adjustierungsmethode ist die geschlossene Testprozedur von Marcus u. a. (1976). Für die Prozedur muß zunächst die bestehende Hypothesenfamilie $\mathcal{H} = \{H_{G_1}, \ldots, H_{G_K}\}$ zu einer Familie $\mathcal{H}^*$ erweitert werden, die *unter Durchschnitt abgeschlossen* ist, das heißt

$$\text{für alle } H_{G_k}, H_{G_l} \in \mathcal{H}^* \text{ muß gelten } H_{G_k} \cap H_{G_l} \in \mathcal{H}^*.$$

Der Durchschnitt $H_{G_k} \cap H_{G_l}$ entspricht der Hypothese $H_{G_k \cup G_l}$ für die vereinigten Genmengen. Das bedeutet, daß alle möglichen Vereinigungsmengen von GO Gruppen gebildet

werden müssen. Man kann nun eine Hypothese H_{G_k} genau dann ablehnen, wenn alle Hypothesen $\{H_{G_\nu} \in \mathcal{H}^* : G_k \subseteq G_\nu\}$ zum Niveau α abgelehnt werden können. Obwohl alle Tests einzeln zum Niveau α durchgeführt werden, wird durch die Prozedur die family-wise error rate für die gesamte Hypothesenfamilie kontrolliert ($FWER \leq \alpha$). Abbildung 5.3 verdeutlicht beispielhaft das Prinzip der geschlossenen Testprozedur. Zwar ist es für die Bildung der geschlossenen Familie günstig, wenn wie im Falle der GO ohnehin starke Überschneidungen zwischen den Gruppen bestehen. Dennoch wird bei der enormen Größe der Gene Ontology schnell die Schwierigkeit für die praktische Umsetzung der Prozedur klar. Im *R* Paket `GlobalAncova` ist die geschlossene Testprozedur implementiert. Diese ist allerdings nur für kleine Familien von Gengruppen, keinesfalls aber für die gesamte GO, zu empfehlen.

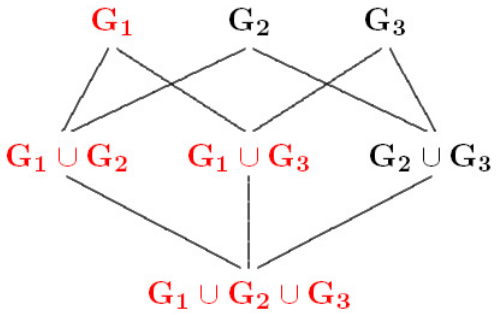


Abbildung 5.3: Beispiel für die Bildung einer durchschnittsabgeschlossenen Familie, ausgehend von den Gruppen G_1, G_2 und G_3. Um in der geschlossenen Testprozedur zum Beispiel die Hypothese G_1 ablehnen zu können, müssen die Hypothesen, die den rot markierten Gruppen entsprechen, zum Niveau α abgelehnt worden sein.

Die focus level Methode kombiniert die beiden Adjustierungsverfahren und bringt so deren Vor- und Nachteile in Balance. Die Prozedur beginnt mit der Wahl des focus levels, einer Menge von Knoten $F_1, \ldots, F_{K_F}$ innerhalb des GO Graphen, die gegenseitig nicht verwandt sind, also kein Term ein Vor- oder Nachfahre eines anderen Knotens im focus level ist. Befindet sich der focus level nahe der Wurzel des Graphen, so liegt der Schwerpunkt auf allgemeinen GO Begriffen. Wählt man für den focus level dagegen spezifischere Knoten, werden auch eher spezialisierte Begriffe durch die Prozedur gefunden. In der Nähe des focus levels ist die Macht des Verfahrens am größten.

Der GO Graph wird in K_F Subgraphen $\mathcal{F}_1, \ldots, \mathcal{F}_{K_F}$ unterteilt, deren Wurzeln die einzelnen focus level Knoten sind. Für jeden dieser Subgraphen $\mathcal{F}_k$ wird ein durchschnittsabgeschlossener Graph $\mathcal{F}_k^*$ erzeugt. Bezeichne Φ den tatsächlichen GO Graphen und Φ^* den erweiterten Graphen $\Phi^* = \Phi \cup (\bigcup_{k=1}^{K_F} \mathcal{F}_k^*)$. Während der Prozedur werden eine Menge von zu testenden Hypothesen $\tilde{\mathcal{H}}$ und ein Korrekturfaktor h, der *Holm Faktor*, entwickelt. Zu Beginn ist $\tilde{\mathcal{H}}$ die Menge der zu den focus level Knoten gehörenden Hypothesen $\{H_{F_1}, \ldots, H_{F_{K_F}}\}$ und h entspricht der Anzahl der Elemente in $\tilde{\mathcal{H}}$: $h = |\tilde{\mathcal{H}}| = K_F$. Der focus level Algorithmus

ist in der folgenden Box dargestellt.

Algorithmus der focus level Prozedur

1. **Test Phase:** Lehne alle Hypothesen in $\tilde{\mathcal{H}}$ ab, für deren rohe p-Werte gilt $p \leq \alpha/h$.
2. **Aufwärts Phase:** Verwerfe alle Vorfahren in Φ^* von Hypothesen, die in Schritt 1. abgelehnt worden sind.
3. **Abwärts Phase:** Füge die Hypothesen in $\mathcal{F}_1^*, \ldots, \mathcal{F}_{K_F}^*$, für die alle Vorfahren verworfen worden sind, zu $\tilde{\mathcal{H}}$ hinzu.
4. **Holm Phase:** Bestimme h neu als die Anzahl der Subgraphen $\mathcal{F}_1^*, \ldots, \mathcal{F}_{K_F}^*$, in denen noch unverworfene Hypothesen sind.

Wiederhole Schritte 1.–4. bis keine weiteren Hypothesen mehr abgelehnt werden.

Goeman und Mansmann (2008) beweisen, daß diese Prozedur die family-wise error rate kontrolliert.

Das Resultat des Verfahrens sind zusammen hängende signifikante GO Subgraphen, wie das Beispiel aus Goeman und Mansmann (2008) in Abbildung 5.4 zeigt. Durch die upward propagation kann es dazu kommen, daß manche Knoten signifikant werden, obwohl ihre rohen p-Werte größer als α sind. Als Argument dafür kann die eventuell nicht ausreichende Macht des Einzeltests für die betreffende Hypothese aufgeführt werden.

Hierarchisches Testen

Die hierarchische Testprozedur von Meinshausen (2008) stammt aus dem Gebiet der Variablenselektion. Wird der Einfluß sehr vieler Variablen einzeln überprüft und für multiples Testen korrigiert, so ist die Macht meist relativ gering, vor allem, wenn die Variablen untereinander korrelieren. Es wird daher vorgeschlagen, die Variablen hierarchisch in Cluster einzuteilen und mit geeigneten Methoden nacheinander die so entstandenen Variablencluster zu testen. Das heißt zunächst wird die globale Hypothese überprüft, daß keine der Variablen einen Einfluß hat. Kann diese Hypothese, nach geeigneter Korrektur, abgelehnt werden, so betrachtet man als nächstes zwei Cluster, die durch geeignete *vollständige* und *disjunkte* Einteilung der Variablen definiert wurden. Entsprechend läuft die Prozedur weiter bis man schließlich ganz unten in der Hierarchie zu den Einzelhypothesen gelangt. Für die Tests der Cluster bieten sich förmlich globale Tests wie von Goeman u. a. (2004) und Mansmann und Meister (2005); Hummel u. a. (2008a) an.

Meinshausen (2008) schlägt eine hierarchische Adjustierung vor, durch die die FWER kontrolliert wird. Ist m die Anzahl der Variablen und p_G der rohe p-Wert des globalen

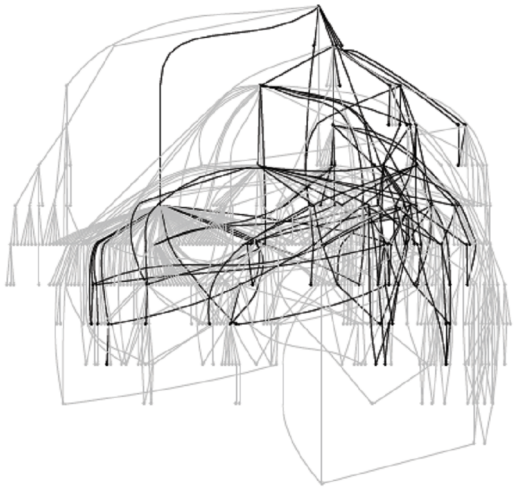


Abbildung 5.4: Beispiel für ein Ergebnis der focus level Methode. Der signifikante Subgraph (schwarz) erstreckt sich von der Wurzel her innerhalb des gesamten (Zelluläre Komponente) Gene Ontology Graphen (grau).

Tests für Cluster G, wird der adjustierte p-Wert definiert als

$$p_{G,adj} = p_G \cdot \frac{m}{|G|}.$$

Die übliche Bonferroni Korrektur $p_G \cdot m$ wird also entsprechend der Größe des Clusters „abgeschwächt“. Beim Testen der großen Cluster ist die Korrektur folglich nicht so streng. Auf dem Niveau der Einzelhypothesen liegt Bonferroni Adjustierung vor. In der hierarchischen Testprozedur werden nun alle Hypothesen H_G abgelehnt,

(a) deren adjustierter p-Wert $p_{G,adj} \leq \alpha$

(b) deren Eltern-Cluster abgelehnt wurde.

Bedingung (b) spiegelt die logische Struktur der Hypothesenhierarchie wider: H_G ist immer wahr, wenn die Hypothese des Eltern-Clusters $H_{pa(G)}$, $G \subset pa(G)$, wahr ist. Demnach kann man bei solchen Hypothesen die Prozedur stoppen, die nicht abgelehnt werden können.

Im Bereich der Variablenselektion wird man beispielsweise mit Hilfe von hierarchischen Cluster Algorithmen binäre Bäume für die oben beschriebene Prozedur erstellen. Ebenso können aber auch bestehende Hierarchien verwendet werden. Ein Beispiel hierfür wäre die Gene Ontology. Bei der GO besteht allerdings das Problem, daß die Aufteilung der Gene über den Graphen weder vollständig noch disjunkt ist. Das Verfahren von Meinshausen (2008) müßte entsprechend angepaßt werden, um weiterhin FWER Kontrolle zu gewährleisten. Denkbar wäre zum Beispiel eine neue Korrektur der Gestalt

$$p_{G,adj} = \begin{cases} p_G \cdot \frac{m}{|G|-|G^*|}, & |G^*| < |G| \\ p_G \cdot m, & |G^*| = |G| \end{cases}$$

wobei G^* alle Gene aus G enthält, die ebenfalls zu Knoten gehören, die nicht mit G verwandt, also weder Vor- noch Nachfahren von G sind. Leider kommt es oft vor, daß alle Gene eines Knotens auch „außerhalb“ annotiert sind und dadurch $|G| = |G^*|$. Die Korrektur für multiples Testen würde somit sehr streng ausfallen.

Weiterhin wäre es interessant, die Ideen von Yekutieli (2008), nach denen hierarchisch getestet und dabei die weniger konservative false discovery rate kontrolliert wird, für die Gene Ontology anzupassen.

5.2.3 Weitere spezielle Gene Ontology Methoden

Die Liste der beschriebenen Verfahren für die Gene Ontology Analyse ist natürlich nicht vollständig. Methoden, die sich mit ähnlichen Fragestellungen wie in den obigen Abschnitten beschäftigen, werden zunehmend entwickelt. Desweiteren gibt es eine ganze Bandbreite von Verfahren, die andere Aspekte der Gene Ontology untersuchen oder zusätzliche Informationen nutzen. Hier seien nur einige dieser Methoden genannt. Vêncio u. a. (2006) entwickeln eine bayesianische GO Analyse. Mit dieser Methode kann berücksichtigt werden, daß mit einem Microarray Experiment nicht alle relevanten Gene beobachtet worden sein könnten. Maglietta u. a. (2007) untersuchen den Zusammenhang zwischen den GO Gruppen und der interessierenden klinischen Information, indem sie die Güte einer Prädiktion des Phänotyps nur auf der Basis der Gene in der jeweiligen GO Gruppe bewerten. Die Methode von Stanley u. a. (2006) bezieht Länge und Position der Gene in die GO Analyse mit ein. Lewin und Grieve (2006) schlagen für die Verbesserung der Gene Set Enrichment Analyse vor, GO Begriffe zuvor geeignet zu gruppieren. Nam u. a. (2006) erzeugen zusätzliche Gengruppen für die Analyse, indem sie Durchschnitte zwischen den drei Ontologien (BP, MF, CC) bilden. Weitere Verfahren beschäftigen sich unter anderem mit der Problematik und Darstellung der GO Annotationen (Dolan u. a., 2005; Ye u. a., 2006), der semantischen Ähnlichkeit zwischen GO Begriffen (Wang u. a., 2007) und der Validierung von Gen-Clustern mit Hilfe der GO (Steuer u. a., 2006).

Kapitel 6

Vergleich von Methoden für die Gengruppenanalyse

6.1 Unterschiede und Zusammenhänge

6.1.1 Gene Set Enrichment und holistische Verfahren

In Abschnitt 3.3 wurden bereits kurz die wesentlichen Unterschiede zwischen Gene Set Enrichment und holistischen Verfahren erwähnt. In Goeman und Bühlmann (2007) wird detailliert auf die theoretischen Grundlagen der beiden Strategien eingegangen. Man denke im Folgenden an das klassische Gene Set Enrichment, das auf 2×2 Kontingenztafeln beruht, und an holistische Methoden wie zum Beispiel den globaltest.

Kompetitive versus in sich geschlossene Teststrategie

Bei Gene Set Enrichment Verfahren für Kontingenztafeln wird eine Gengruppe G stets mit ihrem Komplement $\bar{G}$ verglichen. Die Nullhypothese lautet, daß G nicht mehr mit interessanten Genen angereichert ist als $\bar{G}$. Man kann also sagen, daß die Signifikanz von G „bestraft“ wird in Bezug auf die Signifikanz von $\bar{G}$. Goeman und Bühlmann (2007) sprechen daher von einer *kompetitiven* (*competitive*) Teststrategie. Bei den holistischen Verfahren dagegen wird nur die Gengruppe G per se betrachtet und die Nullhypothese 'kein Gen in G ist differentiell exprimiert' untersucht. Alle übrigen Gene spielen für die Bewertung von G im Grunde keine Rolle. Es handelt sich um eine *in sich geschlossene* (*self-contained*) Teststrategie.

Die in sich geschlossene Nullhypothese H_0^{self} ist viel restriktiver als die kompetitive H_0^{comp}: bei H_0^{self} genügt bereits ein deutlich differentiell exprimiertes Gen für ein signifikantes Gruppenergebnis, während bei H_0^{comp} dafür viele differentielle Gene benötigt werden. Dadurch ergibt sich automatisch, daß holistische Verfahren mehr Macht besitzen, interessante Gengruppen zu detektieren. Dies wird besonders deutlich bei Betrachtung des extremen Beispiels, bei dem alle Gene eines Experiments (gleichermaßen) differentiell ex-

primiert seien. Die kompetitive Hypothese H_0^{comp} wird für *keine* Gengruppe G abgelehnt werden können, da $\bar{G}$ immer genauso stark mit differentiellen Genen angereichert ist wie G selbst. Dagegen wird *jede* in sich geschlossene Hypothese H_0^{self} abgelehnt werden, weil ja tatsächlich jede Gruppe differentielle Gene enthält. Natürlich läßt sich hier diskutieren, welches Ergebnis für die Interpretation als sinnvoller erachtet wird. Die große Macht der holistischen Verfahren könnte in manchen Fällen problematisch sein, da es auch in realen Datensätzen vorkommt, daß fast alle Gengruppen als signifikant deklariert werden.

Beim Gene Set Enrichment wird ein einzelnes Gen ganz anders bewertet als eine Gengruppe, die nur aus einem Gen besteht. Der p-Wert des einzelnen Gens entspricht dem der genweisen Statistik, zum Beispiel der t-Statistik, während der p-Wert für die Gengruppe mit nur einem Gen über die hypergeometrische Verteilung berechnet wird und von der Signifikanz aller übrigen Gene abhängt. Es wäre allerdings wünschenswert, daß der Test hinsichtlich differentieller Expression für ein einzelnes Gen identisch ist zum Test einer Gengruppe der Größe eins. Dies ist bei holistischen Verfahren der Fall: der Gruppentest ist eine direkte Verallgemeinerung des Einzelgentests. Desweiteren ist es mit der kompetitiven Strategie nicht möglich, das komplette Set der Gene zu testen, einfach weil es für die gesamte Genmenge kein Komplement gibt. Das Testen der globalen Hypothese 'kein Gen im ganzen Experiment ist differentiell exprimiert' mit holistischen Methoden ist dagegen möglich und kann als erste grobe Überprüfung der Daten angewendet werden.

Ein weiteres Problem der kompetitiven Strategie besteht darin, daß die p-Werte dazu tendieren negativ zu korrelieren. Dies geschieht dadurch, daß eine große Anzahl an interessanten Genen in einer Gruppe zu höheren p-Werten anderer Gruppen führt. Solche Korrelationen stellen ein Problem dar, wenn für multiples Testen adjustiert werden soll.

Genrandomisierung versus Permutation der Beobachtungen

Gene Set Enrichment beruht auf dem Urnenmodell: die Nullhypothese für eine Gruppe G entspricht zufälligem Ziehen von $|G|$ Einheiten aus der Gesamtheit aller Gene eines Experiments. Die Stichprobeneinheiten sind folglich die Gene. Demnach würde eine neue Stichprobe aus *neuen Genen* bestehen, die für die *selben Beobachtungen* (Patienten) gemessen werden. Die klassischen Rollen von Variablen und Beobachtungen sind in diesem Ansatz vertauscht. Anders verhält es sich bei den holistischen Verfahren: die Gene entsprechen einem festen Set von Variablen, die für jede Beobachtung ausgewertet werden. Eine neue Stichprobe besteht aus *neuen Patienten*, für die die Expression der *gleichen Gene* ermittelt wird. Die Nullhypothese kann hier durch Permutationen der Beobachtungen simuliert werden.

Aufgrund dieser völlig verschiedenen Modelle müssen auch die resultierenden p-Werte unterschiedlich interpretiert werden. Wenn die jeweils zugrunde liegende Nullhypothese wahr ist, so werden bei Wiederholungen des Experiments erwartungsgemäß nur in $(100 \cdot \alpha)\%$ der Fälle p-Werte kleiner α beobachtet werden. Es ist also von entscheidender Bedeutung,

was eine 'Wiederholung des Experiments' bedeutet. Beim Permutationsmodell stehen die p-Werte in Bezug zu Expressionsmessungen der gleichen Gene für neue Patienten. Dies entspricht realen biologischen Replikaten des Experiments. Ein kleiner p-Wert bedeutet, daß man in zukünftigen Patienten eine ähnliche Assoziation zwischen der Genexpression und der untersuchten Phänotypstruktur erwarten kann. Beim Genrandomisierungsmodell dagegen sind die p-Werte der Wiederholung des Urnenexperiments zugeordnet, bei dem eine Stichprobe neuer Gene gezogen werden müßte. Signifikante p-Werte weisen darauf hin, daß ein ähnlicher Zusammenhang zwischen den Eigenschaften 'Zugehörigkeit zur Gengruppe' und 'differentielle Expression' für neue Genstichproben gefunden würde.

Beim Permutationsmodell entspricht die Stichprobengröße der Anzahl n der prozessierten Microarrays, also der Anzahl der biologischen Replikate. Bei der Genrandomisierung ist die Stichprobengröße gleich der Anzahl m der Gene. Dies erscheint verlockend, da üblicherweise $m >> n$. Tatsächlich kann man durch die Randomisierung der Gene sehr kleine p-Werte erhalten, selbst wenn im Extremfall nur zwei Microarrays zur Verfügung stehen. Diese Ergebnisse sind aber „künstlich“ und irreführend, da die reale biologische Stichprobengröße natürlich trotzdem nur zwei bleibt.

Ein weiterer problematischer Aspekt der Genrandomisierung ist die zugrunde liegende Annahme der Unabhängigkeit – das Urnenmodell geht davon aus, daß die Gene unabhängig voneinander „gezogen“ werden können. Dies ist eine höchst unrealistische Annahme für Expressionsdaten, gerade wenn funktionelle Gruppen interagierender Gene betrachtet werden. Korrelationen beeinflussen die Anzahl differentiell exprimierter Gene. Das hat zur Folge, daß die wahre Nullverteilung des Tests für die 2×2 Kontingenztafel nicht hypergeometrisch ist, sondern mehr Masse in den Randbereichen besitzt. Daher kann es zu einem anti-konservativen Verhalten des hypergeometrischen Tests kommen, das heißt zu mehr falsch positiven Ergebnissen als erwartet.

Zusammenfassend sprechen sich Goeman und Bühlmann (2007) klar gegen das Genrandomisierungsmodell und die kompetitive Teststrategie aus. Sie geben Vorschläge, wie ein auf 2×2 Tafeln basierendes Verfahren als in sich geschlossener Test und mit zugrunde liegendem Permutationsmodell spezifiziert werden kann. Um letzteres zu erreichen, müßte man beispielsweise anstatt die Gengruppe B mal zufällig zusammen zu stellen (Genrandomisierung) einfach B mal die Beobachtungen permutieren und die genweisen Statistiken jeweils neu berechnen. Dadurch ergibt sich für jede Permutation eine andere Liste differentieller Gene und entsprechend auch eine andere Aufteilung der Genanzahlen in der Kontingenztafel. Liegt tatsächlich eine deutliche Anhäufung von differentiellen Genen in der betrachteten Gruppe vor, so wird man in den Kontingenztafeln der Permutationen in der Regel weniger Gene beobachten, die gleichzeitig differentiell und innerhalb der Gengruppe sind. Es muß also nicht gleich automatisch lauten Gene Set Enrichment = kompetitiver Ansatz mit Genrandomisierung und holistisch = in sich geschlossener Ansatz mit Permutation der Beobachtungen. Die GSEA Methode von Mootha u. a. (2003) und Subramanian u. a. (2005) ist beispielsweise ein kompetitiver Test, dessen Signifikanz meist über Permutationen der

Beobachtungen bewertet wird, ähnlich wie soeben für den Ansatz mit Kontingenztafeln beschrieben. Die globalen Tests GlobalAncova und globaltest haben die von den Autoren bevorzugten Modelleigenschaften. Ebenso gibt es aber auch bei den holistischen Verfahren „Mischformen“ hinsichtlich der Teststrategie und des sampling Modells, zum Beispiel das Category Verfahren (Gentleman und Falcon, 2007) und den Restandardisierungsansatz (Efron und Tibshirani, 2007).

Gene Set Enrichment und holistische Verfahren für die GO

Verwendet man für die GO Analyse wie in Abschnitt 5.2.2 globale Tests, entspricht jede GO Gruppe der Nullhypothese, daß die zugehörigen Gene nicht differentiell exprimiert sind. Ist der Knoten G ein Nachfahre von G', so gilt aufgrund der Vererbungsregel $G \subseteq G'$. Daraus ergibt sich für die Hypothesen H_G und $H_{G'}$ eine Hierarchie: die Wahrheit der „größeren Hypothese“ $H_{G'}$ impliziert die Wahrheit der „kleineren Hypothese“ H_G. Beziehungsweise gilt umgekehrt, ist H_G falsch, so kann automatisch auch $H_{G'}$ abgelehnt werden. Gemäß dieser logischen Überlegungen kann das Resultat einer GO Analyse nur ein zusammenhängender signifikanter Subgraph innerhalb der Gene Ontology sein, der von der Wurzel der GO ausgeht.

Beim Gene Set Enrichment dagegen wird für jeden Knoten G die kompetitive Nullhypothese betrachtet, daß G nicht mehr mit differentiellen Genen angereichert ist wie sein Komplement $\bar{G}$. Die Nullhypothese des Wurzelknotens, der alle Gene enthält, ist somit immer wahr und die oben beschriebene logische Struktur gilt nicht. Demnach stehen Intention und Resultat der in Abschnitt 5.2.1 beschriebenen Verfahren für die spezielle GO Analyse mit Hilfe von Gene Set Enrichment in deutlichem Gegensatz zu Methoden wie der focus level Prozedur: bei ersteren wird versucht, einzelne relevante Knoten innerhalb von Verwandten heraus zu finden, bei letzteren sucht man nach signifikanten Subgraphen.

6.1.2 Vergleiche einzelner Verfahren

Tests für 2×2 Kontingenztafeln und GSEA

Rivals u. a. (2007) vergleichen auf theoretischer Basis diverse GO Analyse Pakete (unter anderen Al-Shahrour u. a., 2004; Zeeberg u. a., 2003; Beissbarth und Speed, 2004; Zhong u. a., 2004a; Draghici u. a., 2003), die alle auf 2×2 Kontingenztafeln wie in Tabelle 3.3 beruhen, also den Zusammenhang zwischen den Eigenschaften 'Gen ist in der interessanten Liste' und 'Gen ist in der funktionellen Gengruppe' untersuchen. Bei diesen Verfahren werden Fisher's exakter Test, Binomialtests, χ^2-Tests, hypergeometrische Tests und Tests auf die Gleichheit zweier Wahrscheinlichkeiten verwendet. In den Beschreibungen zu den Analysepaketen finden sich teilweise widersprüchliche Aussagen über die Vor- und Nachteile der verschiedenen Tests. Rivals u. a. (2007) zeigen dagegen, daß unter der Nullhypothese die exakte Verteilung der Anzahl der Gene, die gleichzeitig differentiell exprimiert sind und der jeweiligen Gengruppe angehören, stets die hypergeometrische Verteilung ist. Bei

genügend großer Stichprobe ist eine Approximation über die Binomial-, Normal- oder χ^2-Verteilung möglich.

Die Anzahlen in den 2×2 Kontingenztafeln hängen davon ab, wie viele Gene als differentiell exprimiert deklariert werden, also von der Wahl eines Schwellenwertes (*cutoff*). Wie dieser Wert optimal zu wählen ist, ist unklar. Außerdem ist die dichotome Einteilung der Gene in differentielle und nicht-differentielle eher künstlich. Als Abhilfe könnte man mehrere verschiedene cutoffs betrachten oder aber auf das GSEA Verfahren (Mootha u. a., 2003; Barry u. a., 2005; Subramanian u. a., 2005) zurückgreifen, welches mit einer Kolmogorov-Smirnov-Rang-Statistik arbeitet und so das gesamte Kontinuum an differentieller Expression nutzt. Abbildung 6.1 soll den Zusammenhang zwischen der Fisher-Test Strategie und GSEA verdeutlichen. Die Linien zeigen, wie viele unter den jeweils x am stärksten differentiell exprimierten Genen (x-Achse) zu einer betrachteten Gengruppe gehören (y-Achse). Die durchgezogene Linie entspricht einer Gruppe, die viele differentielle Gene enthält. Die gestrichelte Linie würde man unter der Nullhypothese, das heißt bei zufälliger Zusammenstellung der Gengruppe, erwarten. Bei GSEA werden die kompletten Verläufe der beiden Linien verglichen. Bei den 2×2 Tafeln dagegen betrachtet man nur die Anzahlen an einem bestimmten Punkt (cutoff), der durch die senkrechte Linie angedeutet ist.

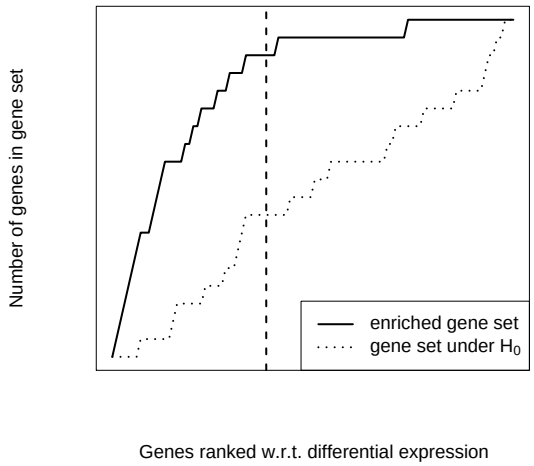


Abbildung 6.1: Skizze zu GSEA. Für jeden Rang innerhalb der geordneten Genliste wird die Anzahl der Gene angegeben, die zur betrachteten Gengruppe gehören. Es werden die beiden Situationen mit (durchgezogene Linie) und ohne (gestrichelte Linie) Anreicherung der Gengruppe mit differentieller Expression gezeigt. Beim Fisher-Test erfolgt die Einteilung in differentielle und nicht-differentielle Gene anhand eines Schwellenwertes (vertikale Linie).

GSEA und holistische Ansätze

Die Gene Set Enrichment Analyse von Mootha u. a. (2003) und Subramanian u. a. (2005) vergleicht die Genränge innerhalb der betrachteten Gengruppe mit den Rängen aller übrigen Gene. Es handelt sich demnach um eine kompetitive Teststrategie. Zur Bewertung der Signifikanz können die Phänotypzugehörigkeiten permutiert werden. Somit bildet das GSEA Verfahren einen interessanten Sonderfall im Vergleich zu klassischen Gene Set Enrichment Methoden auf der einen und holistischen Ansätzen auf der anderen Seite. Wie Dinu u. a. (2007) berichten hat GSEA gegenüber holistischen Verfahren einige Nachteile. Zum einen werden öfters Gengruppen als signifikant detektiert, deren Gene nicht mit der relevanten Phänotypvariablen assoziiert sind. Dies läßt sich dadurch erklären, daß die Kolmogorov-Smirnov Teststatistik immer dann einen extremen Wert annimmt, wenn die Gene innerhalb der interessierenden Gruppe in der Rangfolge aller Gene nahe beieinander liegen. Dies kann auch dann der Fall sein, wenn die Gene alle zwar nicht differentiell exprimiert sind, aber sehr ähnliche Ränge besitzen. Zum anderen hat GSEA wohl nur eine geringe Macht, Gengruppen zu detektieren, die sich sowohl aus differentiell exprimierten als auch aus nicht differentiellen Genen zusammen setzen. Solche „gemischten“ Gengruppen scheinen aber biologisch plausibel. Man wird selten erwarten, daß alle Gene innerhalb einer funktionellen Gruppe auf eine phänotypische Veränderung reagieren. Durch die Mischung an hohen und niedrigen Rängen fällt das Maximum der laufenden Rangsumme meist nicht groß genug aus, um eine solche Gruppe signifikant werden zu lassen. GSEA nutzt durch die Verwendung von Genrängen mehr Information als die auf Genanzahlen in Kontingenztafeln basierenden Verfahren. Dennoch liefert die Rangbildung nur ein *relatives* Maß der Assoziation der Gengruppe mit dem Phänotypen, und zwar relativ zu allen übrigen Genen. Der tatsächliche Grad der Assoziation wird nicht beachtet, wie es im Gegensatz dazu bei holistischen Verfahren der Fall ist.

Dekorrelierung der GO und parent-child Ansatz

Die Dekorrelierung der GO (Alexa u. a., 2006) und der parent-child Ansatz (Grossmann u. a., 2007) sind Modifikationen des klassischen hypergeometrischen Tests angepaßt auf die spezielle Struktur der Gene Ontology. Bei ersterem wird die Signifikanz eines GO Knotens auf der Basis der Signifikanz seiner Kinder (elim Algorithmus) beziehungsweise seiner Kinder und Eltern (weight Algorithmus) bemessen. Man kann sagen, daß der Schwerpunkt hier auf der Detektion spezifischer Begriffe liegt. Im parent-child Ansatz wird dagegen versucht falsch positive Ergebnisse zu vermeiden, die durch das „Vererbungsproblem“ hervorgerufen werden, indem jeder Knoten in Bezug auf dessen Eltern bewertet wird. Hier liegt der Fokus auf den eher allgemeineren Begriffen. Die Intentionen der beiden Ansätze scheinen also gegensätzlich zu sein. Grossmann u. a. (2007) vergleichen ihr Verfahren mit dem klassischen Ansatz und den Methoden von Alexa u. a. (2006) durch Simulationen. Sie geben aber auch zu bedenken, daß es sich um recht verschiedene Maße für die Anreicherung der GO Knoten handelt und es deshalb nicht klar ist, wie die Verfahren „gerechterweise“ verglichen werden können (vergleiche dazu auch Abschnitt 6.2).

Vergleiche mit Category

Die Idee des Category Ansatzes (Gentleman und Falcon, 2007) liegt in der geeigneten Zusammenfassung der genweisen Statistiken x zu Gruppenstatistiken $z = f(A, x)$. Die $K \times m$ Matrix $A = (a_{k,j})_{\substack{k=1,\dots,K \\ j=1,\dots,m}}$ gibt die Zugehörigkeit der m Gene zu den K Gruppen (Kategorien) an. Die Autoren schlagen für Kategorie G_k eine standardisierte Summe derjenigen genweisen Statistiken vor, die zu G_k gehören

$$z_k = \frac{1}{\sqrt{m_k}} \sum_{j=1}^{m} I(a_{k,j} = 1) x_j.$$

Dabei ist $I(\cdot)$ die Indikatorfunktion und somit $m_k = \sum_{j=1}^{m} I(a_{k,j} = 1)$ die Anzahl der Gene in G_k. Ebenso sind andere Zusammenfassungen der genweisen Statistiken denkbar, wie zum Beispiel das arithmetische Mittel (Tian u. a., 2005), die maxmean-Statistik (Efron und Tibshirani, 2007) oder SAM-GS (Dinu u. a., 2007).

Bei solchen Formen der Gruppenstatistik handelt es sich um eine in sich geschlossene Teststrategie, da für die jeweilige Statistik nur Gene innerhalb der Gruppe in Betracht gezogen werden. Ebenso könnte man aber auch eine kompetitive Statistik wählen, zum Beispiel

$$z_k = \frac{1}{\sqrt{m_k}} \sum_{j=1}^{m} I(a_{k,j} = 1) x_j - \frac{1}{\sqrt{\bar{m}_k}} \sum_{j=1}^{m} I(a_{k,j} = 0) x_j,$$

mit $\bar{m}_k = \sum_{j=1}^{m} I(a_{k,j} = 0)$. Die Summe für Gene aus G_k wird verglichen mit der Summe für alle übrigen Gene. Ganz ähnlich wäre hier die Verwendung einer Zwei-Stichproben t-Statistik statt der obigen Differenz. Damit gelangt man zum in Kapitel 3.4.1 beschriebenen t-Test für Verteilungen der genweisen Statistiken, den wir zu den Gene Set Enrichment Ansätzen zählen.

Auch globale Tests können in die Form eines Category Ansatzes gebracht werden. Für GlobalAncova kann man zum Beispiel als genweise Statistik eine zweidimensionale Funktion wählen

$$x_j = (F_{j,1}, F_{j,2}) = (RSS_{j,extra} \cdot df, RSS_{j,VM}).$$

$RSS_{j,extra}$ und $RSS_{j,VM}$ sind dabei die genweisen Residuenquadratsummen aus (4.1) und (4.2) und df der Quotient der Freiheitsgrade. Die Category Funktion $f(A, x)$ muß dann der Summation der $F_{j,1}$ und $F_{j,2}$ über die jeweiligen Gene und anschließender Division entsprechen, um die GlobalAncova Statistik zu erhalten (vergleiche (4.4))

$$z_k = \frac{\sum_{j=1}^{m} I(a_{k,j} = 1) F_{j,1}}{\sum_{j=1}^{m} I(a_{k,j} = 1) F_{j,2}} = \frac{\sum_{j=1}^{m} I(a_{k,j} = 1) RSS_{j,extra}}{\sum_{j=1}^{m} I(a_{k,j} = 1) RSS_{j,VM}} \cdot df = F_{G_k}.$$

Der Category Ansatz kann als allgemeines Konzept betrachtet werden, durch das fast alle Methoden der Gengruppenanalyse dargestellt werden können. Wenn im Folgenden

vom Category Verfahren die Rede ist, ist meistens die von Gentleman und Falcon (2007) vorgeschlagene Version mit genweisen t-Statistiken und deren standardisierter Summation gemeint.

GlobalAncova und globaltest

Dem globaltest (Goeman u. a., 2004) liegt die Idee der *Prädiktion* einer klinischen Variablen Y anhand der Genexpression X zugrunde. Es wird also der Einfluß von X auf den Phänotypen Y untersucht. Eine dem globaltest eigene Anwendungsmöglichkeit ist die Überlebenszeitanalyse: Was sagt das aktuelle Expressionsprofil über das zukünftige Überleben der Patienten aus? GlobalAncova (Mansmann und Meister, 2005; Hummel u. a., 2008a) betrachtet dagegen Y als Regressor(en) für die beobachtete Genexpression. Es basiert auf einem *Mittelwertevergleich*: Können für verschiedene Ausprägungen von Y unterschiedliche Expressionsprofile erwartet werden? Dabei ist Y nicht auf klinische Gruppen beschränkt, sondern kann eine beliebige „Struktur“ innerhalb des Patientengutes widerspiegeln. Die Nullhypothesen von globaltest und GlobalAncova sind also „invers“

$$\begin{aligned} \text{globaltest}: &\quad H_0 : P(Y|X,C) = P(Y|C) \\ \text{GlobalAncova}: &\quad H_0 : P(X|Y,C) = P(X|C), \end{aligned}$$

wobei C Kovariablen darstellt, für die gegebenenfalls adjustiert werden kann. Wenn tatsächlich kein Zusammenhang zwischen X und Y vorliegt, so sind die beiden Hypothesen äquivalent.

Simulationen zeigen, daß die Permutationsansätze von globaltest und GlobalAncova sehr ähnliche Ergebnisse liefern. Die asymptotische Version des globaltest ist dagegen konservativer. Es werden die gleichen Szenarien verwendet wie in Abschnitt 4.2.2 für den Vergleich der approximativen und permutationsbasierten p-Werte von GlobalAncova. Abbildung 6.2 zeigt für das Szenario der Nullhypothese die permutationsbasierten beziehungsweise approximativen p-Werte von globaltest gegenüber den permutationsbasierten p-Werten von GlobalAncova. Für die linken Grafiken wurden die Gene unabhängig voneinander simuliert, im Szenario für die rechten Grafiken wurde eine Korrelation von $\rho = 0.2$ zwischen je zwei Genen vorgegeben. Zwischen den permutationsbasierten p-Werten der beiden Verfahren zeigt sich eine gute Übereinstimmung. Die asymptotischen globaltest p-Werte sind meist etwas größer und zum Niveau von 5% werden fast keine Hypothesen abgelehnt. Mit GlobalAncova ergeben sich dagegen mehr signifikante Ergebnisse, wobei das α-Niveau aber eingehalten wird. Entsprechend wird in Szenarien mit differentieller Expression für GlobalAncova eine größere Macht beobachtet als für den asymptotischen globaltest. Mit den approximativen GlobalAncova p-Werten wird im Falle abhängiger Gene das α-Niveau überschritten, wie bereits in Abschnitt 4.2.2 erläutert wurde. Bei Korrelationen zwischen den Genen ist die Macht der Tests generell geringer als bei unabhängigen Genen. Die α-Fehler und die Macht der Tests bei verschiedenen Szenarien sind in Tabelle 6.1 zusammen gefaßt. Ähnliche Simulationsergebnisse für den Vergleich von globaltest und GlobalAncova werden von Liu u. a. (2007) berichtet.

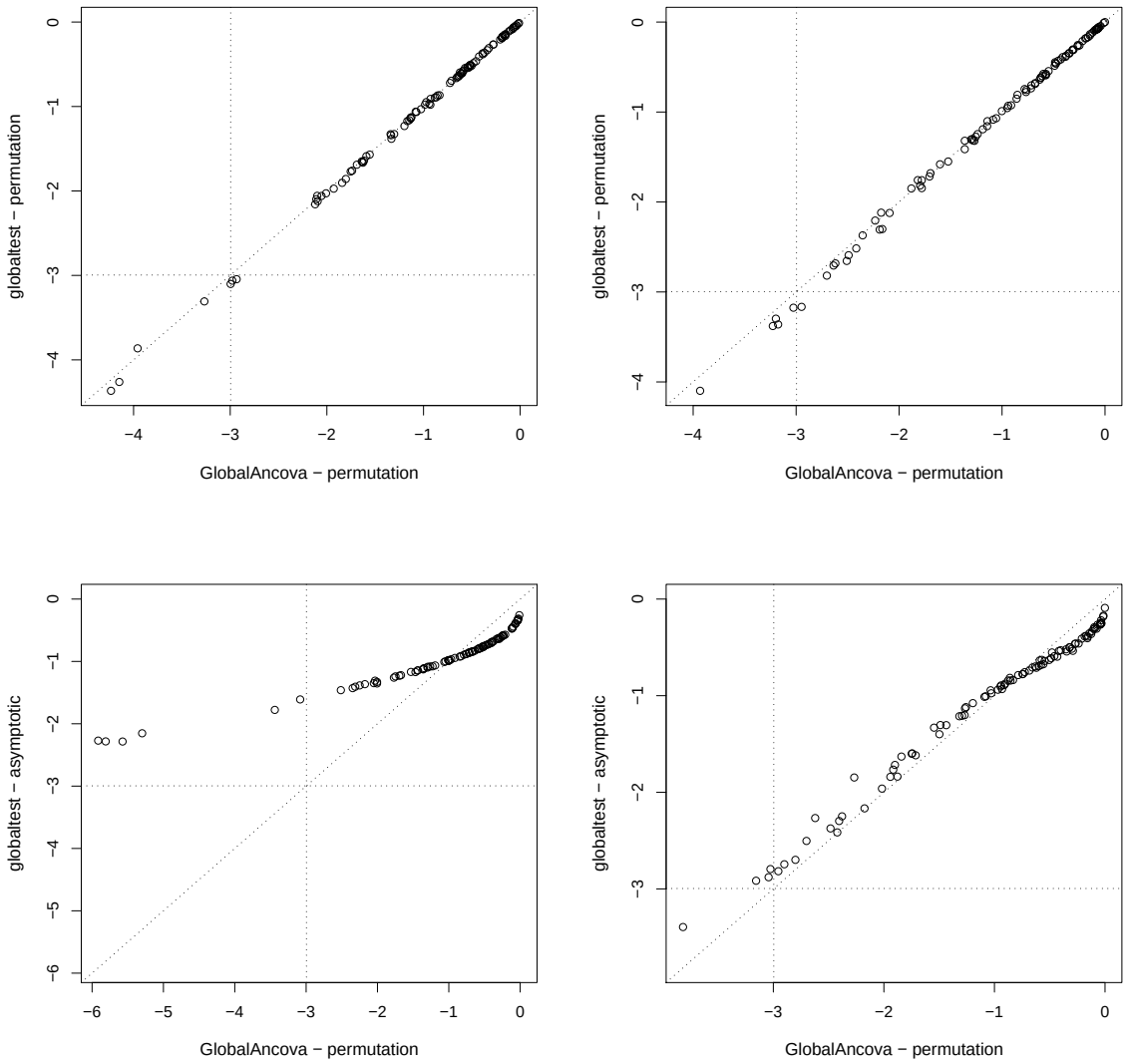


Abbildung 6.2: Vergleich der logarithmierten (permutationsbasierten) GlobalAncova (x-Achse) und der globaltest (y-Achse) p-Werte. Oben: permutationsbasierte globaltest p-Werte, unten: asymptotische globaltest p-Werte. Berechnet für 100 simulierte Datensätze á 200 Gene und 40 Beobachtungen; ohne differentielle Expression. Links: ohne Abhängigkeiten zwischen den Genen, rechts: mit gleicher Korrelation $\rho = 0.2$ zwischen je zwei Genen. Vertikale und horizontale Linien markieren das 5%-Niveau.

Globale Tests und weitere holistische Verfahren

Hotelling's T^2-Test (Kong u. a., 2006; Song und Black, 2007), das Verfahren basierend auf Singulärwertzerlegung (Tomfohr u. a., 2005) und die SAM Gengruppenanalyse (SAM-GS, Dinu u. a., 2007) haben gegenüber den globalen Tests GlobalAncova und globaltest den

	α-Fehler		Macht	
	ohne Korr.	mit Korr.	ohne Korr.	mit Korr.
GlobalAncova – Permutation	0.04	0.05	1	0.3
GlobalAncova – Asymptotisch	0.05	0.12	1	0.5
globaltest – Permutation	0.07	0.06	1	0.29
globaltest – Asymptotisch	0	0.01	0.68	0.17

Tabelle 6.1: Ergebnisse der Simulationsstudie zum Vergleich von globaltest und GlobalAncova, jeweils mit permutationsbasierter und asymptotischer Berechnung der p-Werte. Bei einem Signifikanzniveau von 5% sind α-Fehler für Szenarien unter der Nullhypothese und die Macht für Szenarien unter der Alternativhypothese dargestellt. Es wurden jeweils 100 Datensätze á 200 Gene und 40 Beobachtungen simuliert. In den Szenarien zur Schätzung der Macht wurden jeweils zehn der 200 Gene differentiell exprimiert. Es wurden jeweils sowohl unabhängige Daten als auch Daten mit gleicher Korrelation $\rho = 0.2$ zwischen je zwei Genen simuliert.

entscheidenden Nachteil, daß mit ihnen nur einfache Vergleiche zwischen zwei klinischen Gruppen durchgeführt werden können. Mit den beiden globalen Tests kann man dagegen den Zusammenhang zwischen Genexpression und mehrkategorialen oder stetigen Variablen, sowie mit globaltest auch Überlebenszeiten analysieren. Ein weiterer wichtiger Vorteil der globalen Tests ist die Möglichkeit, Kovariablen mit zu berücksichtigen. Dieses Thema wird bei Analysen von Microarrays leider häufig vernachlässigt. Dabei haben beispielsweise das Alter oder Geschlecht oder krankheitsspezifische Faktoren oft einen deutlichen Einfluß auf die Genexpression.

Kong u. a. (2006) vergleichen Hotelling's T^2-Test mit dem globaltest anhand von realen Daten. Letzterer detektiert etwas mehr Gengruppen, scheint also die größere Macht zu besitzen. Die Übereinstimmung zwischen den Ergebnissen der beiden Verfahren ist für verschiedene Datensätze unterschiedlich gut. An einem Beispiel wird argumentiert, daß einige wichtige pathways nur durch Hotelling's T^2-Test entdeckt werden, nicht aber mit globaltest. Ein Vergleich mit Hilfe von simulierten Daten würde eventuell mehr Aufschluß bringen. Laut Liu u. a. (2007) besitzt die Methode von Tomfohr u. a. (2005) weniger Macht als die globalen Tests. Bei der Singulärwertzerlegung kann das Problem auftreten, daß die Richtung der ersten Hauptkomponente nicht der Richtung entspricht, die die beiden Phänotypgruppen trennt. Somit kann die differentielle Expression in manchen Fällen nicht richtig erfaßt werden. Ebenfalls in Liu u. a. (2007) werden die globalen Tests mit dem Verfahren SAM-GS verglichen, das genweise SAM Statistiken zu Gruppenstatistiken zusammenfaßt. Die Methode SAM-GS zeigt dabei eine größere Macht als die globalen Tests. Es wird eine Standardisierung der Expressionsdaten vor Verwendung der globalen Tests vorgeschlagen, die die Gleichheit der Varianzen über die Gene hinweg gewährleisten soll. Nach einer solchen Standardisierung sind die Ergebnisse der globalen Tests denen von

SAM-GS sehr ähnlich und der Vorteil der letzteren Methode hinsichtlich der Macht ist nur noch geringfügig.

6.1.3 Verallgemeinerung der Assoziation zwischen Annotation und Expression

Bei der in dieser Arbeit beschriebenen Analyse von Gengruppen geht es im Grunde immer darum, die Genexpression, die über die Population hinweg variabel ist, mit fest vordefinierten funktionellen Gengruppen in Verbindung zu bringen. Dudoit u. a. (2006) bieten ein allgemeines System für die Darstellung jener Assoziation zwischen variabler Expression und fixer Annotation der Gene zu Gruppen. Alle bisher beschriebenen Verfahren können als Varianten dieses Systems angesehen werden. Das Konzept ist demnach dem Category Prinzip (Jiang und Gentleman, 2007) recht ähnlich.

Zunächst werden die Gene hinsichtlich ihrer Expression bewertet, zum Beispiel mit genweisen t-Tests. Es ergibt sich daraus ein *Genparameterprofil* $\lambda = (\lambda_1, \ldots, \lambda_m)$. Dieses ist unbekannt und wird mit Hilfe von n Microarrays geschätzt. Im Gegensatz dazu ist das *Genannotationsprofil* $A = (a_{j,k})_{\substack{j=1,\ldots,m \\ k=1,\ldots,K}}$ bekannt und identisch für alle Beobachtungseinheiten. „Bekannt“ bezieht sich dabei auf den aktuellen Kenntnisstand; über die Zeit hinweg ändern sich Annotationsprofile natürlich. Die $m \times K$ Matrix A ist meist eine binäre Inzidenzmatrix, die Auskunft über die Zugehörigkeit von Gen j zu Gruppe k gibt. Es sind aber auch stetige Annotationsprofile möglich, wenn zum Beispiel die Position eines Gens innerhalb der Gruppe angegeben wird. Ob auffällige Verbindungen zwischen den Genparameter- und den Genannotationsprofilen bestehen, wird mit Hilfe eines Assoziationsmaßes $\psi = (\psi_1, \ldots, \psi_K) = \rho(A, \lambda)$ bewertet. Wie ψ gewählt werden sollte, hängt von A, λ und der spezifischen Fragestellung ab. Es kann ein univariates Maß definiert werden, das heißt für die Bestimmung von ψ_k wird nur das k-te Annotationsprofil $a_{\cdot,k}$ verwendet. Dudoit u. a. (2006) schlagen die folgenden univariaten Assoziationsmaße vor.

- **Stetiges Genannotationsprofil und stetiges Genparameterprofil**
  Als Assoziationsmaß ψ_k bietet sich der Pearson Korrelationskoeffizient zwischen den beiden m-Vektoren λ und $a_{\cdot,k}$ an.

- **Binäres Genannotationsprofil und binäres Genparameterprofil**
  Ein binäres Genparameterprofil entspricht der Einteilung in „interessante“ und „nicht interessante“ Gene. Diese Einteilung kann durch Auswahl der „L besten“ Gene oder durch geeignete Adjustierung für multiples Testen und einen cutoff erfolgen. Als Assoziationsmaß wird die χ^2-Statistik vorgeschlagen. Dies entspricht folglich der klassischen kompetitiven Teststrategie für 2×2 Kontingenztafeln.

- **Binäres Genannotationsprofil und stetiges Genparameterprofil**

  (i) Das einfachste Assoziationsmaß lautet

  $$\psi = A^t\lambda, \qquad \psi_k = \sum_{j=1}^{m} I(a_{j,k} = 1)\lambda_j.$$

  Neben der Summe der entsprechenden genweisen Statistiken sind auch standardisierte Summen (Gentleman und Falcon, 2007), Mittelwerte (Tian u. a., 2005), maxmean-Statistiken (Efron und Tibshirani, 2007) etc. denkbar. Es handelt sich hier um in sich geschlossene Tests, da ψ_k nur von den λ_j innerhalb der k-ten Gengruppe abhängt.

  (ii) In ihren Anwendungsbeispielen verwenden Dudoit u. a. (2006) aber meist Statistiken der kompetitiven Strategie, wie zum Beispiel den zwei-Stichproben t-Test zum Vergleich der λ_j innerhalb und außerdhalb der k-ten Gruppe (siehe auch Tian u. a., 2005; Alexa und Rahnenführer, 2007).

Noch allgemeiner könnte man auch multivariate Assoziationsmaße $\rho(A, \lambda)$ in Betracht ziehen, das heißt der k-te Assoziationsparameter ψ_k könnte von der ganzen Annotationsmatrix A abhängen. Beispielsweise könnten Linearkombinationen von Assoziationsparametern mehrerer Gruppen betrachtet werden. Im Falle der Gene Ontology Analyse wäre es von Vorteil, die Struktur des GO Graphen zu berücksichtigen, indem man Annotationsinformationen der Verwandten eines Knotens mit einbezieht.

Der Test auf die Assoziation zwischen Genparameter- und Annotationsprofilen ist allgemein ein Test für die Hypothesenpaare

$$H_{0,k} : \psi_k = \psi_k^0 \quad \text{versus} \quad H_{1,k} : \psi_k \neq \psi_k^0.$$

In manchen Fällen kann auch die einseitige Testsituation

$$H_{0,k} : \psi_k \leq \psi_k^0 \quad \text{versus} \quad H_{1,k} : \psi_k > \psi_k^0$$

gefragt sein. Eine entsprechende Teststatistik kann lauten

$$T_k = \sqrt{n}(\psi_k - \psi_k^0).$$

Im Falle stetiger Genparameterprofile handelt es sich um das zweiseitige Testproblem und wenn für ψ beispielsweise ein t-Test gewählt wird, ist $\psi_k^0 = 0$. Bei binären Genparameterprofilen und der Verwendung des χ^2-Tests liegt das einseitige Testproblem vor mit $\psi_k^0 = 1$. Die Verteilung der Teststatistik T unter der Nullhypothese kann durch Permutationsansätze geschätzt werden, wobei die Beobachtungen permutiert werden.

Analysen von Dudoit u. a. (2006) mit einem realen Datensatz zeigen, daß die Ergebnisse unter Verwendung von binären Genparameterprofilen einerseits und stetigen andererseits

nur schlecht übereinstimmen. Bei den stetigen Parametern wird mehr Information genutzt. Dies macht sich in einer größeren Macht bemerkbar. Bei den binären Genparameterprofilen besteht außerdem das Problem, daß deren Definition, also die Einteilung in „interessante“ und „nicht interessante“ Gene, einen deutlichen Einfluß auf die Ergebnisse hat. Diese Form der Gengruppenanalyse ist also nicht sehr robust. Die Autoren sprechen sich aus diesen Gründen eher gegen die derzeit meist verwendete Strategie der 2×2 Tafeln aus.

6.2 Simulationsstudie für die Gene Ontology

Die wachsende Anzahl an Verfahren für die Analyse von Gengruppen macht die Entwicklung von Vergleichskriterien und Entscheidungshilfen für die Wahl der jeweils geeigneten Methoden unentbehrlich. In einer solchen Situation werden üblicherweise Simulationsstudien durchgeführt. Für den allgemeinen Vergleich einiger Gruppentests gibt es bereits einzelne Simulationsergebnisse (Dinu u. a., 2007; Liu u. a., 2007; Ackermann, 2008). In diesem Abschnitt wird der Versuch einer Simulationsstudie speziell für die Analyse von Gene Ontology Gengruppen gezeigt. Ähnliche Ansätze finden sich bei Alexa u. a. (2006) und Grossmann u. a. (2007).

6.2.1 Simulationsansatz

Es ist bei weitem nicht immer eine leichte Angelegenheit, eine Simulationsstudie sinnvoll, möglichst realistisch und objektiv zu gestalten. Stets besteht die Gefahr, genau das zu entdecken, was man gerne entdecken möchte. Gerade im Falle der Gene Ontology ist es sehr schwierig einen geeigneten Simulationsansatz zu finden, da nicht einmal eindeutig ist, welche Strukturen innerhalb der GO überhaupt detektiert werden sollen. Manche Verfahren zielen darauf ab, einzelne bedeutende Begriffe heraus zu filtern, während andere nach signifikanten Subgraphen suchen (vergleiche Abschnitt 5.2). Bei der Auswertung der Simulationsergebnisse werden hier deshalb beide Ansätze betrachtet.

In den meisten Simulationsstudien werden Daten künstlich erzeugt. Genexpressionsdaten werden in der Regel als multivariat normalverteilte Zufallszahlen simuliert. Differentielle Expression wird dadurch erzeugt, daß die entsprechenden Gene in einer der Phänotypgruppen deutlich erhöht oder erniedrigt werden. Ein solches Szenario ist nicht besonders realistisch. Zum einen sind eventuell manche statistischen Tests im Vorteil, da die geforderten Verteilungsannahmen erfüllt sind. Zum anderen wird differentielle Expression eher als Kontinuum beobachtet und nicht als klare Trennung zwischen differentiell und nicht-differentiell. Aus diesen Gründen basiert unsere Simulationsstudie auf realen Expressionsdaten, nämlich dem Brustkrebsdatensatz von van’t Veer u. a. (2002). Auch Dinu u. a. (2007) verwenden reale Daten für eine Simulationsstudie. Sie stellen zufällige Gruppen von Genen zusammen, die sich in einem Mausexperiment entweder als nicht differentiell oder als stark differentiell exprimiert erwiesen haben. Die entstehenden Gengruppen werden dann auf der Basis derselben Expressionsdaten mit Gruppentests analysiert.

In unserem Fall sind die Gengruppen allerdings durch die Gene Ontology vorgegeben. Um dennoch zum Zweck der Simulation bestimmen zu können, welche GO Gruppen mit differentieller Expression angereichert sein sollen, werden die Expressiondaten der einzelnen Gene vertauscht. Wird ein GO Knoten ausgewählt, so bekommen dessen Gene die Expressionsprofile der im reellen Datensatz differentiell exprimierten Gene. Letzteren werden im Gegenzug die Expressionswerte zugeteilt, die in Realität zur ausgesuchten GO Gruppe gehören. Somit wird gewährleistet, daß die gewählten „interessanten“ GO Gruppen tatsächlich differentielle Expression aufweisen. Die Daten können noch etwas verrauscht werden, indem nicht allen Genen in den interessanten Gruppen ein differentielles Expressionsprofil zugewiesen wird, sondern nur einem gewissen Anteil der Gene. Auch die Expressionswerte aller übrigen Gene werden zufällig neu verteilt, damit nicht systematisch manche Strukturen zwischen Genen erhalten bleiben und andere nicht. Abbildung 6.3 verdeutlicht nochmals die Strategie, mit der die gewählten GO Gruppen mit differentieller Genexpression angereichert werden. Bei der Auswahl interessanter Knoten kann man sich auf solche bestimmter Größe beschränken, zum Beispiel GO Begriffe mit 10-500 annotierten Genen. In verschiedenen Szenarien werden unterschiedlich große GO Gruppen selektiert. Für den durch das Tauschen von Genen entstandenen neuen Datensatz wird sodann eine GO Analyse mit den einzelnen Verfahren durchgeführt. Die ganze Prozedur wird $S = 100$ mal wiederholt. Der Simulationsansatz kann wie folgt zusammen gefaßt werden.

Simulationsansatz

Für die s-te Simulation, $s = 1, \ldots, S$:

1. Wähle zufällig g GO Gruppen (mit eingeschränkter Größe). Diese sollen mit differentieller Expression angereichert werden, stellen also die interessanten Knoten, bzw. falschen Nullhypothesen dar.

2. Ermittle die Menge der M Gene, die zu den g gewählten Gruppen gehören.

3. Ermittle im realen Datensatz über genweise Statistiken, z.B. fold changes, eine Liste L der $M = |L|$ am stärksten differentiell exprimierten Gene.

4. Vertausche die Expressionsprofile von $M \cdot (1 - \gamma)$ Genen in den gewählten Gruppen mit $M \cdot (1 - \gamma)$ Expressionsprofilen aus der Liste L der differentiell exprimierten Gene. Der Parameter γ steuert, wie stark die Daten verrauscht werden.

5. Teste alle K GO Gruppen auf der Basis der in 1.-4. simulierten Daten mit den zu vergleichenden Gruppentests.

Fasse die Ergebnisse aus den S Simulationen zusammen.

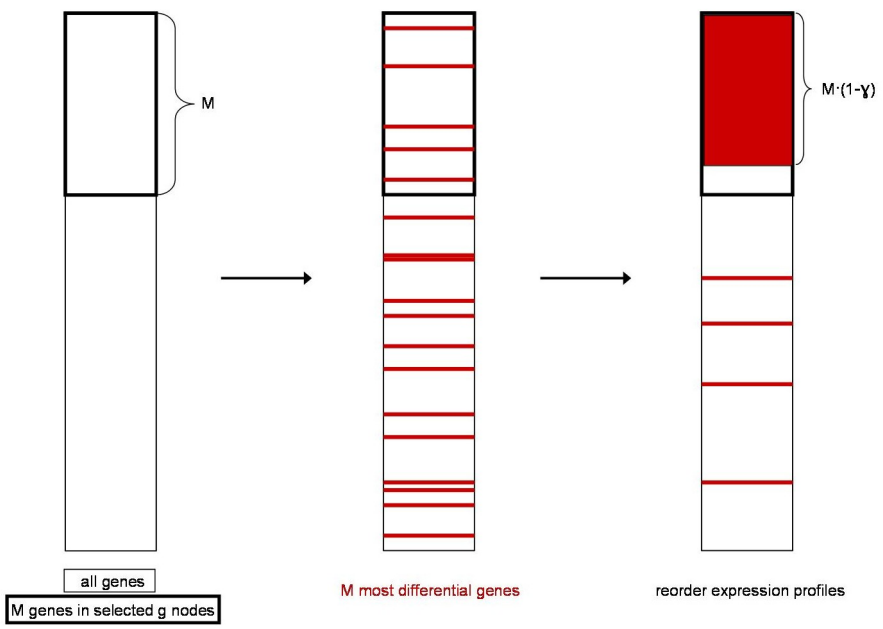


Abbildung 6.3: Schema zur Anreicherung ausgewählter GO Gruppen mit differentieller Expression. Die Struktur der GO bleibt unverändert, aber die Expressionsmatrix wird entsprechend umsortiert.

Wir verwenden für die Simulation die größte der drei Ontologien, nämlich die der Biologischen Prozesse. Die Gene aus dem vorliegenden Datensatz konnten mit insgesamt 3450 GOBP Begriffen assoziiert werden. Es werden in jedem Simulationsdurchlauf $g = 50$ Gruppen ausgewählt. Wir betrachten dabei drei Szenarien, nämlich kleine (10-100 Gene), kleine bis mittlere (10-500 Gene) und große Knoten (500-2000 Gene). Für die Bestimmung der differentiell exprimierten Gene werden fold changes berechnet. Die Gene mit den größten absoluten fold changes werden auf die gewählten Knoten verteilt. Als Anteile, wie viele der Gene in den gewählten Knoten nicht differentiell exprimiert sein sollen verwenden wir $\gamma = 0$, $\gamma = 0.2$ und $\gamma = 0.5$. Für alle Verfahren, die auf genweisen scores für differentielle Expression beruhen, werden t-Test Statistiken berechnet. Wir verwenden bewußt ein anderes Maß für differentielle Expression als für die Generierung der Daten, damit diese Verfahren keinen eventuellen Vorteil genießen gegenüber Methoden, deren Gruppenstatistiken sich nicht aus den genweisen zusammen setzen. In der Simulation werden die folgenden Verfahren verglichen

- **Fisher-Test:** Alle GO Gruppen werden unabhängig voneinander mit dem klassischen Test auf Gene Set Enrichment überprüft. Die hierfür benötigte Liste differentieller Gene wird durch t-Tests und FDR-Adjustierung bestimmt (cutoff der adjustierten p-Werte: 0.05). Die p-Werte der GO Gruppen werden auf der Basis der hypergeometrischen Verteilung berechnet.

- **Kolmogorov-Smirnov-Test (KS):** Die KS Statistik entspricht dem GSEA Ansatz. Die Gene werden nach absoluten t-Statistiken sortiert. Aus Gründen der Rechenzeit wird die Signifikanz über die asymptotische Verteilung und nicht über einen Permutationsansatz bestimmt.

- **t-Test für genweise Statistiken:** Der mittlere Unterschied der absoluten genweisen t-Teststatistiken zwischen je einer GO Gruppe und allen übrigen Genen wird mit t-Tests überprüft (mit Annahme der Varianzengleichheit). Die p-Werte werden gemäß der t-Verteilung berechnet.

- **elim:** Beim klassischen elim Ansatz werden die GO Gruppen mit Fisher-Tests bewertet. Die Liste differentieller Gene wird wie oben definiert. Die Gene eines Knotens werden dann aus den Vorfahren entfernt, wenn der Knoten einen Fisher-Test p-Wert $< 10^{-6}$ hat.

- **elim + KS-Test:** Der elim Algorithmus wird wie üblich durchgeführt. Die Bewertung der GO Gruppen erfolgt über den Kolmogorov-Smirnov-Test für Genränge gemäß absoluter t-Statistiken. Die Gene eines Knotens werden dann aus den Vorfahren entfernt, wenn der Knoten einen KS p-Wert $< 10^{-6}$ hat.

- **elim + t-Test:** Der elim Algorithmus wird wie üblich durchgeführt. Die Bewertung der GO Gruppen erfolgt über einen t-Test für den Vergleich absoluter t-Statistiken. Die Gene eines Knotens werden dann aus den Vorfahren entfernt, wenn der Knoten einen t-Test p-Wert $< 10^{-6}$ hat.

- **weight:** Der weight Algorithmus basiert auf Fisher-Tests. Die Liste differentieller Gene wird wie oben definiert.

- **Category:** Für den klassischen Category Ansatz werden genweise t-Statistiken innerhalb der GO Gruppen aufsummiert und durch die Wurzel der entsprechenden Genanzahlen geteilt. Aus Gründen der Rechenzeit wird die Signifikanz über die Normalverteilungsapproximation und nicht über einen Permutationsansatz bestimmt.

- **globaltest:** Für alle GO Gruppen werden asymptotische globaltest p-Werte berechnet.

- **GlobalAncova:** Für alle GO Gruppen werden permutationsbasierte GlobalAncova p-Werte berechnet.

- **focus level:** Die GO Gruppen werden mit globaltest getestet und gemäß der focus level Methode adjustiert.

Die Verfahren von Tomfohr u. a. (2005) und Grossmann u. a. (2007) werden nicht untersucht, da sie nur als web-tools und nicht als *R* Pakete verfügbar sind. Weiter Methoden aus Kapitel 3.4 fehlen hier, weil sie numerisch sehr aufwendig sind, zum Beispiel Hotelling's T^2 (Kong u. a., 2006; Song und Black, 2007) für sehr große Gengruppen oder der

Restandardisierungsansatz (Efron und Tibshirani, 2007) wegen der benötigten zweifachen Permutation.

6.2.2 Ergebnisse

Raten der Richtig und Falsch Positiven

Für die Bewertung der Simulationsergebnisse betrachten wir zwei verschiedene Ansätze. Die erste naheliegende Vorgehensweise besteht darin, nur die zur Simulation ausgewählten GO Knoten als die tatsächlich interessanten zu betrachten. Die Verfahren werden dahingehend bewertet, wie gut sie diese Knoten detektieren. Diese Strategie entspricht der Sichtweise des Gene Set Enrichment: die am meisten mit differentiellen Genen angereicherten Gruppen sollen entdeckt werden. Aus der Perspektive der globalen Tests dagegen, die jegliche differentielle Expression innerhalb einer Gengruppe aufspüren sollen, muß man aufgrund der hierarchischen Struktur der Gene Ontology automatisch die Vorfahren der gewählten Knoten zu der Menge der tatsächlich interessanten hinzu fügen. Für jeweils die selbe Simulation betrachten wir also zwei verschiedene Mengen von falschen Nullhypothesen

1. Nur die ausgewählten 50 Knoten sind die zu detektierenden Gengruppen (H1).

2. Die ausgewählten Knoten mitsamt all ihren Vorfahren im GO Graphen sind die zu detektierenden Gengruppen. Die Anzahl der falschen Nullhypothesen variiert somit über die Simulationsdurchläufe, je nachdem welche Knoten gewählt wurden (H2).

In beiden Fällen ist zu beachten, daß man aufgrund der Verwendung von realen Daten nicht eindeutig zwischen wahren und falschen Nullhypothesen trennen kann. Je nachdem wie viele Gene im Datensatz tatsächlich differentiell exprimiert sind, kann es weniger oder auch mehr tatsächlich interessante Knoten geben.

Zum Vergleich der einzelnen Verfahren werden ROC-ähnliche Kurven gezeichnet. Zur Schätzung der *Rate der Richtig Positiven* (*True Positive Rate, TPR*) wird für jede Anzahl k an abgelehnten Knoten die Anzahl der detektierten interessanten Knoten, also der Richtig Positiven, ermittelt und durch die Anzahl aller falschen Nullhypothesen geteilt. Die Anzahlen sind wie gesagt unterschiedlich für die oben beschriebenen Ansätze (H1) und (H2). Die berechneten Anteile werden über die 100 Simulationen gemittelt

$$\widehat{TPR}_k = \frac{1}{S}\sum_{s=1}^{S} \frac{\#\text{ Richtig Positive unter } k \text{ Detektierten in Simulation } s}{\#\text{ falsche Nullhypothesen in Simulation } s}, \qquad k = 1, \ldots, K.$$

Analog berechnet sich die geschätzte *Rate der Falsch Positiven* (*False Positive Rate, FPR*) als Anteil der fälschlicherweise detektierten an den wahren Nullhypothesen

$$\widehat{FPR}_k = \frac{1}{S}\sum_{s=1}^{S} \frac{\#\text{ Falsch Positive unter } k \text{ Detektierten in Simulation } s}{\#\text{ wahre Nullhypothesen in Simulation } s}, \qquad k = 1, \ldots, K.$$

Die Werte $\widehat{TPR}_k$ und $\widehat{FPR}_k$ werden gegeneinander geplottet. Wie bei ROC-Kurven ist ein Verfahren umso besser je eher sich seine Kurve in die linke obere Ecke schmiegt, das heißt gleichzeitig eine hohe Rate Richtig Positiver und eine niedrige Rate Falsch Positiver erreicht wird. Bei der sehr großen Anzahl an getesteten GO Gruppen und dabei der relativ großen Menge an wahren Nullhypothesen wird man unter einer angemessenen Menge an detektierten Knoten generell sehr kleine Falsch Positiv Raten erwarten. Deshalb werden jeweils zusätzlich auch die „ROC-Kurven“ nur bis zu einer Falsch Positiv Rate von 0.05 gezeichnet. Dies entspricht bereits ca. 200 abgelehnten Hypothesen. Abbildungen 6.4 - 6.8 zeigen die Kurven für alle Simulationsszenarien.

Der Kolmogorov-Smirnov und t-Test und besonders der elim Algorithmus in Verbindung mit diesen beiden erweisen sich für die erste Strategie (H1), für die nur die 50 gewählten Knoten detektiert werden sollen, als vorteilhaft. Interessiert man sich aber auch für die Vorfahren der ausgesuchten GO Gruppen (H2), so bekommt man für die Kombinationen mit elim deutlich schlechtere Ergebnisse. Das läßt sich dadurch erklären, daß bei elim jeweils die Gene der signifikanten Knoten aus den Vorfahren gelöscht werden. Somit sind diese in den meisten Fällen nicht mehr genug mit differentiellen Genen angereichert. KS und t-Test aber erscheinen für die zweite Strategie (H2) sinnvoll. Im Szenario mit großen GO Knoten schneiden sie besonders gut ab. Bei einer generellen Bevorzugung großer Gengruppen erklärt sich, warum auch die Vorfahren der angereicherten Knoten recht gut detektiert werden.

Der klassische Fisher-Test erweist sich in dieser Simulationsstudie als recht passabel. Erstaunlicherweise gilt das sogar für die Strategie (H2). Für (H1) ist er dem klassischen elim Ansatz, der ebenfalls den Fisher-Test benutzt, recht ähnlich. In Situationen in denen der Fisher-Test Schwierigkeiten hat, nämlich in den Szenarien mit sehr großen gewählten Gengruppen und wenn die Anreicherung mit differentiell exprimierten Genen stark verrauscht ist, fallen Fisher-Test und elim quasi zusammen (deshalb ist die Linie für elim in den Abbildungen 6.4 - 6.8 teilweise nicht zu sehen). Wenn nämlich der Fisher-Test keine signifikanten Ergebnisse liefert, kommt der elim Algorithmus nicht zum Tragen, da ja nur die Gene von signifikanten Knoten aus deren Vorfahren entfernt werden. Für die Strategie (H2) ist das elim Verfahren aus den oben erwähnten Gründen gegenüber dem Fisher-Test im Nachteil.

Die Leistung der globalen Tests ist in etwa vergleichbar mit der des Fisher-Tests. Die Unterschiede zwischen globaltest und GlobalAncova lassen sich durch die unterschiedliche Berechnung der p-Werte erklären. Würden auch für globaltest permutationsbasierte p-Werte verwendet, könnte man wohl sehr ähnliche Ergebnisse erwarten. In schwierigeren Situationen (große Knoten oder stark verrauschte Daten) scheint die größere Macht des Permutationsansatzes von Vorteil zu sein, da hier GlobalAncova (Permutationsansatz) besser abschneidet als globaltest (asymptotischer Ansatz). Für die Strategie (H2) sind die globalen Tests etwas besser geeignet als für (H1).

Das Category Verfahren liefert allgemein sehr schlechte Ergebnisse. Dies mag zum Groß-

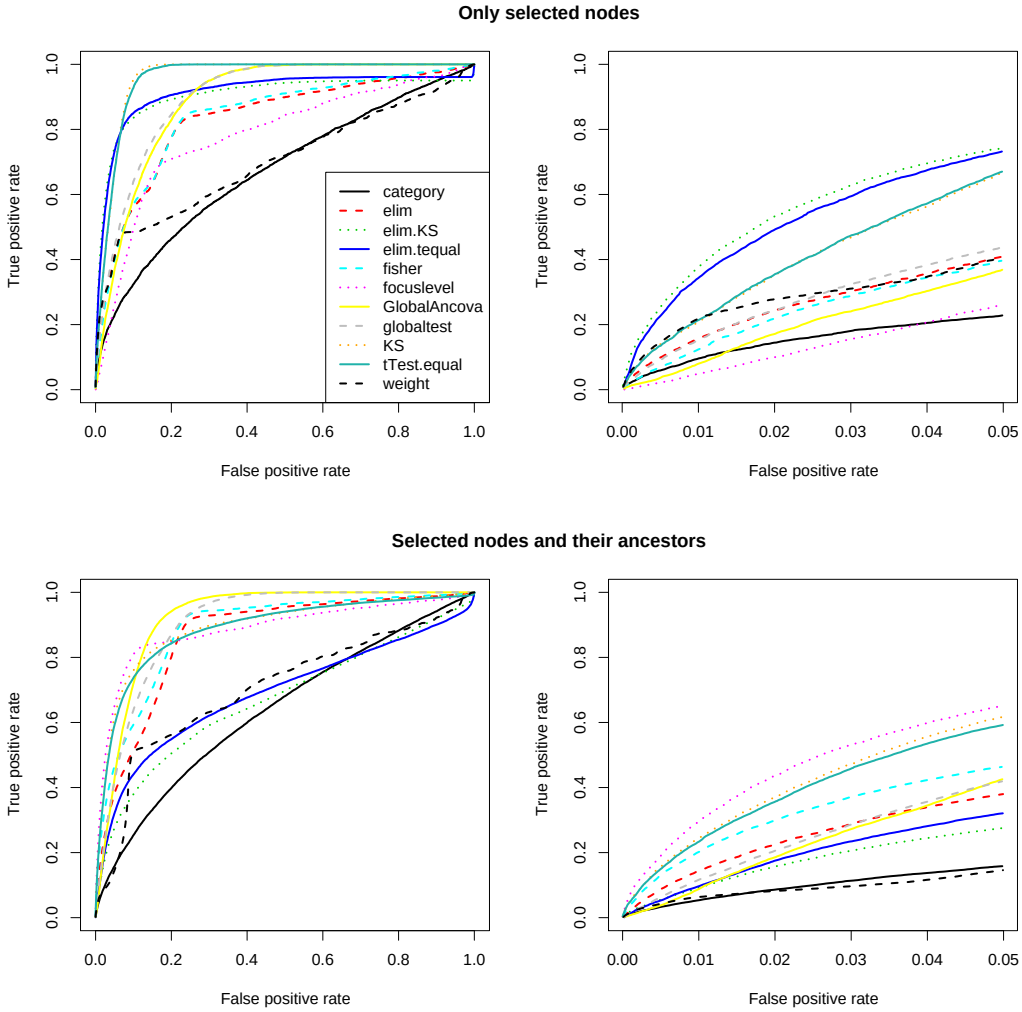


Abbildung 6.4: Simulationsergebnisse für das Szenario 'kleine GO Gruppen': Auswahl von 50 GO Gruppen mit 10 bis 100 Genen, $\gamma = 0$, 100 Simulationen. Richtig Positiv Raten sind gegen Falsch Positiv Raten aufgetragen, links für FPR bis 1, rechts für FPR bis 0.05. Oben: nur die ausgewählten 50 Gruppen sind falsche Nullhypothesen (H1), unten: die ausgewählten 50 Gruppen und all ihre Vorfahren sind falsche Nullhypothesen (H2).

teil daran liegen, daß sich in der Category Statistik starke Effekte in entgegen gesetzte Richtungen gegenseitig „aufheben“ können. Wenn allgemein nach differentieller Expression und nicht nach „gleichgerichteter“ differentieller Expression gefragt ist, müßte man eher eine andere Gruppenstatistik wählen, wie zum Beispiel Summe oder Mittelwert der *absoluten* genweisen Statistiken, die maxmean-Statistik (Efron und Tibshirani, 2007) oder SAM-GS (Dinu u. a., 2007). Dafür müßten die p-Werte über einen Permutationsansatz berechnet werden, da die Normalverteilungsapproximation dann nicht gilt. Womöglich ist die Verwendung der Normalverteilung auch im Falle der klassischen Category Statistik problematisch aufgrund von Abhängigkeiten zwischen den Genen. Dies kann ein zusätzlicher

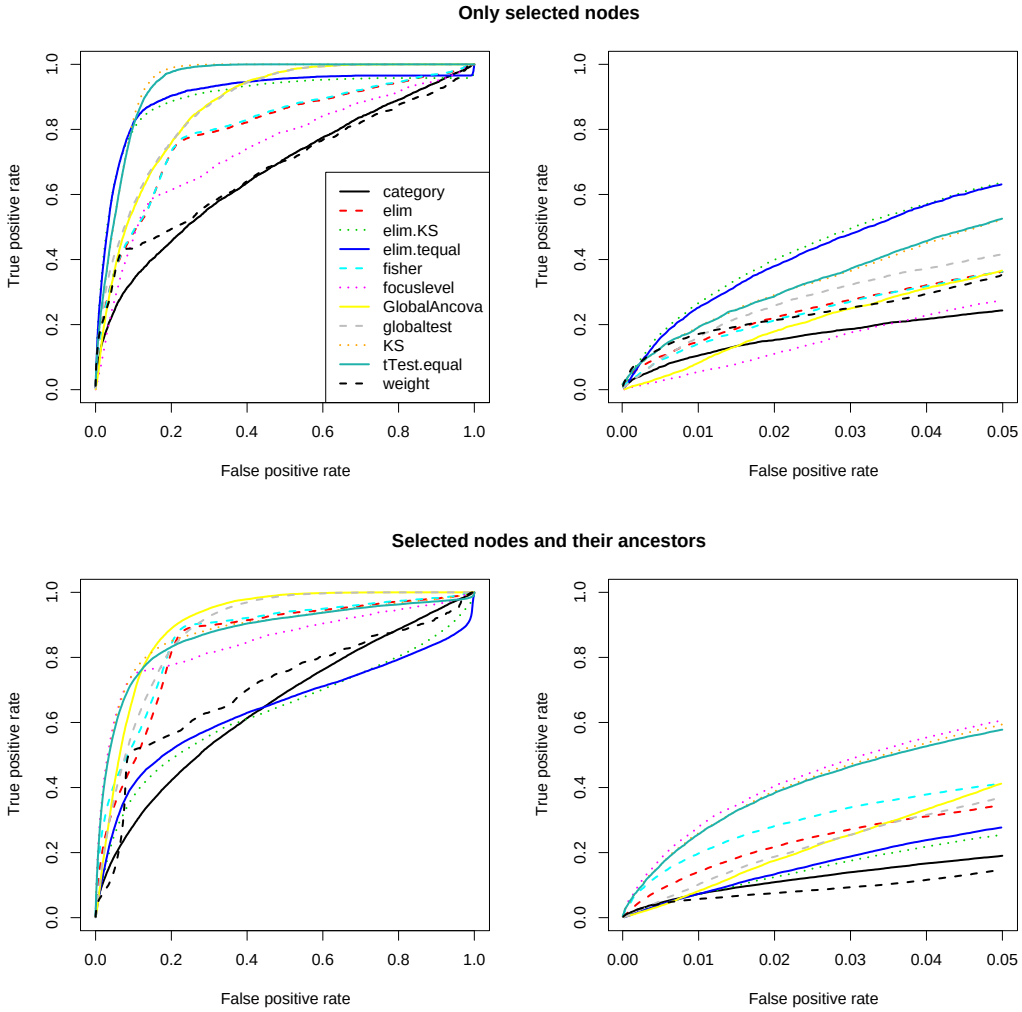


Abbildung 6.5: Simulationsergebnisse für das Szenario 'kleine bis mittelgroße GO Gruppen': Auswahl von 50 GO Gruppen mit 10 bis 500 Genen, $\gamma = 0$, 100 Simulationen. Richtig Positiv Raten sind gegen Falsch Positiv Raten aufgetragen, links für FPR bis 1, rechts für FPR bis 0.05. Oben: nur die ausgewählten 50 Gruppen sind falsche Nullhypothesen (H1), unten: die ausgewählten 50 Gruppen und all ihre Vorfahren sind falsche Nullhypothesen (H2).

Grund sein für das schlechte Abschneiden des Category Ansatzes.

Desweiteren liefert auch der weight Algorithmus in dieser Studie keine guten Ergebnisse. Die Wirkungsweise dieses Verfahrens müßte noch weiter untersucht werden.

Die focus level Methode ist für Strategie (H1) wie erwartet nicht geeignet, am ehesten noch für das Szenario mit großen Knoten. Die Stärke des focus level Verfahrens liegt eindeutig bei (H2), wenn signifikante Subgraphen innerhalb der GO detektiert werden sollen.

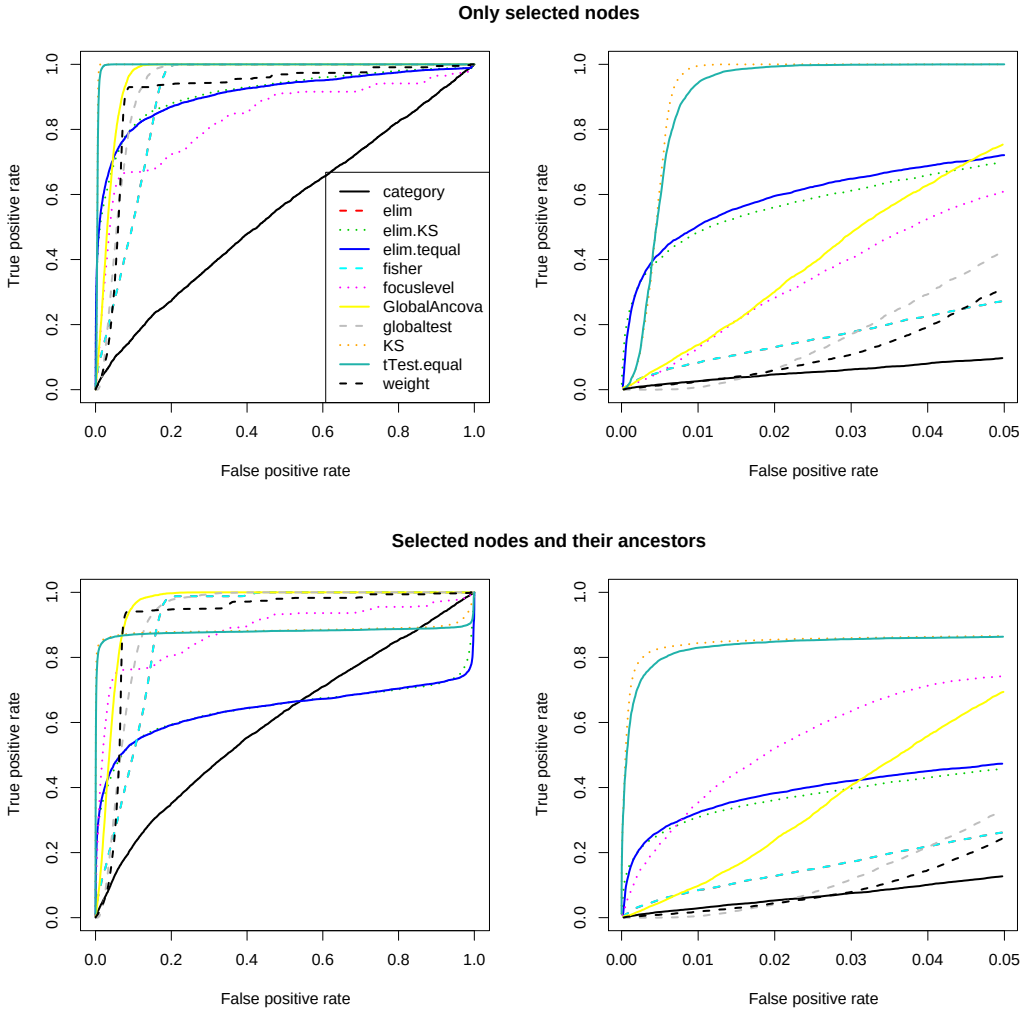


Abbildung 6.6: Simulationsergebnisse für das Szenario 'große GO Gruppen': Auswahl von 50 GO Gruppen mit 500 bis 2000 Genen, $\gamma = 0$, 100 Simulationen. Richtig Positiv Raten sind gegen Falsch Positiv Raten aufgetragen, links für FPR bis 1, rechts für FPR bis 0.05. Oben: nur die ausgewählten 50 Gruppen sind falsche Nullhypothesen (H1), unten: die ausgewählten 50 Gruppen und all ihre Vorfahren sind falsche Nullhypothesen (H2).

Hier liefert es im Bereich kleiner Falsch Positiv Raten meist von allen Verfahren die besten Ergebnisse. Betrachtet man die Kurven für alle FPR Werte bis eins, so fällt auf, daß die Linie von focus level bei (H2) zunächst stark ansteigt und den anderen Verfahren überlegen ist, bis sie dann einen „Knick“ macht und sehr flach weiter verläuft. Dies liegt daran, daß focus level im Gegensatz zu den anderen Verfahren eine (echte) Adjustierung für multiples Testen darstellt. Demnach haben viele p-Werte den Wert eins und die entsprechenden Knoten können deshalb auch nicht weiter „sinnvoll“ angeordnet werden. Nach diesem Punkt ist es also Zufall, ab der wie vielten abgelehnten Hypothese die verbleibenden, zuvor nicht detektierten interessanten Knoten hinzu kommen.

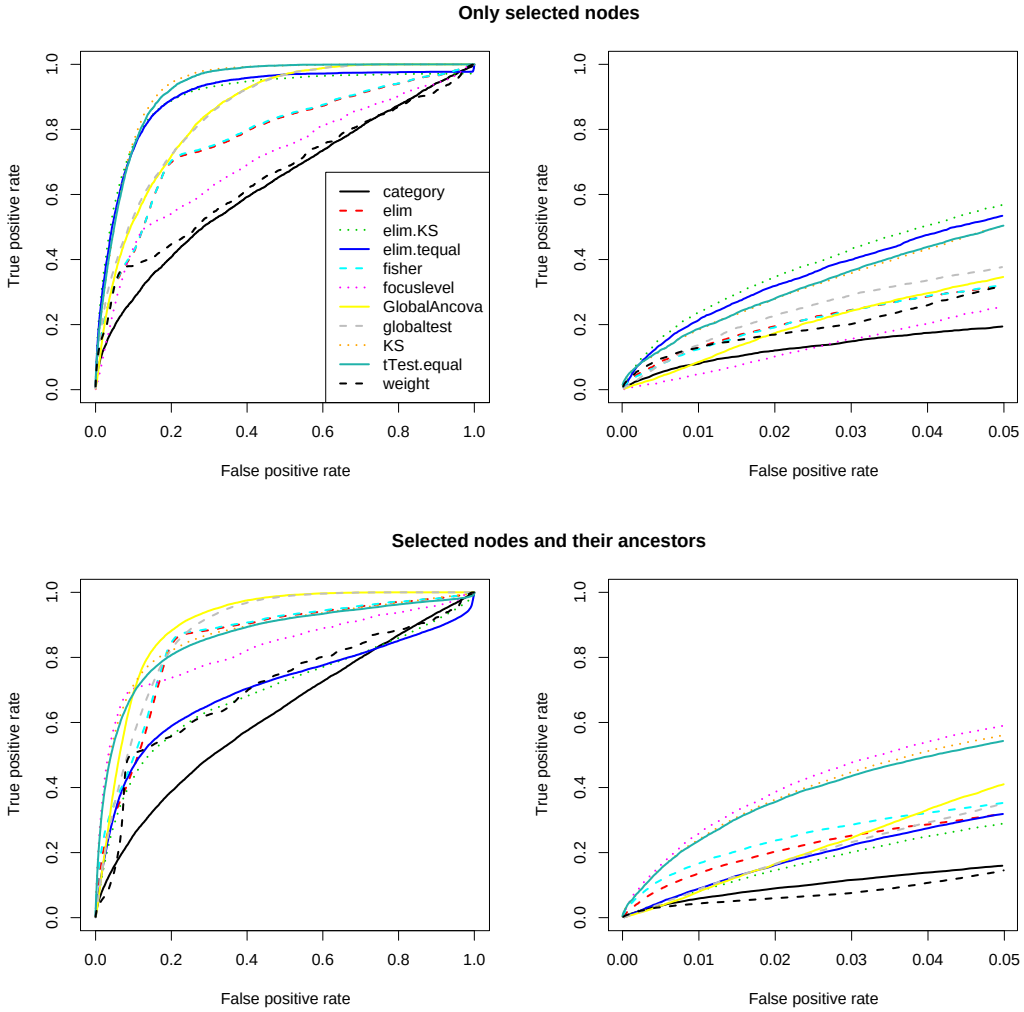


Abbildung 6.7: Simulationsergebnisse für das Szenario 'etwas verrauscht': Auswahl von 50 GO Gruppen mit 10 bis 500 Genen, $\gamma = 0.2$, 100 Simulationen. Richtig Positiv Raten sind gegen Falsch Positiv Raten aufgetragen, links für FPR bis 1, rechts für FPR bis 0.05. Oben: nur die ausgewählten 50 Gruppen sind falsche Nullhypothesen (H1), unten: die ausgewählten 50 Gruppen und all ihre Vorfahren sind falsche Nullhypothesen (H2).

True Discovery Rate

Anstatt die Verfahren auf der Basis der jeweils „besten“ k Knoten zu vergleichen, kann man auch fragen wie viele Gengruppen mit den verschiedenen Methoden tatsächlich als signifikant beurteilt würden. Bei der großen Anzahl an Tests muß über eine Adjustierung für multiples Testen nachgedacht werden. Wir verwenden beispielhaft für alle Verfahren eine Holm Korrektur. Ausgenommen werden davon nur die Ergebnisse der focus level Methode, da diese bereits die FWER kontrolliert, und die elim und weight Algorithmen. Letztere stellen zwar keine Adjustierung für multiples Testen im eigentlichen Sinne dar, dennoch werden

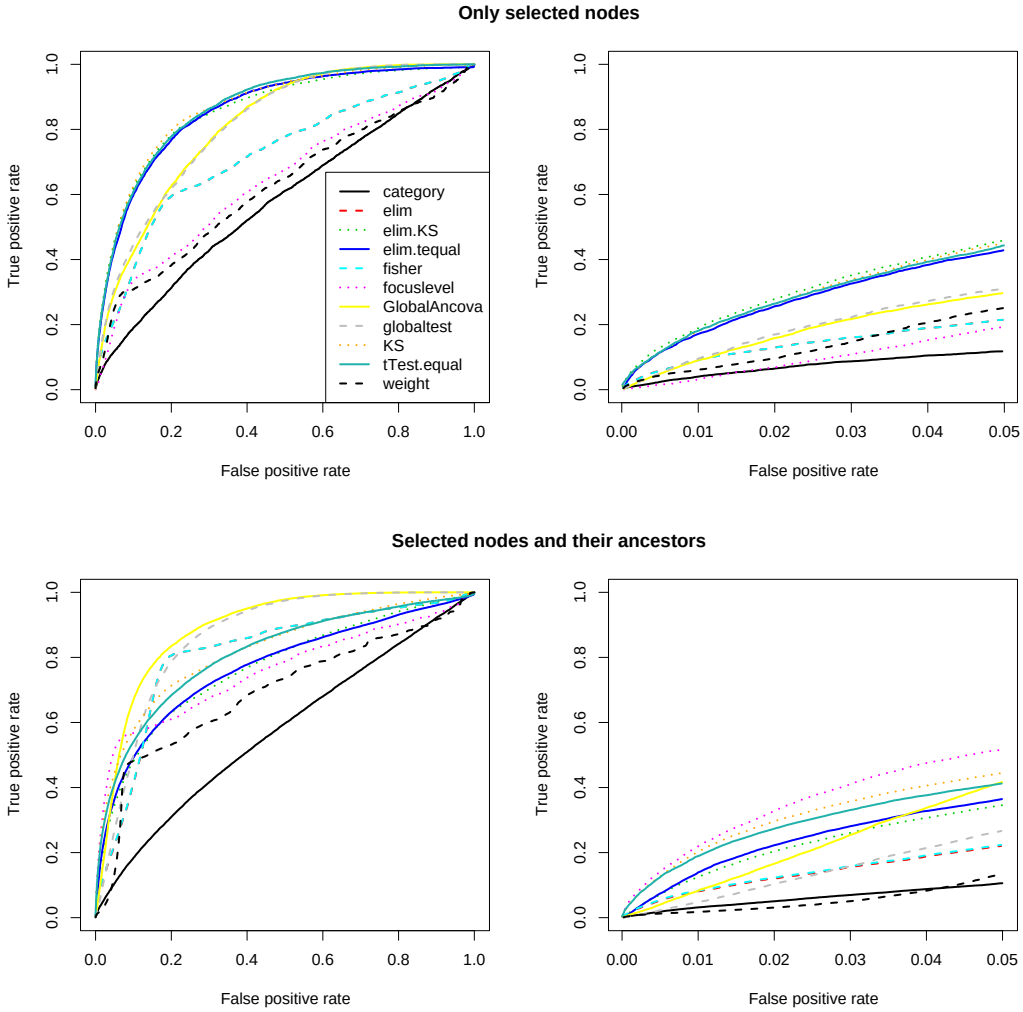


Abbildung 6.8: Simulationsergebnisse für das Szenario 'stark verrauscht': Auswahl von 50 GO Gruppen mit 10 bis 500 Genen, $\gamma = 0.5$, 100 Simulationen. Richtig Positiv Raten sind gegen Falsch Positiv Raten aufgetragen, links für FPR bis 1, rechts für FPR bis 0.05. Oben: nur die ausgewählten 50 Gruppen sind falsche Nullhypothesen (H1), unten: die ausgewählten 50 Gruppen und all ihre Vorfahren sind falsche Nullhypothesen (H2).

durch die Prozeduren bereits Signifikanzen abgeschwächt, beziehungsweise hervorgehoben. Für die jeweilige Menge der detektierten Knoten kann die Richtig Positiv Rate (entspricht der Macht) ermittelt werden. Allerdings variieren die Verfahren sehr stark bezüglich der Anzahlen abgelehnter Hypothesen. Eine Methode, die sehr viele Gruppen für signifikant befindet, wird leicht eine relativ große Macht besitzen. Die große Menge an falsch positiven Ergebnissen geht in die Berechnung der Macht nicht ein. Für die Praxis ist an dieser Stelle eher von Bedeutung, welcher Anteil an den abgelehnten Hypothesen tatsächlich interessanten GO Gruppen entspricht. Wir definieren hierfür eine Art *True Discovery Rate*

Szenario	**Kleine Knoten**			**Kleine – mittlere Kn.**			**Große Knoten**		
Strategie		**(H1)**	**(H2)**		**(H1)**	**(H2)**		**(H1)**	**(H2)**
interessant		50	268.9		50	255.2		50	83.7
	signif.	**TDR**	**TDR**	**signif.**	**TDR**	**TDR**	**signif.**	**TDR**	**TDR**
Fisher	47.4	0.16	**0.70**	33.3	0.19	**0.74**	0.1	0.00	0.01
KS	263.8	0.15	0.55	324.9	0.12	0.48	191.8	0.27	0.39
t-Test	438.1	0.11	0.43	460.9	0.10	0.39	149.9	**0.34**	0.49
elim + Fisher	424.9	0.07	0.32	347.7	0.07	0.31	93.7	0.09	0.14
elim + KS	802.6	0.06	0.17	890.8	0.05	0.15	722.0	0.06	0.07
elim + t-Test	892.7	0.05	0.17	960.3	0.05	0.15	698.3	0.06	0.07
weight	115.8	0.14	0.22	98.6	0.12	0.22	33.4	0.04	0.05
Category	112.4	0.08	0.25	109.7	0.09	0.29	15.6	0.04	0.08
globaltest	67.3	**0.17**	0.50	42.7	**0.20**	0.45	1.0	0.00	0.00
GlobalAncova	276.4	0.09	0.42	233.3	0.09	0.40	45.3	0.13	0.16
focus level	200.3	0.07	0.64	157.7	0.08	0.62	27.2	0.12	**0.51**

Tabelle 6.2: True Discovery Rates gemäß Strategien (H1) und (H2) für die Szenarien 'kleine GO Gruppen', 'kleine bis mittelgroße GO Gruppen' und 'große GO Gruppen', jeweils Auswahl von 50 Knoten, $\gamma = 0$, 100 Simulationen. Die mittleren Anzahlen zu detektierender Gruppen (Zeile 'interessant'), sowie mittlere Anzahlen detektierter Gruppen (Spalten 'signif.') sind gegeben. Die besten TDR Werte je Szenario und Strategie sind fett markiert.

(TDR)

$$\widehat{TDR} = \frac{1}{S}\sum_{s=1}^{S} \frac{\#\text{ Richtig Positive unter den Detektierten in Simulation } s}{\#\text{ Detektierte in Simulation } s}.$$

Tabellen 6.2 und 6.3 zeigen die TDR für die laut der einzelnen Verfahren nach Adjustierung signifikanten GO Gruppen. Zusätzlich sind die jeweiligen Anzahlen zu detektierender und detektierter Knoten, gemittelt über die 100 Simulationen, angegeben. Erstere Anzahlen sind für Strategie (H1) immer 50 (die 50 ausgewählten Knoten). Für Strategie (H2) kommen noch jeweils die Vorfahren dazu, so daß die Anzahlen von Simulation zu Simulation schwanken und davon abhängen, ob kleine oder große Knoten gewählt werden. Die besten TDR Werte je Szenario und Strategie sind fett markiert.

Als erstes fällt auf, daß die Anzahlen signifikanter GO Gruppen zwischen den einzelnen Verfahren stark variieren, zum Beispiel werden in dem Szenario mit kleinen bis mittelgroßen Knoten mit dem Fisher-Test im Mittel 33 Gruppen detektiert, mit elim in Kombination mit dem t-Test ca. 960. Die Verfahren, die viele Hypothesen ablehnen wie elim mit KS oder elim mit t-Test, sind im Vorteil in Bezug auf die Macht. Betrachtet man dagegen die True Discovery Rate, stehen diese Methoden schlecht da, weil unter den vielen detektierten Knoten meist auch viele Falsch Positive sind. Im Gegensatz dazu haben eher strenge

Szenario	$\gamma = \mathbf{0}$			$\gamma = \mathbf{0.2}$			$\gamma = \mathbf{0.5}$		
Strategie		**(H1)**	**(H2)**		**(H1)**	**(H2)**		**(H1)**	**(H2)**
interessant		50	255.2		50	249.8		50	259.0
	signif.	**TDR**	**TDR**	**signif.**	**TDR**	**TDR**	**signif.**	**TDR**	**TDR**
Fisher	33.3	0.19	**0.74**	15.3	**0.23**	**0.76**	1.8	0.13	0.34
KS	324.9	0.12	0.48	213.1	0.13	0.53	52.3	**0.20**	0.67
t-Test	460.9	0.10	0.39	319.7	0.11	0.44	96.4	0.16	0.58
elim + Fisher	347.7	0.07	0.31	274.7	0.07	0.32	167.7	0.06	0.29
elim + KS	890.8	0.05	0.15	745.2	0.06	0.19	455.0	0.07	0.28
elim + t-Test	960.3	0.05	0.15	811.4	0.06	0.18	505.2	0.07	0.27
weight	98.6	0.12	0.22	78.2	0.11	0.19	50.7	0.08	0.12
Category	109.7	0.09	0.29	72.2	0.09	0.28	26.9	0.06	0.22
globaltest	42.7	**0.20**	0.45	24.0	0.20	0.41	4.9	0.11	0.18
GlobalAncova	233.3	0.09	0.40	175.9	0.10	0.38	87.8	0.10	0.39
focus level	157.7	0.08	0.62	116.8	0.07	0.64	58.9	0.05	**0.69**

Tabelle 6.3: True Discovery Rates gemäß Strategien (H1) und (H2) für die Szenarien 'nicht verrauscht' ($\gamma = 0$), 'etwas verrauscht' ($\gamma = 0.2$) und 'stark verrauscht' ($\gamma = 0.5$), jeweils Auswahl von 50 Knoten mit 10 bis 500 Genen, 100 Simulationen. (Das Szenario 'nicht verrauscht' entspricht dem Szenario 'kleine bis mittelgroße GO Gruppen' aus Tabelle 6.2.) Die mittleren Anzahlen zu detektierender Gruppen (Zeile 'interessant'), sowie mittlere Anzahlen detektierter Gruppen (Spalten 'signif.') sind gegeben. Die besten TDR Werte je Szenario und Strategie sind fett markiert.

Verfahren wie globaltest und Fisher-Test nur eine geringe Macht, dafür aber hohe TDRs, insofern unter den wenigen signifikanten Knoten relativ viele „sinnvolle“ zu finden sind. Dementsprechend haben Fisher-Test und globaltest in den Tabellen 6.2 und 6.3 in der Regel die besten, elim mit KS und elim mit t-Test die schlechtesten Werte. Unter schwierigen Bedingungen, also bei großen Knoten oder stark verrauschten Daten, haben globaltest und Fisher-Test allerdings zu wenig Macht. Es werden fast keine GO Gruppen für signifikant befunden und demnach ist in diesen Fällen auch die TDR nur gering.

Bei Strategie (H2) sind die Werte in jedem Fall besser als bei Strategie (H1). Da jeweils die selben Simulationsergebnisse verwendet werden, betrachtet man die selben signifikanten GO Gruppen. Bei Strategie (H2) ist die Menge der interessanten Knoten größer und enthält die Menge interessanter Knoten aus (H1). Somit hat ein Verfahren bei (H2) niemals weniger und meistens mehr Richtig Positive unter den Detektierten und damit eine höhere TDR. Am geringsten ist die Verbesserung bei (H2) im Vergleich zu (H1) für die verschiedenen elim Verfahren. Dies deckt sich mit den Erkenntnissen, die aus den Abbildungen 6.4 - 6.8 gewonnen wurden. Ebenso wie in den Abbildungen zeigt sich bei (H2) auch hier der deutlichste Vorteil für die focus level Methode. Sie hat bei (H2) stets die beste oder

zweitbeste TDR.

GlobalAncova gehört zu den eher sensitiven Methoden, die viele Hypothesen ablehnen. Aufgrund der vielen Falsch Negativen sind die TDR Werte nicht sehr gut. Für Strategie (H2) ergeben sich allerdings deutlich bessere Ergebnisse. In den Situationen, in denen globaltest wegen zu geringer Macht (wegen Verwendung der asymptotischen p-Werte) zu wenige GO Gruppen detektiert, erzielt GlobalAncova größere TDRs.

Die Verfahren Category und weight zeigen sich auch in den Tabellen 6.2 und 6.3 wie in den Abbildungen 6.4 - 6.8 als nicht sehr vorteilhaft.

Generell erscheinen die präsentierten True Discovery Rates eher niedrig. Hierbei gilt wieder zu bedenken, daß man bei dem vorliegenden Design der Simulationsstudie mit realen Daten nicht wirklich von „Richtig“ und „Falsch“ Positiven sprechen kann. In den echten Daten kann tatsächlich noch mehr oder weniger differentielle Expression vorhanden sein und somit kann es auch mehr oder weniger tatsächlich falsche Nullhypothesen geben.

Übereinstimmungen zwischen den Verfahren

Schließlich ist es noch interessant zu untersuchen, wie stark die Listen der detektierten GO Gruppen zwischen den einzelnen Verfahren überlappen, beziehungsweise welche Verfahren sich diesbezüglich am ähnlichsten sind. Zu diesem Zweck zeigt Abbildung 6.9 die Anzahlen der gemeinsam detektierten Gruppen unter den ersten ein bis 200 Knoten für das Szenario mit kleinen bis mittelgroßen GO Gruppen. Die Ergebnisse der einzelnen Verfahren werden jeweils mit denen eines anderen verglichen. Als solche „Basisverfahren“ wählen wir beispielhaft je einmal den Fisher-Test, Kolmogorov-Smirnov-Test, elim mit KS, weight, Category, globaltest, GlobalAncova und focus level.

Einige Verfahren sind sich recht ähnlich in Bezug auf die detektierten GO Gruppen, nämlich Fisher-Test und elim mit Fisher-Test, KS-Test und t-Test, elim mit t-Test und elim mit KS. Dies konnte bereits aus den Ergebnissen in den vorigen Abschnitten vermutet werden. Die Listen der laut weight oder Category besten k Knoten zeigen sehr wenige Überschneidungen mit denen der anderen Verfahren. GlobalAncova und globaltest liefern relativ ähnliche Listen interessanter Gruppen. Die Übereinstimmung wäre sicherlich noch deutlich größer, wenn für den globaltest ebenfalls permutationsbasierte p-Werte berechnet worden wären. Interessanterweise sind die Überschneidungen zwischen den globalen Tests und dem t-Test und damit auch dem KS-Test relativ hoch. Auch der focus level Methode sind t-Test und KS-Test am ähnlichsten.

Für die anderen Simulationsszenarien ergeben sich mit Abbildung 6.9 weitestgehend vergleichbare Grafiken.

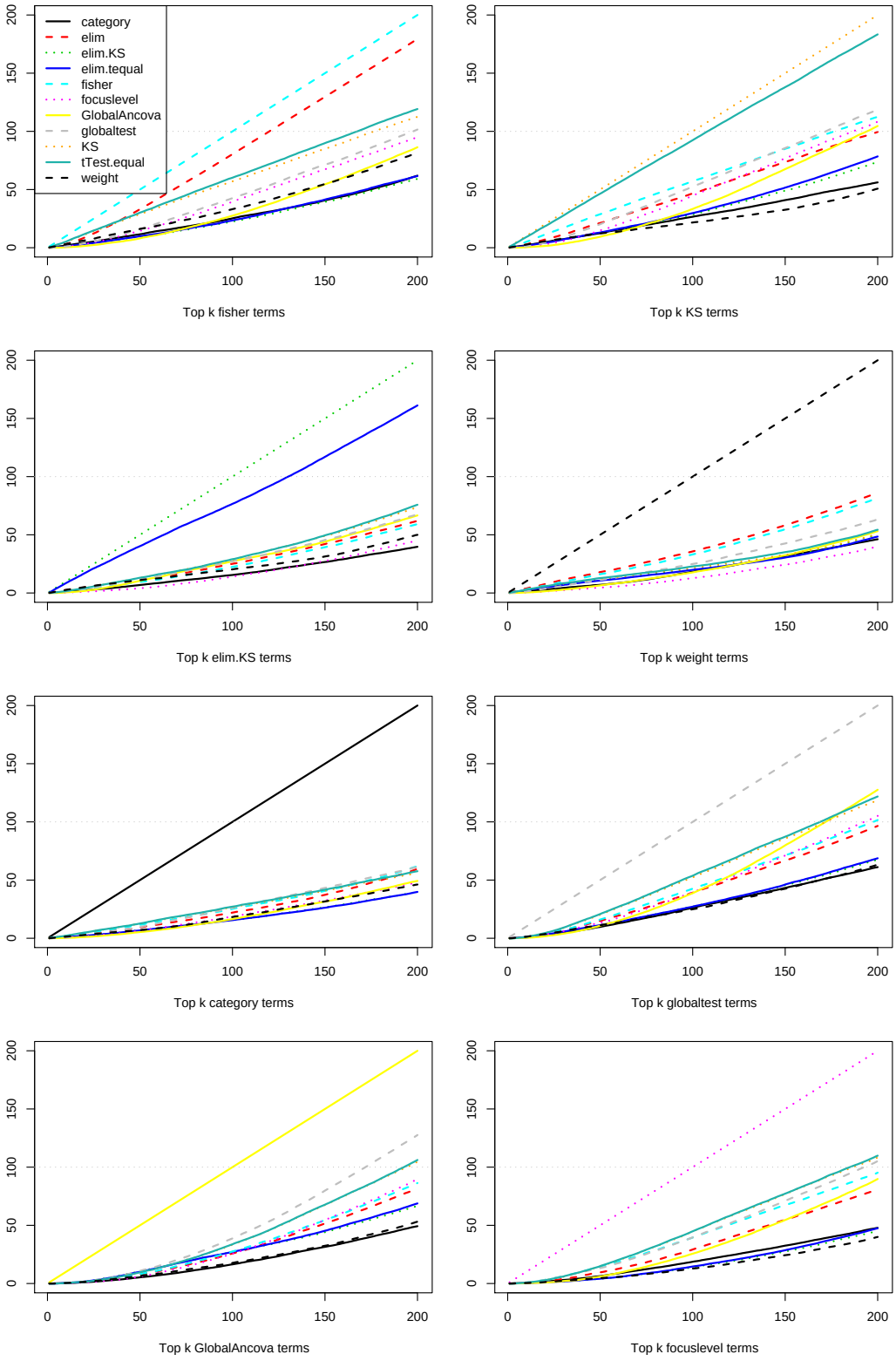


Abbildung 6.9: Anzahlen gemeinsam detektierter GO Gruppen unter den ersten $k = 1, \ldots, 200$ Gruppen zwischen jedem Verfahren und jeweils einem anderen Basisverfahren. Als Basis werden in dieser Reihenfolge verwendet: Fisher, KS, elim + KS, weight, Category, globaltest, GlobalAncova, focus level. Die horizontale Linie bei 100 dient der Orientierung. Ergebnisse für das Szenario 'kleine bis mittelgroße GO Gruppen'.

6.2.3 Probleme des Simulationsansatzes

Das Hauptproblem der beschriebenen Simulationsstudie liegt wohl darin, daß man aufgrund der Verwendung realer Daten nicht hunderprozentig die „Wahrheit“ kennt. Damit fehlt an sich eine sehr wichtige Voraussetzung zum Simulieren. Wir verteilen zwar die Gene so über die Gene Ontology, daß die gewählten GO Gruppen mit den am stärksten differentiell exprimierten Genen angereichert sind. Dennoch kann man nicht sicher sagen, ob sich dadurch „genug“ differentielle Expression in allen gewählten Knoten befindet, um diese tatsächlich interessant zu machen. Anders herum kann es auch sein, daß zusätzlich zu den Genen in den gewählten Gruppen noch weitere mehr oder weniger stark differentiell exprimiert sind. Somit kann es neben den selektierten noch andere interessante Knoten geben. Wir versuchen Anreicherungen von zusätzlichen differentiellen Genen in nicht-selektierten Knoten zu vermeiden, indem wir alle Gene des gesamten Datensatzes neu verteilen. Dennoch muß bei den gezeigten Ergebnissen stets beachtet werden, daß die Anzahlen Richtig und Falsch Positiver hier nicht exakt zu bestimmen sind.

Durch die Verwendung realer Daten als Basis für die Simulation wird versucht, ein möglichst realistisches Szenario zu erstellen. Allerdings werden die Expressionsprofile sämtlicher Gene permutiert, um die differentielle Expression auf bestimmte GO Knoten konzentrieren zu können. Damit gehen wichtige Strukturen der Daten, nämlich die Korrelationen zwischen Genen innerhalb der funktionellen GO Gruppen, verloren. Zum einen werden die simulierten Daten dadurch doch wieder unrealistischer. Zum anderen könnte es sein, daß Gene Set Enrichment Verfahren, deren Signifikanz auf der Basis von Genpermutationen bewertet wird, hier bevorzugt werden, da die natürlichen Korrelationen zwischen Genen durch das Simulationsdesign sowieso bereits „zerstört“ sind. Der womögliche Vorteil globaler Ansätze, die die Korrelationsstruktur berücksichtigen, kommt hier nicht zur Geltung.

Die Begünstigung der Gene Set Enrichment Methoden wird noch dadurch verstärkt, daß die ausgewählten GO Gruppen mit Genen besetzt werden, die möglichst gute genweise Statistiken aufweisen. Genau auf solche Anreicherungen von stark differentiellen Genen zielt natürlich das Gene Set Enrichment ab. Holistische Verfahren dagegen detektieren auch andere interessante Konstellationen, zum Beispiel Gruppen deren Gene alle schwach und womöglich gleich gerichtet differentiell exprimiert sind. Solche Gengruppen werden allerdings in der vorliegenden Simulationsstudie nie die zu detektierenden sein. Deshalb schneidet beispielsweise der Category Ansatz hier sehr schlecht ab.

Aus Zeitgründen wurden bei der Durchführung der Simulation für viele Verfahren p-Werte auf der Basis asymptotischer Verteilungen berechnet (Kolmogorov-Smirnov, t-Test, Category, globaltest, focus level), obwohl in manchen Fällen ein Permutationsansatz geeigneter wäre. Womöglich könnten die Ergebnisse für diese Verfahren mit permutationsbasierten p-Werten besser ausfallen.

Die Simulationsstudie müßte ausgeweitet werden auf andere Datensätze, um generell noch

mehr Erfahrungen für die einzelnen Methoden sammeln zu können und vor allem um nicht nur auf der Basis eines, möglicherweise mit Artefakten belasteten, Datensatzes Schlüsse zu ziehen. Wahrscheinlich wäre es zusätzlich sinnvoll, doch auch Daten künstlich zu simulieren. Damit könnte man besser steuern, welche Gruppen die interessanten sein sollen. Außerdem könnten mit simulierten Daten spezielle Korrelationsstrukturen innerhalb der GO Gruppen untersucht werden. Grundsätzlich bleibt bei einer Simulationsstudie für die Gene Ontology weiterhin die Frage offen, welche Art von Begriffen eigentlich als interessant befunden werden soll – eher spezifische oder eher allgemeine, einzelne über die GO verstreute Begriffe oder zusammen hängende Subgraphen?

Kapitel 7

Anwendungsbeispiele

In den folgenden Abschnitten werden einige reale Datenauswertungen vorgestellt, die die Nützlichkeit der Gengruppenanalyse für die praktische Auswertung und Interpretation von Microarray Experimenten verdeutlichen sollen. Es handelt sich jeweils um komplexe Fragestellungen, die mit GlobalAncova (siehe Kapitel 4) bearbeitet werden können. In den Abschnitten 7.1 und 7.3 wird unter anderem die spezialisierte Analyse von Gene Ontology Gruppen behandelt.

7.1 Differentielle zeitliche Expressionsverläufe in Gene Ontology Gruppen

Die Daten des ersten Beispiels werden in Xiang u. a. (2007) beschrieben und wurden bereits in Abschnitt 3.1 kurz vorgestellt. Es handelt sich um eine experimentelle Studie an Mäusen zur Untersuchung von dynamischen Transkriptionsveränderungen im zentralen Nervensystem während der Krankheitsgeschichte nach Infektion mit Prionen. Hierfür wurden Mäuse mit dem ME7-Prion infiziert. Nach 90, 120 und 150 Tagen wurden jeweils drei Tiere getötet und Probenmaterial gewonnen. Die Gruppe der infizierten Mäuse wurde mit einer Kontrollgruppe verglichen. Um Effekte der Anästhesie, des Vorgangs der Infektion oder von altersbedingten Veränderungen im Gehirn ausschließen zu können, wurden gleichaltrige Kontrollmäuse mit gesundem Hirnhomogenat „infiziert“ (*mock-Infektion*) und zu den gleichen Zeitpunkten nach der Infektion untersucht. Tabelle 7.1 zeigt das Studiendesign. Für jede einzelne Maus wurde ein Affymetrix Microarray des Typs MOE 430a erstellt.

Das Ziel des Experiments war die Identifikation von Genen, deren Expressionsprofil sich zwischen den beiden Studiengruppen über die Zeit hinweg unterschiedlich entwickelt. Zu diesem Zweck haben wir lineare Modelle für die einzelnen Gene betrachtet, wobei das Hauptinteresse in der Interaktion zwischen Studiengruppe und Zeit lag. Nach Adjustierung für multiples Testen ergab sich eine Liste von 449 probesets mit signifikantem Interaktionseffekt. Mit Hilfe des üblichen Gene Set Enrichments (Fisher-Tests) wurde diese Genliste dann auf Anreicherung in den Kategorien der Ontologie 'Biologischer Prozeß'

	Tag 90	Tag 120	Tag 150
Kontrollen	3 Mäuse	3 Mäuse	3 Mäuse
Prion-Infizierte	3 Mäuse	3 Mäuse	3 Mäuse

Tabelle 7.1: Experimentelles Design der Mausdaten mit den beiden Faktoren 'Studiengruppe' und 'Zeit'.

untersucht. Es wurde also für die Datenanalyse die klassische zweistufige Vorgehensweise gewählt, bei der auf die genweise Auswertung die Enrichment Analyse folgt. Nach Holm-Adjustierung der Gene Set Enrichment Ergebnisse bleibt keine der 4,472 getesteten GO Gruppen signifikant angreichert mit interessanten Genen (kleinster adjustierter p-Wert: 0.13). In Tabelle 7.2 werden lediglich die fünf „besten" Kategorien berichtet. (Die Daten wurden re-analysiert mit aktuellen Annotationen. Deshalb bestehen Unterschiede zu den veröffentlichten Ergebnissen.)

ID	Term
GO:0006911	phagocytosis, engulfment
GO:0030889	negative regulation of B cell proliferation
GO:0002526	acute inflammatory response
GO:0051186	cofactor metabolic process
GO:0006954	inflammatory response

Tabelle 7.2: Die fünf am deutlichsten mit signifikanten Genen angereicherten GO Kategorien (Mausdaten).

Alternativ zum Gene Set Enrichment wird in Hummel u. a. (2008a) eine holistische Analyse mit den Daten durchgeführt. Die Auswertung basiert nicht auf der zuvor definierten Liste interessanter Gene. Stattdessen werden mit GlobalAncova direkt alle GO Gruppen hinsichtlich differentieller zeitlicher Expressionsverläufe getestet. Wie bei der genweisen Analyse wird die Interaktion zwischen experimenteller Gruppe und Zeit überprüft. Für GlobalAncova werden hierfür die beiden Modelle

$$\text{Volles Modell:} \quad E(x) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot time + \beta_{int}(group \cdot time)$$
$$\text{Reduziertes Modell:} \quad E(x) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot time$$

verglichen. Der GO Term 'quinone cofactor metabolic process' (GO:0042375) hat den kleinsten (approximativen) GlobalAncova p-Wert ($p = 0.00003$), das heißt diese Gruppe enthält Gene, deren zeitliche Expressionsverläufe sich zwischen infizierten und Kontrollmäusen unterscheiden. Abbildung 7.1 links zeigt die je Zeitpunkt gemittelten Expressionswerte der Gene aus GO:0042375, getrennt für die beiden experimentellen Gruppen. Bei einigen Genen ist ein differentieller Verlauf zu erkennen. So bleibt beispielsweise die Expression des

schwarz markierten Gens in der Kontrollgruppe über die Zeit relativ konstant, während sie bei Infektion zwischen dem zweiten und dritten Meßzeitpunkt deutlich ansteigt. Im entsprechenden gene plot auf der rechten Seite wird ersichtlich, daß dieses Gen den größten Einfluß auf die GlobalAncova Statistik hat.

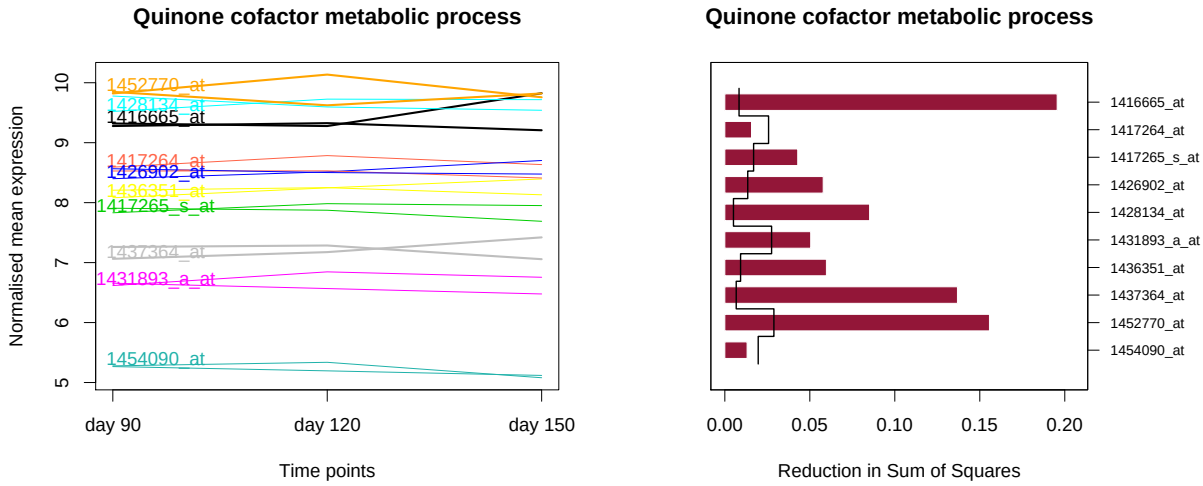
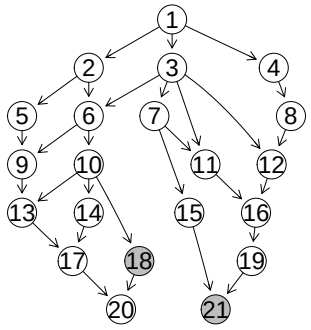


Abbildung 7.1: Links: Mittlere Genexpressionswerte der Gene in der GO Kategorie 'quinone cofactor metabolic process' zu den drei Beobachtungszeitpunkten im Mausdatensatz. Die jeweils zwei Linien einer Farbe entsprechen der Expression eines Gens in der Prion-infizierten und in der Kontrollgruppe. Rechts: Entsprechender gene plot der GO Kategorie 'quinone cofactor metabolic process'.

Da mehrere tausend GO Gruppen getestet werden, muß natürlich auch bei der holistischen Analyse eine Adjustierung für multiples Testen durchgeführt werden. Nach einer Holm-Korrektur ist wie bei der Gene Set Enrichment Auswertung kein GO Term signifikant (kleinster adjustierter p-Wert: 0.15). Im Vergleich zur Holm-Adjustierung bietet die focus level Methode (Goeman und Mansmann, 2008) eine mächtigere Alternative, bei der außerdem die hierarchische Struktur der GO berücksichtigt wird. Zu einem Niveau von $\alpha = 0.1$ erhält man den in Abbildung 7.2 gezeigten signifikanten Subgraphen. Der zuvor auffälligste GO Term 'quinone cofactor metabolic process' wird auch durch die focus level Prozedur detektiert. Die Überschneidung mit den fünf signifikantesten Begriffen der GSE Analyse aus Tabelle 7.2 ist lediglich der Term 'cofactor metabolic process' (GO:0051186).

7.2 Korrelation einer prognostischen Gensignatur für Brustkrebs mit dem Zellzyklus pathway

Im folgenden Beispiel wird der Datensatz von van't Veer u. a. (2002) betrachtet. Die Autoren haben anhand dieser Daten eine Signatur von 68 Genen erstellt, die eine Prognose für



	ID	Term		ID	Term
1	8150	biological process	12	50794	regulation of cellular proc.
2	8152	metabolic process	13	51188	cofactor biosynthetic proc.
3	9987	cellular process	14	6732	coenzyme metabolic process
4	65007	biological regulation	15	278	mitotic cell cycle
5	9058	biosynthetic process	16	9108	coenzyme biosynthetic proc.
6	44237	cellular metabolic proc.	17	9108	coenzyme biosynthetic proc.
7	7049	cell cycle	18	42375	quinone cofactor metabol. proc.
8	50789	regulation of biol. proc.	19	74	regul. of progr. thr. cell cycle
9	44249	cellular biosynth. proc.	20	45426	quinone cofactor biosynth. proc.
10	51186	cofactor metabolic proc.	21	7346	regul. of progr. thr. mitotic c.c.
11	22402	cell cycle process			

Abbildung 7.2: Ergebnis der focus level Methode mit $\alpha = 0.1$ für die 'Biologischer Prozeß' Ontologie im Mausdatensatz. Die gezeigten GO Gruppen weisen eine signifikante Interaktion zwischen Zeit nach Infektion und experimenteller Gruppe auf. Die GO Gruppen innerhalb des focus levels sind grau markiert.

das Überleben von Brustkrebspatienten ermöglichen soll. Um ein homogenes Patientenkollektiv zu erhalten, betrachten wir nur die 96 Patientinnen, die nicht die spezielle Mutation 'Brca1' tragen. Bei 45 von diesen Patientinnen haben sich innerhalb von fünf Jahren nach Operation Metastasen gebildet. Sie bilden die Gruppe mit schlechter Prognose. Die übrigen 51 Patientinnen blieben mindestens fünf Jahre frei von Metastasen und werden somit in die Gruppe mit guter Prognose eingeordnet.

Das Erstellen und Validieren einer prognostischen Signatur ist eines der großen Ziele und eine schwierige Aufgabe innerhalb der Expressionsanalyse. Etwas genauer wird dieser Bereich am Beispiel aus Abschnitt 7.3 erläutert. Ebenso ist es von Bedeutung, die biologischen Zusammenhänge und Wirkungsweisen der gefundenen Signaturgene beziehungsweise der aus der Signatur abgeleiteten Klassifikationsregel zu verstehen. Die Exaktheit der Prognose auf der Basis einer Signatur reicht noch lange nicht aus, damit sie tatsächlich Eingang findet in die medizinische Entscheidungsfindung. Erst wenn die Verallgemeinerbarkeit und klinische Relevanz eines prognostischen Faktors ausreichend belegt ist, wird er möglicherweise in der klinischen Praxis zu Rate gezogen werden (Wyatt und Altman, 1995). Es ist demnach unabdingbar, die molekulare Struktur einer Signatur näher zu ergründen, um letztlich ihre Auswirkungen auf das phänotypische Erscheinungsbild verstehen zu können. Dieses Ziel wird hier verfolgt, indem Zusammenhänge der Signatur mit funktionellen Gengruppen untersucht werden. Yu u. a. (2007) verfolgen eine ähnliche Strategie. Sie testen mit Hilfe von Gene Set Enrichment Methoden, welche pathways mit Signaturgenen angereichert sind. Die Assoziation der detektierten pathways mit dem jeweiligen klinischen outcome wird daraufhin durch den globaltest (Goeman u. a., 2004) überprüft. Allerdings sind für die mit Signaturgenen angereicherten Gruppen signifikante Testergebnisse nicht besonders erstaunlich. Die Signaturgene wurden schließlich hinsichtlich ihrer Korrelation mit der klinischen Variablen ausgesucht.

Die funktionellen Gengruppen, die nach der Strategie von Yu u. a. (2007) gefunden werden, sind solche die die Signatur am besten *repräsentieren*. Wir dagegen sind auf der Suche nach Gengruppen, die mit der Signatur *interagieren*. Es wird eine mögliche *Co-Expression* zwischen der Signatur und krebsrelevanten pathways analysiert. Für eine solche Fragestellung bietet sich GlobalAncova an, wobei die Expressionswerte der Signaturgene als lineare Regressoren in das Modell eingehen. Beispielhaft wählen wir den durch Literaturrecherche zusammen gestellten pathway 'cell cycle control' aus. Wir stellen nun die Frage, ob beziehungsweise welche Gene aus der Signatur einen signifikanten Einfluß auf das Expressionsprofil des pathways haben. Zusätzlich ist es sinnvoll, die Analyse hinsichtlich der beiden prognostischen Patientengruppen zu adjustieren. Dadurch wird gewährleistet, daß auch Co-Expression zwischen Signatur und pathway innerhalb der prognostischen Gruppen detektiert wird. Zunächst wählen wir das Signaturgen 'cyclin E2' aus, das auch in der Publikation von van't Veer u. a. (2002) benannt wird, und überprüfen dessen Einfluß auf den Zellzyklus pathway. Wir vergleichen also die Modelle

$$\begin{aligned} &\text{Volles Modell:} && E(x) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot cyclinE2 \\ &\text{Reduziertes Modell:} && E(x) = \beta_0 + \beta_1 \cdot group. \end{aligned}$$

Es kann ein hoch signifikanter Effekt β_2 des Signaturgens festgestellt werden ($p < 0.0001$). Abbildung 7.3 soll verdeutlichen, wie Co-Expression zwischen einem Gen und einem ganzen pathway zu verstehen ist. Es werden einzelne Effekte des Signaturgens auf jeweils eines der pathway Gene durch Regressionsgeraden dargestellt. Dabei sind signifikante positive Effekte blau gekennzeichnet. Eine hohe Expression von cyclin E2 bewirkt auch eine hohe Expression dieser Zellzyklus Gene. Ein signifikanter negativer Zusammenhang, das heißt Herunterregulation des entsprechenden Gens bei starker cyclin E2 Expression, ist rot markiert. „Signifikant“ bezieht sich hier auf genweise Tests, die nur für die Darstellung berechnet wurden. Da das Signaturgen auf mehrere Gene des pathways' einen starken, in den meisten Fällen positiven, Einfluß hat, erklärt sich das hochsignifikante Ergebnis des GlobalAncova Tests. Die Effekte in Abbildung 7.3 wurden außerdem getrennt für die beiden prognostischen Patientengruppen berechnet, um mögliche Unterschiede in der Wirkungsweise von cyclin E2 zwischen Patienten mit guter und schlechter Prognose identifizieren zu können. Es zeigt sich allerdings in beiden Gruppen ein sehr ähnliches Bild. Dies spiegelt sich auch in einem nicht-signifikanten Ergebnis des Tests auf *differentielle Co-Expression* wider, bei dem die Interaktion zwischen cyclin E2 und klinischer Gruppe getestet wird ($p = 0.2$).

Man ist nicht darauf beschränkt, den Einfluß eines einzelnen Signaturgens auf einen pathway zu überpüfen. Ebenso kann der Zusammenhang zwischen der gesamten Signatur und dem pathway getestet werden, indem alle Signaturgene als Kovariablen in das volle Modell eingehen. In dem vorliegenden Beispiel erhält man für diesen Test ein hoch signifikantes Ergebnis ($p < 0.0001$). Problematisch ist hierbei allerdings, daß ein Modell mit 68 Variablen für einen Datensatz mit nur 96 Beobachtungen aufgestellt wird. Zudem detektiert

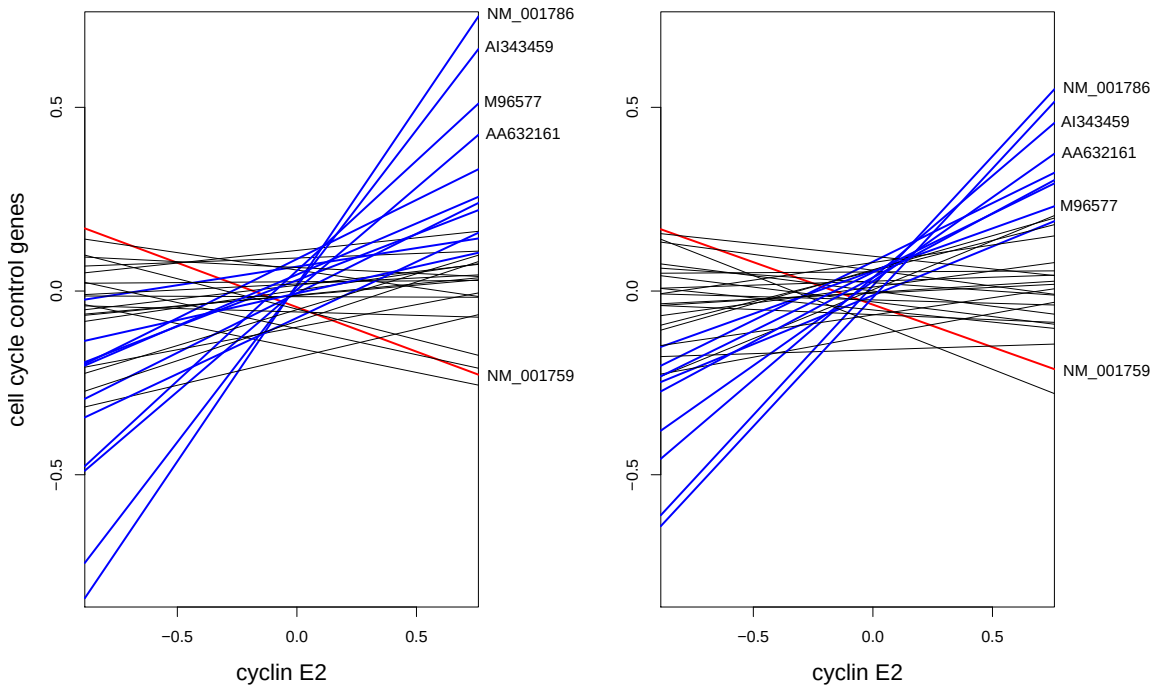


Abbildung 7.3: Lineare Effekte des Gens 'cyclin E2' aus der van't Veer Signatur auf die einzelnen Gene des Zellzyklus pathways'; getrennt berechnet für die gute (links) und schlechte (rechts) Prognosegruppe. Starke positive (negative) Korrelationen sind blau (rot) gekennzeichnet. Die Effekte unterscheiden sich nicht wesentlich zwischen den beiden Prognosegruppen, was auch die einzelnen angegebenen Gennamen verdeutlichen sollen.

GlobalAncova jegliche Art von Korrelationen zwischen Signatur- und pathway-Genen. Ein signifikantes Ergebnis für den obigen Test ist somit nicht sehr außergewöhnlich. Außerdem wäre es ein interessanteres Ergebnis zu wissen, welche der Signaturgene oder welche Cluster von ähnlichen Signaturgenen einen signifikanten Einfluß auf den pathway haben. Man könnte nun für jedes Signaturgen einen separaten Test rechnen und die resultierenden p-Werte für multiples Testen korrigieren. Diese Vorgehensweise ist allerdings bekanntermaßen nicht sehr effizient. Stattdessen verwenden wir die Idee des hierarchischen Testens (Meinshausen, 2008) für eine Variablenselektion. Dabei werden hierarchisch geschachtelte Cluster von Signaturgenen mit ähnlichen Expressionsmustern hinsichtlich ihres Einflusses auf den pathway getestet. Die Korrelationsstrukturen innerhalb der Signatur werden somit berücksichtigt. Die Prozedur ist so konzipiert, daß die FWER kontrolliert wird. Die letztendlich selektierten Signaturgene haben also einen korrigiert signifikanten Einfluß auf den pathway.

Für die hierarchische Testprozedur müssen die Variablen, in diesem Falle die $s = 68$ Signaturgene, zunächst in hierarchisch gegliederte Cluster eingeteilt werden. Wir verwenden

hierfür ein übliches hierarchisches Clustern mit korrelationsbasiertem Abstandsmaß, bei dem sukzessiv die jeweils ähnlichsten Cluster zusammen gefaßt werden. Dadurch entsteht ein binärer Baum, dessen Wurzel alle Variablen enthält und dessen Blätter die einzelnen Variablen sind. Abbildung 7.4 zeigt den hierarchischen Baum der van't Veer Signatur. Wie in Abschnitt 5.2.2 beschrieben werden die Cluster ausgehend von der Wurzel nacheinander prozessiert. Für einen Cluster C wird mit GlobalAncova der Einfluß der entsprechenden Signaturgene aus C auf den Zellzyklus pathway getestet. Der resultierende p-Wert p_C wird adjustiert

$$p_{C,adj} = p_C \cdot \frac{s}{|C|}.$$

Ein Cluster gilt dann als signifikant, wenn

(a) sein adjustierter p-Wert $p_{C,adj} \leq \alpha$

(b) seine Vorfahren-Cluster abgelehnt wurden, das heißt für alle Cluster $D \supset C$ gilt $p_{D,adj} \leq \alpha$.

Das Verfahren wird fortgesetzt so lange es signifikante Cluster gibt, bis man schließlich auf das Niveau der Einzelvariablen gelangt. In Abbildung 7.4 ist das Ergebnis der hierarchischen Variablenselektion für die van't Veer Signatur und den Zellzyklus pathway dargestellt, indem die resultierenden Signaturgene mit signifikantem Einfluß rot gekennzeichnet sind. Neben der Selektion der interessanten Größen gewinnt man durch dieses Verfahren auch einen Einblick in die Zusammenhänge innerhalb der Signatur. In der Abbildung sind Sub-Cluster erkennbar, die keinen Einfluß auf den pathway ausüben. Ebenso gibt es Sub-Cluster von Genen, die in ihrer Expression recht ähnlich sind und deshalb auch einen ähnlichen Effekt auf den pathway haben.

7.3 Analyse einer prognostischen Signatur für AML

Im Folgenden wird eine Gensignatur für die Prognose bei Akuter Myeloischer Leukämie (AML) analysiert, die in Zusammenarbeit mit dem Institut für Innere Medizin III des Klinikums Großhadern erstellt wurde (Metzeler u. a., 2008). Wie im vorigen Abschnitt liegt das Hauptinteresse in der Untersuchung der biologischen Bedeutung der Signatur. Zu diesem Zweck wird ihr Zusammenhang mit bekannten pathways und Gene Ontology Kategorien überprüft. Wir verwenden wieder die hierarchische Variablenselektion in Verbindung mit GlobalAncova, um Korrelationen zwischen einzelnen Signaturgenen und einer großen Anzahl verschiedener KEGG pathways zu entdecken. Die Zusammenhänge zwischen der Signatur und ausgewählten pathways wird über die Schätzung von Geninteraktionsnetzwerken (Schäfer und Strimmer, 2005a,b) noch genauer betrachtet. Nicht nur die Signaturgene selbst, sondern auch der prognostische score, der basierend auf ihrer Expression für die Patienten erstellt wird, kann dabei helfen, die Wirkungsweise der Signatur zu verstehen. Der score stellt eine „Zusammenfassung“ des Expressionsprofils der kompletten Signatur über einen Klassifikationsalgorithmus dar. Alternativ zur obigen Strategie mit den einzelnen Signaturgenen kann man somit auch den Zusammenhang zwischen dem score und

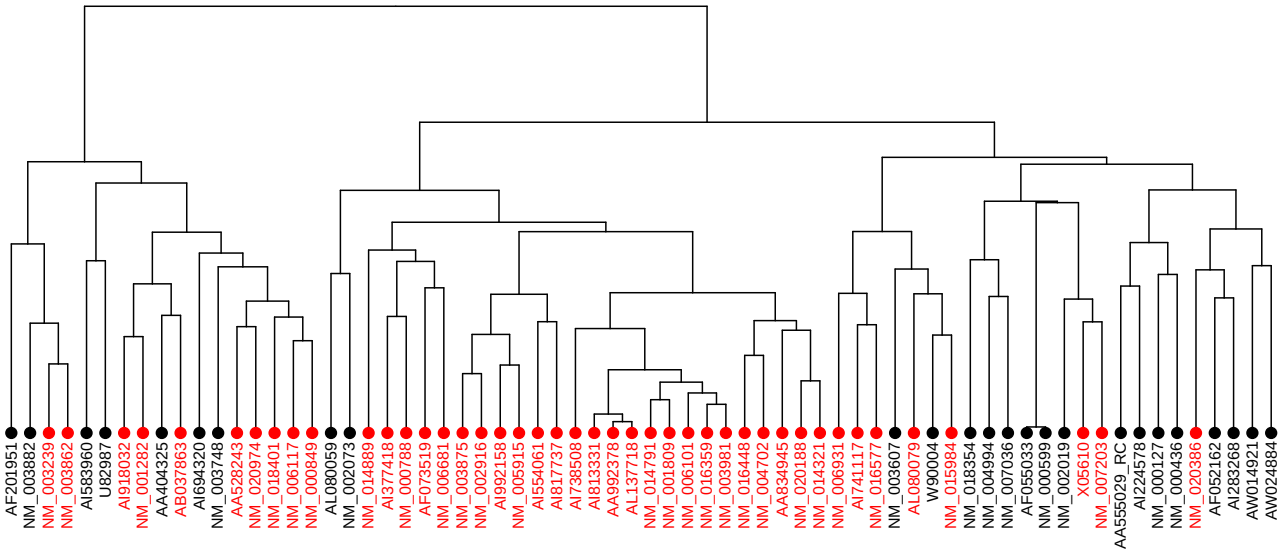


Abbildung 7.4: Hierarchisches Clustern der 68 Gene der van't Veer Signatur. Gene mit laut hierarchischer Variablenselektion signifikantem Einfluß auf den Zellzyklus pathway sind rot gekennzeichnet.

funktionellen Gengruppen untersuchen. Hierfür werden KEGG pathways und durch die Gene Ontology definierte biologische Prozesse (GOBP) sowohl mit GlobalAncova als auch mit globaltest getestet. Für die GO Analyse wird dabei die focus level Methode (Goeman und Mansmann, 2008) angewendet. Von besonderem Interesse ist außerdem, wie der score mit einem weiteren wichtigen prognostischen Faktoren in Zusammenhang steht. Dies wird über das Überprüfen der Interaktion zwischen score und Faktor zu erklären versucht. Ein solcher Interaktionseffekt kann nur mit GlobalAncova und nicht mit globaltest getestet werden. Die Ergebnisse zu diesem Kapitel sind in Hummel u. a. (2008b) beschrieben.

Erstellen der Genexpressionssignatur

Die erstellte Signatur bezieht sich auf AML Patienten mit normalem Karyotypen. Chromosomale Aberrationen erlauben die Einteilung von AML Patienten in verschiedene Prognosegruppen. Aber auch das hier vorliegende Patientengut mit normalem Karyotypen, welches fast die Hälfte aller AML Fälle ausmacht, ist in sich sehr heterogen. Üblicherweise wird diese Gruppe hinsichtlich des Vorliegens von Mutationen in spezifischen Genen weiter in prognostische Subgruppen unterteilt. Zwei wichtige prognostische Faktoren sind 'fms-like tyrosine kinase 3' (FLT3) und 'nucleophosmin 1' (NPM1). Eine Mutation in FLT3 führt zu einer schlechten Prognose, während eine Mutation in NPM1 mit besserer Progno-

se assoziiert ist (Marcucci u. a., 2005). Mittels Genexpressionsanalyse wird nun versucht, alternativ oder ergänzend zu den oben genannten Faktoren noch genauere prognostische Gruppen zu definieren, um letztlich speziell angepaßte Therapieformen entwickeln und anwenden zu können.

Für die Erstellung des Genexpressions-scores an einem Datensatz mit 163 AML Patienten (Trainingsdaten) wurde die Methode *supervised principal components* (Bair und Tibshirani, 2004; Bair u. a., 2006; Bair und Tibshirani, 2007) verwendet. Dabei werden zunächst einzelne Gene selektiert, deren Expressionsprofile den größten Zusammenhang mit der Überlebenszeit der Patienten aufweisen. Die 67 ausgewählten Gene (86 Affymetrix probesets) stellen die Signatur dar. Anschließend wird auf der Basis der gewählten Gene über Hauptkomponentenanalyse ein stetiger score erstellt, der mit der Überlebenszeit korreliert. Je höher der Wert des scores für einen Patienten ist, desto schlechter ist seine Prognose. Aus der Hauptkomponentenanalyse lassen sich Gewichte für die Signaturgene ermitteln, mit denen man für neue Patienten, nach Standardisierung der entsprechenden Expressionsdaten, den Prognose-score über eine einfache Linearkombination berechnen kann. Dies wurde von Metzeler u. a. (2008) zur Validierung der Signatur an einem internen Testdatensatz mit 79 Patienten und einem externen Validierdatensatz mit 64 Patienten aus der *Cancer and Leukemia Group B* (CALGB) 9621 Studie (Radmacher u. a., 2006) durchgeführt. Der erstellte Genexpressions-score erweist sich in beiden Datensätzen auch unter Berücksichtigung der anderen prognostischen Faktoren Alter, FLT3 und NPM1 Mutationsstatus als relevante Einflußgröße für die Überlebenszeit. Zur Untersuchung der molekularbiologischen Eigenschaften der Signatur werden hier, wie in Hummel u. a. (2008b) beschrieben, nur die 163 Patienten aus dem Trainingsdatensatz betrachtet. Dies ist nicht problematisch, da es in diesem Fall nicht um die Einschätzung der prognostischen Güte der Signatur geht.

Alle Analysen werden auf Gen- und nicht auf probeset Ebene durchgeführt, da die probesets zu ein und demselben Gen oft stark korrelieren. Dadurch würden bei einer Auswertung auf probeset Basis häufig in erster Linie diese Korrelationen entdeckt statt der eigentlich interessierenden Zusammenhänge zwischen Genen der Signatur und der pathways. Wenn mehrere probesets einem Gen entsprechen, wird jeweils beliebig eines der probesets als Repräsentant des Gens ausgewählt. Gerade weil die probesets eines Gens meist stark korrelieren erscheint dieses Vorgehen unproblematisch. Außerdem ist man sich derzeit nicht über eine optimale Strategie zur Auflösung der probeset Redundanz einig. Probesets, die keinem bekannten Gen entsprechen, werden in der Analyse beibehalten. Diese sind in den folgenden Grafiken mit ihren probeset Namen bezeichnet, während die übrigen Gene mit den offiziellen Gensymbolen benannt sind.

7.3.1 Assoziation zwischen Signatur und pathways

Hierarchische Variablenselektion für die Assoziation zwischen Signatur und pathways

Wie in Abschnitt 7.2 wird die Gensignatur mit bekannten pathways in Verbindung gebracht. Eine Co-Expression zwischen Signatur- und pathway Genen kann Aufschluß über die Wirkungsweise der Signatur liefern. Im vorliegenden Beispiel ist von besonderem Interesse, wie die Signatur mit dem KEGG pathway 'Acute myeloid leukemia' (KEGG ID hsa05221) interagiert. Das AML Onkogen RUNX1 gehört sowohl zum AML pathway als auch zur Signatur. Wie im vorigen Kapitel beschrieben kann man mit GlobalAncova den Einfluß eines einzelnen Signaturgens oder auch der ganzen Signatur auf den gewählten pathway testen. Der Zusammenhang zwischen einem Gen und einem kompletten pathway kann wie in Abbildung 7.3 dargestellt werden. Der Test für die Verbindung der ganzen Signatur mit dem pathway liefert ein signifikantes Ergebnis ($p < 0.0001$). Um den Sachverhalt zu visualisieren, wird eine heatmap verwendet, deren Farbgebung die genweisen Korrelationen kodiert. In Abbildung 7.5 entsprechen die Spalten der heatmap den 67 Signaturgenen und die Zeilen den 53 Genen des AML pathways. Die Grafik zeigt, daß die meisten Signaturgene positiv (rot markiert) mit einigen wichtigen Onkogenen der AML, unter anderen FLT3 und c-KIT korrelieren. Abbildung 7.6 zeigt die KEGG Darstellung des AML pathways (http://www.genome.jp/dbget-bin/www_bget?pathway+hsa05221). Die entsprechenden Gene in dieser Grafik, die mit mindestens einem der Signaturgene signifikant korrelieren (nach einfacher Bonferroni-Korrektur), sind mit roten Rahmen markiert. Man sieht, daß die Signatur mit großen Teilen des AML pathways interagiert. Neben dieser explorativen Analyse wird auch wie in Abschnitt 7.2 eine hierarchische Variablenselektion (Meinshausen, 2008) durchgeführt, um diejenigen Signaturgene heraus zu filtern, die auf das globale Expressionsprofil des AML pathways einen signifikanten Effekt haben. Diese sind 55 der 67 Signaturgene. Die entsprechenden Gennamen sind in Abbildung 7.5 rot markiert. Zusammenfassend kann man also feststellen, daß zum einen viele Komponenten des AML pathways von der Signatur beeinflußt werden und zum anderen tatsächlich der Großteil der Signaturgene einen Einfluß auf den pathway hat. Dies kann eine Erklärung auf molekularer Ebene sein für die Relevanz der Signatur für den Krankheitsverlauf und somit die Prognose bei Akuter Myeloischer Leukämie.

Die hierarchische Variablenselektion wird im Folgenden auf 200 pathways angewendet. Dies sind alle derzeit beschriebenen KEGG pathways für das menschliche Genom, die mindestens ein Gen aus dem vorliegenden Datensatz enthalten. Das Ergebnis ist in Form von heatmaps dargestellt (Abbildungen 7.7 und 7.8), bei denen die Spalten den Signaturgenen und die Zeilen den einzelnen pathways entsprechen. Krebsspezifische pathways sind in Abbildung 7.7 dargestellt, die übrigen 185 pathways finden sich in Abbildung 7.8. Hat ein Signaturgen einen signifikanten Einfluß auf einen pathway, so wird die entsprechende Stelle in der heatmap dunkel markiert, ansonsten hellgrau. Ein signifikanter Einfluß liegt in diesem Fall nur vor, wenn der rohe p-Wert dem bei 10,000 Permutationen kleinst mögli-

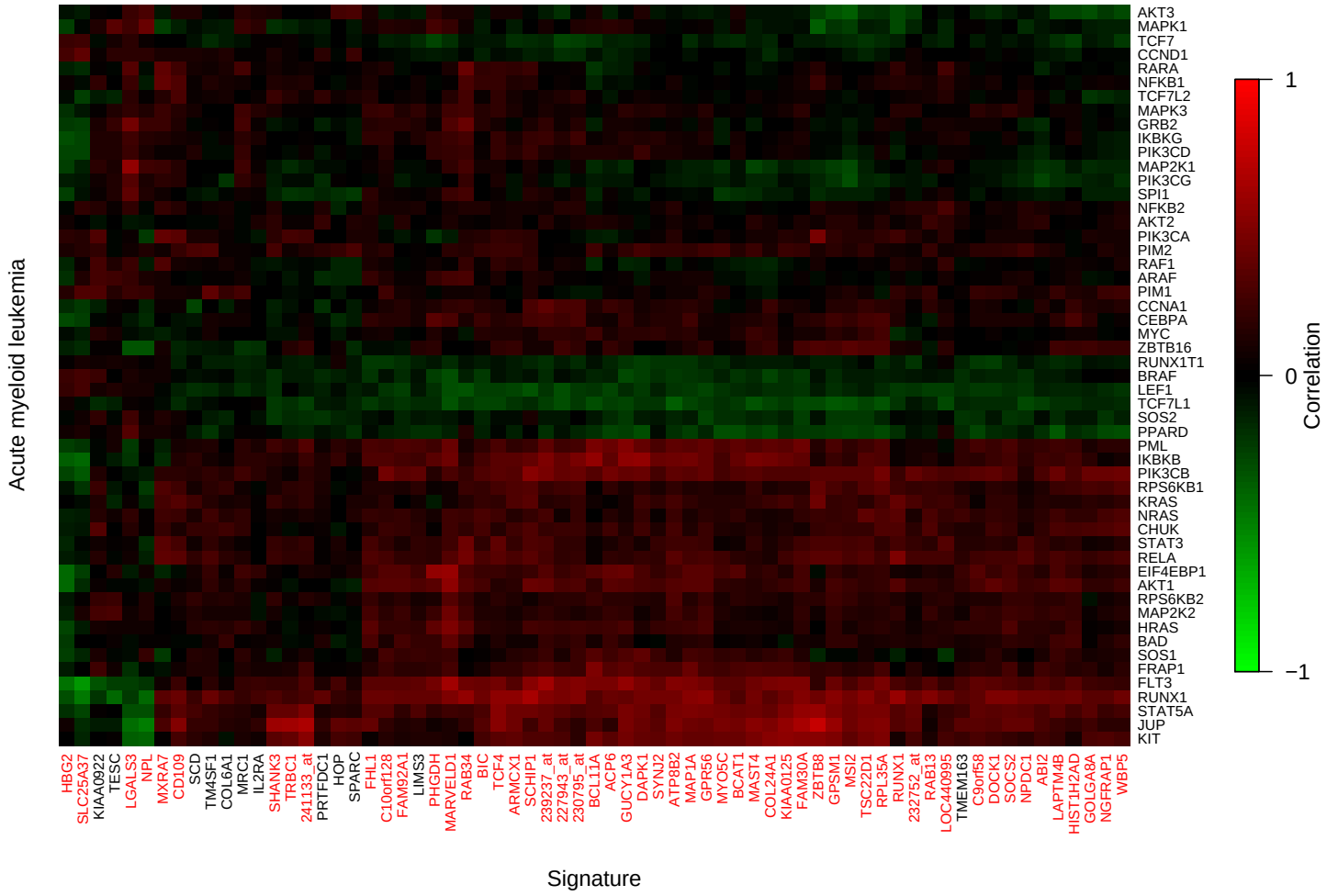


Abbildung 7.5: Genweise Korrelationen zwischen der prognostischen AML Signatur und dem KEGG pathway 'Acute myeloid leukemia'. Die Beschriftung der Signaturgene ist rot, falls sich das entsprechende Gen in der hierarchischen Variablenselektion als signifikant erweist; ansonsten schwarz.

chen entspricht. Die Korrektur gemäß der hierarchischen Testprozedur entspricht auf der Ebene der Einzelvariablen einer Bonferroni-Korrektur. Zusätzlich müßte an sich noch für das Testen der 200 pathways korrigiert werden. Die Tests mit kleinst möglichem p-Wert müßten nach der zusätzlichen Korrektur demnach entweder alle als signifikant oder alle als nicht signifikant bewertet werden. Sinnvollerweise müßte die gesamte Prozedur mit noch deutlich mehr Permutationen durchgeführt werden, um eine feinere Auflösung zu erhalten. Aufgrund der sehr langen Rechenzeit wurde dies bisher allerdings nicht verwirklicht. Wir verzichten deshalb auf die zusätzliche Korrektur und bedenken im Folgenden, daß mit einem gewissen Anteil falsch positiver Ergebnisse zu rechnen ist. Es zeigt sich, daß große Teile der Signatur mit vielen pathways interagieren. Jedes Signaturgen ist mit mindestens

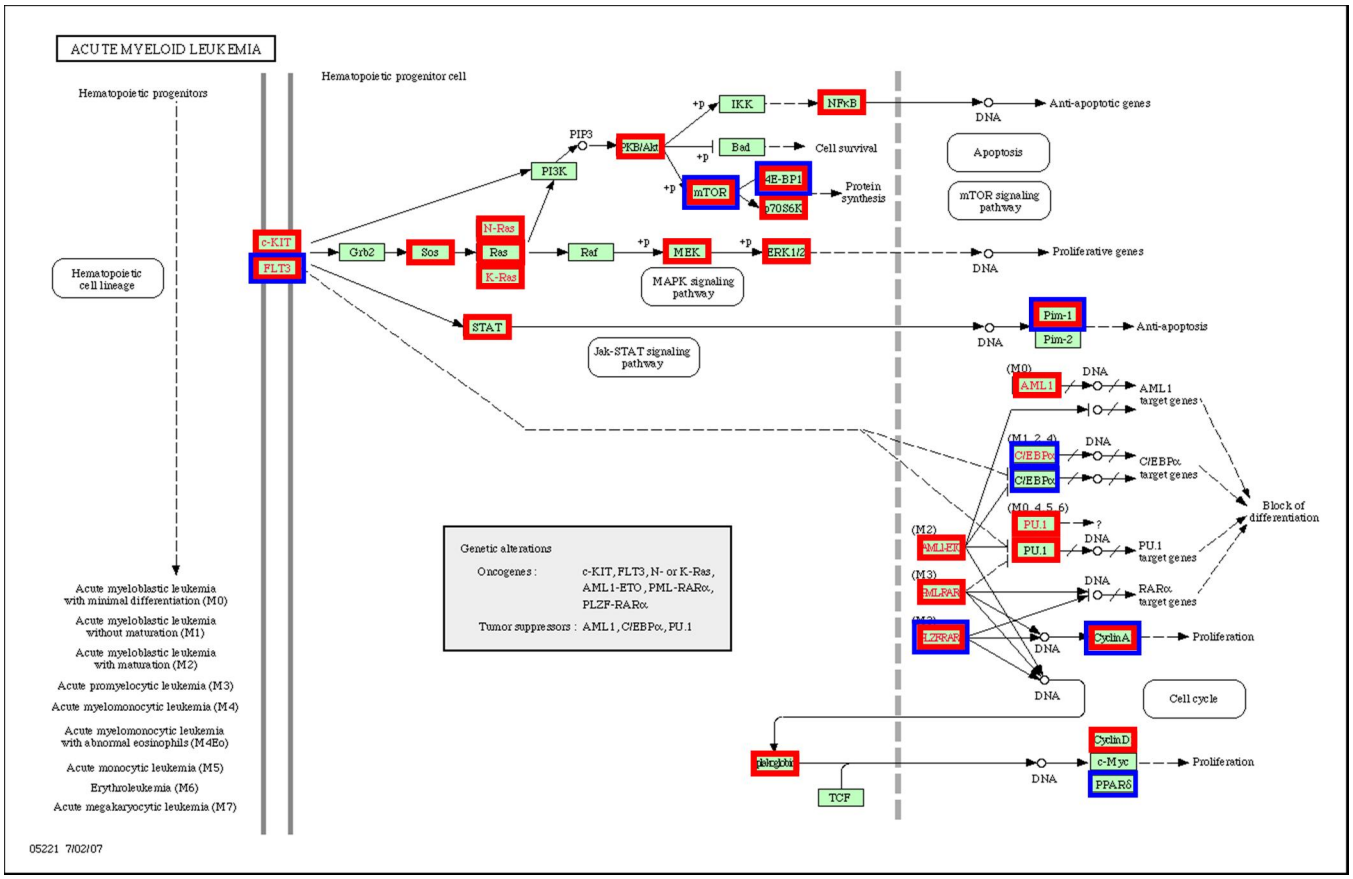


Abbildung 7.6: KEGG pathway 'Acute myeloid leukemia'. Komponenten, die mit mindestens einem Signaturgen signifikant korrelieren, sind mit roten Rahmen markiert. Komponenten mit signifikanter partieller Korrelation (im Gen-Assoziationsnetzwerk) sind blau umrahmt.

einem der pathways korreliert. Im Mittel beeinflußt ein Gen 122 der 200 pathways. Es gibt nur einen pathway, der mit keinem der Signaturgene assoziiert ist. Im Mittel korreliert ein pathway mit 41 der 67 Signaturgene. Besonders die krebsspezifischen pathways stehen mit großen Teilen der Signatur in Verbindung. Die Gene der Signatur, die für die Prognose bei AML erstellt wurde, scheinen in der Pathogenese vieler Arten von Krebs involviert zu sein. Dies unterstreicht die Ähnlichkeiten verschiedener Krebserkrankungen auf molekularer Ebene.

Genregulatorische Netzwerke für die Assoziation zwischen Signatur und pathways

Alternativ zum Ansatz im vorigen Abschnitt, bei dem jegliche Korrelation zwischen der Signatur und dem pathway gemessen wird, kann man sich auch für die spezifischen Gen-Gen-Interaktionen interessieren. Eine Möglichkeit zur Untersuchung, welches Signaturgen

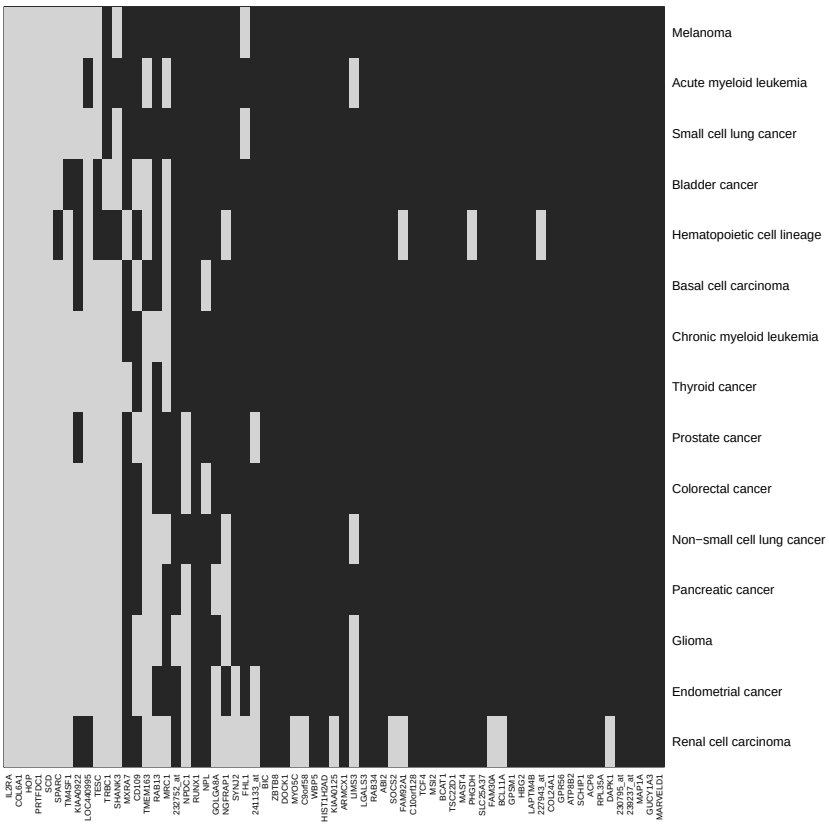


Abbildung 7.7: Ergebnis der hierarchischen Variablenselektion für 15 krebsspezifische pathways. Zeilen: pathways, Spalten: Signaturgene. Liegt ein signifikanter Einfluß eines Signaturgens auf einen pathway vor, ist ein dunkles Rechteck gezeichnet, andernfalls ein helles.

welches pathway Gen direkt beeinflußt, bieten Gennetzwerke. Verwendet man hohe Korrelation als Kriterium für eine Kante im Graphen zwischen zwei Genen, werden sowohl direkte als auch indirekte Assoziationen erfaßt. Folglich wird man in den meisten Situationen eine große Anzahl an Kanten erwarten. Um dagegen nur direkte Abhängigkeiten zu detektieren und somit womöglich einen detaillierteren Einblick zu gewinnen, welches Signaturgen tatsächlich mit welchem Gen des pathways interagiert, schlagen wir das Schätzen von Gen-Assoziationsnetzwerken wie in Schäfer und Strimmer (2005a) vor. Dieser Ansatz basiert auf *partiellen Korrelationen*, also der Korrelation zwischen je zwei Genen, wobei der Einfluß sämtlicher weiterer Gene berücksichtigt wird. Zwei Variablen sind genau dann partiell unabhängig, wenn sie bedingt auf alle übrigen Variablen unabhängig sind. Die Matrix aller partiellen Korrelationen wird mit Hilfe eines shrinkage Schätzers ermittelt, der auch in der $n << p$ Situation adäquat ist (Schäfer und Strimmer, 2005b). Um zu entscheiden, ob eine Kante im Graphen vorhanden sein soll oder nicht, wird die Situation als multiples Testproblem betrachtet. Für jede mögliche Kante wird die *lokale false discovery rate (fdr)* geschätzt. Eine Kante wird hier wie von den Autoren vorgeschlagen als „signifikant“ ange-

sehen, wenn ihre fdr kleiner ist als 0.2. Die Netzwerke werden mit dem *R* Paket `GeneNet` geschätzt (Opgen-Rhein u. a., 2006).

Wir erstellen ein Netzwerk für die Menge aller Signaturgene zusammen mit allen Genen das KEGG pathways 'Acute myeloid leukemia'. Verglichen mit der potenziellen Größe des Netzwerks wird eine relativ kleine Menge an Kanten detektiert. Abbildung 7.9 zeigt nur die Gene, die mit mindestens einem anderen Gen verbunden sind. Kanten zwischen Signatur- und pathway Genen, die für unsere Fragestellung die interessantesten Verbindungen darstellen, kommen in zwölf Fällen vor. Die Ergebnisse der hierarchischen Variablenselektion aus vorigem Abschnitt sind nicht direkt mit dem geschätzten Gen-Assoziationsnetzwerk vergleichbar. Zum einen finden wir mit dem graphischen Modell deutlich weniger Assoziationen zwischen der Signatur und dem AML pathway. Zweitens gibt es Kanten im geschätzten Netzwerk, während die entsprechenden Verbindungen im hierarchischen Ansatz nicht als signifikant detektiert werden. Der Grund für diese Unterschiede sind die unterschiedlichen Konzepte der üblichen und der partiellen Korrelation. Zwei Gene A und B können eine signifikante partielle Korrelation haben, während die (übliche) Korrelation von A oder B zu irgend einem anderen Gen stärker sein kann als die Korrelation zwischen A und B selbst. Die Korrelationen zu den „anderen" Genen könnten ihre Ursache in indirekten Einflüssen haben. Dies mag erklären, warum man in manchen Fällen eine signifikante partielle Korrelation zwischen zwei Genen beobachtet aber keine signifikante Korrelation, und anders herum. Zudem ist zu bedenken, daß die Rekonstruktion von Netzwerken extrem schwierig ist. Beispielsweise entsprechen die Verbindungen innerhalb des AML pathways in Abbildung 7.9, die durch das graphische Modell postuliert werden, meist nicht direkt den gemäß des derzeitigen Kenntnisstandes bekannten Gen-Gen-Interaktionen (vergleiche Abbildung 7.6).

In Abbildung 7.6 werden die Stellen im AML pathway mit blauen Boxen hervor gehoben, die im Gen-Assoziationsnetzwerk (Abbildung 7.9) Interaktionen zwischen der Signatur und dem pathway entsprechen. Vergleicht man die roten und blauen Markierungen, so erkennt man eine relativ gute Übereinstimmung, wobei beim Ansatz mit üblichen Korrelationen wie gesagt mehr Assoziationen zwischen Signatur und pathway gefunden werden. Somit ähneln sich das graphische Modell und der hierarchische Ansatz trotz der offensichtlichen Unterschiede im Endergebnis dennoch einigermaßen gut.

7.3.2 Assoziation zwischen prognostischem score und funktionellen Gengruppen

Die vorgestellte Signatur wurde dazu verwendet, einen stetigen prognostischen score für die einzelnen Patienten zu erstellen. Dieser score stellt somit eine Art Zusammenfassung der Information dar, die im Expressionsprofil der gesamten Signatur enthalten ist. Im Gegensatz zu den zuvor beschriebenen Analysen, bei denen die einzelnen Signaturgene mit funktionellen Gengruppen in Verbindung gebracht wurden, wird nun der Zusammenhang

des prognostischen scores mit pathways und Gene Ontology Gruppen untersucht. Bei der Validierung der Gensignatur wird überprüft, ob der resultierende score tatsächlich einen Rückschluß auf die klinische Zielgröße, in diesem Beispiel auf die Überlebenszeit bei AML, zuläßt. Dabei ist es wichtig, daß weitere bekannte prognostische Faktoren berücksichtigt werden. Ansonsten könnte es sein, daß der genexpressionsbasierte score zwar mit dem Überleben assoziiert ist, er aber nur ein Surrogat für einen anderen Parameter darstellt. Im Falle der AML Signatur ist der Zusammenhang zum Mutationsstatus des Gens FLT3 am stärksten. Wenn nun die Assoziation zwischen dem score und den Expressionsprofilen von Gengruppen überprüft wird, wird dabei zusätzlich für den FLT3 Status adjustiert, um ein komplettes confounding zwischen score und FLT3 ausschließen zu können.

Mit dem GlobalAncova Ansatz wird dazu der Effekt des scores auf die Gengruppen getestet. Der FLT3 Status geht als Kovariable in das Modell ein. Ebenso ist auch die Sichtweise des globaltest möglich, bei dem der Einfluß der Expression in der Gengruppe auf den score bewertet wird. Auch beim globaltest kann FLT3 als weitere Kovariable berücksichtigt werden. Wir testen mit den beiden globalen Tests die 200 KEGG pathways, die bereits zuvor betrachtet wurden, und 4779 biologische Prozesse der Gene Ontology (GOBP). Die multiplen Tests bei der pathway Analyse werden nach Bonferroni korrigiert. Für die Korrektur bei der Gene Ontology Analyse verwenden wir das focus level Verfahren (Goeman und Mansmann, 2008), das die spezielle Struktur der GO berücksichtigt. Tabelle 7.4 zeigt jeweils für globaltest und GlobalAncova die Anzahlen der signifikanten pathways und GOBP Gruppen, sowie die Anzahlen der Gruppen, die mit beiden globalen Tests detektiert werden. Von den 200 KEGG pathways erhält man 168 mit adjustiertem GlobalAncova p-Wert < 0.01 und 167 mit globaltest p-Wert < 0.01, wobei 166 der detektierten pathways zwischen beiden Tests übereinstimmen. In der GO Analyse ergeben sich, ebenfalls bei einem Signifikanzniveau von 1%, 1781 (GlobalAncova), beziehungsweise 1751 (globaltest) signifikante Gruppen von den insgesamt 4779 getesteten, mit einer Überschneidung von 1730 Gruppen. Es zeigt sich also wie schon in Mansmann und Meister (2005) und Liu u. a. (2007) eine gute Übereinstimmung zwischen GlobalAncova und globaltest.
Aus der großen Anzahl an detektierten Gengruppen schließen wir zum einen, daß AML Patienten mit verschiedenen Werten des prognostischen scores stark variieren hinsichtlich einer großen Anzahl an biologischen Prozessen und pathways. Dies unterstreicht die biologische Relevanz des prognostischen scores. Dabei gilt natürlich zu bedenken, daß die beschriebene Analyse am selben Datensatz durchgeführt wurde, in dem der score auch erstellt wurde. Zum zweiten bestätigt die große Anzahl an signifikanten Ergebnissen trotz der Adjustierung für den FLT3 Status, daß der score nicht nur ein reines Surrogat des prognostischen Faktors FLT3 darstellt. Auch innerhalb von Patientengruppen, die hinsichtlich FLT3 Status eingeteilt werden, sollte noch eine verbesserte Prognose anhand des Gen scores möglich sein.

Dennoch besteht ein Zusammenhang zwischen dem expressionsbasierten score und dem FLT3 Status. Um diesen Zusammenhang genauer zu untersuchen kann man den Interaktionseffekt zwischen score und FLT3 auf die Expression der Gengruppen überprüfen.

Ein solcher Interaktionseffekt kann nur mit GlobalAncova getestet werden, weil dabei im Gegensatz zu globaltest sowohl der score als auch FLT3 Kovariablen darstellen. Man vergleicht dazu das volle Modell mit den Haupteffekten des scores und des FLT3 Status' sowie dem Interaktionseffekt der beiden Variablen mit dem reduzierten Modell, das lediglich die Haupteffekte enthält. Da nicht so starke Interaktionseffekte erwartet werden, wählen wir in diesem Fall ein Signifikanzniveau von 5%. Wiederum werden die 200 KEGG pathways und knapp 5000 GOBP Gruppen mit den gleichen Adjustierungen für multiples Testen wie zuvor analysiert. Das pathway scoring liefert kein signifikantes Ergebnis. Bei der GO Analyse mit der focus level Prozedur wird ein signifikanter Subgraph von zehn Gruppen detektiert (siehe Abbildung 7.10). Von meistem Interesse sind wohl die spezifischsten GO Begriffe 'notochord development' (GO:0030903) und 'embryonic organ development' (GO:0048568). Ebenso wird mit der focus level Methode auch der Effekt von FLT3 auf alle GOBP Gruppen getestet. Dies resultiert in einer großen Anzahl an signifikanten Tests. Die beiden eben genannten Begriffe mit signifikantem Interaktionseffekt sind allerdings nicht darunter. Veränderungen in diesen entwicklungsspezifischen biologischen Prozessen bei einer AML werden folglich nicht durch den FLT3 Status alleine erfaßt, sondern nur in Kombination mit dem expressionsbasierten score.

Abbildung 7.11 verdeutlicht die Bedeutung eines signifikanten Interaktionseffektes. Es werden die linearen Effekte des scores auf die Gene der GO Gruppe GO:0048568 ('embryonic organ development') gezeigt. Dabei gibt es je eine Grafik für Patienten mit und ohne die FLT3 Mutation. Signifikante Effekte, auf genweiser Basis, sind farbig markiert. Das Gen MLL zeigt ein ähnliches Verhalten in den beiden FLT3 Gruppen. Abgesehen davon werden die Gene recht unterschiedlich von dem score beeinflußt. Die Grafik bestätigt die detektierte Interaktion zwischen score und FLT3 Status. Die vorliegende Analyse ist ein erster Schritt für die Erklärung der prognostischen Information, die der score zusätzlich zu anderen Risikofaktoren wie FLT3 beiträgt.

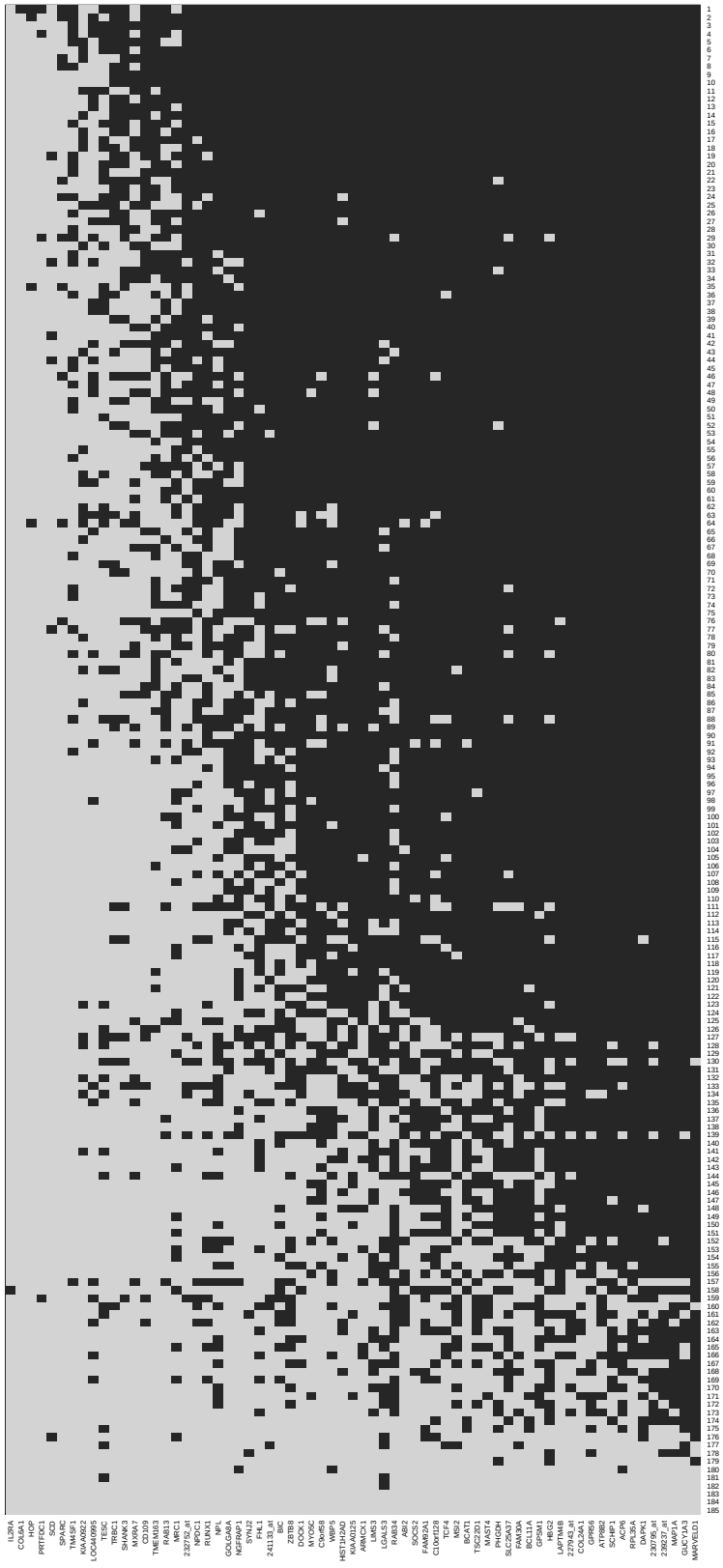


Abbildung 7.8: Ergebnis der hierarchischen Variablenselektion für die übrigen 185 KEGG pathways (vergleiche Abbildung 7.7). Legende für die pathways, siehe Tabelle 7.3.

	ID	Name	ID	Name
1	01430	Cell Communication	94	00790	Folate biosynthesis
2	00960	Alkaloid biosynthesis II	95	00970	Aminoacyl-tRNA biosynthesis
3	00760	Nicotinate and nicotinamide metabolism	96	00626	Naphthalene and anthracene degradation
4	00361	gamma-Hexachlorocyclohexane degradation	97	00561	Glycerolipid metabolism
5	02010	ABC transporters - General	98	00730	Thiamine metabolism
6	04540	Gap junction	99	00450	Selenoamino acid metabolism
7	04530	Tight junction	100	00640	Propanoate metabolism
8	00590	Arachidonic acid metabolism	101	04130	SNARE interactions in vesicular transport
9	04630	Jak-STAT signaling pathway	102	00520	Nucleotide sugars metabolism
10	04020	Calcium signaling pathway	103	00440	Aminophosphonate metabolism
11	04350	TGF-beta signaling pathway	104	00565	Ether lipid metabolism
12	04916	Melanogenesis	105	00930	Caprolactam degradation
13	04340	Hedgehog signaling pathway	106	00290	Valine, leucine and isoleucine biosynthesis
14	04910	Insulin signaling pathway	107	00710	Carbon fixation
15	04670	Leukocyte transendothelial migration	108	00280	Valine, leucine and isoleucine degradation
16	00051	Fructose and mannose metabolism	109	00252	Alanine and aspartate metabolism
17	04070	Phosphatidylinositol signaling system	110	00030	Pentose phosphate pathway
18	04115	p53 signaling pathway	111	04610	Complement and coagulation cascades
19	00340	Histidine metabolism	112	00040	Pentose and glucuronate interconversions
20	00521	Streptomycin biosynthesis	113	03030	DNA polymerase
21	04912	GnRH signaling pathway	114	00563	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
22	04512	ECM-receptor interaction	115	03010	Ribosome
23	00052	Galactose metabolism	116	00310	Lysine degradation
24	04080	Neuroactive ligand-receptor interaction	117	00071	Fatty acid metabolism
25	04120	Ubiquitin mediated proteolysis	118	05030	Amyotrophic lateral sclerosis (ALS)
26	04060	Cytokine-cytokine receptor interaction	119	03020	RNA polymerase
27	00530	Aminosugars metabolism	120	00400	Phenylalanine, tyrosine and tryptophan biosynthesis
28	04010	MAPK signaling pathway	121	00140	C21-Steroid hormone metabolism
29	00627	1,4-Dichlorobenzene degradation	122	00272	Cysteine metabolism
30	00562	Inositol phosphate metabolism	123	04110	Cell cycle
31	00740	Riboflavin metabolism	124	00410	beta-Alanine metabolism
32	00903	Limonene and pinene degradation	125	05120	Epithelial cell signaling in Helicobacter pylori infection
33	04730	Long-term depression	126	00360	Phenylalanine metabolism
34	04510	Focal adhesion	127	00910	Nitrogen metabolism
35	05010	Alzheimer's disease	128	00533	Keratan sulfate biosynthesis
36	00220	Urea cycle and metabolism of amino groups	129	00480	Glutathione metabolism
37	05131	Pathogenic Escherichia coli infection - EPEC	130	00791	Atrazine degradation
38	05130	Pathogenic Escherichia coli infection - EHEC	131	00300	Lysine biosynthesis
39	04740	Olfactory transduction	132	00600	Sphingolipid metabolism
40	04210	Apoptosis	133	04320	Dorso-ventral axis formation
41	00120	Bile acid biosynthesis	134	00860	Porphyrin and chlorophyll metabolism
42	00642	Ethylbenzene degradation	135	04140	Regulation of autophagy
43	00251	Glutamate metabolism	136	00950	Alkaloid biosynthesis I
44	00100	Biosynthesis of steroids	137	00362	Benzoate degradation via hydroxylation
45	00230	Purine metabolism	138	00401	Novobiocin biosynthesis
46	00532	Chondroitin sulfate biosynthesis	139	04950	Maturity onset diabetes of the young
47	01031	Glycan structures - biosynthesis 2	140	00130	Ubiquinone biosynthesis
48	04150	mTOR signaling pathway	141	05020	Parkinson's disease
49	00980	Metabolism of xenobiotics by cytochrome P450	142	03050	Proteasome
50	00591	Linoleic acid metabolism	143	00190	Oxidative phosphorylation
51	00240	Pyrimidine metabolism	144	04650	Natural killer cell mediated cytotoxicity
52	04810	Regulation of actin cytoskeleton	145	00670	One carbon pool by folate
53	04370	VEGF signaling pathway	146	00062	Fatty acid elongation in mitochondria
54	01030	Glycan structures - biosynthesis 1	147	00630	Glyoxylate and dicarboxylate metabolism
55	05040	Huntington's disease	148	00628	Fluorene degradation
56	00500	Starch and sucrose metabolism	149	00020	Citrate cycle (TCA cycle)
57	04720	Long-term potentiation	150	00351	1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation
58	00260	Glycine, serine and threonine metabolism	151	00720	Reductive carboxylate cycle (CO2 fixation)
59	04330	Notch signaling pathway	152	04620	Toll-like receptor signaling pathway
60	04520	Adherens junction	153	05110	Cholera - Infection
61	04660	T cell receptor signaling pathway	154	00780	Biotin metabolism
62	03320	PPAR signaling pathway	155	00604	Glycosphingolipid biosynthesis - ganglioseries
63	04920	Adipocytokine signaling pathway	156	05060	Prion disease
64	01510	Neurodegenerative Disorders	157	04710	Circadian rhythm
65	00624	1- and 2-Methylnaphthalene degradation	158	00770	Pantothenate and CoA biosynthesis
66	04310	Wnt signaling pathway	159	00602	Glycosphingolipid biosynthesis - neo-lactoseries
67	00564	Glycerophospholipid metabolism	160	04514	Cell adhesion molecules (CAMs)
68	00350	Tyrosine metabolism	161	00430	Taurine and hypotaurine metabolism
69	00010	Glycolysis / Gluconeogenesis	162	00601	Glycosphingolipid biosynthesis - lactoseries
70	00150	Androgen and estrogen metabolism	163	00920	Sulfur metabolism
71	00271	Methionine metabolism	164	01032	Glycan structures - degradation
72	00643	Styrene degradation	165	05050	Dentatorubropallidoluysian atrophy (DRPLA)
73	00061	Fatty acid biosynthesis	166	03060	Protein export
74	00625	Tetrachloroethene degradation	167	00531	Glycosaminoglycan degradation
75	00380	Tryptophan metabolism	168	00031	Inositol metabolism
76	04360	Axon guidance	169	00471	D-Glutamine and D-glutamate metabolism
77	00900	Terpenoid biosynthesis	170	00511	N-Glycan degradation
78	00512	O-Glycan biosynthesis	171	00750	Vitamin B6 metabolism
79	04664	Fc epsilon RI signaling pathway	172	00603	Glycosphingolipid biosynthesis - globoseries
80	00623	2,4-Dichlorobenzoate degradation	173	00902	Monoterpenoid biosynthesis
81	00620	Pyruvate metabolism	174	00660	C5-Branched dibasic acid metabolism
82	00330	Arginine and proline metabolism	175	00072	Synthesis and degradation of ketone bodies
83	04742	Taste transduction	176	00053	Ascorbate and aldarate metabolism
84	00632	Benzoate degradation via CoA ligation	177	04612	Antigen processing and presentation
85	04662	B cell receptor signaling pathway	178	00785	Lipoic acid metabolism
86	03022	Basal transcription factors	179	00460	Cyanoamino acid metabolism
87	00363	Bisphenol A degradation	180	00830	Retinol metabolism
88	04614	Renin-angiotensin system	181	04940	Type I diabetes mellitus
89	04930	Type II diabetes mellitus	182	00550	Peptidoglycan biosynthesis
90	00510	N-Glycan biosynthesis	183	00940	Phenylpropanoid biosynthesis
91	04012	ErbB signaling pathway	184	00680	Methane metabolism
92	00534	Heparan sulfate biosynthesis	185	00472	D-Arginine and D-ornithine metabolism
93	00650	Butanoate metabolism		

Tabelle 7.3: Legende zu Abbildung 7.8.

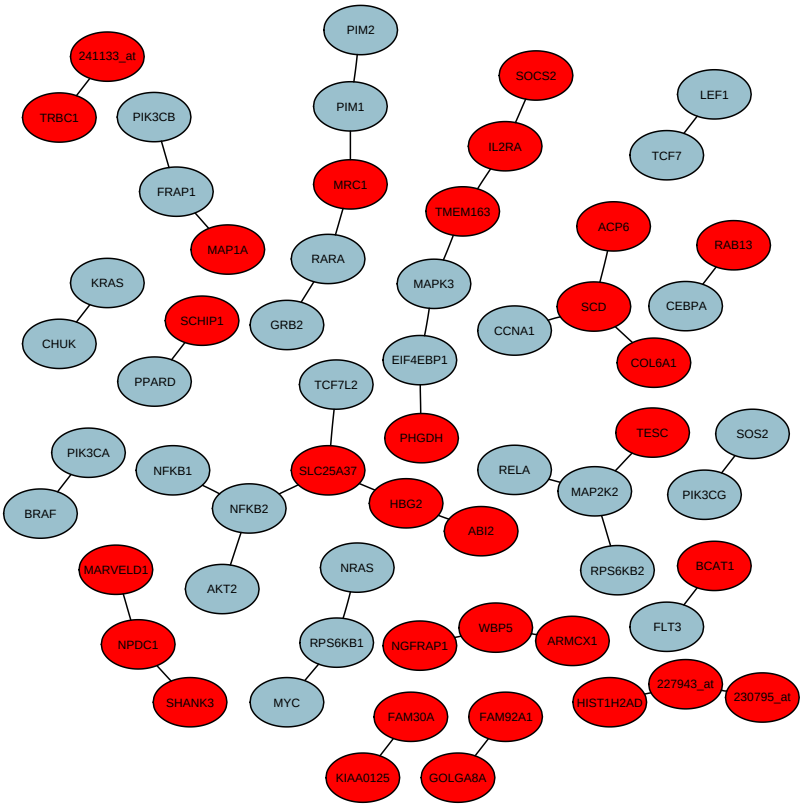
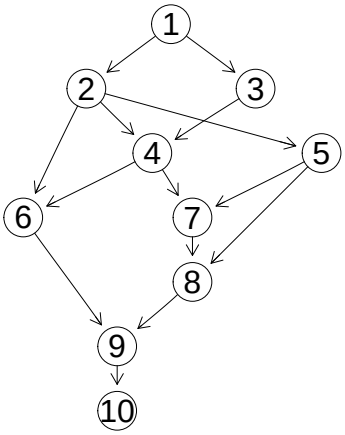


Abbildung 7.9: Gen-Assoziationsnetzwerk, das Verbindungen zwischen Genen des pathways 'Acute myeloid leukemia' (blau) und den Signaturgenen (rot) zeigt. Es sind nur Gene dargestellt, die mit mindestens einem weiteren Gen über eine Kante verbunden sind. Kanten entsprechen deutlich von 0 verschiedenen partiellen Korrelationen (mit fdr adjustierten p-Werten $p < 0.2$).

	KEGG	GO
GlobalAncova	168	1,781
globaltest	167	1,751
Übereinstimmung	166	1,730
Insgesamt	200	4,779

Tabelle 7.4: Ergebnis der pathway und GO Analyse, bei denen der Zusammenhang zum prognostischen score, adjustiert für FLT3, getestet wird. Gezeigt sind die Anzahlen der signifikanten Tests laut GlobalAncova und globaltest, die Anzahlen der Gruppen, die laut beiden Verfahren detektiert werden und die Anzahlen der insgesamt getesteten Gengruppen.



	ID	Terms	genes
1	GO:0008150	biological process	12,224
2	GO:0032502	developmental process	2,774
3	GO:0032501	multicellular organismal process	2,892
4	GO:0007275	multicellular organismal develop.	1,949
5	GO:0048856	anatomical structure development	1,772
6	GO:0009790	embryonic development	223
7	GO:0048731	system development	1,454
8	GO:0048513	organ development	1,021
9	GO:0048568	embryonic organ development	17
10	GO:0030903	notochord development	3

Abbildung 7.10: Ergebnis des GO scorings mit der focus level Prozedur. Es werden GOBP Kategorien mit signifikanter Interaktion (adjustierte p-Werte $p < 0.05$) zwischen AML Prognosescore und FLT3 Status gezeigt.

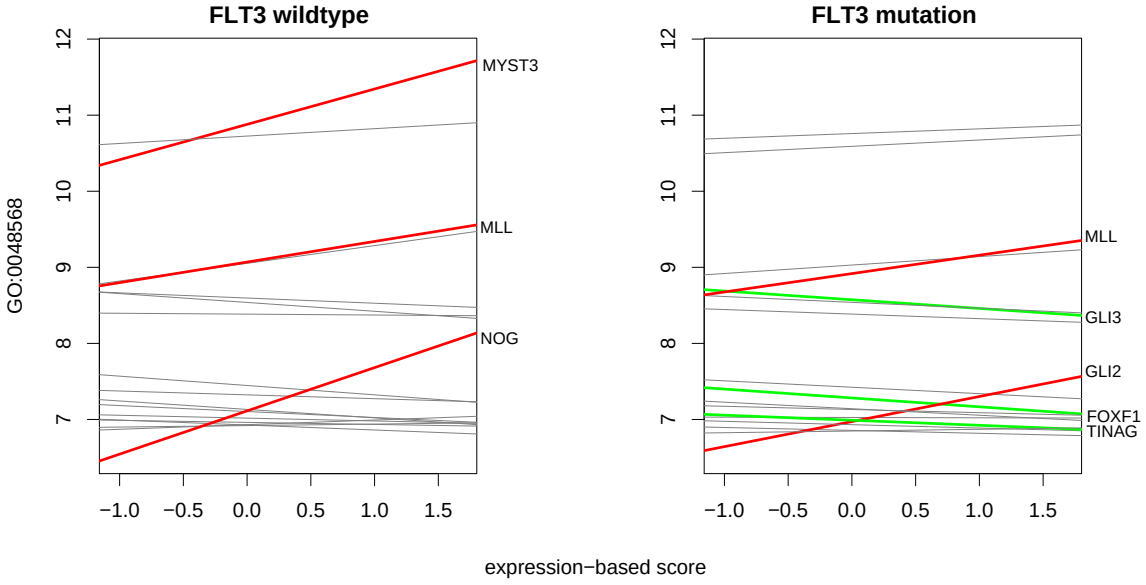


Abbildung 7.11: Lineare Effekte des prognostischen scores auf die GOBP Kategorie 'embryonic organ development' (GO:0048568, Knoten 9 in Abbildung 7.10). Signifikante (Bonferroni-adjustierte p-Werte $p < 0.05$) positive (negative) genweise Effekte sind mit dicken roten (grünen) Linien und entsprechenden Gennamen gekennzeichnet. Die Effekte sind getrennt berechnet für Patienten mit (rechts) bzw. ohne (links) FLT3 Mutation.

Kapitel 8

Zusammenfassung und Ausblick

Die Analyse von funktionellen Gengruppen wurde in der vorliegenden Arbeit als sehr sinnvolle Ergänzung, beziehungsweise Alternative zur üblichen genweisen Auswertungsstrategie vorgestellt. Bei einer statistischen Analyse ist es von Vorteil, vorhandenes Vorwissen über die betrachteten Parameter einfließen zu lassen. Gengruppen repräsentieren ein solches biologisches Vorwissen – sie setzen sich aus Genen zusammen, die ähnlichen Funktionen zugeordnet werden können, ähnliche Positionen im Genom einnehmen, co-reguliert sind oder gemeinschaftlich mit einem klinischen Phänomen assoziiert werden können. Neben der besseren biologischen Interpretierbarkeit liefert die Betrachtung von Gengruppen im Gegensatz zur genweisen Analyse weitere Vorteile wie die erhöhte Stabilität der Ergebnisse und die kleinere Dimension des multiplen Testproblems.

Ein Schwerpunkt lag in Darstellung und Vergleich von Verfahren für die Analyse von funktionellen Gengruppen hinsichtlich differentieller Expression. Wir unterschieden dabei zwischen dem Gene Set Enrichment Ansatz, der in der Regel durch die kompetitive Teststrategie und das Genrandomisierungsmodell charakterisiert ist, und holistischen Verfahren, die sich durch die in sich geschlossene Teststrategie und die Permutation von Beobachtungen auszeichnen. Wie Goeman und Bühlmann (2007) dargelegt haben, ist das Genrandomisierungsmodell aus statistischer Sicht sehr kritisch zu sehen. Will man allerdings Gengruppen untereinander vergleichen, so muß man verschiedene Zusammenstellungen von Gengruppen bewerten und somit eine Art Genrandomisierung durchführen. Tian u. a. (2005) schlagen generell verschiedene Teststrategien vor, je nachdem ob kompetitive oder in sich geschlossene Hypothesen überprüft werden sollen. Efron und Tibshirani (2007) dagegen kombinieren das Permutieren von Genen und von Beobachtungen, um von beiden Strategien profitieren zu können. Auch im Manual des *R* Pakets `globaltest` (Goeman und Oosting, 2007) wird ein Genpermutationstest (*comparative p*) vorgeschlagen, der mißt wie außergewöhnlich eine Gengruppe im Vergleich zu zufällig zusammen gesetzten Gruppen ist. Ein solcher Test ist vor allem in Situationen sinnvoll, in denen sehr viele oder gar alle Gengruppen gemäß der in sich geschlossenen Teststrategie ein signifikantes Ergebnis liefern.

Besonderes Interesse lag in dieser Arbeit auf den durch die Gene Ontology definierten

Gengruppen. Mittlerweile gibt es einige Verfahren, die beim Testen der einzelnen GO Gruppen die spezielle Struktur der GO berücksichtigen. Sie verfolgen deutlich unterschiedliche Strategien, was bei der Anwendung natürlich bedacht werden sollte. In der kleinen Simulationsstudie in Kapitel 6.2 haben sich theoretischen Bedenken zum Trotz einige Gene Set Enrichment basierte und heuristische Verfahren als sinnvoll erwiesen, um einzelne GO Begriffe aus verschiedenen Bereichen im GO Graphen zu detektieren. Besonders der elim Algorithmus (Alexa u. a., 2006) in Kombination mit verschiedenen Gruppenteststatistiken erscheint hierfür vielversprechend. Verfolgt man dagegen eher den globalen Ansatz, so wird man konsequenterweise nach signifikanten Subgraphen innerhalb der GO suchen. Für diese Strategie treten die globalen Tests und die focus level Methode vorteilhaft hervor. Insgesamt kann man sagen, daß die Frage nach der optimalen Gene Ontology Analyse noch nicht ausreichend geklärt ist. Die beschriebene Simulationsstudie war ein erster Versuch, die verschiedenen Ansätze auf der Basis unterschiedlicher Fragestellungen zu vergleichen und so einen Überblick zu gewinnen, in welchen Situationen welche Methode sinnvoll sein könnte.

Am ausführlichsten wurde der im Rahmen dieser Arbeit weiter entwickelte GlobalAncova Ansatz beschrieben. GlobalAncova ist ein sensitiver Test, der in manchen Datensätzen, in denen starke differentielle Expression zu erwarten ist, sehr viele Gengruppen detektiert. Deshalb empfehlen wir die Verwendung von GlobalAncova hauptsächlich für Daten, die nur schwache Expressionsunterschiede aufweisen. Allgemein sind globale Tests vielleicht nicht so gut geeignet für das „large scale setting“, bei dem aus einer großen Menge an Hypothesen nur ein kleiner Anteil an interessanten Fällen heraus gefiltert werden soll. Sie passen eher zu dem klassischen Ein-Testszenario, bei dem man mit großer Macht die Nullhypothese ablehnen möchte (Efron, 2004). Werden viele Tests durchgeführt, sollte eine stringente Adjustierung für multiples Testen durchgeführt werden, wie zum Beispiel bei der focus level Methode (Goeman und Mansmann, 2008) oder beim hierarchischen Testen (Meinshausen, 2008). Zudem bietet sich eventuell zusätzlich ein Vergleich der Gengruppen untereinander an, wie im vorigen Absatz erwähnt.

Verbesserungswürdig ist noch die Asymptotik, mit der p-Werte für GlobalAncova berechnet werden. Wie Simulationsstudien gezeigt haben, liefert die derzeit verwendete asymptotische Verteilung in manchen Situationen deutlich zu antikonservative Ergebnisse. Die shrinkage Schätzung der Gen-Kovarianzmatrix müßte angepaßt werden, um allgemeinere Korrelationsstrukturen in Betracht ziehen zu können. Desweiteren wäre eine Optimierung der numerischen Berechnung der asymptotischen p-Werte vorteilhaft, da sie derzeit für große Gengruppen sehr zeitaufwendig oder gar unmöglich ist. In Bezug auf die Verteilungsannahmen von GlobalAncova und auch globaltest ist außerdem über eine Standardisierung der Daten vor Anwendung der Tests nachzudenken. Liu u. a. (2007) zentrieren die Expressionsdaten und bringen sie auf gleiche genweise Varianzen. Sie berichten, daß sie dadurch in Simulationsstudien bessere Ergebnisse für die beiden globalen Tests erzielen.

Die große Stärke von GlobalAncova ist dessen Flexibilität für Analysen, die weit über

die einfache Fragestellung der differentiellen Genexpression zwischen zwei klinischen Gruppen hinaus reichen. Hierfür wurden ausführliche Anwendungsbeispiele gebracht. Komplexe Studiendesigns, zum Beispiel mit mehreren Faktoren, zeitlichen Trends und entsprechenden Interaktionen, können hinsichtlich ihres Zusammenhangs mit globalen Expressionsprofilen in Gengruppen untersucht werden. Sehr interessant erscheint auch das Gebiet der Co-Expression. Besonderes Augenmerk lag hier auf der Analyse des Zusammenspiels zwischen Gensignaturen und funktionellen Gengruppen. Diese Art der Untersuchung einer Signatur hinsichtlich ihrer biologischen Grundlagen stellt eine recht neue Richtung im Bereich der Genexpressionsanalyse dar.

Anhang A

Vignette des *R* Paketes `GlobalAncova`

Global testing of differential gene expression

Manuela Hummel, Reinhard Meister, Ulrich Mansmann

Abstract

In studies about differential gene expression between different clinical diagnoses the main interest may often not be in single genes but rather in groups of genes that are associated with a pathway or have a common location in the genome. In such cases it may be better to perform a global test because the problems of multiple testing can be avoided. The approach presented here is an ANCOVA global test on phenotype effects and gene–phenotype interaction.

Testing many pathways simultaneously is also possible. This, of course, causes again need for correction for multiple testing. Besides the standard approaches for correction we introduce a closed testing procedure in which the experiment–wise error rate equals the required level of confidence of the overall test.

This document was created using R version 2.6.0 and versions 3.5.0 and 4.8.0 of the packages *GlobalAncova* and *globaltest*, respectively.

Changes to Previous Versions

Version 3.3.3

- The permutation approach is now implemented in *C* and therefore faster.
- If the number of possible phenotype permutations is smaller than the number specified in *perm* (i.e. in very small sample sizes), all possible permutations are considered for the permutation test.

- Some more error messages are included.
- In `Plot.genes` and `Plot.subjects` bar labels can be manipulated with the argument *bar.names*.

Version 3.x.x

- Besides the permutation-based p-values also asymptotic p-values based on an approximation of the distribution of the test statistic are provided. Theoretical F-test p-values are no longer displayed since they are not valid in case of correlations or non-normality.
- The focus level procedure for finding interesting Gene Ontology subgraphs from Goeman and Mansmann (2007) was adapted for the use with `GlobalAncova`.
- Sequential and type III decompositions of the residual sum of squares, adjustment for global covariates and pair-wise comparisons of different levels of a categorical factor are implemented. These functionalities are described in the additional vignette *GlobalAncovaDecomp.pdf*.
- Now the parameter *test.genes* allows for specifying the gene group which a graph shall be based on in the plotting functions.

Version 2.5.1

- Testing several groups of genes is now more efficient and less time consuming.
- In the gene and subjects plots bar heights (gene-wise reduction in sum of squares and subject-wise reduction in sum of squares, respectively) can be returned.
- Plots are more flexible regarding graphical parameters like specification of colors, titles, axis labels and axis limits.

Version 2.x.x

- The major modification in the new version is the transfer from simple two group comparisons to a general linear model framework where arbitrary clinical variables (in especially with more groups or also continuous ones), time trends, gene–gene interactions, co–expression and so forth can be analysed.
- According to the new framework also the diagnostic plots are more flexible. The variable defining the coloring of bars can now be specified by the user, see section *Diagnostic Plots* for details.
- A bug was fixed concerning testing only a single gene for differential expression with the global ANCOVA F–test.
- Within the closed testing procedure a bug was fixed concerning testing non–disjunct groups of genes.

Introduction

The ANCOVA global test is a test for the association between expression values and clinical entities. The test is carried out by comparison of linear models via the extra sum of squares principle. If the mean expression level for at least one gene differs between corresponding models the global null hypothesis, which is the intersection of all single gene null hypotheses, is violated. As our test is based on the sum of gene-wise reduction in sum of squares due to phenotype, all systematic differences in gene expression between phenotypes equally contribute to the power of the test.

Single genes are not, in general, the primary focus of gene expression experiments. The researcher might be more interested in relevant pathways, functional sets or genomic regions consisting of several genes. Most of the current methods for studying pathways analyse differential expression of single genes. In these methods pathways where many genes show minor changes in their expression values may not be identified. Goeman's global test and the ANCOVA global test were designed to address this issue.

Applying global tests for differential expression in pathways substantially reduces the number of tests compared to gene-wise multiple testing. The amount of correction for multiple testing decreases. Function (KEGG, GO) or location (chromosome, cytoband) could be used as grouping criteria, for example.

We want to compare our method with the global test of Goeman et al. (2004). Therefore text and examples in this document follow to a certain extent the vignette presented in the R-package *globaltest*. Our function `GlobalAncova` tests whether the expectation of expression levels differs between biological entities for a given group of genes. This vignette has its focus on the practical use of the test. For more details about the mathematical background and the interpretation of results, we refer to the papers by Mansmann and Meister (2005) and Hummel, Meister and Mansmann (2008).

This document shows the functionality of the R-package *GlobalAncova*. The datasets, all necessary R–packages and our package *GlobalAncova* are available from the Bioconductor website (*www.bioconductor.org*).

First we load the packages and data we will use.

```
> library(GlobalAncova)
> library(globaltest)
> library(golubEsets)
> library(hu6800)
> library(vsn)
> library(multtest)
> data(Golub_Merge)
> golubX <- justvsn(Golub_Merge)
```

This creates a dataset `golubX`, which is of the format *ExpressionSet*, the standard format for gene expression data in BioConductor. It consists of 7129 genes and 72 samples (the data are from Golub et al., 1999). We used *vsn* to normalize the data. Other appropriate normalization methods may be used as well. From several phenotype variables we

use “ALL.AML” as the clinical diagnoses of interest. ALL and AML are two types of acute leukemia. There are 47 patients with ALL and 25 with AML.

Global Testing of a Single Pathway

Golub Data and Cell Cycle Pathway

Suppose we are interested in testing whether AML and ALL have different gene expression patterns for certain pathways, for example from the KEGG database. With *globaltest* we answer the question whether the expression profile has prognostic power with respect to diagnosis of AML or ALL. *GlobalAncova* asks for differences in mean expression between the two clinical groups.

Testing all Genes

We start by applying our test to all genes in the Golub dataset so that differences in the overall gene-expression pattern can be demonstrated.

```
> gr <- as.numeric(golubX$ALL.AML == "ALL")
> ga.all <- GlobalAncova(xx = exprs(golubX), group = gr,
+     covars = NULL, perm = 100)
```

The first input *xx* is a 7129 × 72 matrix that contains the expression values of all genes and samples. Missing values in the expression matrix *xx* are not allowed because otherwise gene-wise linear models could not be summarized adequately to a global group statement. If missing values occur we propose either leaving out the genes with missing values (i.e. the corresponding rows in the gene expression matrix), or imputing the data before applying *GlobalAncova*. An easy way to do the latter would be for example to calculate linear models for each gene using the available model variables (e.g. phenotype group labels). Missing values can then be estimated based on the resulting model parameters and the actual values of phenotype variables of the corresponding samples. The use of more sophisticated imputation methods Rubin (1987) would be computationally expensive and is not implemented in *GlobalAncova*. Note that we did not yet evaluate how data imputation affects *GlobalAncova* results and whether the easier imputation methods described above yield similar results as the more complex approaches. The second input *group* in the `GlobalAncova` function is a vector that defines the clinical diagnosis for the 72 patients.

Note that *GlobalAncova* is not restricted to the analysis of dichotomous phenotype groups. More complex tasks like variables with more groups or also continuous ones, time trends, gene–gene interactions and co–expression can be performed as well. Some examples will be given in section *van't Veer Data and p53-Signalling Pathway*. The realization of such tasks is done by definition of two linear models that shall be compared via the extra sum of squares principle. Hence model formulas for the full model containing all parameters and the reduced model, where the terms of interest are omitted, have to be given. An alternative

is to provide the formula for the full model and a character vector naming the terms of interest. Those names can be chosen by previous output of the `GlobalAncova` function. Consequently we could run the same analysis as above with two possible further function calls shown below (output is omitted). In both cases a data frame with information about all variables for each sample is required. In the case of microarray data this can be the corresponding `pData` object.

```
> GlobalAncova(xx = exprs(golubX), formula.full = ~ALL.AML,
+     formula.red = ~1, model.dat = pData(golubX), perm = 100)
> GlobalAncova(xx = exprs(golubX), formula.full = ~ALL.AML,
+     test.terms = "ALL.AMLAML", model.dat = pData(golubX),
+     perm = 100)
```

To avoid alpha–inflation due to correlated data and effects of non–normality of the data tests for significance of the resulting F–ratios are performed using a permutation test approach. We apply permutation of samples which is equivalent to permuting rows of the full design matrix. Note that permutation is only conducted for such columns of the design matrix that correspond to the variables of interest. Values of additional covariates remain in the original order. This prevents us for destroying covariate effects. Still the permutation approach is not optimal since residuals may be correlated. However, this does not seem to be a severe problem. The argument *perm* defines the number of permutations, which is 10,000 for default. Here we set *perm* to just 100 or 1000 so that creating this vignette will not last too long. For getting more reliable results one should recompute the examples with more permutations.
As an alternative to the permutation approach an approximation of the F–statistic nominator according to Robbins and Pitman (1949) yields asymptotic p–values. Note that the approximation is not feasible for very large gene groups since the huge gene expression covariance matrix has to be estimated, which is not possible for too many genes. The default value for group size (*max.group.size*) is 2500, groups above this size are treated by the permutation approach. When using work stations with good working memory this number may be increased. The estimation of the covariance matrix is carried out with the R package *corpcor* from Schaefer and Strimmer (2006).
Whether the permutation–based or the asymptotic p–values or both should be calculated is controlled by the argument *method*.

The result of the `GlobalAncova` function is a typical ANOVA table with information about sums of squares, degrees of freedom and mean sums of squares for the effect and error term, respectively. Besides F–statistics there are given either p–values from the permutation test or the asymptotic p–values or both. The names of all involved parameters are displayed as well as the name(s) of the tested effect(s).

```
> ga.all

$effect
```

```
[1] "group"

$ANOVA
             SSQ     DF        MS
Effect  28577.04   7129 4.0085628
Error  338961.37 499030 0.6792405

$test.result
            [,1]
F.value 5.901537
p.perm  0.000000

$terms
[1] "(Intercept)" "group"
```

From this result we conclude that the overall gene expression profile for all 7129 genes is associated with the clinical outcome. This means that samples with different AML/ALL status tend to have different expression profiles. We expect most pathways (especially the ones containing many genes) also to be associated with the phenotype groups.

If we apply Goeman's global test we get

```
> gt.all <- globaltest(golubX, "ALL.AML")
> gt.all

Global Test result:
Data: 72 samples with 7129 genes; 1 gene set
Model: logistic
Method: Asymptotic distribution

    Genes Tested Statistic Q Expected Q sd of Q    P-value
all  7129   7129      55.982         10  2.5609 5.4272e-11
```

Both tests show that the data contain overwhelming evidence for differential gene expression between AML and ALL.

Testing the Cell Cycle Pathway

Now we ask the more specific question of whether there is evidence for differential gene expression between both diagnoses restricted to genes belonging to the cell cycle pathway. First we load all KEGG pathways.

```
> kegg <- as.list(hu6800PATH2PROBE)
```

The list `kegg` consists of 195 pathways. Each pathway is represented by a vector of gene names. We are mainly interested in the cell cycle pathway which has the identifier "04110" in the KEGG database. It corresponds to 104 probe sets on the hu6800 chip.

```
> cellcycle <- kegg[["04110"]]
```

We apply the global test to this pathway using the option *test.genes*.

```
> ga.cc <- GlobalAncova(xx = exprs(golubX), group = gr,
+     test.genes = cellcycle, method = "both", perm = 1000)
> ga.cc

$effect
[1] "group"

$ANOVA
             SSQ   DF        MS
Effect  481.8479  104 4.6331525
Error  4523.9501 7280 0.6214217

$test.result
                [,1]
F.value  7.45573e+00
p.perm   0.00000e+00
p.approx 7.52741e-12

$terms
[1] "(Intercept)" "group"
```

Also with *globaltest* we get a very small p–value

```
> gt.cc <- globaltest(X = golubX, Y = "ALL.AML", genesets = cellcycle)
> gt.cc

Global Test result:
Data: 72 samples with 7129 genes; 1 gene set
Model: logistic
Method: Asymptotic distribution

     Genes Tested Statistic Q Expected Q sd of Q    P-value
[1,]   104    104      64.705     9.3361  3.3609 1.8155e-08
```

The test results clearly indicate that the expression pattern of the cell cycle pathway is different between the two clinical groups.

Adjusting for Covariates

Covariate information can be incorporated by specifying the *covars* option.

For example if we want to adjust for whether samples were taken from bone marrow or from peripheral blood (BM.PB), we can do this by

```
> ga.cc.BMPB <- GlobalAncova(xx = exprs(golubX), group = gr,
+     covars = golubX$BM.PB, test.genes = cellcycle, method = "both",
+     perm = 1000)
> ga.cc.BMPB

$effect
[1] "group"

$ANOVA
             SSQ   DF        MS
Effect  472.4872  104 4.5431463
Error  4381.0204 7176 0.6105101

$test.result
                 [,1]
F.value  7.441558e+00
p.perm   0.000000e+00
p.approx 1.292588e-11

$terms
[1] "(Intercept)" "group"       "covarsPB"
```

With the more general function call we would simply adjust the definitions of model formulas, namely $formula.full = \sim ALL.AML + BM.PB$ and $formula.red = \sim BM.PB$.

The source of the samples does not seem to have an explanatory effect on the outcome since F–statistics and p–values are very similar to the model without adjustment.

With the *globaltest* we get a higher p–value.

```
> gt.cc.BMPB <- globaltest(X = golubX, Y = ALL.AML ~ BM.PB,
+     genesets = cellcycle)
> gt.cc.BMPB

Global Test result:
Data: 72 samples with 7129 genes; 1 gene set
Model: logistic, ALL.AML ~ BM.PB
Adjusted: 99.8 % of variance of Y remains after adjustment
Method: Asymptotic distribution

     Genes Tested Statistic Q Expected Q sd of Q    P-value
[1,]   104    104      64.917     9.2512  3.2773 8.1185e-09
```

Permutation based p–values can also be obtained with Goeman's test, however only when covariates are absent.

van't Veer Data and p53–Signalling Pathway

We present another example from a study on breast cancer from van't Veer et al. (2002). This example illustrates how more complex tasks than comparing just two clinical groups can be performed with *GlobalAncova*. A subset of the data consisting of the expression values for 96 patients without *BRCA1* or *BRCA2* mutations is available with the package. The dataset (`vantVeer`) is restricted to 1113 genes associated with 9 cancer related pathways that are provided as a list named (`pathways`), too. We take one gene from the original data additionally to the expression set, namely 'AL137718'. This gene is part of the original van't Veer prognosis signature. We will later use it to demonstrate how signature genes can be related to pathways. Information about some of the originally surveyed covariates is stored in `phenodata`. The tumour suppressor protein p53 contributes as a transcription factor to cell cycle arrest and apoptosis induction. Therefore, first the p53-signalling pathway is selected as a candidate, where differential expression between relevant prognostic groups, defined by the development of distant metastases within five years, was expected.

```
> data(vantVeer)
> data(phenodata)
> data(pathways)
> metastases <- phenodata$metastases
> p53 <- pathways$p53_signalling
```

We get a significant result with the global ANCOVA.

```
> vV.1 <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = p53, method = "both", perm = 1000)
> vV.1

$effect
[1] "metastases"

$ANOVA
             SSQ   DF         MS
Effect  2.893417   33 0.08767929
Error  97.424573 3102 0.03140702

$test.result
                [,1]
F.value  2.791710175
p.perm   0.006000000
```

```
p.approx 0.009093267

$terms
[1] "(Intercept)" "metastases"
```

Analysis of Various Clinical Groups

In the new version of the package also clinical variables with more than two groups can be considered. For demonstration we investigate differential expression for the three ordered levels of tumour grade.

```
> vV.3 <- GlobalAncova(xx = vantVeer, formula.full = ~grade,
+     formula.red = ~1, model.dat = phenodata, test.genes = p53,
+     method = "both", perm = 1000)
> vV.3

$effect
[1] "grade.L" "grade.Q"

$ANOVA
             SSQ   DF         MS
Effect  3.638565   66 0.05512977
Error  96.679425 3069 0.03150193

$test.result
               [,1]
F.value  1.75004422
p.perm   0.04100000
p.approx 0.03463237

$terms
[1] "(Intercept)" "grade.L"     "grade.Q"
```

Gene–Gene Interaction

Now we want to go into the matter of other interesting biological questions. For example one might ask if there exists interaction between the expression of genes which van't Veer et al. (2002) presented as signature for prediction of cancer recurrence and the expression of genes in a certain pathway. This question can be answered by viewing the expression values of the signature genes as linear regressors and to test their effects on the expression pattern of the pathway genes. For demonstration we pick the signature gene 'AL137718', which is not part of any of the pathways, and test its effect on the p53–signalling pathway. Assume that we also want to adjust for the Estrogen receptor status. The analysis can be carried out in the following way.

```
> signature.gene <- "AL137718"
> model <- data.frame(phenodata, signature.gene = vantVeer[signature.gene,
+     ])
> vV.4 <- GlobalAncova(xx = vantVeer, formula.full = ~signature.gene +
+     ERstatus, formula.red = ~ERstatus, model.dat = model,
+     test.genes = p53, method = "both", perm = 1000)
> vV.4

$effect
[1] "signature.gene"

$ANOVA
             SSQ   DF         MS
Effect  2.667014   33 0.08081859
Error  89.867452 3069 0.02928232

$test.result
               [,1]
F.value  2.75997881
p.perm   0.01500000
p.approx 0.01387833

$terms
[1] "(Intercept)"    "signature.gene" "ERstatuspos"
```

Assuming a significance level of 0.05 we get a significant effect of the signature gene on the p53–signalling pathway.

Co–Expression

Next we want to analyse co–expression regarding the clinical outcome of building distant metastases within five years. This can be done by simply adding the variable `metastases` to the full and reduced model, respectively. Such layout corresponds to testing the linear effect of the signature gene stratified not only by Estrogen receptor status but also by `metastases`.

```
> vV.5 <- GlobalAncova(xx = vantVeer, formula.full = ~metastases +
+     signature.gene + ERstatus, formula.red = ~metastases +
+     ERstatus, model.dat = model, test.genes = p53, method = "both",
+     perm = 1000)
> vV.5

$effect
[1] "signature.gene"
```

```
$ANOVA
             SSQ   DF         MS
Effect  2.284391   33 0.06922396
Error  87.463681 3036 0.02880885

$test.result
               [,1]
F.value  2.40287099
p.perm   0.03200000
p.approx 0.02876237

$terms
[1] "(Intercept)"    "metastases"     "signature.gene"
[4] "ERstatuspos"
```

Again we get a significant result.

Supposably the most interesting question in this case concerns differential co–expression. Differential co–expression is on hand if the effect of the signature gene behaves different in both `metastases` groups. In a one dimensional context this would become manifest by different slopes of the regression lines. Hence what we have to test is the interaction between `metastases` and `signature.gene`.

```
> vV.6 <- GlobalAncova(xx = vantVeer, formula.full = ~metastases *
+     signature.gene + ERstatus, formula.red = ~metastases +
+     signature.gene + ERstatus, model.dat = model, test.genes = p53,
+     method = "both", perm = 1000)
> vV.6

$effect
[1] "metastases:signature.gene"

$ANOVA
             SSQ   DF         MS
Effect  2.520643   33 0.07638311
Error  84.943038 3003 0.02828606

$test.result
               [,1]
F.value  2.70038011
p.perm   0.02600000
p.approx 0.01829782
```

```
$terms
[1] "(Intercept)"               "metastases"
[3] "signature.gene"            "ERstatuspos"
[5] "metastases:signature.gene"
```

We observe a significant differential co-expression between the chosen signature gene and the p53–signalling pathway.

With *globaltest* we can also test gene-gene interaction, also adjusted for phenotype groups. But it is not possible to test for differential co-expression or the influence of more than one signature gene on a pathway. On the other hand globaltest is able to deal with survival times as clinical outcome.

Testing Several Pathways Simultaneously

Systems biology involves the study of mechanisms underlying complex biological processes as integrated systems of many diverse interacting components, often referred to as pathways.

We regard the possibility to investigate differential gene expression simultaneously for several of those pathways as a contribution towards understanding biological relevant relations.

The user can apply `GlobalAncova` to compute p–values for a couple of pathways with one call by specifying the *test.genes* option. The members of each pathway to be tested must belong to genes in the expression–matrix. Afterwards a suitable correction for multiple testing has to be applied. An alternative based on the closed testing approach is described later.

Suppose for example that for sake of simplicity we want to test the first four of the cancer related pathways with the van't Veer data. We proceed as follows.

```
> metastases <- phenodata$metastases
> ga.pw <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = pathways[1:4], method = "both", perm = 1000)
> ga.pw

                            genes  F.value p.perm     p.approx
androgen_receptor_signaling    72 2.389837  0.013 3.045062e-03
apoptosis                     187 1.968467  0.012 7.694809e-04
cell_cycle_control             31 4.639853  0.000 2.958656e-05
notch_delta_signalling         34 1.497222  0.123 8.537110e-02
```

The result is a matrix whose rows correspond to the different pathways.

With the `globaltest` we get a similar matrix.

```
> gt.pw <- globaltest(X = vantVeer, Y = metastases, genesets = pathways[1:4])
> gt.pw
```

```
Global Test result:
Data: 96 samples with 1113 genes; 4 gene sets
Model: logistic
Method: Asymptotic distribution

                            Genes Tested Statistic Q Expected Q
androgen_receptor_signaling    72     72      22.197     9.3260
apoptosis                     187    187      16.454     8.3560
cell_cycle_control             31     31      51.241    11.3470
notch_delta_signalling         34     34      12.711     8.4455
                            sd of Q    P-value
androgen_receptor_signaling  3.8650 0.01059600
apoptosis                    2.8665 0.01664700
cell_cycle_control           5.7933 0.00036812
notch_delta_signalling       4.1647 0.13406000
```

Simultaneous Adjustment of p–values

Next we show how to extract p–values for correction for multiple testing. Note however that due to the extremely high correlations between these tests, many procedures that correct for multiple testing here are inappropriate. An appropriate way of adjusting would be for example the method of Holm (1979). An alternative to such adjustments that is not affected by correlations between tests is a closed testing procedure. For this approach you need a family of null hypotheses that is closed under intersection. Then a single hypothesis can be rejected at level α if it is rejected along with all hypotheses included in it (Marcus et al., 1976).

For the adjustment according to Bonferroni and Holm we build a vector of the raw p–values. The function `mt.rawp2adjp` provides several adjusting methods. We here display only the raw and "Holm" adjusted p–values. To obtain the original order of the pathways we order the result of `mt.rawp2adjp` according to `index`.

```
> ga.pw.raw <- ga.pw[, "p.perm"]
> ga.pw.adj <- mt.rawp2adjp(ga.pw.raw)
> ga.result <- ga.pw.adj$adjp[order(ga.pw.adj$index), c("rawp",
+     "Holm")]
> rownames(ga.result) <- names(pathways)[1:4]
> ga.result

                             rawp  Holm
androgen_receptor_signaling 0.013 0.036
apoptosis                   0.012 0.036
cell_cycle_control          0.000 0.000
notch_delta_signalling      0.123 0.123
```

```
> gt.pw.raw <- p.value(gt.pw)
> gt.pw.adj <- mt.rawp2adjp(gt.pw.raw)
> gt.result <- gt.pw.adj$adjp[order(gt.pw.adj$index), c("rawp",
+     "Holm")]
> rownames(gt.result) <- names(pathways)[1:4]
> gt.result

                                  rawp        Holm
androgen_receptor_signaling 0.0105958152 0.031787446
apoptosis                   0.0166473874 0.033294775
cell_cycle_control          0.0003681155 0.001472462
notch_delta_signalling      0.1340573473 0.134057347
```

Allowing a family–wise error rate of 0.05 all but one pathways remain significant for both methods.

Closed Testing Procedure

Closed testing procedures (Marcus et al., 1976) offer a versatile and powerful approach to the multiple testing problem. Implementation is non–trivial, therefore, the program given in this version should be regarded as a prototype.

In order to apply the closed testing procedure we first have to create the required family of hypotheses by building all intersections between the "natural" hypotheses tested above and all intersections of those new hypotheses and so on.

The resulting family of hypotheses can be illustrated in a directed graph (figure A.1). If we just for the sake of illustration assume that we have only four hypotheses named "1", ... "4" then the node "1-2-3-4" for example stands for the global hypothesis that the genes of all four pathways are not differentially expressed. Now the interesting hypothesis "1" for example can be rejected if also the hypotheses "1-2-3-4", "1-2-3", "1-2-4", "1-3-4", "2-3-4", "1-2", ..., "1-4" are rejected. These relationships are represented by the edges of the graph.

We can compute the closed testing procedure using the function

```
> ga.closed <- GlobalAncova.closed(xx = vantVeer, group = metastases,
+     test.genes = pathways[1:4], previous.test = ga.pw,
+     level = 0.05, method = "approx")
```

where *test.genes* is again a list of pathways. In order to shorten computing time we can provide the results of the previous application of `GlobalAncova` for the pathways of interest. The option *level* allows to manipulate the level of significance. There is again the possibility to choose between permutation and asymptotic p-values via the option *method*. Note that if you provide results of previous computation, the type of p-values has to correspond, i.e. if we now want to use *method = 'approx'* in the previous test we

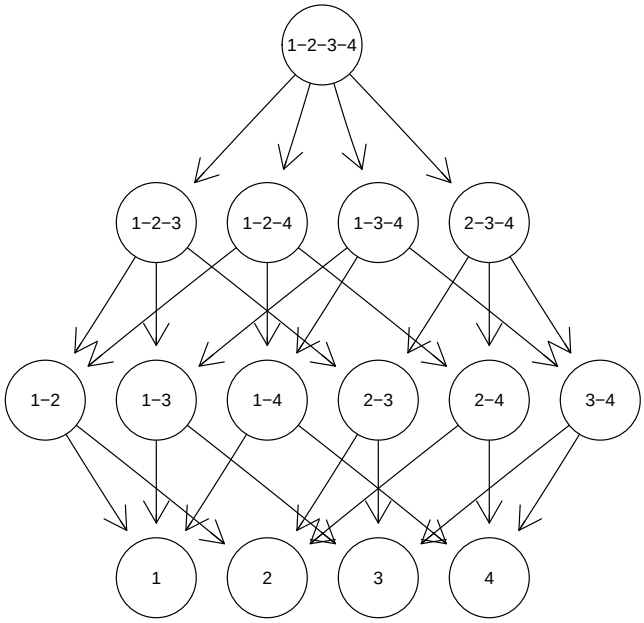


Abbildung A.1: Closed family of hypotheses.

should have used *method = 'approx'* or *method = 'both'* such that asymptotic p-values are available.

Also for `GlobalAncova.closed` all three different function calls as for `GlobalAncova` itself are possible.

The function `GlobalAncova.closed` provides the formed null hypotheses (this means lists of genes to be tested simultaneously), the test results for each pathway of interest and the names of significant and non significant pathways. Names for the intersections of hypotheses are built by simply coercing the names of the respective pathways. If for a pathway one single hypothesis can not be rejected there is no need to test all the remaining hypotheses. That is why in test results of non significant pathways lines are filled with NA's after a p–value $> \alpha$ occured. Here only test results for the first pathway are displayed.

```
> names(ga.closed)

[1] "new.data"        "test.results"    "significant"
[4] "not.significant"

> rownames(ga.closed$test.results[[1]])

[1] "androgen_receptor_signaling"
[2] "androgen_receptor_signaling.apoptosis"
[3] "androgen_receptor_signaling.cell_cycle_control"
```

```
[4] "androgen_receptor_signaling.notch_delta_signalling"
[5] "apoptosis.androgen_receptor_signaling.cell_cycle_control"
[6] "apoptosis.androgen_receptor_signaling.notch_delta_signalling"
[7] "cell_cycle_control.androgen_receptor_signaling.notch_delta_signalling"
[8] "cell_cycle_control.apoptosis.androgen_receptor_signaling.notch_delta_signalling"
```

```
> rownames(ga.closed$test.results[[1]]) <- NULL
> ga.closed$test.results[1]
```

```
$androgen_receptor_signaling
     genes  F.value    p.approx
[1,]    72 2.389837 0.003045062
[2,]   258 2.096100 0.000400000
[3,]   100 3.040900 0.000200000
[4,]   106 2.120700 0.002400000
[5,]   286 2.381000 0.000100000
[6,]   292 2.027300 0.000400000
[7,]   134 2.686400 0.000200000
[8,]   320 2.290200 0.000100000
```

```
> ga.closed$significant
```

```
[1] "androgen_receptor_signaling" "apoptosis"
[3] "cell_cycle_control"
```

```
> ga.closed$not.significant
```

```
[1] "notch_delta_signalling"
```

We get the same significant and non significant pathways as before.

Multiple testing on the Gene Ontology graph

When testing gene sets defined by the Gene Ontology it is of special interest to incorporate the hierarchical structure of the GO graph. Goeman and Mansmann (2008) developed the *focus level* method, a multiple testing approach on the GO that combines the correction method of Holm (1979) and the closed testing procedure from Marcus et al. (1976) (also used in section *Closed Testing Procedure*). The method is originally implemented in package *globaltest*. We adapted the corresponding functions such that the procedure now is available also with `GlobalAncova`. For details see the vignette of *globaltest*.

First, the GO graph must be prepared for the data set at hand. We use again the data of Golub et al.

```
> bp <- makeGOstructure(golubX, "hu6800")
```

This creates a *GOstructure* object (class from *globaltest*) for the Golub data, using the annotation package *hu6800*.

The focus level procedure requires definition of a *focus level*. This is a set of GO terms reflecting the level of specificity that is of most interest. The focus level can be specified as an arbitrary vector of GO identifiers. However, there is also an automatic way for defining the focus level via the function `getFocus`.

```
> focusBP <- getFocus(bp, maxatoms = 5)
> str(focusBP)

 chr [1:1059] "GO:0000002" "GO:0000012" "GO:0000018" ...
```

The argument *maxatoms* determines the maximum complexity of the subgraphs of all descendants of focus level nodes. We choose a lower value (default 10) which leads to more specific focus level terms and less time consuming processing.

The significant subgraph of predefined size (*stopafter = 30*) can be calculated with

```
> go30 <- GAGO(exprs(golubX), group = pData(golubX)$ALL.AML,
+     focus = focusBP, GO = bp, stopafter = 30)

> str(go30)

 Named num [1:31] 1.24e-09 2.97e-09 1.24e-09 2.97e-09 2.97e-09 ...
 - attr(*, "names")= chr [1:31] "GO:0008150" "GO:0002252" "GO:0002376" "GO:0002526" ...
```

All arguments for specifying the linear model used in `GlobalAncova` can be given here. Only the parameter *method* is not available because the focus level procedure does only work with the asymptotic test. Note however, that still a number of permutations can be specified (*perm*, default 10,000) since very large GO terms (with more annotated genes than defined by parameter *max.group.size*, default 2500) are tested permutation based.

See the *globaltest* vignette for examples how to visualize the results of the focus level procedure.

Diagnostic Plots

There are two types of diagnostic plots available supporting communication and interpretation of results of the global ANCOVA. The `Plot.genes` visualizes the influence of individual genes on the test result while the `Plot.subjects` visualizes the influence of individual samples. Both plots are based on the decomposition of sums of squares.

We use again the van't Veer data constricted to the genes of the p53–signalling pathway for demonstration of the plot functions.

Gene Plot

The influence of each gene on the outcome of the test can be assessed and visualized with a diagnostic plot generated by our function `Plot.genes`. It corresponds to the function `geneplot` in the *globaltest* package. The function `Plot.genes` gives a graphical display of single gene-wise analysis for all genes. Bars are always positive as a reduction of sum of squares is always achieved in this case. The bar height indicates the influence of the respective gene on the test statistic. The added reference line is the residual mean square error per gene and corresponds to the expected height of the bars under the null hypothesis which says that the gene is not associated with the clinical outcome. The actual and expected bar heights also correspond to the nominator and denominator of gene-wise F-statistics. Hence the ratio of the two values is a measure for the association of the respective gene with the phenotype. Bar heights for all genes can be returned by setting the option *returnValues* to `TRUE`. This helps to detect genes with most influence on the global statistics. Note however that comparisons between different gene groups can not easily be done by means of these values directly since different group sizes have an impact on global significance.

The bars can be colored according to a variable of interest with the option *Colorgroup* in order to show in which of the groups a gene has the highest expression values. The automatically chosen bar labels can be manipulated with the parameter *bar.names*.

The commands for creating gene plots in the *GlobalAncova* and the *globaltest* are as follows. Note that for the former one again three alternatives for function calls are provided, see section *Global Testing of a Single Pathway* for details.

The two approaches show almost the same results (figures A.2 and A.3). We prefer plotting horizontal bars rather than vertical because we think it is easier to read off the bar heights this way.

```
> Plot.genes(xx = vantVeer, group = metastases, test.genes = p53)
> gt.vV <- globaltest(X = vantVeer, Y = metastases, genesets = p53)
> geneplot(gt.vV)
```

In this case where only the influence of one variable is of interest (and therefore the easiest version of possible function calls is chosen), the same variable is assumed to be relevant for coloring. However one is free to specify another coloring. For example for the same plot we could ask which genes are higher expressed in samples with either positive or negative Estrogen receptor status, see figure A.4.

```
> Plot.genes(xx = vantVeer, formula.full = ~metastases,
+     formula.red = ~1, model.dat = phenodata, test.genes = p53,
+     Colorgroup = "ERstatus")
```

Subjects Plot

The function `Plot.subjects` visualizes the influence of the individual samples on the test result and corresponds to the `sampleplot` of Goeman. The function `Plot.subjects` gives

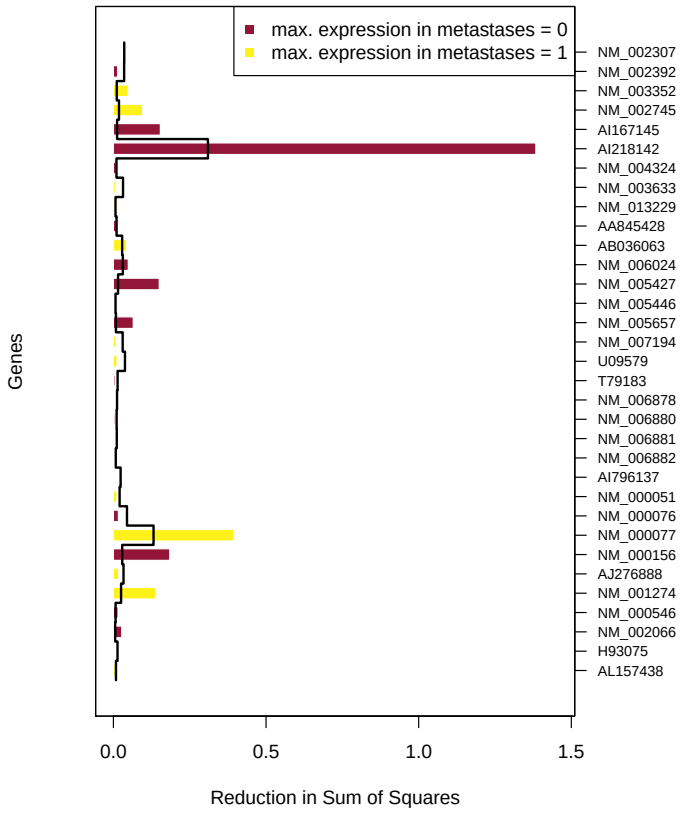


Abbildung A.2: Gene Plot for the van't Veer data with *GlobalAncova*. Shown are the genes of the p53–signalling pathway. The bar height indicates the influence of the respective gene on the test statistic. The color shows in which of the phenotype groups the gene has higher expression values. The reference line is the residual mean square error per gene.

information on the reduction of sum of squares per subject. Here we sum over genes. Large reduction demonstrates a good approximation of a subject's gene expressions by the corresponding group means. If an individual does not fit into the pattern of its phenotype, negative values can occur. A small p–value will therefore generally coincide with many positive bars. If there are still tall negative bars, these indicate deviating samples: removing a sample with a negative bar would result in a lower p-value. The bars are colored to distinguish samples of different clinical entities that can again be specified by the user through the option *Colorgroup*. With the option *sort* it is also possible to sort the bars with respect to the phenotype groups. Bar labels can be changed with the argument *bar.names*. Also in the subjects plot bar heights can be returned by setting the option *returnValues* to `TRUE`. That may help to detect, not only visually, samples which do not fit into their respective clinical groups.

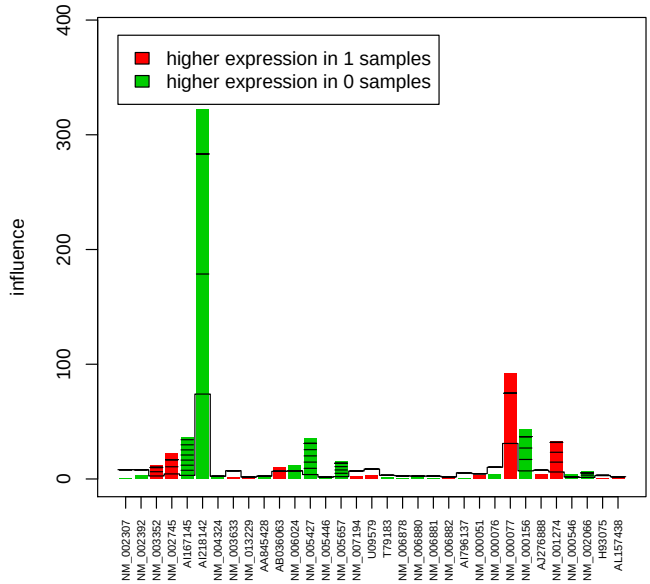


Abbildung A.3: Gene Plot for the van't Veer data with *globaltest*. Shown are the genes of the p53–signalling pathway. The bar height indicates the influence of the respective gene on the test statistic. The colour shows in which of the phenotype groups the gene has higher expression values. The reference line gives the expected height of the bar under the null hypothesis. Marks indicate with how many standard deviations the bar exceeds the reference line.

We compare again the different approaches (figures A.5 and A.6):

```
> Plot.subjects(xx = vantVeer, group = metastases, test.genes = p53,
+     legendpos = "bottomright")
> sampleplot(gt.vV)
```

The function `Plot.subjects` can be invoked by the three alternative function calls (see section *Global Testing of a Single Pathway*) and hence also plots corresponding to more complex testing challenges can be produced as well. To give just one example we consider again the influence of the tumour grade, which can take three possible values, on gene expression (figure A.7).

```
> Plot.subjects(xx = vantVeer, formula.full = ~grade, formula.red = ~1,
+     model.dat = phenodata, test.genes = p53, Colorgroup = "grade",
+     legendpos = "topleft")
```

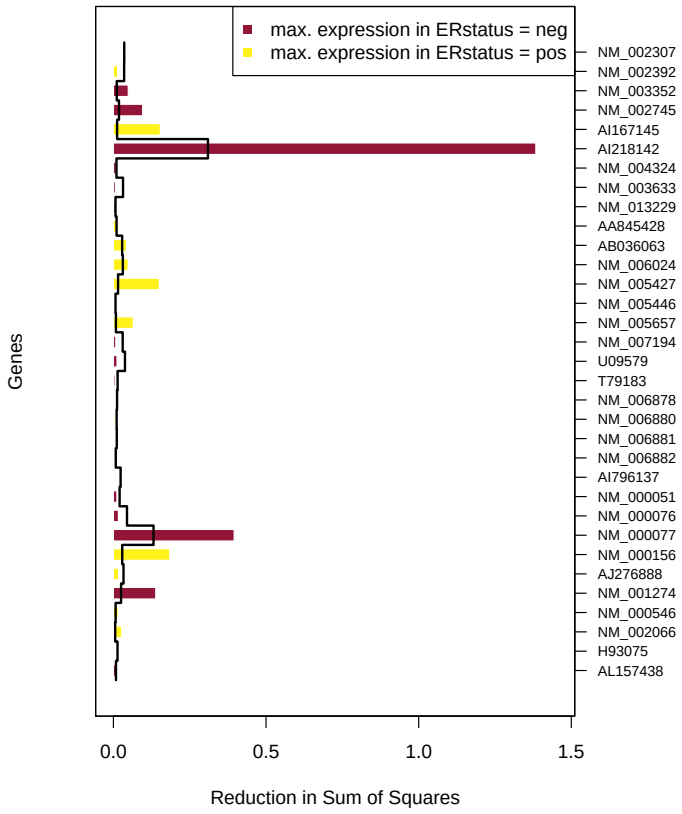


Abbildung A.4: Gene Plot for the van't Veer data with *GlobalAncova*. Shown are the genes of the p53–signalling pathway. The bar height indicates the influence of the respective gene on the test statistic. The color shows in which of the specified phenotype groups, in this case Estrogen receptor status, the gene has higher expression values. The reference line is the residual mean square error per gene.

Acknowledgements

We thank Sven Knüppel who took the initiative and contributed a C-code implementation of the permutation test to our package.
This work was supported by the NGFN project 01 GR 0459, BMBF, Germany.

References

1. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; 20: 93-99.

2. Goeman JJ, Mansmann U. Multiple testing on the directed acyclic graph of Gene Ontology. *Bioinformatics* 2008; 24(4): 537-44.

3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286 (5439): 531-537.

4. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist* 1979; 6: 65-70.

5. Hummel M, Meister R and Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 2008; 24 (1): 78-85.

6. Mansmann U, Meister R. Testing differential gene expression in functional groups. *Methods Inf Med* 2005; 44(3): 449-453.

7. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; 63 (3): 655-660.

8. Robbins H, Pitman EJG. Application of the Method of Mixtures to Quadratic Forms in Normal Variates. *The Annals of Mathematical Statistics* 1949; 20 (4): 552-560.

9. J. Schaefer, R. Opgen-Rhein, and K. Strimmer. corpcor: Efficient Estimation of Covariance and (Partial) Correlation, 2006. R package version 1.4.4.

10. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; 415: 530-536.

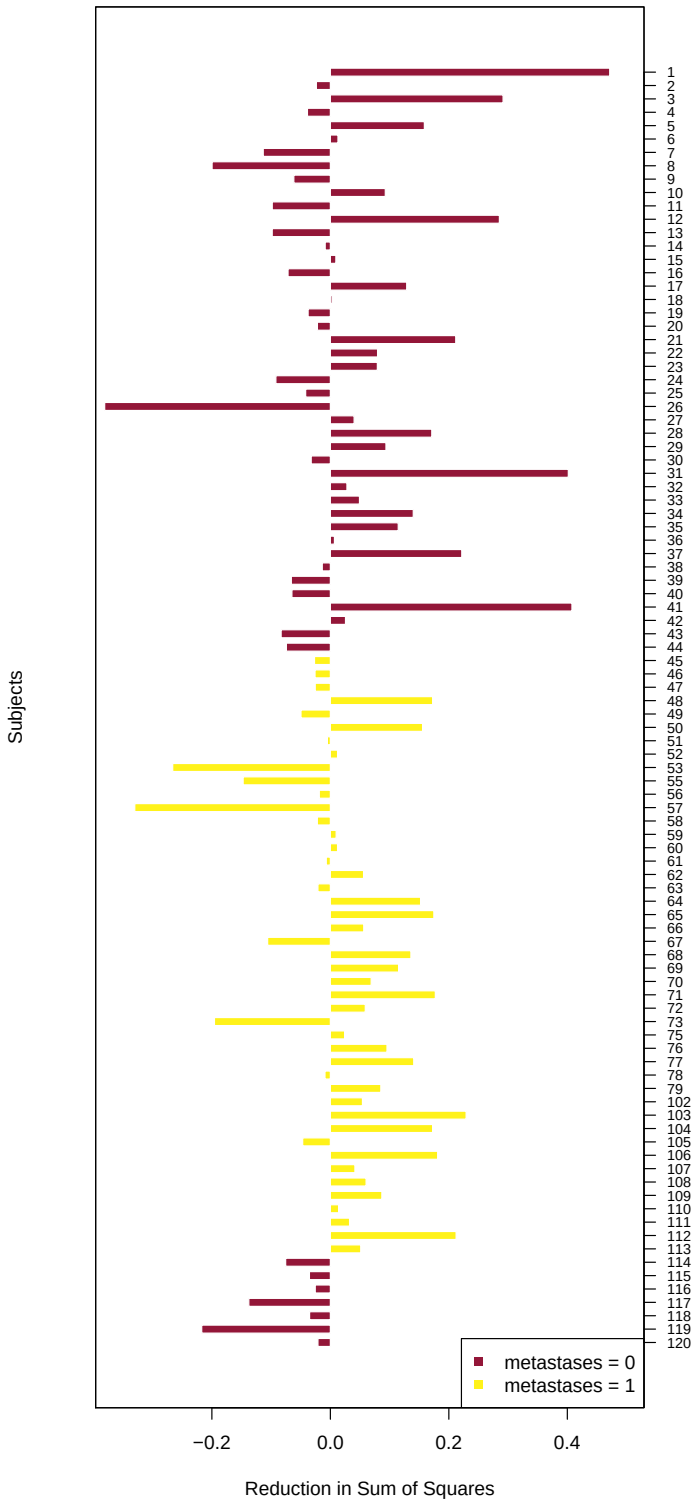


Abbildung A.5: Subjects Plot for the van't Veer data with *GlobalAncova*. The bar height indicates the influence of the respective sample on the test result. If an individual does not fit into the pattern of its phenotype, negative values can occur. Bars are colored corresponding to phenotype groups.

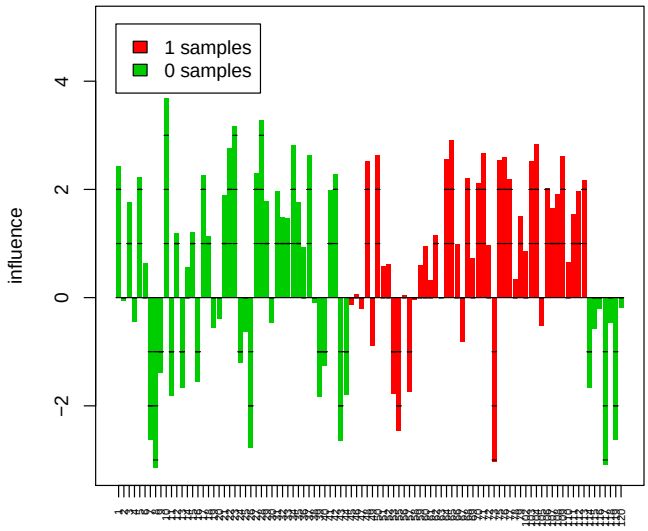


Abbildung A.6: Subjects Plot for the van't Veer data with *globaltest*. The bar height indicates the influence of the respective sample on the test result. If an individual does not fit into the pattern of its phenotype, negative values can occur. Bars are colored corresponding to groups. The reference line shows the expected influence of the samples under the null hypothesis. Marks on the bars indicate the standard deviation of the influence of the sample under the null hypothesis.

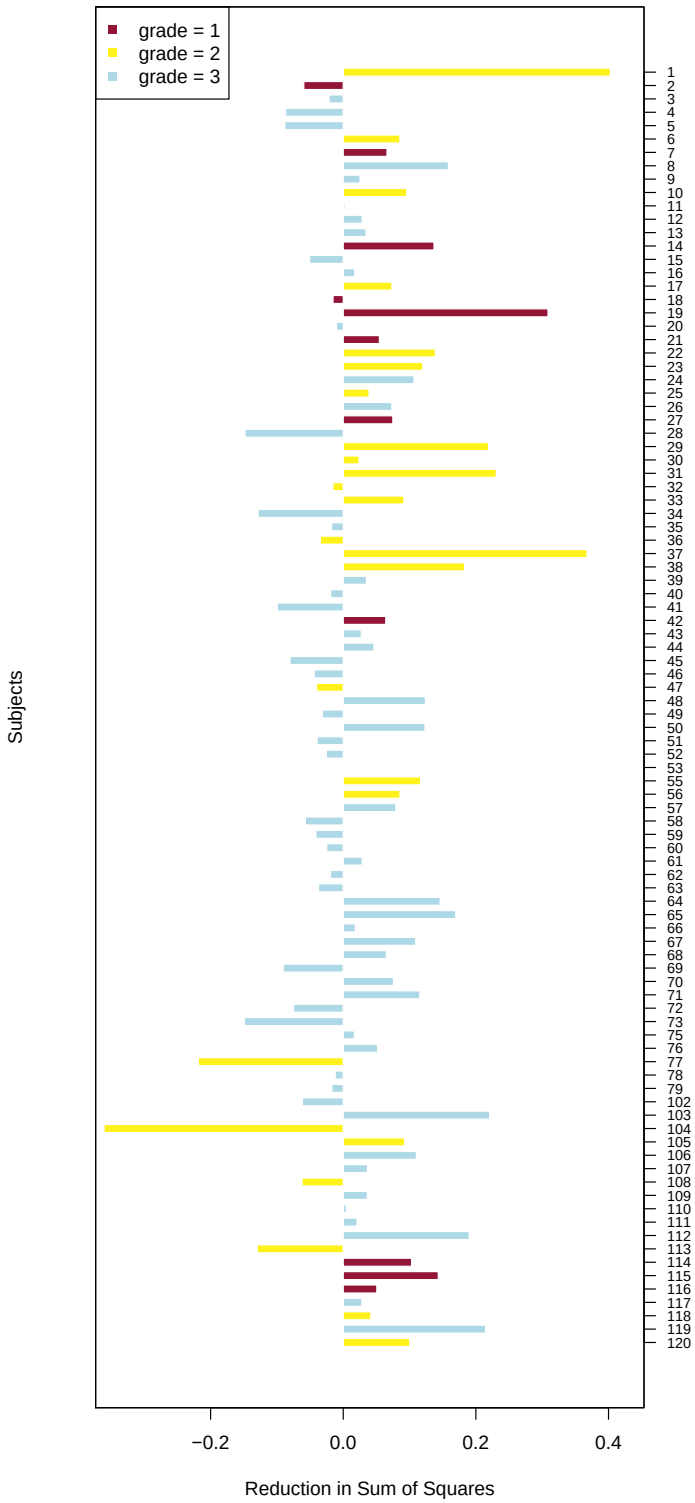


Abbildung A.7: Subjects Plot for the van't Veer data with *GlobalAncova*. Tumour grade is the clinical variable of interest. The bar height indicates the influence of the respective sample on the test result. If an individual does not fit into the pattern of its phenotype, negative values can occur. Bars are colored corresponding to phenotype groups.

Literaturverzeichnis

[Ackermann 2008] Ackermann, M.: *A comparison of statistical methods for gene set enrichment analysis*. 2008. – Diploma thesis, Department of Statistics, TU Dortmund

[Al-Shahrour u. a. 2004] Al-Shahrour, F. ; Diaz-Uriarte, R. ; Dopazo, J.: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. In: *Bioinformatics* 20 (4) (2004), S. 578–580

[Alexa und Rahnenführer 2007] Alexa, A. ; Rahnenführer, J.: *topGO: Enrichment analysis for Gene Ontology*, 2007. – R package version 1.3.0

[Alexa u. a. 2006] Alexa, A. ; Rahnenführer, J. ; Lengauer, T.: Improved significance scoring of functional groups from gene expression data by decorrelating GO graph structure. In: *Bioinformatics* 22 (13) (2006), S. 1600–1607

[Ashburner u. a. 2000] Ashburner, M. ; Ball, C.A. ; Blake, J.A. ; Botstein, D. ; Butler, H. ; Cherry, J.M. ; Davis, A.P. ; Dolinski, K. ; Dwight, S.S. ; Eppig, J.T. ; Harris, M.A. ; Hill, D.P. ; Issel-Tarver, L. ; Kasarskis, A. ; Lewis, S. ; Matese, J.C. ; Richardson, J.E. ; Ringwald, M. ; Rubin, G.M. ; Sherlock, G.: Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. In: *Nature Genetics* 25 (1) (2000), S. 25–29. – URL `http://www.geneontology.org/`

[Bair u. a. 2006] Bair, E. ; Hastie, T. ; Paul, D. ; Tibshirani, R.: Prediction by Supervised Principal Components. In: *Journal of the American Statistical Association* 101 (473) (2006), S. 119–137

[Bair und Tibshirani 2004] Bair, E. ; Tibshirani, R.: Semi-supervised methods to predict patient survival from gene expression data. In: *PLOS Biology* 2 (4) (2004), S. 511–522

[Bair und Tibshirani 2007] Bair, Eric ; Tibshirani, R.: *superpc: Supervised principal components*, 2007. – URL `http://www-stat.stanford.edu/~tibs/superpc`. – R package version 1.03

[Barry u. a. 2005] Barry, W.T. ; Nobel, A.B. ; Wright, F.A.: Significance analysis of functional categories in gene expression studies: a structured permutation approach. In: *Bioinformatics* 21 (9) (2005), S. 1943–9

[Beissbarth und Speed 2004] Beissbarth, T. ; Speed, T.P.: GOstat: find statistically overrepresented Gene Ontologies within a group of genes. In: *Bioinformatics* 20 (9) (2004), S. 1464–1465

[Benjamini und Hochberg 1995] Benjamini, Y. ; Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. In: *Journal of the Royal Statistical Society Series B* 57 (1) (1995), S. 289–300

[Benjamini und Yekutieli 2001] Benjamini, Y. ; Yekutieli, D.: The control of the false discovery rate in multiple hypothesis testing under dependency. In: *The Annals of Statistics* 29 (4) (2001), S. 1165–1188

[Campbell 2000] Campbell, N.A.: *Biologie*. Heidelberg, Berlin, Oxford : Spektrum Akademischer Verlag, 2000. – 324–352 S. – 2. korrigierter Nachdruck

[Dinu u. a. 2007] Dinu, I. ; Potter, J.D. ; Mueller, T. ; Liu, Q. ; Adewale, A.J. ; Jhangri, G.S. ; Einecke, G. ; Famulski, K.S. ; Halloran, P. ; Yasui, Y.: Improving gene set analysis of microarray data by SAM-GS. In: *BMC Bioinformatics* 8 (2007), S. 242

[Dolan u. a. 2005] Dolan, M.E. ; Ni, L. ; Camon, E. ; Blake, J.A.: A procedure for assessing GO annotation consistency. In: *Bioinformatics* 21 Suppl 1 (2005), S. i136–143

[Doniger u. a. 2003] Doniger, S.W. ; Salomonis, N. ; Dahlquist, K.D. ; Vranizan, K. ; Lawlor, S.C. ; Conklin, B.R.: MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. In: *Genome Biol* 4 (1) (2003), S. R7

[Draghici u. a. 2003] Draghici, S. ; Khatri, P. ; Martins, R.P. ; Ostermeier, G.C. ; Krawetz, S.A.: Global functional profiling of gene expression. In: *Genomics* 81 (2003), S. 98–104

[Draper und Smith 1998] Draper, N.R. ; Smith, H.: *Applied Regression Analysis*. New York : Wiley-Interscience, 1998. – 3. Auflage

[Dudoit u. a. 2006] Dudoit, S. ; Keles, S. ; Laan, M.J. van der: Multiple tests of association with biological annotation metadata. In: *UC Berkeley Division of Biostatistics Working Paper Series* Working Paper 202 (2006)

[Dudoit u. a. 2003] Dudoit, S. ; Popper Shaffer, J. ; Boldrick, J.C.: Multiple hypothesis testing in microarray experiments. In: *Statistical Science* 18 (1) (2003), S. 71–103

[Efron 2004] Efron, B.: Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. In: *Journal of the American Statistical Association* 99 (2004), S. 96–104

[Efron und Tibshirani 2007] Efron, B. ; Tibshirani, R.: On testing the significance of sets of genes. In: *Annals of Applied Statistics* 1 (1) (2007), S. 107–129

[Ewens und Grant 2005] Ewens, W.J. ; Grant, G.R.: *Statistical Methods in Bioinformatics. An Introduction.* New York : Springer, 2005. – 430–474 S

[Falcon und Gentleman 2007] Falcon, S. ; Gentleman, R.: Using GOstats to test gene lists for GO term association. In: *Bioinformatics* 23 (2) (2007), S. 257–258

[Gentleman u. a. 2005] Gentleman, R. ; Carey, V. ; Huber, W. ; Irizarry, R. ; Dudoit, S.: *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* New York : Springer, 2005

[Gentleman und Falcon 2007] Gentleman, R. ; Falcon, S.: *Category: Category Analysis*, 2007. – R package version 2.1.30

[Gentleman u. a. 2004] Gentleman, R.C. ; Carey, V.J. ; Bates, D.M. ; Bolstad, B. ; Dettling, M. ; Dudoit, S. u. a.: Bioconductor: open software development for computational biology and bioinformatics. In: *Genome Biology* 5 (2004), S. R80. – URL `http://www.bioconductor.org/`

[Goeman und Bühlmann 2007] Goeman, J.J. ; Bühlmann, P.: Analyzing gene expression data in terms of gene sets: methodological issues. In: *Bioinformatics* 23 (8) (2007), S. 980–987

[Goeman u. a. 2006] Goeman, J.J. ; Geer, S.A. van de ; Houwelingen, H.C. van: Testing against a high-dimensional alternative. In: *J R Statist Soc B* 68 (3) (2006), S. 477–493

[Goeman u. a. 2004] Goeman, J.J. ; Geer, S.A. van de ; Kort, F. de ; Houwelingen, H.C. van: A global test for groups of genes: testing association with a clinical outcome. In: *Bioinformatics* 20 (1) (2004), S. 93–99

[Goeman und Mansmann 2008] Goeman, J.J. ; Mansmann, U.: Multiple testing on the directed acyclic graph of Gene Ontology. In: *Bioinformatics* 24 (4) (2008)

[Goeman und Oosting 2007] Goeman, J.J. ; Oosting, J.: *Globaltest: testing association of a group of genes with a clinical variable*, 2007. – R package version 4.6.0

[Gold u. a. 2007] Gold, D.L. ; Coombes, K.R. ; Wang, J. ; Mallick, B.: Enrichment analysis in high-throughput genomics – accounting for dependency in the NULL. In: *Briefings in Bioinformatics* 8 (2) (2007), S. 71–77

[Grond-Ginsbach u. a. 2008] Grond-Ginsbach, C. ; Hummel, M. ; Wiest, T. ; Horstmann, S. ; Pfleger, K. ; Hergenhahn, M. ; Hollstein, M. ; Mansmann, U. ; Grau, A.J. ; Wagner, S.: Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. In: *Journal of Neurology* 255 (5) (2008), S. 723–31

[Grossmann u. a. 2007] Grossmann, S. ; Bauer, S. ; Robinson, P.N. ; Vingron, M.: Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. In: *Bioinformatics* 23 (22) (2007), S. 3024–3031

[Holm 1979] Holm, S.: A simple sequentially rejective multiple test procedure. In: *Scand. J. Statist.* 6 (1979), S. 65–70

[Huber u. a. 2002] Huber, W. ; Heydebreck, A. von ; Sultmann, H. ; Poustka, A. ; Vingron, M.: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. In: *Bioinformatics* 18 Suppl 1 (2002), S. 96–104

[Hummel u. a. 2008a] Hummel, M. ; Meister, R. ; Mansmann, U.: GlobalANCOVA: exploration and assessment of gene group effects. In: *Bioinformatics* 24 (1) (2008), S. 78–85

[Hummel u. a. 2008b] Hummel, M. ; Metzeler, K.H. ; Bohlander, S.K. ; Buske, C. ; Mansmann, U.: Association between a Prognostic Gene Signature and Functional Gene Sets. In: *Bioinformatics and Biology Insights* 2 (2008), S. 335–347

[Irizarry u. a. 2003a] Irizarry, R.A. ; Bolstad, B.M. ; Collin, F. ; Cope, L.M. ; Hobbs, B. ; Speed, T.P.: Summaries of Affymetrix GeneChip probe level data. In: *Nucleic Acids Res* 31 (4) (2003), S. e15

[Irizarry u. a. 2003b] Irizarry, R.A. ; Hobbs, B. ; Collin, F. ; Beazer-Barclay, Y.D. ; Antonellis, K.J. ; Scherf, U. ; Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. In: *Biostatistics* 4 (2) (2003), S. 249–264

[Jiang und Gentleman 2007] Jiang, Z. ; Gentleman, R.: Extensions to Gene Set Enrichment. In: *Bioinformatics* 23 (3) (2007), S. 306–313

[Kanehisa u. a. 2006] Kanehisa, M. ; Goto, S. ; Hattori, M. ; Aoki-Kinoshita, K.F. ; Itoh, M. ; Kawashima, S. ; Katayama, T. ; Araki, M. ; Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. In: *Nucleic Acids Res* 34 (2006), S. D354–357

[Khatri und Draghici 2005] Khatri, P. ; Draghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. In: *Bioinformatics* 21 (18) (2005), S. 3587–395

[Kong u. a. 2006] Kong, S.W. ; Pu, W.T. ; Park, P.J.: A multivariate approach for integrating genome-wide expression data and biological knowledge. In: *Bioinformatics* 22 (19) (2006), S. 2373–2380

[Kotz u. a. 1967] Kotz, S. ; Johnson, N.L. ; Boyd, D.W.: Series representations of distributions of quadratic forms in normal variables. I. Central case. In: *The Annals of Mathematical Statistics* 38 (3) (1967), S. 823–837

[Lamb u. a. 2003] Lamb, J. ; Ramaswamy, S. ; Ford, H.L. ; Contreras, B. ; Martinez, R.V. ; Kittrell, F.S. ; Zahnow, C.A. ; Patterson, N. ; Golub, T.R. ; Ewen, M.E.: A mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer. In: *Cell* 114 (2003), S. 323–334

[Ledoit und Wolf 2004] Ledoit, O. ; Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. In: *Journal of Multivariate Analysis* 88 (2004), S. 365–411

[Lewin und Grieve 2006] Lewin, A. ; Grieve, I.C.: Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. In: *BMC Bioinformatics* 7 (2006), S. 426

[Liu u. a. 2007] Liu, Q. ; Dinu, I. ; Adewale, A. ; Potter, J. ; Yasui, Y.: Comparative evaluation of gene-set analysis methods. In: *BMC Bioinformatics* 8 (1) (2007), S. 431

[Maglietta u. a. 2007] Maglietta, R. ; Piepoli, A. ; Catalano, D. ; Licciulli, F. ; Carella, M. ; Liuni, S. ; Pesole, G. ; Perri, F. ; Ancona, N.: Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data. In: *Bioinformatics* 23 (16) (2007), S. 2063–2072

[Manoli u. a. 2006] Manoli, T. ; Gretz, N. ; Groene, H.J. ; Kenzelmann, M. ; Eils, R. ; Brors, B.: Group testing for pathway analysis improves comparability of different microarray data sets. In: *Bioinformatics* 22 (20) (2006), S. 2500–2506

[Mansmann und Meister 2005] Mansmann, U. ; Meister, R.: Testing differential gene expression in functional groups. In: *Methods Inf Med* 44 (3) (2005), S. 449–453

[Marcucci u. a. 2005] Marcucci, G. ; Mrózek, K. ; Bloomfield, C.D.: Molecular heterogeneity and prognostic biomarkers in adults with acute myeloid leukemia and normal cytogenetics. In: *Curr Opin Hematol* 12 (1) (2005), S. 68–75

[Marcus u. a. 1976] Marcus, R. ; Peritz, E. ; Gabriel, K. R.: On closed testing procedures with special reference to ordered analysis of variance. In: *Biometrika* 63 (3) (1976), S. 655–660

[Martens u. a. 2006] Martens, J.H. ; Kzhyshkowska, J. ; Falkowski-Hansen, M. ; Schledzewski, K. ; Gratchev, A. ; Mansmann, U. ; Schmuttermaier, C. ; Dippel, E. ; Koenen, W. ; Riedel, F. ; Sankala, M. ; Tryggvason, K. ; Kobzik, L. ; Moldenhauer, G. ; Arnold, B. ; Goerdt, S.: Differential expression of a gene signature for scavenger/lectin receptors by endothelial cells and macrophages in human lymph node sinuses, the primary sites of regional metastasis. In: *J Pathol* 208 (4) (2006), S. 574–589

[Meinshausen 2008] Meinshausen, N.: Hierarchical testing of variable importance. In: *Biometrika* 95 (2) (2008), S. 265–278

[Metzeler u. a. 2008] Metzeler, K.H. ; Hummel, M. ; Bloomfield, C.D. ; Spiekermann, K. ; Braess, J. ; Sauerland, M.-C. ; Heinecke, A. ; Radmacher, M. ; Marcucci, G. ; Whitman, S.P. ; Maharry, K. ; Paschka, P. ; Larson, R.A. ; Berdel, W.E. ; Buechner, T. ; Woermann, B. ; Mansmann, U. ; Hiddemann, W. ; Bohlander, S.K. ; Buske, C.: An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. In: *Blood* 112 (10) (2008), S. 4193–201

[Mootha u. a. 2003] Mootha, V.K. ; Lindgren, C.M. ; Eriksson, K.F. ; Subramanian, A. ; Sihag, S. ; Lehar, J. ; Puigserver, P. ; Carlsson, E. ; Ridderstrale, M. ; Laurila, E. ; Houstis, N. ; Daly, M.J. ; Patterson, N. ; Mesirov, J.P. ; Golub, T.R. ; Tamayo, P. ; Spiegelman, B. ; Lander, E.S. ; Hirschhorn, J.N. ; Altshuler, D. ; Groop, L.C.: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. In: *Nature Genetics* 34 (3) (2003), S. 267–273

[Nam u. a. 2006] Nam, D. ; Kim, S.B. ; Kim, S.K. ; Yang, S. ; Kim, S.Y. ; Chu, I.S.: ADGO: analysis of differentially expressed gene sets using composite GO annotation. In: *Bioinformatics* 22 (18) (2006), S. 2249–53

[Opgen-Rhein u. a. 2006] Opgen-Rhein, R. ; Schäfer, J. ; Strimmer, K.: *GeneNet: Modeling and Inferring Gene Networks*, 2006. – URL `http://www.strimmerlab.org/software/genenet/`. – R package version 1.0.1

[R Development Core Team 2006] R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (Veranst.), 2006. – URL `http://www.R-project.org`. – ISBN 3-900051-07-0

[Radmacher u. a. 2006] Radmacher, M.D. ; Marcucci, G. ; Ruppert, A.S. ; Mrozek, K. ; Whitman, S.P. ; Vardiman, J.W. ; Paschka, P. ; Vukosavljevic, T. ; Baldus, C.D. ; Kolitz, J.E. ; Caligiuri, M.A. ; Larson, R.A. ; Bloomfield, C.D.: Independent confirmation of a prognostic gene-expression signature in adult acute myeloid leukemia with a normal karyotype: a Cancer and Leukemia Group B study. In: *Blood* 108 (5) (2006), S. 1677–1683

[Rivals u. a. 2007] Rivals, I. ; Personnaz, L. ; Taing, L. ; Potier, M.C.: Enrichment or depletion of a GO category within a class of genes: which test? In: *Bioinformatics* 23(4) (2007), S. 401–407

[Robbins und Pitman 1949] Robbins, H. ; Pitman, E.J.G.: Application of the Method of Mixtures to Quadratic Forms in Normal Variates. In: *The Annals of Mathematical Statistics* 20 (4) (1949), S. 552–560

[Rubin 1987] Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley and Sons, Inc, 1987

[Schäfer u. a. 2006] Schäfer, J. ; Opgen-Rhein, R. ; Strimmer, K.: *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*, 2006. – URL `http://www.strimmerlab.org/software/corpcor/`. – R package version 1.4.4

[Schäfer und Strimmer 2005a] Schäfer, J. ; Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks. In: *Bioinformatics* 21 (6) (2005a), S. 754–764

[Schäfer und Strimmer 2005b] Schäfer, J. ; Strimmer, K.: A shrinkage approach to large-scale covariance estimation and implications for functional genomics. In: *Statist Appl Genet Mol Biol* 4 (1) (2005b)

[Scholtens und von Heydebreck 2005] Scholtens, D. ; Heydebreck, A. von: *Analysis of differential gene expression studies*. S. 229–248. In: Gentleman, R. (Hrsg.) ; Carey, V. (Hrsg.) ; Dudoit, S. (Hrsg.) ; Irizarry, R. (Hrsg.) ; Huber, W. (Hrsg.): *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York : Springer, 2005

[Searle 1971] Searle, S.R.: *Linear Models*. Wiley, 1971

[Smyth 2005] Smyth, G.K.: *Limma: linear models for microarray data*. S. 397–420. In: Gentleman, R. (Hrsg.) ; Carey, V. (Hrsg.) ; Dudoit, S. (Hrsg.) ; Irizarry, R. (Hrsg.) ; Huber, W. (Hrsg.): *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York : Springer, 2005

[Song und Black 2007] Song, S. ; Black, M.: *pcot2: Principal Coordinates and Hotelling's T-Square method*, 2007. – R package version 1.6.0

[Stanley u. a. 2006] Stanley, S.M. ; Bailey, T.L. ; Mattick, J.S.: GONOME: measuring correlations between GO terms and genomic positions. In: *BMC Bioinformatics* 7 (2006), S. 94

[Steuer u. a. 2006] Steuer, R. ; Humburg, P. ; Selbig, J.: Validation and functional annotation of expression-based clusters based on gene ontology. In: *BMC Bioinformatics* 7 (2006), S. 380

[Subramanian u. a. 2005] Subramanian, A. ; Tamayo, P. ; Mootha, V.K. ; Mukherjee, S. ; Ebert, B.L. ; Gillette, M.A. ; Paulovich, A. ; Pomeroy, S.L. ; Golub, T.R. ; Lander, E.S. ; Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In: *PNAS* 102 (43) (2005), S. 15545–15550

[Tian u. a. 2005] Tian, L. ; Greenberg, S.A. ; Kong, S.W. ; Altschuler, J. ; Kohane, I.S. ; Park, P.J.: Discovering statistically significant pathways in expression profiling studies. In: *Proc Natl Acad Sci USA* 102 (38) (2005), S. 13544–9

[Tomfohr u. a. 2005] Tomfohr, J. ; Lu, J. ; Kepler, T.B.: Pathway level analysis of gene expression using singular value decomposition. In: *BMC Bioinformatics* 6 (2005), S. 225

[Tusher u. a. 2001] Tusher, V.G. ; Tibshirani, R. ; Chu, C.: Significance analysis of microarrays applied to the ionizing radiation response. In: *Proc Natl Acad Sci USA* 98 (2001), S. 5116–5121

[van't Veer u. a. 2002] Veer, L. J. van't ; Dai, H. ; Vijver, M.J. van de ; He, Y.D. ; Hart, A.A.M. ; Mao, M. ; Peterse, H.L. ; Kooy, K. van der ; Marton, M.J. ; Witteveen, A.T. ; Schreiber, G.J. ; Kerkhoven, R.M. ; Roberts, C. ; Linsley, P.S. ; Bernards, R. ; Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. In: *Nature* 415 (2002), S. 530–536

[Vêncio u. a. 2006] Vêncio, R.Z. ; Koide, T. ; Gomes, S.L. ; Pereira, C.A.: BayGO: Bayesian analysis of ontology term enrichment in microarray data. In: *BMC Bioinformatics* 7 (2006), S. 86

[Wang u. a. 2007] Wang, J.Z. ; Du, Z. ; Payattakool, R. ; Yu, P.S. ; Chen, C.F.: A new method to measure the semantic similarity of GO terms. In: *Bioinformatics* 23 (10) (2007), S. 1274–1281

[Westfall und Young 1993] Westfall, P.H. ; Young, S.S.: *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York : Wiley, 1993

[Wu u. a. 2004] Wu, Z. ; Irizarry, R. ; Gentleman, R. ; Martinez Murillo, F. ; Spencer, F.: A model based background adjustment for oligonucleotide expression arrays. In: *Journal of the American Statistical Association* 99 (468) (2004), S. 909–917

[Wyatt und Altman 1995] Wyatt, J.C. ; Altman, D.G.: Commentary: Prognostic models: clinically useful or quickly forgotten? In: *BMJ* 311 (1995), S. 1539–1541

[Xiang u. a. 2007] Xiang, W. ; Hummel, M. ; Mitteregger, G. ; Pace, C. ; Windl, O. ; Mansmann, U. ; Kretzschmar, H.A.: Transcriptome analysis reveals altered cholesterol metabolism during the neurodegeneration in mouse scrapie model. In: *Journal of Neurochemistry* 102 (3) (2007), S. 834–847

[Ye u. a. 2006] Ye, J. ; Fang, L. ; Zheng, H. ; Zhang, Y. ; Chen, J. ; Zhang, Z. ; Wang, J. ; Li, S. ; Li, R. ; Bolund, L. ; Wang, J.: WEGO: a web tool for plotting GO annotations. In: *Nucleic Acids Res* 34 (2006), S. W293–297

[Yekutieli 2008] Yekutieli, D.: Hierarchical False Discovery Rate-controlling methodology. In: *Journal of the American Statistical Association* 103 (481) (2008), S. 309–316(8)

[Yu u. a. 2007] Yu, J.X. ; Sieuwerts, A.M. ; Zhang, Y. ; Martens, J.W. ; Smid, M. ; Klijn, J.G. ; Wang, Y. ; Foekens, J.A.: Pathway analysis of gene signatures

predicting metastasis of node-negative primary breast cancer. In: *BMC Cancer* 7 (1) (2007), S. 182

[Zeeberg u. a. 2003] Zeeberg, B.R. ; Feng, W. ; Wang, G. ; Wang, M.D. ; Fojo, A.T. ; Sunshine, M. ; Narasimhan, S. ; Kane, D.W. ; Reinhold, W.C. ; Lababidi, S. ; Bussey, K.J. ; Riss, J. ; Barrett, J.C. ; Weinstein, J.N.: GoMiner: a resource for biological interpretation of genomic and proteomic data. In: *Genome Biol* 4 (4) (2003), S. R28

[Zhong u. a. 2004a] Zhong, S. ; Storch, K.F. ; Lipan, O. ; Kao, M.C. ; Weitz, C.J. ; Wong, W.H.: GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. In: *Appl Bioinformatics* 3 (4) (2004), S. 261–264

[Zhong u. a. 2004b] Zhong, S. ; Tian, L. ; Li, C. ; Storch, K.F. ; Wong, W.H.: Comparative analysis of gene sets in the Gene Ontology space under the multiple hypothesis testing framework. In: *Proc IEEE Comput Syst Bioinform Conf* (2004), S. 425–435

Danksagung

Diese Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der Ludwig-Maximilians-Universität München. Sie wurde durch das Nationale Genomforschungsnetz (NGFN) gefördert (Fördernummer 01 GR 0459).

Mein Dank gilt in erster Linie meinem Doktorvater Prof. Dr. Ulrich Mansmann für die kompetente und verständnisvolle Betreuung. Die Zusammenarbeit war äußerst lehrreich und hat maßgeblich zum Gelingen dieser Arbeit beigetragen. Es war mir eine Freude, am IBE an verschiedensten Projekten zu aktuellen molekularbiologischen und medizinischen Fragestellungen mitzuwirken.

In das Themengebiet Microarrays und Bioinformatik habe ich mich dank der Unterstützung durch das „NGFN Team“, insbesondere Jörg Rahnenführer, Adrian Alexa, Markus Ruschhaupt und Achim Tresch bestens eingefunden. Ebenso gebührt Jelle Goeman an dieser Stelle Dank für seine Anregungen. Für die freundliche und konstruktive Zusammenarbeit am „Projekt GlobalAncova“ danke ich Prof. Dr. Reihnhard Meister, Ramona Scheufele und Sven Knüppel. Desweiteren bin ich unseren Hilfswissenschaftlerinnen Renate Effner, Vindi Jurinović und Esmeralda Vicedo, die mir viel Arbeit abgenommen haben, zu Dank verpflichtet. Sehr gefreut habe ich mich auch über die erfolgreichen Kooperationen mit Kollegen aus den Bereichen der Medizin und Biologie. Hervorheben möchte ich hier Klaus Metzeler, Stefan Bohlander, Wei Xiang und Caspar Grond-Ginsbach. Ich bedanke mich ebenfalls bei allen Kollegen am IBE für die entspannte und freundschaftliche Atmosphäre und vor allem bei Markus Schmidberger für seine Vorschläge hinsichtlich meiner Arbeit.

Ganz besonders will ich auch meinen Freunden danken, die mich stets motiviert haben, allen voran Klaus Hechenbichler, Ursula Gerhardinger, Christiane Belitz, Ingrid Kreuzmair, Florian Leitenstorfer, Anna Töller und Alexander Koschke. Sehr hilfreich war der Erfahrungsaustausch mit den ebenfalls Promovierten bzw. Promovierenden unter ihnen. Ebenso dankbar bin ich meiner ganzen Familie, die mich immer unterstützt und scheinbar nie an mir zweifelt. Zuguterletzt danke ich Florian Stigloher, der auch diesen Lebensabschnitt mit mir geteilt hat.

Lebenslauf

Manuela Benita Hummel

11. April 1980	Geburt in Aschaffenburg
1986 - 1990	Besuch der Grundschulen Obernburg und Weilbach
1990 - 1999	Besuch der Gymnasien Amorbach und Bad Aibling
25. Juni 1999	Allgemeine Hochschulreife (Note 1,3)
1999 - 2005	Diplomstudium Statistik mit Anwendungsgebiet Biologie (Schwerpunkt Anthropologie und Humangenetik) an der Ludwig-Maximilians-Universität München
9. Mai 2005	Diplom (Note „ausgezeichnet“ (1,22))
01.04.2005 - 31.05.2008	Wissenschaftliche Assistentin am Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der Ludwig-Maximilians-Universität München
Seit 01. Juli 2008	Bioinformatikerin am Centre for Genomic Regulation (CRG) in Barcelona, Spanien