Aus der Klinik mit Schwerpunkt Nephrologie und internistische Intensivmedizin der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Entwicklung und Vergleich biostatistischer Methoden zur Auswertung von Microarray Experimenten

zur Erlangung des akademischen Grades
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Djork-Arné Clevert

aus Berlin

Gutachter: 1. Prof. Dr. med. P. Reinke

2. Prof. Dr. rer. nat. P. Nürnberg

3. Prof. Dr. med. N. Hübner

Datum der Promotion: 7. September 2012

Inhaltsverzeichnis

Kurzzusammenfassung

Motivation: Kosteneffiziente Microarrays, wie der *Affymetrix SNP 6.0* und der *Human Gene 1.1 ST*, sind noch immer die vorherrschende Hochdurchsatztechnologie, um DNA-*Copy Number Variations* (CNVs), bzw. die Aktivität der Genexpression zu untersuchen. Microarraymessungen werden durch eine Vielzahl von Störgrößen (Kreuzhybridisierung, RNA/DNA-Degradation, Effizienz der Transkription, Hybridisierungstemperatur, usw.) beeinflusst und weisen daher ein hohes Messrauschen auf. Diese zufällige Variabilität der Microarraydaten ist in zweifacher Hinsicht problematisch, erstens können schwache Signale im Messrauschen nicht mehr detektiert werden und zweitens können Fluktuationen im Messrauschen zu Scheinkorrelationen mit dem beobachteten Phänotyp führen. Für die quantitative Analyse der Genexpression, bzw. von CNVs bedeutet dies unter anderem, dass die Aktivität der Genexpression bzw. die Kopienzahl einer CNV-Region fehlerhaft geschätzt wird. Diese fehlerhaften Ergebnisse werden jedoch beim Korrigieren für multiples Testen (z.B. Bonferoni-Korrektur) berücksichtigt, und schwächen somit die Power einer Studie. Die Essenz von fünf wissenschaftlichen Artikeln, über die eingangs erwähnte Problematik bei der Analyse von Microarraydaten, ist in dieser Publikationspromotion zusammengetragen.

Methode: Im Rahmen dieser Arbeit wurden vier neue Methoden zur Analyse von Microarraydaten entwickelt und validiert. In (1) „*Factor Analysis for Robust Microarray Summarization*“ (FARMS) wurde ein multivariates Maximum-a-posteriori Faktorenanalysemodell entwickelt, welches die quantitative Genexpressionsanalyse verbessert. (2) „*Informative/Non-Informative-calls*“ (I/NI-calls) beschreibt ein informationstheoretisches Filterverfahren, welches es ermöglicht, die für ein Experiment relevanten Gene zu identifizieren. In (3) Copy-number FARMS (cn.FARMS) wurden Methoden zur Normalisierung, Summarization und Segmentierung von SNP-Arraydaten entwickelt und mit bestehenden Methoden verglichen. Während in (4) „*Factor Analysis for Bicluster Acquisition*“ (FABIA) eine neue Methode zum Biclustern von Genexpressionsdaten entwickelt wurde.

Ergebnisse: Der FARMS-Algorithmus wurde hinsichtlich seiner Sensitivität und Spezifität rigoros ausgetestet und ist seit 2006 die führende Summarizationmethode im internationalen Affycomp Benchmark. Das Filterkriterium I/NI-calls wurde auf mehr als 30 Microarrayexperimenten evaluiert und konnte durchschnittlich 84% (±1,5%), bzw. in Spike-in Experimenten sogar über 99,5%, der irrelevanten Gene herausfiltern, ohne dabei ein relevantes Signal zu verlieren. Für die genomweite Genexpressionsanalyse kann durch den I/NI-call die Anzahl der Nullhypothesen von ca. 28.000 auf ca. 4.500 reduziert werden und führt nach Korrektur für multiples Testen zu ca. 6-fach kleineren p-Werten. cn.FARMS wurde auf HapMap-Daten mit den gängigsten CNV-Analysemethoden verglichen und konnte den Type-1-Fehler um ca. 20%, bzw. bei geringerer Auflösung um ca. 60% reduzieren.

Konklusion: Die Ergebnisse zeigen, dass Methoden des maschinellen Lernens zur Auswertung von Microarraydaten sehr gut geeignet sind, um die quantitative Genexpressionsanalyse zu verbessern, um die statistische Power einer Studie zu erhöhen, um CNV-Variationen zuverlässig zu entdecken und um biologisch plausible Bicluster zu identifizieren.

Abstract

Motivation: Cost-effective oligonucleotide arrays like the Affymetrix SNP 6.0 and the Human Gene 1.1 ST are still the predominant technique to measure DNA copy number variations (CNVs) and gene expression, respectively. However, microarray data are characterized by high levels of noise induced by DNA preparation, staining, hybridization or measurement processes. This obscuring variation can blur out the signal of interest and, even worse, lead to spurious correlations which misguide the researcher. Thus, methods for detecting CNVs overestimate both the number and the size of CNV regions, while methods for gene expression summarization are imprecise regarding the quantitative analysis. As a consequence suffer both techniques from a high false discovery rate (FDR), meaning that many findings are wrongly detected and therefore not associated with the tested condition. These false positives decrease furthermore the study's discovery power due to the correction for multiple testing. The core of this thesis consist of five peer-reviewed scientific publications which treat the before mentioned issues.

Methods: Four probabilistic latent variable models for processing of -omnic data were developed to tackle the problem of false discoveries and accuracy of the estimates. (1) Factor Analysis for Robust Microarray Summarization (FARMS) was developed to provide more accurate gene expression estimates and is based on a Bayesian maximum a posteriori factor analysis model. (2) The FARMS algorithm provides further the Informative/Non-Informative (I/NI) call, which is an unsupervised filtering technique which allows the researcher to identify those genes that are informative for the interpretation of the experiment. (3) Copy-number FARMS (cn.FARMS) was purposed to correct for allele specific cross-hybridization in genotyping data and to estimate DNA copy numbers from genotyping array data. Whereas (4) „Factor Analysis for Bicluster Acquisition“ (FABIA) was developed for biclustering of -omnics data.

Results: FARMS has been rigorous evaluated on all public available spike-in data sets and at the international Affycomp benchmark, where it outperformed all preexisting summarization methods both with respect to sensitivity and specificity. Furthermore, I/NI calls excluded the non-informative probe sets without loss of sensitivity and specificity. The exclusion rates were in average 84% (±1.5%) on 30 real world data sets and on spiked-in data set even up to 99.5% while never losing a spiked-in gene.
On HapMap data, cn.FARMS clearly outperformed the two methods which performed best in other comparative studies on copy number estimation. For single-locus and for 4-loci estimates on SNP 6.0 arrays, cn.FARMS had about 20% less false positives (56,145 FP) than the second best method (68,593 FP) and about 3.5 times less false positives (366 FP) than next best method (1338 FP), respectively.

Conclusion: The results show, that FARMS-based array preprocessing methods for gene expression analysis as well as for CNV-detection outperformed its competitors both with respect to FDR and sensitivity. They further provide a both statistical sound and objective feature reduction criterion that offers a critical solution to the curse of high-dimensionality in the analysis of microarray data.

Zusammenfassung

1. Einleitung

In den letzten 15 Jahren sind Microarrays zu einem der wichtigsten diagnostischen Werkzeuge der molekularbiologischen Forschung herangereift. Sie werden routiniert eingesetzt, um beispielsweise die Genexpression einer Zelle, die epigenetische Regulation der Genexpression durch DNA-Methylierung, die Nukleotidvarianten von Punktmutationen (Single Nucleotide Polymorphismen - SNPs) oder die DNA-Kopienzahl eines chromosomalen Segmentes zu bestimmen.
Ein Affymetrix Microarray besteht aus einem ca. 1 cm^2 großem Glaskörper, auf dem durch photolithographische Herstellung DNA-Oligos in situ synthetisiert sind. Je nach Anwendungszweck unterscheidet sich das Arraydesign. So wird für die Genexpressionsanalyse jedes mRNA-Transskript durch 11-60 unterschiedliche Oligos gemessen, während die Genotypisierung der Allelevariante durch drei gleiche Oligios erfolgt. Unabhängig von dem Design des Arrays führen eine Vielzahl von Störgrößen wie Kreuzhybridisierung, RNA/DNA-Degradation, Effizienz der Transkription oder die Hybridisierungstemperatur zu hohem Messrauschen. Diese zufällige Variabilität der Microarraydaten ist in zweifacher Hinsicht problematisch, erstens können schwache Signale im Messrauschen nicht mehr erkannt werden und zweitens können Fluktuationen im Messrauschen zu Scheinkorrelationen mit dem beobachteten Phänotyp führen. Für die quantitative Analyse der Genexpression, bzw. der CNV-Analyse bedeutet dies unter anderem, dass die Genexpressionsstärke bzw. die Kopienzahl einer CNV-Region fehlerhaft geschätzt wird. Diese fehlerhaften Ergebnisse werden jedoch bei der p-Wertkorrektur für multiples Testen (z.B. Bonferoni-Korrektur) berücksichtigt und schwächen infolgedessen die Power der Studie.

2. Zielsetzung

Die vorliegende Publikations-Dissertation thematisiert die Entwicklung robuster biostatistischer Methoden zur Analyse von Microarraydaten und deren Vergleich mit bestehenden Verfahren. Die zusammengefassten Arbeiten umfassen vier neue Analysemethoden, um

- die Genexpression bei hoher Sensitivität und Spezifität genauer quantitativ bestimmen zu können.
- die Type-1-Fehlerrate bei der Detektierung differentiell exprimierter Gene in Microarrayexperimenten zu reduzieren.
- numerische Chromosomenaberrationen in SNP-Arraydaten mit geringer False-Discovery-Rate (FDR) zu detektieren.
- biologisch plausible Bicluster (wie z.B. transkriptionelle Module) in genomischen Daten zu identifizieren.

Alle Methoden wurden als frei nutzbare *open source* Programme in der Statistiksoftware **R** entwickelt und sind als Pakete über das Softwareprojekt *Bioconductor* frei verfügbar.

3. Methodik

Die in dieser Arbeit entwickelten Methoden basieren auf dem linearen Faktorenanalysemodell, einem generativen, latenten Variablenmodell der multivariaten Statistik. Die Faktorenanalyse ist ein dimensionsreduzierendes Verfahren das 1904 von Spearman zur Erklärung der Ergebnisse von Intelligenztests entwickelt wurde. Ziel seiner Überlegung war es, die linearen Abhängigkeiten (Korrelationen) zwischen einer großen, unübersichtlichen Menge von beobachteten Variablen durch eine kleinere Menge aussagekräftiger latenter Variablen zu beschreiben. Mathematisch lässt sich das Faktorenanalysemodell für p Faktoren formulieren als

$$X \;=\; \sum_{i=1}^{p} \lambda_i \; z_i^T \;+\; \Upsilon \;=\; \Lambda \; Z \;+\; \Upsilon \;,$$

wobei $\boldsymbol{\Upsilon} \in \mathbb{R}^{n \times l}$ den additiven Fehleranteil darstellt; $\boldsymbol{\lambda}_i \in \mathbb{R}^n$ der Faktorladungsvektor und $\boldsymbol{z}_i \in \mathbb{R}^l$ der Faktorvektor des i-ten Faktors sind. Die Kernidee der Faktorenanalyse besteht darin, statistisch unabhängige Faktoren zu finden, welche die multidimensionale Korrelationsstruktur der beobachteten Messdaten erklären können. Dazu werden die beobachteten Messdaten, wie in Abbildung 1 gezeigt, durch das dyadische Produkt der inhärenten, nicht direkt messbaren Variablen z und der Faktorenla-

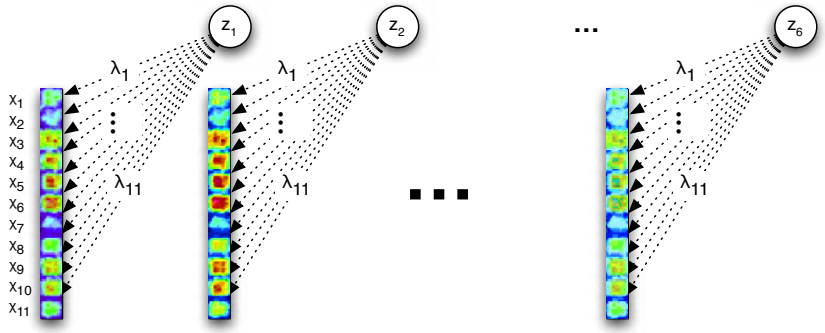

Abbildung 1. Pfaddiagramm des FARMS-Modells ohne Rauschanteil. Im Fall der Genexpressionsanalyse modelliert die Faktorladung λ probespezifische Charakteristika wie z.B. Kreuzhybridisierung, während die latente Variable z die vorliegende mRNA-Konzentration des betrachteten Genes beschreibt.

dung $\boldsymbol{\lambda}$ generiert. Da die latente Variable ein unmessbarer Modellparameter ist, wird auf Basis der beobachteten Daten auf sie geschlossen. Hierzu wird die Kovarianzsstruktur der gemessenen Daten analysiert und in ihre Signal- und Fehlervarianz dekomponiert. Im konventionellen Faktorenanalysemodell wird die Zerlegung durch ein Maximum-Likelihood-Ansatz optimiert.

In den folgenden Arbeiten floss zusätzliches Domänen-Wissen über die zu untersuchenden Daten in Form von a priori Wahrscheinlichkeiten ein. Nach den Regeln der Bayesschen Statistik kann mit diesem zusätzlichen Wissen die Zerlegung der Kovarianzstruktur über einen Maximum-a-posteriori-Lösungsansatz optimiert werden. In den folgenden Abschnitten werden die Unterschiede zwischen den entwickelten Modellen und deren mathematische Formulierung beschrieben. Des Weiteren wird auf

die unterschiedlichen Priorannahmen, deren biologische, bzw. experimentelle Bedeutung und deren Einfluss auf die Modellierung eingegangen.

3.1 FARMS - Genexpressionsanalyse von Affymetrix Microarrays

Kosteneffiziente Microarrays, wie der *Human Gene 1.1 ST*, werden zur Hochdurchsatzbestimmung relativer Änderungen in der Genexpression eingesetzt und ermöglichen z.B. die simultane Analyse des gesamten Transkriptoms einer Zelle. Zur Gewinnung dieser Information, wird zunächst die RNA der eukaryotischen Zelle extrahiert, durch reverse Transkriptasen zu Doppelstrang cDNA katalysiert und dann in vitro zu cRNA transkribiert. Nach einer Reinigungsprozedur wird die fragmentierte cRNA dem Array zugeführt, wo sie sich kovalent an den fixierten Oligos des Arrays bindet. Ist für ein Gen mehr cRNA vorhanden, so hybridisiert auch eine größere Menge cRNA an den korrespondierenden Oligos. Abschließend wird der Array von ungebundener Rest-cRNA gereinigt, mit Fluoreszenzfarbe koloriert,

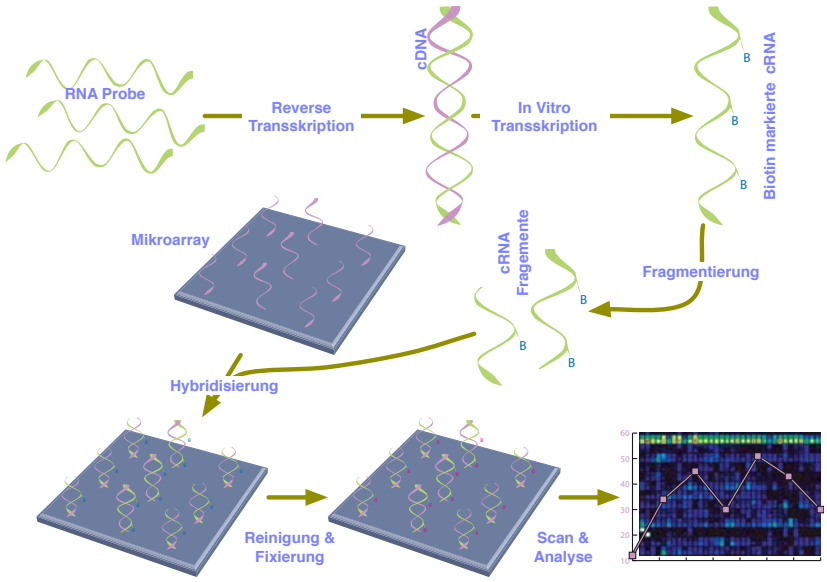

Abbildung 2. Ablaufdiagramm der Genexpressionsanalyse für Affymetrix GeneChip Arrays.

erneut gereinigt und mittels eines Laserscanners ausgelesen. Aufgrund unspezifischer Hybridisierung an Affymetrixs 25-mer Oligos wird jedes RNA-Transskript durch Probeset von 11-60 nicht überlappender Oligos repräsentiert. Nach dem optischen Auslesen werden die unterschiedlichen Intensitätswerte des Probesets zu einem der RNA-Konzentration proportionalen Genexpressionssignal kombiniert. Anfangs wurde die Genexpression über das arithmetische Mittel des Probesets [1], später durch einen robusten Mittelwert [2] und schließlich durch modellbasierte Ansätze bestimmt [3-4].
Das Ziel dieser Publikation war es, eine neue Methode zur quantitativen Bestimmung der Genexpression zu entwickeln. Zu diesem Zweck wurde der FARMS-Algorithmus (**F**actor **A**nalysis for **R**obust **M**icroarray **S**ummarization) [4] entwickelt. Die mathematische Formulierung des FARMS-Algorithmus lautet

$$\boldsymbol{X} = \boldsymbol{\lambda}\, \boldsymbol{z} + \boldsymbol{\epsilon}\,,$$

wobei $\boldsymbol{\epsilon}$ die additiven $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$-verteilten residuellen Varianzen des Probesets darstellt; $\boldsymbol{\lambda}$ der Faktorladungsvektor und $\boldsymbol{z}$ der extrahierte Faktor ist. Im Gegensatz zum klassischen Faktorenanalysemodell extrahiert der FARMS-Algorithmus nur einen Faktor und ist daher rotationsinvariant bezüglich seiner Lösung.

Als Priorverteilung wurde der positive Teil einer bei Null abgeschnittenen Gaussverteilung angenommen, welcher Werte um Null bevorzugt und negative Werte ausschließt. Die Wahl des Prior berücksichtig die Tatsache, dass

- negative Korrelationen zwischen Proben, die das gleiche Transskript messen, biologisch nicht plausibel sind.
- die beobachtete Datenvarianz oftmals klein ist. Was dazu führt, dass hohe Werte von $\boldsymbol{\lambda}$ unwahrscheinlich sind.
- ein Array typischerweise viel mehr irrelevante Gene (konstantes Signal und $\boldsymbol{\lambda}$ näherungsweise Null) umfasst, als relevante Gene (variables Signal und hohe Werte von $\boldsymbol{\lambda}$).

FARMS wurde in der in der Statistiksoftware *R* entwickelt und ist als R-Paket über *Bioconductor* unter http://www.bioconductor.org/packages/release/bioc/html/farms.html verfügbar.

3.2 I/NI-call - ein unüberwachtes Genfilterkriterium

Basierend auf dem FARMS-Algorithmus wurde in [5] das informationstheoretische Filterverfahren „**I**nformation/**N**on-**I**nformation-call“ (I/NI-call) entwickelt, um potentielle Kandidatengene anhand ihrer Signalkorrelation zu identifizieren. Der I/NI-call beruht auf dem Informationsgewinn der a-posteriori-Wahrscheinlichkeitsverteilung im Vergleich zur a priori Wahrscheinlichkeitsverteilung. Wobei a priori die Nullhypothese, ein konstantes Signal mit Faktorenladung $\boldsymbol{\lambda}$ näherungsweise um Null, am wahrscheinlichsten ist. Nur durch ein starkes konsistentes Signal in den Daten kann die latente Variable von der Nullhypothese abweichen. Die mathematische Formulierung für den I/NI-call lautet

$$\mathrm{I/NI} \;=\; -log\left(\left(\mathbf{1} + \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}\right)^{-1}\right),$$

wobei $\boldsymbol{\Psi}$ die Residual- und $\boldsymbol{\lambda}$ die Signalvarianz ist. Aus der Formel ist erkenntlich, dass Probesets mit hoher Residualvarianz und geringer Signalvarianz einen kleinen I/NI-call aufweisen und somit für die biologische Interpretation des Experimentes irrelevant sind. Standardmäßig wird ein Probeset (Gen) als informativ erachtet, wenn dessen I/NI-call gößer als -log(0,5) ist. Gene, die diesen Filter passieren, erzielen in einem späteren statistischen Test höhere p-Werte, da nach dem Filtervorgang die Anzahl der zu testenden Nullhypothesen geringer ist und die p-Wert-Korrektur für multiples Testen somit weniger Gene betrifft.

Der I/NI-Filter ist Bestandteil des FARMS-Paket und kann von der *Bioconductor* Webseite unter http://www.bioconductor.org/packages/release/bioc/html/farms.html heruntergeladen werden.

3.3 cn.FARMS - CNV-Analyse von SNP-Microarrays

Kosteneffiziente Microarrays wie der *Affymetrix SNP 6.0* oder der *Cytogenetics Whole-Genome 2.7M* Array sind noch immer die vorherrschende Hochdurchsatztechnologie, um DNA-*Copy Number Variations* (CNVs) zu untersuchen. Die hohe Anzahl der Proben und deren gleichförmige Verteilung über das Genom ermöglichen es, numerische Chromosomenaberrationen kleiner als 735 Basen, in Krebsgenen sogar kleiner als 280 Basen, zu identifizieren. Das Prinzip beruht auf der quantitativen Analyse der Probesignale, die über ihre Signalstärke Rückschlüsse auf die DNA-Kopienzahl des chromosomalen Segmentes zulassen.
Obwohl SNP-Arrays schon seit einigen Jahren zur CNV-Analyse eingesetzt werden, besteht trotzdem die Notwendigkeit bestehende Methoden zu verbessern. So schreibt Baross et al. [6]: „*The frequency of false positive deletions was substantial with different methods like dChip and CNAG.*“ In Clevert et al. [7] sollte daher geklärt werden, inwieweit neue statistische Methoden aus dem Bereich des maschinellen Lernens und der Bayesschen Statistik geeignet sind, die Anzahl der DNA-Kopien zu schätzen und die False-Discovery-Rate (FDR) bei der Analyse von numerischen Chromosomenaberrationen zu reduzieren. Ferner sollte belegt werden, dass die entwickelten Methoden zu reproduzierbaren und validen Ergebnissen führen.
Vor der **c**opy-**n**umber.FARMS (cn.FARMS) Publikation wurden CNV-Analysen auf Affymetrix SNP-Arrays mit *CNAG* [8], *CNAT* [9], *dChip* [10-11] oder *aroma.affymetrix* [12-13] durchgeführt. Die letzten beiden Verfahren werden am häufigsten verwendet und sind, wie cn.FARMS, ursprünglich Methoden der quantitativen Genexpressionsanalyse. Der *dChip*-Algorithmus bestimmt die DNA-Kopie durch ein lineares Modell pro SNP-Locus. Da die Modellparameter über die kleinste Fehlerquadrat-Schätzung optimiert werden, setzt der *dChip*-Algorithmus einen Gauß-verteilten Fehler voraus. Ferner liefert der Algorithmus keinen p-Wert zur Bestimmung der Modellgüte. *aroma.affymetrix* baut auf dem *RMA*-Algorithmus (*Robust Mircoarray Summarization*), einem linearen additiven Modell das durch Median Politur optimiert wird, auf. Aufgrund der sehr guten Performanz des FARMS-Algorithmus lag es nahe, das Modell so zu erweitern, dass anstelle der mRNA-Kopienzahl pro Transskript nun die DNA-Kopienzahl pro SNP-Locus berechnet wird.

Das cn.FARMS-Modell basiert ebenfalls auf dem Faktorenanalysemodell, wobei hier die latente Variable die Kopienzahl des DNA-Segments beschreibt. cn.FARMS hat gegenüber anderen Verfahren klare Vorteile, da

- experimentell verifizierte Fehlerannahmen modelliert werden können.
- zusätzliche Nebenbedingungen, wie z.B. eine positive Korrelation der SNP-Probes oder deren Signalstärke, berücksichtigt werden können.
- benachbarte SNP-Loci zu einem Meta-Probeset kombiniert werden können.
- der I/NI-call für jede CNV-Region bestimmt werden kann und man somit über ein verlässliches Filterkriterium zur Reduzierung der FDR verfügt.

Bisherige Methoden bestimmen die DNA-Kopienzahl in zwei Schritten, wobei zuerst die Intensitätswerte der einzelnen Probesets bestimmt werden. Dann, im zweiten Schritt wird im *Sliding-Window*-Verfahren über benachbarte Probesets geglättet, um so die Güte der Messung zu verbessern. Entge-

gen anderer Methoden kann der FARMS-Algorithmus den zweiten Schritt in das Modell integrieren. Zieht man die physikalische Position der SNP-Proben auf dem Genom bei der Modellbildung hinzu, so kann man - durch das zusätzliche Wissen - auf dem Genom benachbarte SNP-Proben zu Meta-Probesets zusammenfassen. Dies verbessert die Signalstärke, da somit korrelierte Fehler in den einzelnen Probemessungen minimiert werden. In Abbildung 4 ist das in Clevert et al. [7] publizierte Konzept der Meta-Probesets dargestellt. Zu dessen Umsetzung wurden die Oligonukleotidsequenzen der einzelnen SNP-Proben gegen ein Referenzgenom neu verglichen (*aligned*), um Annotationsfehler bezüglich der physikalische Position der SNP-Proben auf dem Genom zu minimieren. Danach konnte jeder SNP-Probe die wahrscheinlichste physikalische Position auf dem Genom zugeordnet werden, anhand derer Meta-Probesets definiert wurden. Wie links in Abbildung 3 gezeigt, betrachten konventionelle Verfahren die Probesets isoliert, und modellieren daher das Messrauschen der einzelnen Proben nicht explizit.

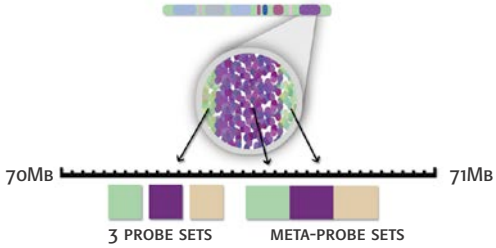

Abbildung 3. Alle gängigen Verfahren bestimmen nach der Berechnung der Kopienzahl den Durchschnitt über benachbarte Probesets, um so die Signalqualität der Messung zu verbessern. Dies erfolgt jedoch nach der CNV-Berechnung und entspricht nicht den Erkenntnissen der Informationstheorie und geht mit einem Informationsverlust einher. Es ist daher von Vorteil, erst die einzelnen Probesets zu einem Meta-Probeset zusammenzufassen und dann die Kopienzahl des Segmentes zu bestimmen.

Nach der Implementation wurde der Algorithmus mit *dChip* und *aroma.affymetrix, CNAG* und *CNAT* auf Affymetrix 500K und SNP6.0 Daten des internationalen HapMap Projekts verglichen. Die im HapMap Projekt identifizierten CNV-Regionen sind allerdings nur bedingt für einen Methodenvergleich geeignet, da keine Grundwahrheit (Goldstandard) über die gefundenen CNV-Regionen existiert. So kann es bspw. sein, dass eine neue Methode eine CNV-Region findet, die mit bisherigen Methoden nicht identifizierbar war. Diese neu gefundene Region würde jedoch immer zu den Falsch-Positiven zählen und somit die Performanz der sensitiveren Methode schmälern. Daher wurde der Methodenvergleich unter Ausschluss pseudoautosomaler und homologer Regionen (PAR1, PAR2, Xq21.3) auf dem X Chromosom durchgeführt. Das so beschnittene X Chromosom weist bei Frauen zwei Kopien und bei Männern eine Kopie auf. Somit kann die geschätzte Kopienzahl mit der wahren Kopienzahl des chromosomalen Segmentes verglichen werden und so Aufschluss über die Präzision der Vergleichsmethoden geben. Die Performanz der Methoden wurde über ihre „*Receiver Operating Characteristic*“ (ROC)-Kurve und durch die Fläche unter der ROC-Kurve (AUC) bewertet. Die ROC-Kurve stellt die Abhängigkeit der Sensitivität von der Spezifität (bzw. 1-Spezifität) dar und ist eine Methode zur Bewertung von Analyseverfahren. Der AUC-Wert ist das Integral der ROC-Kurve und liegt zwischen 0 und 1, wobei ein Wert von 0,5 der zu erwartenden Trefferhäufigkeit eines Zufallsprozesses entspricht. Je größer der AUC-Wert ist, desto größer ist die Sensitivität (Richtig-Positiv-Rate) und desto geringer die

Fehlerquote (Falsch-Positiv-Rate) der verwendeten Methode. Die statistische Signifikanz der Ergebnisse wurde durch einen McNemar-Test bzw. durch den „Wilcoxon signed-rank"-Test bestätigt.

cn.FARMS wurde in R und C entwickelt und ist als *open source* Paket über *Bioconductor* verfügbar unter http://www.bioconductor.org/packages/release/bioc/html/cn.farms.html.

3.4 CNV-Analyse von Formalin-Fixed, Paraffin-Embedded-Präparaten (FFPE)

In Zusammenarbeit mit M. Tuefferd et al. [14] untersuchten wir SNP-Arraydaten von „*Fresh-Frozen*"-(FF) und „*Formalin-Fixed, Paraffin-Embedded*"-Präparaten (FFPE) hinsichtlich der Detektierbarkeit von CN-Variationen. Die Arbeit zeigte, dass herkömmliche Methoden nur begrenzt geeignet sind, um CNV-Regionen in FFPE-Material zu entdecken.
Um cn.FARMS für Untersuchungen von FFPE-Präparaten zu optimieren, finanzierte Merck Serono eine Nachfolgestudie mit FF- und FFPE-Proben. Dazu wurden jeweils 12 passende Proben von Tumor- und Normalgeweben auf Affymetrix SNP 6.0 Arrays hybridisiert. Die in Kürze publizierten Ergebnisse zeigen, dass mit cn.FARMS signifikant weniger Falsch-Positive CNV-Regionen in FFPE-Material gefunden werden als mit anderen Methoden.

3.5 FABIA - Biclustering von Genexpressionsdaten

Klassische Clusteringverfahren wie hierarchisches Clustering sind, seit Alizadeh et al. [15] wegweisender Publikation übers Clustern von B-Zell-Lymphomen, ein beliebtes Werkzeug, um Gene (Zeilen) oder Zelltypen (Spalten) anhand der Ähnlichkeit ihres Expressionsprofiles nach einzuordnen und in Clustern zu gruppieren. Im Gegensatz zu Clusteringmethoden suchen Biclusteringverfahren simultan in Zeilen und Spalten nach Ähnlichkeiten in der Genexpressionsmatrix. Die gleichzeitige Suche in beiden Dimensionen der Expressionsmatrix ist für eine Vielzahl biologischer, medizinischer und pharmakologischer Fragestellungen von großer Relevanz, z.B.

- um Transkriptionsmodule zu finden, d.h. um koregulierte Gene und die Koregulation auslösende Bedingung zu gruppieren.
- um dysregulierte Pathways in Tumorsubtypen zu identifizieren.
- um Wirkstoffkandidaten in *high-throughput-screenings* (HTS) zu priorisieren, d.h. um unter Tausenden von Wirkstoffmolekülen, distinkte Molekülgruppen zu finden, die gleiche Gengruppen aktivieren oder deaktivieren. Diese Information gibt frühzeitig Einblick, wie sich die Biologie in der Zelle unter dem Einfluss der Wirkstoffmoleküle verändert. Somit können schon in der frühen Phase der Medikamentenentwicklung Vorhersagen gemacht werden, welche Molekülmodifikationen nötig wären, um die Potenz eines Pharmakons zu erhöhen, bzw. um dessen Nebenwirkungen zu minimieren.

Biclustermethoden lassen sich in vier Kategorien einteilen: 1) iteratives zweifaches Clustering, 2) Varianz minimierende, 3) motivbasierte und 4) generative Methoden. „**F**actor **A**nalysis for **Bi**cluster **A**cquisition" (FABIA) ist ein multiplikatives Modell und gehört zur Gruppe der generativen Methoden. Abbildung 4 zeigt, wie ein Bicluster aus dem dyadischen Produkt zweier schwachbesetzter Vektoren (*spar-*

se vectors) generiert wird. Dem FABIA-Algorithmus liegt ebenfalls das eingangs beschriebene Faktorenanalysemodell zugrunde. Es unterscheidet sich aber in der Faktorenanzahl und der Verteilungsan-

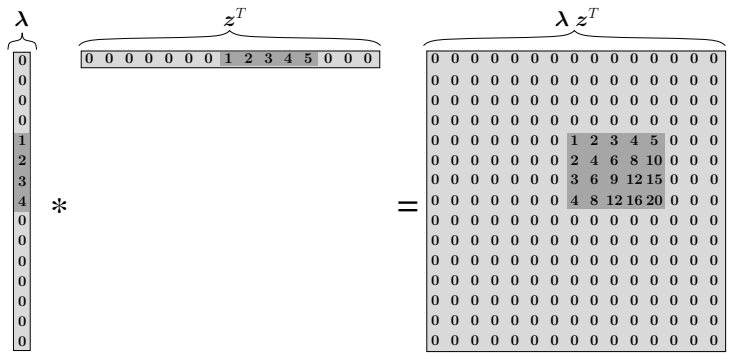

Abbildung 4: FABIA ist ein probabilistisches Modell und generiert Bicluster aus dem dyadische Produkt zweier schwachbesetzter Vektoren. Schwachbesetzte Vektoren (sparse vectors) sind Vektoren, bei denen viele Einträge aus Nullen bestehen.

nahme der latenten Variable deutlich von den in Abschnitt 3.1 und 3.3 entwickelten Modellen. Die mathematische Formulierung des FABIA-Modells für p Bicluster lautet

$$X \;=\; \sum_{i=1}^{p} \lambda_i \; z_i^T \;+\; \Upsilon \;=\; \Lambda \; Z \;+\; \Upsilon \;,$$

wobei $\boldsymbol{\Upsilon} \in \mathbb{R}^{n \times l}$ den additiven Fehleranteil darstellt; $\boldsymbol{\lambda}_i \in \mathbb{R}^n$ der Prototypvektor und $\boldsymbol{z}_i \in \mathbb{R}^l$ der Faktorenvektor des i-ten Biclusters sind. Die Schwachbesetztheit von Faktoren und Faktorenladung wurde durch einen Laplace-Prior auf Null forciert.
Der FABIA-Algorithmus wurde mit den 11 gängigsten Biclustermethoden verglichen. Der Vergleich wurde auf 400 simulierten Genexpressionsdatensätzen mit n=1000 Gene, l=100 Proben und p=10 implantierten Biclustern bzw. auf 3 zuvor vom *Broad Institute* analysierten realen Microarrayexperimenten [16] ausgeführt. Die Performanz der einzelnen Methoden wurde anhand des Jaccard-Index, dem Quotienten der Schnittmenge der gefundenen, implantierten Biclusterelementen und der Vereinigungsmenge der implantierten, gefundenen Biclusterelementen, bestimmt.

FABIA wurde in R und C entwickelt und ist über das *Open Development* Softwareprojekt *Bioconductor* verfügbar unter http://www.bioconductor.org/packages/release/bioc/html/fabia.html.

4. Ergebnisse

Der FARMS-Algorithmus für Affymetrix Genexpressionsarrays ist seit 2006 die führende Methode im internationalen "*Benchmark for Affymetrix GeneChip Expression Measures*" [17]. Der Methodenvergleich wurde an der *Johns Hopkins Bloomberg School of Public Health* vorgenommen (http://affycomp.biostat.jhsph.edu) und zeigte, dass der entwickelte Algorithmus alle anderen 131

Vergleichsmethoden in Bezug auf Sensitivität, Spezifität und FDR übertrifft. Ferner ist der Algorithmus im Durchschnitt viermal schneller als die zweitschnellste Vergleichsmethode.

Das I/NI-call Filterkriterium ist auf mehr als 30 Microarrayexperimenten evaluiert worden und konnte durchschnittlich 84% (±1,5%), bzw. in Spike-in Experimenten sogar über 99,5%, der irrelevanten Gene herausfiltern, ohne dabei ein relevantes Signal zu verlieren. Für die genomweite Genexpressionsanalyse kann durch den I/NI-call die Anzahl der Nullhypothesen von ca. 28.000 auf ca. 4.500 reduziert werden und führt nach p-Wertkorrektur für multiples Testen zu ca. 6-fach kleineren p-Werten.

Die Performanz des cn.FARMS-Algorithmus wurde auf Affymetrix 500K und SNP 6.0 Arrays des internationalen HapMap Projektes mit den vier gängigsten Methoden zur Analyse von CNVs verglichen. Der Vergleich zeigt, dass cn.FARMS sowohl auf 500K, als auch auf SNP6.0 Arrays die DNA-Kopienzahl hochsignifikant (p_{500K} =1,8e-65, $p_{SNP6.0}$ =1,e-1160) besser schätzt als die nächstbeste Vergleichsmethode. Wie die Precision-Recall-Kurven (Figure 5) auf Seite 22 zeigen, detektiert cn.FARMS zudem bei geringer FDR deutlich mehr CNVs als die Vergleichsmethoden.

Die Biclustermethode FABIA konnte auf simulierten Daten signifikant besser (da keine überlappenden Konfidenzintervalle) implantierte Bicluster identifizieren ($Jaccard_{FABIAS}$= 0.58±0.01) als alle 11 anderen Vergleichsmethoden (Jaccard Index der zweitbesten Methode $Jaccard_{ISA\_2}$= 0.33±0.05). Die bessere Performanz wurde zusätzlich auf Microarraydaten vom *Broad Institute* bestätigt, wonach mit FABIA gefundene Bicluster in der *Gene Set Enrichment Analysis* (GSEA) biologisch plausibler waren als jene anderer Vergleichsmethoden.

5. Diskussion

Es konnte gezeigt werden, dass statistische Methoden, aus dem Bereich des maschinellen Lernens und der Bayesschen Statistik, zur Auswertung von Microarraydaten sehr gut geeignet sind.
Beginnend mit FARMS, ist über I/NI-call und FABIA und später cn.FARMS ein statistisch wohl fundierter Framework zur Analyse und Interpretation genomischer Daten entstanden. Dessen Weiterentwicklung konsequenterweise auf „*Next-Generation Sequencing*“ (NGS) Daten ausgedehnt wurde und nun, mit „***C**opy **N**umber estimation by a **M**ixture **O**f Poisson**S**”* (cn.MOPS) zur Untersuchung von numerischen Chromosomenaberrationen, im Peer-Review ist.

6. Literaturverzeichnis

[1] Affymetrix. Microarray Suite User Guide (2001).

[2] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. “Affy -- Analysis of Affymetrix GeneChip data at the probe level” Bioinformatics 20.3 (2004):307-315.

[3] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. “Summaries of Affymetrix GeneChip probe level data” Nucleic Acids Research 31.4 (2003):1-8.

[4] S. Hochreiter, D.-A. Clevert, and K. Obermayer. “A new summarization method for Affymetrix probe level data” Bioinformatics 22.8 (2006):943-949.

[5] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. H. Göhlmann. “I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data” Bioinformatics 23.21 (2007):2897-2902.

[6] A. Baross, A. Delaney, I. H. Li, T. Nayar, S. Flibotte, H. Qian, S. Chan, J. Asano, A. Ally, M. Cao et al.. “Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data” BMC Bioinformatics 8.1 (2007):368.

[7] D.-A. Clevert, A. Mitterecker, A. Mayr, G.Klambauer, M. Tuefferd, A. De Bondt, W. Talloen,H. Göhlmann, and S. Hochreiter. “cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate” Nucleic Acids Research 39.12 (2011): e79.

[8] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D. K. Bailey, G. C. Kennedy et al.. “A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays” Cancer Research 65.14 (2005): 6071-6079.

[9] Affymetrix. CNAT 4.0: Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip Human Mapping 10/50/100/250/500K Array Set (2007).

[10] C. Li, and W. Wong. “Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection” Proceedings of the National Academy of Sciences 98.1 (2001): 31-36.

[11] M. Lin, L.-J. Wei, W. R. Sellers, M. Lieberfarb, W. H. Wong, and C. Li. “dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data” Bioinformatics 20.8 (2004): 1233-1240.

[12] H. Bengtsson, R. Irizarry, B. Carvalho, and T. P. Speed. “Estimation and assessment of raw copy numbers at the single locus level” Bioinformatics 24.6 (2008):759-767.

[13] H. Bengtsson, P. Wirapati, and T. P. Speed. “A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 &6“ Bioinformatics 25.17 (2009):2149-2156.

[14] M. Tuefferd, A. De Bondt, I. Van Den Wyngaert, W. Talloen, T. Verbeke, B. Carvalho, D.-A. Clevert, M. Alifano, N. Raghavan, D. Amaratunga et al.. “Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays” Genes Chromosomes Cancer 47.11 (2008): 957-964.

[15] A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, et al.. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling” Nature 403.6769 (2000):503--511.

[16] Y. Hoshida, J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. “Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets” PLoS ONE 2.11 (2007):e1195.

[17] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu and T. P. Speed. “A benchmark for Affymetrix GeneChip expression measures” Bioinformatics 20.3 (2004):323-331.

Anteilserklärung

1. **Clevert DA,** Mitterecker A, Mayr A, Klambauer G, Tuefferd M, Bondt AD, Talloen W, Göhlmann H, Hochreiter S. **cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate.** Nucleic Acids Res. 2011 http://www.ncbi.nlm.nih.gov/pubmed/21486749

   Die Software basiert auf einer Idee von DA Clevert und wurde von ihm entwickelt und validiert. (Anteil 55 Prozent). Impact Factor 7,85

2. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Van Sanden S, Lin D, Talloen W, Bijnens L, Göhlmann HW, Shkedy Z, **Clevert DA. FABIA: factor analysis for bicluster acquisition.** Bioinformatics. 2010 Jun 15;26(12):1520-7 http://www.ncbi.nlm.nih.gov/pubmed/20418340

   In diese Arbeit flossen Datenanalysen von DA Clevert sowie Vorschläge zur Heransgehensweise ein. (Anteil 20 Prozent). Impact Factor 4,93

3. Tuefferd M, De Bondt A, Van Den Wyngaert I, Talloen W, Verbeke T, Carvalho B, **Clevert DA,** Alifano M, Raghavan N, Amaratunga D, Göhlmann H, Broët P, Camilleri-Broët S. **Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays.** Genes Chromosomes Cancer. 2008 Nov;47(11):957-6 http://www.ncbi.nlm.nih.gov/pubmed/18663747

   In diese Arbeit flossen Datenanalysen von DA Clevert sowie Vorschläge zur Heransgehensweise ein. (Anteil 5 Prozent) Impact Factor 4,53

4. Talloen W*, **Clevert DA***, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Göhlmann HW., **I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data.** Bioinformatics. 2007 Nov 1;23(21):2897-902. Epub 2007 Oct 5. http://www.ncbi.nlm.nih.gov/pubmed/17921172

   Die Software wurde komplett von ihm entwickelt. (Anteil 35 Prozent / * geteilte Erstautorenschaft). Impact Factor 5,04

5. Hochreiter S, **Clevert DA**, Obermayer K. **A new summarization method for Affymetrix probe level data.** Bioinformatics. 2006 Apr 15;22(8):943-9. Epub 2006 Feb 10. http://www.ncbi.nlm.nih.gov/pubmed/16473874

   Die Software wurde von DA Clevert entwickelt und validiert. (Anteil 40 Prozent). Impact Factor 6,02

Dipl.-Inf. Djork-Arné Clevert

Nucleic Acids Research Advance Access published April 12, 2011

Nucleic Acids Research, 2011, ***1–13***
doi:10.1093/nar/gkr197

cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate

Djork-Arné Clevert[1,2,*], Andreas Mitterecker[1], Andreas Mayr[1], Günter Klambauer[1], Marianne Tuefferd[3], An De Bondt[3], Willem Talloen[3], Hinrich Göhlmann[3] and Sepp Hochreiter[1,*]

[1]Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria, [2]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany and [3]Johnson & Johnson Pharmaceutical Research & Development, a Division of Janssen Pharmaceutica, Beerse, Belgium

Received January 13, 2011; Revised March 15, 2011; Accepted March 18, 2011

ABSTRACT

Cost-effective oligonucleotide genotyping arrays like the Affymetrix SNP 6.0 are still the predominant technique to measure DNA copy number variations (CNVs). However, CNV detection methods for microarrays overestimate both the number and the size of CNV regions and, consequently, suffer from a high false discovery rate (FDR). A high FDR means that many CNVs are wrongly detected and therefore not associated with a disease in a clinical study, though correction for multiple testing takes them into account and thereby decreases the study's discovery power. For controlling the FDR, we propose a probabilistic latent variable model, 'cn.FARMS', which is optimized by a Bayesian maximum a posteriori approach. cn.FARMS controls the FDR through the information gain of the posterior over the prior. The prior represents the null hypothesis of copy number 2 for all samples from which the posterior can only deviate by strong and consistent signals in the data. On HapMap data, cn.FARMS clearly outperformed the two most prevalent methods with respect to sensitivity and FDR. The software cn.FARMS is publicly available as a R package at http://www.bioinf.jku.at/software/cnfarms/cnfarms.html.

INTRODUCTION

Copy number varations (CNVs) are one or more kilobases long DNA regions with varying copy numbers between individuals (1). In biology and population genetics, CNVs help to understand the origin and evolution of genomes (1–3). In medicine, associations between CNVs and diseases were discovered, e.g. for systemic autoimmunity (4), HIV (5), Crohn's disease and type 1 diabetes (6), type 2 diabetes (7–9), malaria, breast and prostate cancer, multiple sclerosis and bipolar disorder (10). In most CNV studies, DNA oligonucleotide arrays like the Affymetrix Genome-wide SNP 6.0 arrays are applied. These arrays possess both high coverage and high resolution through their large number of genetic markers (the probes). They are able to detect CNVs in formalin-fixed, paraffin-embedded (FFPE) tissue samples which were stored decades ago (11,12). FFPE samples are attractive because instead of designing new studies, existing biobanks can be utilized, though the measurements are more noisy.

If analyzing CNV data from microarrays, researchers face the serious problem of high false discovery rates (FDRs), i.e. the fraction of wrongly detected or too large CNV regions. CNVs are wrongly detected because of random probe variations through measurement noise. Current array techniques strive steadily to increase the number of probes in order to obtain higher coverage and higher resolution. However, this coverage is traded off against more false discoveries, which increase proportional to the number of probes. Each falsely discovered CNV region may give a false hint for population geneticists or may generate a spurious correlation with a disease and, therefore, misguides the medical expert. More seriously, a high FDR at CNV detection decreases the discovery power of studies and the significance of discoveries after correction for multiple testing. Falsely discovered CNVs are not associated with diseases, though correction for multiple testing takes them into account and reduces the discovery power of the study. Therefore, FDR control

*To whom correspondence should be addressed. Tel: +49 30 6883 5306; Fax: +49 30 6883 5307; Email: okko@clevert.de
Correspondence may also be addressed to Sepp Hochreiter. Tel: +43 732 2468 8880; Fax: +43 732 2468 9511; Email: hochreit@bioinf.jku.at

is a highly desired feature of CNV analysis methods to avoid that the advantage of higher coverage is counteracted by correction for multiple testing. However, current CNV analysis methods do not control the FDR, as Baross *et al*. (13) write 'The frequency of false positive deletions was substantial' with different methods like dChip (14) and CNAG (15). We introduce cn.FARMS for array-based CNV analysis which is designed to control the FDR while ensuring high sensitivity.

Previous array-based CNV analysis methods

We assume that the DNA is first cut by enzymes into fragments which are then amplified by PCR. The PCR products are then mechanically fragmented into smaller pieces before being put on the array. Each CNV region is broken by enzymes into several DNA fragments each of which is targeted by several probes. This gives a copy number hierarchy probes-fragment-region which is depicted in Figure 1. The more copies of the region exist, the more fragment copies exist, the higher are the probe intensities.

As visualized in Figure 2, copy number analysis is, in principle, a three-step pipeline: (i) normalization, (ii) probe-level modeling and (iii) segmentation. We introduce this pipeline to describe previous methods in the following and to describe our cn.FARMS method in section 'MATERIALS AND METHODS' (note, that cn.FARMS neither does segmentation nor integer copy number estimation).

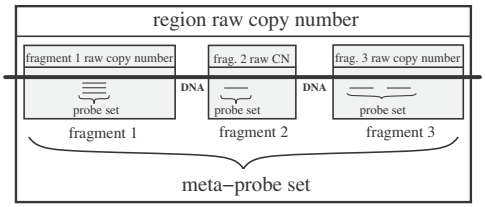

Figure 1. The copy number hierarchy probes-fragment-region. Fragment copy numbers serve as meta-probes used for 'multi-loci modeling' which yields region copy numbers. Inner boxes: the probes which target a fragment (often at a SNP position) are summarized to a raw copy number of this fragment. Note, that instead of fragments a DNA probe loci can be summarized. Outer box: the raw fragment copy numbers are the meta-probes for a DNA region and are summarized to a raw region copy number.

Normalization. Normalization is performed at two levels. It has as 'input' the raw probe intensity values and as 'output' intensity values at chromosome locations which are leveled between arrays and are allele independent. At the 'first level', normalization methods remove technical variations between arrays arising from differences in sample preparation or labeling, array production (e.g. batch effects) or scanning differences. The goal of the first level is to correct for array-wide effects. At the 'second level', alleles are combined to one intensity value at a chromosome location. Optional correction for cross-hybridization between allele A and allele B probes is performed. Cross-hybridization arise due to close sequence similarity between the probes of different alleles, therefore a probe of one allele picks up a signal of the other allele. The optional corrections for differences in PCR yield can be performed at this step or after 'single-locus modeling' (see below). After normalization, arrays have comparable, allele-independent probe intensity values, which measure the copy number of a specific target fragment or DNA probe site.

Modeling. Modeling is also performed at two levels. The 'input' is the probe intensity values which independently measure the copy number of a specific target fragment or DNA probe locus. The 'output' is an estimate for the region copy number. At the 'first level', 'single-locus modeling', the probes which measure the same fragment are combined to a raw fragment copy number ('raw' means that the copy number is still a continuous value; Figure 1). An optional intermediate level corrects for the fragment length and sequence features like the GC content to make raw fragment copy numbers comparable along the chromosome. Nannya *et al*. (15) suggested considering fragment characteristics like sequence patterns and the

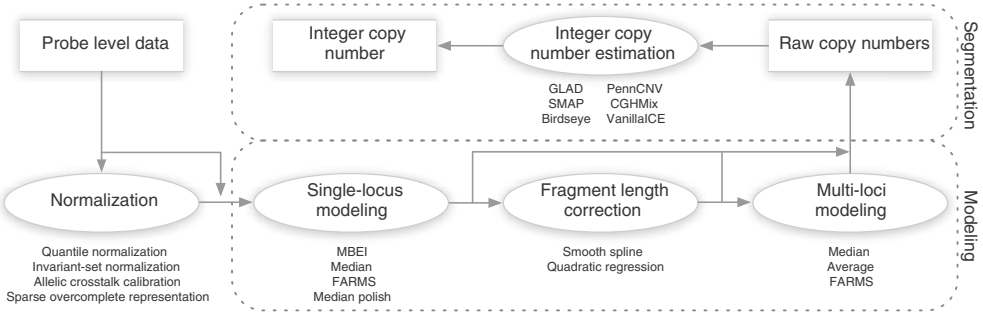

Figure 2. Copy number analysis for (Affymetrix) DNA genotyping arrays as a three-step pipeline: (i) normalization, (ii) modeling and (iii) segmentation. Modeling is divided into 'single-locus modeling' and 'multi-loci modeling' with 'fragment length correction' as an optional intermediate step. As described in subsection 'cn.FARMS: FARMS for CNV Detection', cn.FARMS' pipeline is as follows: normalization by sparse overcomplete representation, single-locus modeling by FARMS, fragment length correction and multi-loci modeling by FARMS.

length because they affect PCR amplification. For example, PCR is usually less efficient for longer fragments, which lead to fewer copies to hybridize and result in weaker probe intensities. At the 'second level', 'multi-loci modeling', the raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined to a 'meta-probe set' which targets a DNA region. Raw fragment copy numbers or DNA probe loci in a region now serve as probes themselves which measure the region's copy number (Figure 1). Multi-loci modeling considerably reduces the FDRs, because raw copy numbers of neighboring fragments or neighboring DNA probe loci must agree to each other on the copy number, which reduces the likelihood of a discovery by chance. However, low FDR is traded against high resolution by the window size for multi-loci modeling, i.e. by how many raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined.

Segmentation. Segmentation is also performed at two levels. It has as 'input' the continuous raw copy numbers and as 'output' integer copy numbers for segments. At the 'first level', segmentation groups together adjacent raw copy numbers with similar intensity values. At the 'second level', integer copy numbers are assigned to the regions. Neighboring regions are separated by breakpoints which indicate a change in the copy number (16). Note, that this step overlaps with the previous modeling step because in both steps single loci can be combined to regions. For example, hidden Markov models automatically assign integer copy numbers (the hidden states) and segment the DNA by runs of the same hidden state.

Using this pipeline, we next categorize existing methods for analyzing copy number variations on microarray data: (i) the first CNV analysis method has been supplied by Affymetrix with the hardware. It is called 'Chromosome Copy Number Analysis Tool' (CNAT) where version 1.0 appeared as early as 2004 but now version 4.0 (17) can be used. (a) Normalization is performed at the first level by quantile normalization (18). The second level is skipped because the alleles are separately modeled. (b) Modeling uses robust multichip average [RMA (18–20)] for allele-specific single-locus modeling. RMA is an additive model fitted by median polish. (ii) Following CNAT, the 'DNA-Chip Analyzer' (dChip) software for transcriptomic data was modified to allow for CNV analysis (21). (a) Normalization at the first level is based on the invariant set method which corresponds to normalize the arrays based on probes with known copy numbers. At the second level, allele A and B probe intensities are added. (b) Modeling is based on model-based expression index [MBEI (14)] for single-locus modeling. MBEI iteratively estimates a linear model that is the product of a raw copy number and a probe pattern by least squares. (c) Segmentation is either performed by computing the median over a region or by a hidden Markov model. (iii) One of the early CNV analysis methods is 'Copy Number Analyser for GeneChip' [CNAG (15)]. (a) Normalization starts with the second level, namely to remove allele-specific probe signals by adding allele A and B probes to give allele-independent fragment probe intensities per array. Next the arrays are normalized to have the same mean signal intensity for all autosomal probes which make fragment probes comparable between arrays. (b) Modeling skips single-locus modeling and directly corrects for fragment length and for the GC content. Both corrections are realized by a quadratic regression which predicts intensities based on GC content and fragment length. (iv) A CNV analysis software, which is broadly used, is Birdsuite's Birdseye (22). (a) Normalization is performed at the first level by quantile normalization like with CNAT. Normalization at the second level is realized by SNP genotyping through the Birdseed method via a mixture clustering. (b) Modeling and (c) Segmentation are performed together at the multi-loci level. The hidden states of a hidden Markov model (HMM) give the copy numbers and its outputs are the probe intensities for the estimated genotype. The HMM reuses the mixture distributions from Birdseed as emission probabilities for copy number 2 while emission probabilities for copy number 0 and 1 are estimated on the X chromosome using the sex information. (v) Most recently 'Copy-number estimation using Robust Multichip Analysis' [CRMA (23), CRMA_v2 (24)] has been proposed as an extension of the RMA model. (a) Normalization at the first and second level are combined by allelic cross-hybridization correction (ACC). ACC performs allele correction array-wise in the 2D space of the allele A and allele B intensity. A cone is fitted to the data such that one border of the cone is a regression line for the AA genotype and the other border for the BB genotype. Similar to the left and right line in Figure 3. The cone fitting allows estimating how much allele A cross-hybridizes at the allele B probe and vice versa. Genotype AA (allele A only) should lead to minimal intensity at the allele B probe and genotype BB (allele B only) to minimal intensity at allele A probe. The genotype AB is assumed to have the same cross-hybridization characteristics as genotypes AA and BB. Finally, the probes are normalized by scaling them to a pre-specified mean intensity value. (b) Modeling for single-locus raw copy numbers is performed via RMA. Then CRMA corrects for the GC pattern and for the fragment length where the former showed little effect and is therefore not recommended by the authors (23). Most CNV analysis methods allow using an arbitrary segmentation algorithm [for an overview see Ref. (25)].

Popular is the Gain and Loss Analysis of DNA (GLAD) model which is a local constant Gaussian regression model (26). Using a weighted maximum likelihood estimator, GLAD estimates regions with constant copy numbers. Other methods like CGHMIX (27) estimate the copy number by a mixture model incorporating spatial information. Spatial information is also utilized by segmentation with an HMM like in Birdseye and in the 'Segmental Maximum A Posteriori' approach [SMAP (28)]. Also 'PennCNV' (29) and 'vanillaICE' (30) apply an HMM to integer copy number estimation using spacial and genotype information.

However, all mentioned methods do not control the FDR and are prone to high FDRs. We will control the

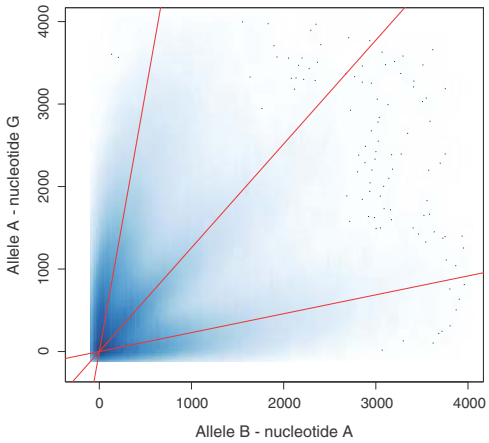

Figure 3. Sparse overcomplete representation of allele A and B probes. The smooth scatter plot for a HapMap Affymetrix 250K_NSP array sample (CEU_NA12878, G/A allele probes). The three clouds going outwards from the origin correspond to genotypes AA (upper left cloud), AB (middle cloud), and BB (lower right cloud). For the genotype AA, allele A probes show a strong signal and allele B probes shows a weak signal due to cross-hybridization (analog for genotype BB). Note, that the middle cloud is closer to the left cloud than to the right (violating CRMA's ACC assumptions). The lines are the estimates of sparse overcomplete representation. They are used to correct for cross-hybridization by moving the left cloud to be vertical, the middle cloud to be at the 45° line and the lower right cloud to be horizontal.

FDR by selecting CNVs based on high information content determined by a latent variable model.

MATERIALS AND METHODS

We propose a novel CNV detection method, called 'cn.FARMS', which is based on our FARMS ['factor analysis for robust microarray summarization' (31)] algorithm for summarizing probe sets of expression arrays. Expression array summarization estimates the expression value of a gene which is basically its mRNA copy number. The expression value of an mRNA is computed from intensity values of all probes targeting it that is the probe intensities are summarized. Since 2006, FARMS is the leading summarization method of the international 'affycomp' competition if sensitivity and specificity are considered simultaneously. We extend FARMS to cn.FARMS for detecting CNVs by moving from mRNA copy numbers to DNA copy numbers.

cn.FARMS: FARMS for CNV detection

cn.FARMS is described by the pipeline depicted in Figure 2: (i) normalization at the first and second level are combined similar as for CRMA (23). However, instead of CRMA's ACC, we propose sparse overcomplete representation in the 2D space of allele A and B intensity. Therefore, we do not only estimate the AA and the BB cross-hybridization like CRMA but also the AB cross-hybridization. The latter takes into account that hybridization and cross-hybridization may be different for the AB genotype, where for both allele probes target fragments are available and compete for hybridization. After allele correction, we follow CRMA and normalize by scaling the probes to a pre-specified mean intensity value. CNV probes which have only one allele are scaled in the same way. (ii) At the first level, 'single-locus modeling', raw fragment copy numbers are estimated by FARMS. The original FARMS was designed to summarize probes which target the same mRNA. This can readily be transfered to CNV analysis where FARMS now summarizes probes which target the same DNA fragment. Either both strands can be summarized together or separately where our default is the former. Following the suggestions in Nannya *et al.* (15), cn.FARMS performs GC and fragment length correction. At the second level, 'multi-loci modeling', the raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined to a 'meta-probe set' which targets a DNA region. The raw fragment copy numbers from single-locus modeling are now themselves probes for a DNA region as depicted in Figure 1. Again, we use FARMS to summarize metaprobes and to estimate a raw copy number for the region. This modeling across samples is novel as previous methods only model along the chromosome. FARMS supplies an informative/non-informative (I/NI) call (32,33) which is used to detect CNVs. Additionally, the I/NI value gives the signal-to-noise-ratio of the estimated raw copy number. (iii) Segmentation and estimation of integer copy numbers is performed by segmentation methods like those which were mentioned at the end of the 'Introduction' section.

In our pipeline, FARMS is used for both single-locus and multi-loci CNV analysis. The more loci are combined, the more the FDR is reduced, because more metaprobes must mutually agree on the region's copy number. The window size for multi-loci modeling is a hyperparameter which trades off low FDR against high resolution. We recommend a window size of 5 as default, 3 for high resolution and 10 for low FDR. Alternatively to a fixed number of CNV or SNP sites, the cn.FARMS software allows defining a window in terms of base pairs. In this case, multi-loci modeling may use a different number of metaprobes at different DNA locations, in particular for less than two metaprobes multi-loci modeling is skipped. Note, however, that controlling the FDR is more difficult because a minimal number of metaprobes cannot be assured for each window and modeling with few metaprobes is prone to false discoveries. cn.FARMS introduces at several steps novel algorithms into the CNV detection pipeline. First, at the normalization step sparse overcomplete representation is used for allele correction. Second, FARMS is used for 'single-locus modeling'. Third, FARMS is used for 'multi-loci modeling' which supplies the raw region copy numbers. Fourth, and most importantly, I/NI calls for controlling the FDR are supplied. In the following subsections,

we describe the methods which are utilized by cn.FARMS and are novel in the CNV detection pipeline.

Sparse overcomplete representation

At the pipeline's step (i), the normalization, cn.FARMS corrects for cross-hybridization between allele A and allele B probes. We generalize the ACC method of CRMA. ACC performs a cone fitting in the 2D space of the allele A and allele B intensity, where cone borders lay at the AA and BB genotype (see left and right line in Figure 3). For each array, the probes are first divided into the allele groups A/T, A/C, A/G, T/C, T/G, C/G to each of which a cone is separately fitted. ACC assumes that cross-hybridization for the AB genotype has the same characteristics as for AA and BB genotypes. Consequently, the AB genotype regression line is supposed to be exactly between the AA and BB genotype regression line (the cone borders), that is the AB regression line divides the cone into two equal halves. However, the assumption on the AB genotype regression line is not always true as shown in Figure 3 for a HapMap Affymetrix 500K array sample. In this example, the AB regression line does not divide the cone into two equal halves, which indicates that cross-hybridization is different for the AB genotype. For the AB genotype, target fragments for both alleles are present and compete for hybridization at the probe's spots. Motivated by such examples, at the ACC step we not only estimate a regression line for the AA and BB genotype but also for the AB genotype. After correction for cross-hybridization, the AA and BB genotypes should lay on the x-axis (allele A) and y-axis (allele B), respectively, because one probe allele is supposed to be zero, while the AB genotype should be on the 45° line. This problem of fitting three lines in a 2D space is solved in the field of machine learning by sparse overcomplete representation (34,35). Data points are described by more vectors than the dimension of the space, therefore the description of a data point is not unique. Sparse overcomplete representations choose the most sparse one from the set of all possible data descriptions. A sparse description is appropriate if each data point is mainly determined by few describing vectors. For allele correction, the sparse description is justified because a data point which is given by the two allele probe intensities can be described by (i) its angle given by the genotype (AA, AB and BB–the genotype determines three main directions) and (ii) its radius given by the copy number. Thus, we represent the 2D vector of allele A and allele B probe intensity by a 3D vector where the components correspond to the genotypes AA, AB and BB. The solution of a sparse overcomplete representation is shown as the lines in Figure 3. A sparse overcomplete representation of 2D data $\boldsymbol{x}_s \in \mathbb{R}^2$ can be modeled as:

$$\boldsymbol{x}_s = \lambda_s \boldsymbol{z}_s + \epsilon_s \quad (1)$$

where $\boldsymbol{z}_s \in \mathbb{R}^3$, $\lambda_s \in \mathbb{R}^{2\times 3}$ and $\epsilon_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_s)$. Here $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_s)$ is the 2D Gaussian distribution with mean vector $\mathbf{0} \in \mathbb{R}^2$ and covariance matrix $\boldsymbol{\Psi}_s \in \mathbb{R}^{2\times 2}$. Sparseness is enforced by assuming a Laplacian prior for $\boldsymbol{z}_s$:

$$p(\boldsymbol{z}_s) = (2)^{-\frac{3}{2}} \prod_{l=1}^{3} \exp\left(\sqrt{2}\,|z_{sl}|\right). \quad (2)$$

Because the likelihood for this model is analytically intractable, we employ a variational approach according to Girolami (36). The Laplacian prior is locally approximated from below by a local Gaussian at the mode of the Laplacian. An expectation–maximization algorithm (37) is used to optimize the parameters $\boldsymbol{\lambda}_s$ and $\boldsymbol{\Psi}_s$. Using these parameters, the maximum of the $\boldsymbol{z}_s$-posterior $\hat{\boldsymbol{z}}_s$ allows back-transforming the data to $\hat{\boldsymbol{x}}_s$ by $\hat{\boldsymbol{x}}_s = \lambda_s \hat{\boldsymbol{z}}_s$.

FARMS algorithm

Overview. The main idea of the FARMS algorithm is to detect a common hidden cause in the measurements assuming independent noise. The probabilistic FARMS model:

- regards that probes measuring the same target (fragment or region) can only be positively correlated,
- estimates (meta-)probe-specific characteristics,
- automatically trades off signal against noise via the z-posterior distribution,
- can adjust the signal/noise tradeoff via the priors on the parameters and
- supplies I/NI calls (32,33).

The I/NI call measures the information gain of the posterior over the prior which can be interpreted as the negative log signal-to-noise ratio. High data information content leads to a low variance of the latent variable's posterior and a high confidence in the copy number estimate. The original FARMS applied to 30 real-life expression data sets could exclude 70–99% of all probe sets because of their low information content while never excluding a gene that was known to be biologically meaningful (32). We want to introduce this I/NI call property into the field of CNV analysis to control the FDR.

Brief review. The vector of n probes $\boldsymbol{x}$ is modeled by probe-effects $\boldsymbol{\lambda}$ and a factor z (latent variable or signal) representing the raw normalized copy number as:

$$\boldsymbol{x} = \lambda z + \epsilon, \quad (3)$$

where $\boldsymbol{x}$, $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $z \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. Here $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is the diagonal noise covariance matrix to address independent measurement noise. ϵ and z are assumed to be statistically independent. Given these assumptions, $\boldsymbol{x}$ is distributed according to the following Gaussian:

$$\boldsymbol{x} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right). \quad (4)$$

The covariance matrix of $\boldsymbol{x}$ is decomposed into signal $\boldsymbol{\lambda}\boldsymbol{\lambda}^T$ and noise $\boldsymbol{\Psi}$. Because $\boldsymbol{\Psi}$ is diagonal, probe correlations are attributed to the signal z via $\boldsymbol{\lambda}$. That means highly correlated probes lead to large $\boldsymbol{\lambda}$ which in turn leads to

low noise because the diagonal of the covariance matrix of $\boldsymbol{x}$ is mainly explained by $\boldsymbol{\lambda}$.

Higher intensity of the probes means more copies and vice versa, therefore noise-free probes must be positively correlated. FARMS ensures the positive correlation of probes by a prior on $\boldsymbol{\lambda}$ which enforces only positive values: $p(\lambda) = \prod_{j=1}^{n} p(\lambda_j)$, where the rectified Gaussian $p(\lambda_j)$ is given by

$$\lambda_j = \max\{y_j, 0\} \text{ with } y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda) . \quad (5)$$

Further, the prior on $\boldsymbol{\lambda}$ prefers small values and, therefore, model selection tends to explain variation by noise instead by a signal. Using μ_λ and σ_λ, the prior's influence on model selection and, therefore, the signal/noise tradeoff can be adjusted. FARMS selects the model parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ by an expectation–maximization algorithm (37) that maximizes the parameter posterior. To ensure data consistency, negative entries in the data covariance matrix are set to zero.

I/NI calls. The I/NI call measures the information gain of the posterior hidden variable distribution compared to its prior distribution where the latter represents the null hypothesis. Therefore, the I/NI call measures the tendency to reject the null hypothesis based on the observed data.

From the model Equation (4) and the Gaussian z-prior $N(0, 1)$, we can compute the z-posterior $p(z \mid \boldsymbol{x})$ as

$$\begin{aligned} z \mid \boldsymbol{x} &\sim \mathcal{N}\left(\mu_{z|\boldsymbol{x}} , \sigma^2_{z|\boldsymbol{x}}\right) \\ \mu_{z|\boldsymbol{x}} &= (\boldsymbol{x})^T \boldsymbol{\Psi}^{-1}\lambda \left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-1} \\ \sigma^2_{z|\boldsymbol{x}} &= \left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-1} . \end{aligned} \quad (6)$$

We see that large $\boldsymbol{\lambda}$ (going with low noise $\boldsymbol{\Psi}$) leads to low variance of $z \mid \boldsymbol{x}$, which means a precise conditional z.

The variance of z is decomposed into a signal and a noise part:

$$\text{var}(z) = \frac{1}{N}\sum_{i=1}^{N} E_{z_i|\boldsymbol{x}_i}\left(z_i^2\right) = \frac{1}{N}\sum_{i=1}^{N} \mu^2_{z_i|\boldsymbol{x}_i} + \sigma^2_{z|\boldsymbol{x}}, \quad (7)$$

where the noise part $\sigma^2_{z|\boldsymbol{x}}$ is independent of $\boldsymbol{x}_i$ according to Equation (6) and serves as I/NI call in FARMS (32).

At the same time $-\log\sigma_{z|\boldsymbol{x}}$ measures the information gain between the prior and the posterior because the prior has unit variance and therefore zero entropy.

cn.FARMS: I/NI calls and FDR control

As the FARMS I/NI call also cn.FARMS' I/NI call measures the information gain of the posterior hidden variable distribution compared to its prior distribution that represents the null hypothesis. The variance across samples of the signal part of maximum posterior hidden variable z given the observation x is cn.FARMS I/NI call. This signal variance is zero for the prior. In contrast to FARMS I/NI call, cn.FARMS I/NI call also includes the signal strength. This reflects the assumption that data from null hypotheses produce only spurious signals that are low. Such spurious signals are more likely to be observed for cn.FARMS at multi-loci modeling with few metaprobes than for FARMS on expression arrays with larger probe sets.

First, we compute the signal strength S. The data $\{\boldsymbol{x}_i\}$ has been probewise standardized to variance 1 and mean zero, where $\text{std}(\boldsymbol{x}^{\text{raw}})$ is the probes' SD vector of the raw data $\boldsymbol{x}^{\text{raw}}$. We reintroduce the signal strength S as the median of $\boldsymbol{\lambda}$ scaled by std $\boldsymbol{x}^{\text{r}}$:

$$S = \text{median}(\lambda \cdot \text{std}(\boldsymbol{x}^{\text{raw}})), \quad (8)$$

where '$\cdot$' is the element-wise product.

Second, we extract the variance of the maximum a posterior hidden variable z given the observation $\boldsymbol{x}$:

$$\begin{aligned} \text{sigvar}(z) &= \frac{1}{N}\sum_{i=1}^{N} \mu^2_{z_i|\boldsymbol{x}_i} \\ &= \lambda^T\boldsymbol{\Psi}^{-1}\,\text{covar}(\boldsymbol{x})\,\boldsymbol{\Psi}^{-1}\lambda\left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-2}, \end{aligned} \quad (9)$$

which is between 0 (no signal, only noise) and 1 (only signal, no noise). Note, that sigvar(z) is one minus FARMS' I/NI call squared and corresponds to the part of the variance in the data explained by the signal.

cn.FARMS' I/NI call is signal variance multiplied by the signal strength squared:

$$\text{I/NI} = \text{sigvar}(z)\, S^2. \quad (10)$$

Note, that I/NI calls allow comparing two data sets with respect to common CNVs. In this case, the model is selected on one data set $\{\boldsymbol{x}_i\}$ and the calls are made on the other data set $\{\boldsymbol{y}_i\}$ using covar($\boldsymbol{y}$).

The I/NI call value considers both the signal strength and the information gain. If the true copy numbers vary, then probe intensities are consistent (correlated) and large (high signal-to-noise-ratio) and therefore lead to a large $\boldsymbol{\lambda}$ and a small $\boldsymbol{\Psi}$ which in turn gives a large sigvar(z) (close to 1). In contrast to these true positives, false positives come from random independent Gaussian noise variations, which are unlikely to produce consistent and large probe intensities. Thus, the larger the I/NI call, the less likely it was caused by noise. Consequently, the ratio of false positives decreases with increasing I/NI call values. A CNV is detected by an I/NI call value exceeding a detection threshold, therefore the threshold controls the FDR. The effect of the detection threshold can be seen in Figure 5, where precision–recall curves on HapMap SNP 6.0 arrays are shown for cn.FARMS. Note that the precision is 1−FDR, thus the distance of the curve to the upper limit gives the FDR. Therefore, the curve shows the FDR as a function of the threshold where indeed higher thresholds (more to the left) result in smaller FDRs. The detection threshold for a desired FDR can either be estimated at chromosome locations where CNVs are unlikely or at reference data sets.

RESULTS

We compare the new cn.FARMS algorithm with the two methods which performed best in other comparative studies on raw copy number estimation (23), namely the

dChip software for CNV analysis (21) and CRMA (23,24) (see the 'Introduction' section for a brief description of these and the following methods). In Bengtsson *et al.* (23), it was shown that both CRMA and dChip perform better than CNAG and CNAT . Therefore, these two methods are not regarded in our experiments. Other methods like Birdseye do not estimate raw copy numbers and incorporate segmentation and integer copy number estimation. The latter methods can still be applied on the output of cn.FARMS for single-locus or multi-loci modeling.

Because true copy numbers are in general not known, we use two benchmark data sets from 'The International HapMap Project' where the sex must be determined by the raw copy numbers at the X chromosome. (i) We first use the 250K Affymetrix array benchmark data set from Bengtsson *et al.* (23). Even if these arrays are outdated, they allow comparisons to other CNV analysis methods like CNAT and CNAG investigated in Bengtsson *et al.* (23). (ii) Next, this benchmark was upgraded to Affymetrix SNP 6.0 arrays, to allow further assessment on recent arrays. (iii) Finally, we assess the FDR at CNV detection on the HapMap phase 2 data set with Affymetrix SNP 6.0 arrays. To estimate the FDR, we define as true CNVs those which were multiple confirmed by other techniques and reported in Conrad *et al.* (38).

250K array benchmark

The first data set is from Bengtsson *et al.* (23). It comprises the 90 CEU founders (30 triplets of father, mother, child) from 'The International HapMap Project' (phase 2) where the children are removed to avoid biases due to inherited CNVs. For these 60 CEU founders, their DNA has been analyzed by Affymetrix Mapping250K_NSP arrays. Female NA12145 had too low copy number level on chromosome X and has been excluded (23) which leads to the final data set of 59 CEU founders. The X chromosome serves as ground truth to assess the performance of CNV detection methods because there males possess one copy and females two. At every location on the X chromosome, raw fragment copy numbers (single-loci) and raw region copy numbers (multi-loci) are used to classify the sex of the person the sample stems from. To allow multi-loci classification for dChip and CRMA_v2, adjacent raw fragment copy numbers are averaged within a region to give a raw region copy number. However, not all locations on the X chromosome can distinguish the sex based on the copy numbers. At the pseudo-autosomal regions (PAR1 and PAR2), the copy numbers of males and females match. Besides PAR1 and PAR2, there are segmental duplications on chromosome Y which match regions at chromosome X (obtained from 'Segmental duplication DB' at http://humanparalogy.gs.washington.edu/build36/). Further chromosome X has CNV regions (1,2). All loci in pseudo-autosomal, segmental duplications in Y and CNV regions are excluded in our classification task. Finally, 5557 single loci on the X chromosome for distinguishing males from females were kept which gives 327 863 ($=59\times5557$) single loci sex classification tasks. The performance of the methods is measured by receiver operating characteristic (ROC) curves. The ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1−specificity). Methods with ROC curves at the upper left corner indicate better performance of the corresponding method—a method's ROC curve above another method's ROC curve shows that the former method performs better than the latter. The classification results are shown as ROC curves (A and B) in Figure 4. The ROC curves are summarized by the area under the ROC (AUC) in Table 1. Further, we give the false positives (males classified as females) where the numbers of false positives and false negatives are equal—that is the false positives in the largest 161 153 (number of true female loci = 29×5557) raw copy numbers. To evaluate the statistical significance of the method's differences in performance, we use McNemar's χ^2 test under the null hypothesis that the compared algorithms should have the same error rate (39). The results show that cn.FARMS performs significantly better than dChip and CRMA_v2 and has much fewer false discoveries—confirming that cn.FARMS yields low FDRs.

SNP 6.0 array benchmark

Because the Affymetrix 250K arrays are outdated, we perform the same benchmark test as in the previous

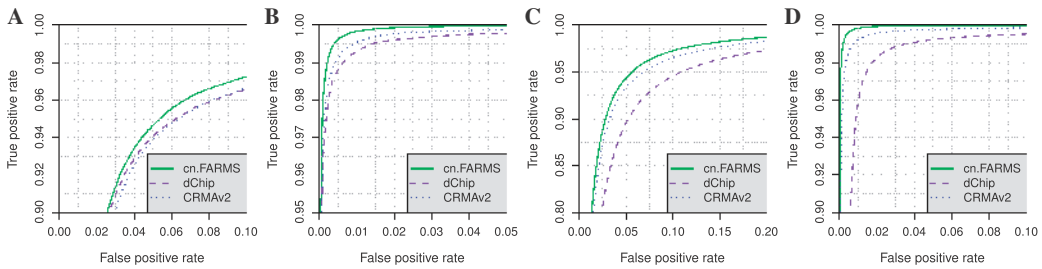

Figure 4. ROC curves for cn.FARMS, CRMA_v2 and dChip at the sex classification task for 59 HapMap CEU founders based on the X chromosome copy numbers. The panels show (**A**) single-locus and (**B**) three-loci modeling of Affymetrix Mapping250K_NSP arrays. While panels show (**C**) single-locus and (**D**) three-loci modeling of Affymetrix SNP 6.0 arrays. ROC curves more at the upper left indicate better performing methods (AUC values for Affymetrix Mapping250K_NSP and Affymetrix SNP 6.0 are given in Table 1). cn.FARMS performs better than CRMA_v2 and dChip.

Table 1. AUC values for cn.FARMS, CRMA_v2 and dChip at the sex classification task for 59 HapMap CEU founders based on the X chromosome copy numbers measured by Affymetrix 250K and Affymetrix SNP 6.0 arrays

Loci	Criteria	Affymetrix Mapping250K_NSP			Affymetrix SNP 6.0		
		cn.FARMS	CRMA_v2	dChip	cn.FARMS	CRMA_v2	dChip
1	AUC	**0.9852**	0.9820	0.9819	**0.9838**	0.9807	0.9721
	FP	**8472**	9106	9018	**56145**	68593	77438
	p-value	–	1.8e-65	3.1e-26	–	1e-1160	1e-6949
2	AUC	**0.9983**	0.9974	0.9969	**0.9983**	0.9963	0.9894
	FP	**1375**	1449	1611	**9777**	11705	18039
	p-value	–	2.7e-4	2.5e-12	–	1e-317	1e-3713
3	AUC	**0.9998**	0.9995	0.9992	**0.9998**	0.9990	0.9953
	FP	**240**	366	440	**1573**	3462	6625
	p-value	–	2.6e-38	7.2e-58	–	1e-896	1e-3455
4	AUC	**1.0000**	0.9999	0.9998	**0.9999**	0.9995	0.9976
	FP	**49**	95	153	**366**	1338	2985
	p-value	–	2.8e-10	1.9e-48	–	1e-594	1e-2023

The first column gives the number of combined loci, where '1' means single-locus modeling. The second column gives (i) area under the receiver operating curve given in Figure 4 ('AUC'), (ii) false positives ('FP' – females are classified as males) and (iii) the *P*-value of McNemar's χ^2 test for difference to the cn.FARMS ('*p*-value'). False positives are counted in the lowest 166 710 and 1 075 680 (number of true male loci) raw copy numbers for Affymetrix 250K and Affymetrix SNP 6.0 arrays, respectively. The last six columns give the values for the according array types and methods, where significant better performance is indicated by boldface numbers. cn.FARMS clearly outperforms CRMA_v2 and dChip.

subsection but now with up-to-date Affymetrix SNP 6.0 arrays. The SNP 6.0 data set comprises again the same 59 CEU founders as for the 250K array benchmark. Note, in contrast to Affymetrix 250K arrays, for Affymetrix SNP 6.0 arrays we model single SNP and CNV loci instead of fragments because for Affymetrix SNP 6.0 the fragment that is targeted by a probe is ambiguous as both Sty and Nsp fragments can hybridize to a probe. We again excluded regions which are pseudo-autosomal, have segmental duplications or have been reported as CNV regions and kept 35 856 single loci for the classification task which sums up to 2 115 504 ($=59\times35\,856$) sex classification tasks. ROC curves (C and D) in Figure 4 show the results for which the AUCs and McNemar significance tests are given in Table 1. Again we report the false positives (females classified as males) while equalizing the numbers of false positives and false negatives. By doing this, we give the number of false positives in the largest 1 039 824 raw copy numbers, which is the number of true female loci $=29\times35\,856$. Again cn.FARMS significantly outperforms CRMA_v2 and dChip and has fewer false discoveries. The absolute improvement in terms of the AUC values seem to be marginal. However, for single locus modeling, we obtained *P*-values of 1.8e-65 and 3.1e-26 by the McNemar test for 250K, even going down to 1e-1160 and 1e-6949 for SNP 6.0 arrays. Clearly, these *P*-values indicate significant performance improvement of cn.FARMS over its competitors. For 250K arrays, cn.FARMS has 8472 false positives and the second best method (dChip) has 9018, which is about 6.5% more false positives. For SNP 6.0 arrays, cn.FARMS has 56 145 false positives and the second best method (CRMA) has 68 593, which is about 21% more false positives. For 250K arrays and multi-loci modeling with 4 loci, the number of 49 false positives almost doubles if we look at the next best method with 95 false positives. For SNP 6.0 arrays and multi-loci modeling with 4 loci, the number of 366 false positives increases by a factor of 3.5 if we look at the next best method with 1338 false positives.

CNV Detection on HapMap

In this subsection, we want verify that cn.FARMS can indeed control the FDR. In the previous two subsections, we classified male/female based on raw copy numbers at X chromosome locations. The majority of loci have a CNV as half of the samples are male with copy number one and the other half are female with copy number two. Therefore, false discoveries can only appear at the few pseudo-autosomal or CNV regions. In CNV association studies, however, false discoveries are much more likely because true CNVs are rather rare. Therefore, we define rare true CNV regions in this experiment where we use again 'The International HapMap Project' phase 2 data set with Affymetrix SNP 6.0 arrays. The goal is now to identify true rare CNV regions with a low FDR.

We define as 'true CNV regions' those regions which were detected and verified by different biotechnologies in Conrad *et al.* (38). In Conrad *et al.* (38), first, CNV candidate regions were identified by NimbleGen tiling arrays with 2.1 million long oligonucleotide probes covering the genome with a median probe spacing of 56 bp. From the identified CNVs, random control samples were selected and successfully verified by quantitative PCR. The CNV regions identified by NimbleGen tiling arrays served to design CNV-typing Agilent CGH arrays comprising 105 000 long oligonucleotide probes. With these Agilent arrays, 4978 CNVs were detected on 450 HapMap phase 3 samples and then completed by 59 CNV regions from McCarroll *et al.* (40). The third platform, Illumina Infinium genotyping (Human660W), found CNVs of which 87% were already genotyped by the Agilent CGH arrays. Almost all CNVs from Conrad *et al.* (38) were confirmed by at least two different platforms (NimbleGen tiling arrays, Agilent CGH or Illumina

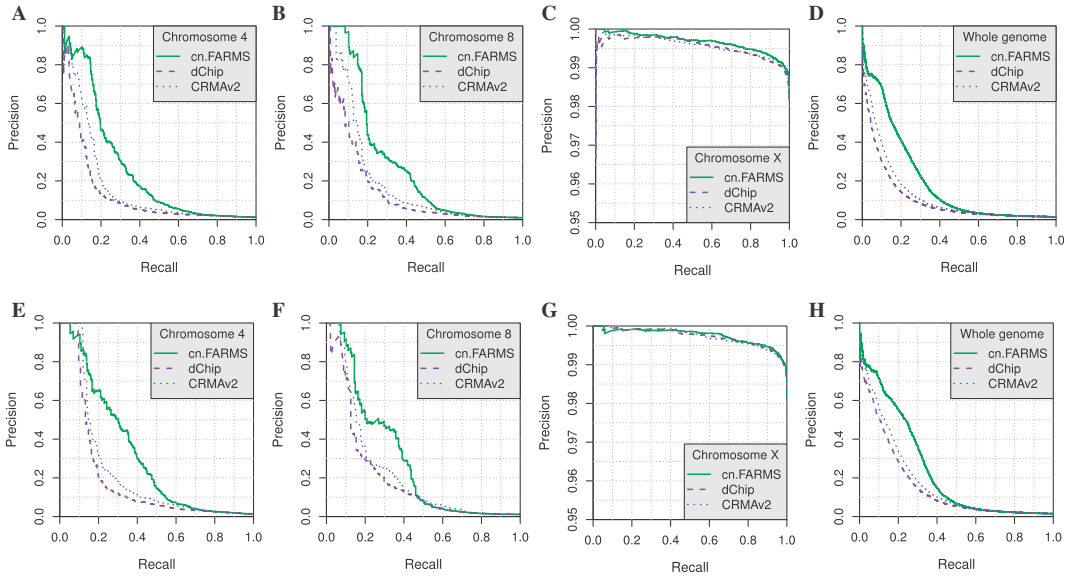

Figure 5. Precision-recall curves (PRCs) on HapMap SNP 6.0 arrays for cn.FARMS, CRMA_v2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad *et al.* (38). cn.FARMS detection criteria is the I/NI call, whereas CRMA_v2 and dChip use the variance of raw copy numbers. A PRC more in the upper-right hand corner indicates better performance. Note, that precision is (1−FDR) thus the FDR is the distance of the curve to the upper limit. Panels (**A–D**) give the PRC for chromosome 4, 8, chromosome X and the whole genome for 3 loci. Panels (**E–H**) show the same for 5 loci. cn.FARMS (solid green) has a clear advantage over dChip (dashed purple) and CRMA_v2 (dotted blue). cn.FARMS has a considerable lower FDR compared to the other methods.

Human660W). Of these 5037 CNV regions, we only selected CNV regions from the 60 CEU HapMap phase 2 samples (CEU trios without children). Finally, we obtained 2515 true CNV regions as reference for our experiment.

For detecting CNV regions, cn.FARMS uses its I/NI calls. However for CRMA_v2 and dChip, we have to define a CNV calling criterion. We tested different criteria of which the variance of the raw copy numbers on the samples gave the best results. This variance calling criterion is like I/NI call independent of the test statistic, thus correction for multiple testing is still valid (33,41).

Using the true CNVs, we can assess the FDR. Instead of reporting the FDR for a fixed classification threshold, we present the CNV detection results as precision–recall curves (PRCs). PRCs plot the precision (which is 1−FDR) as a function of the true positive rate (recall or sensitivity). Thus, a PRC that is more in the upper-right hand corner performs better. A larger *y*-value of the PRC means a lower FDR for a given sensitivity. Figure 5 shows the PRC plots where cn.FARMS has indeed lower FDRs compared to the other methods. The corresponding areas under the precision–recall curves are listed in Table 2. A larger value means that the method has lower FDR averaged over different given recall values. We observed that for some chromosomes increasing the window size also increases the FDR because of the reduced resolution and an overestimate of CNV regions. cn.FARMS performed significantly better than CRMA_v2 and dChip. The significance was obtained by a one sample *t*-test under the hypothesis that the differences between values of the area under the PRC for two methods have a mean equal to 0. The Gaussian assumption of the *t*-test was verified beforehand by a Shapiro–Wilk test. The *P*-values of the *t*-test were smaller than 3.9e-7 for CRMA_v2 and smaller than 7.1e-8 for dChip. Figure 6 shows CNV calling plots across chromosome 4 for 3-loci and 5-loci regions. The *y*-axis gives cn.FARMS' I/NI call and for both CRMA_v2 and dChip the raw copy number variance across samples. Calling values are scaled such that the maximum is one. Local calling densities are encoded by blue color shades. True CNVs (reported in 38) are marked as light-rose bars and calls at these loci by red circles. A perfect calling method would call all true CNVs (red circles at 1) and would not call others (dark blue background at 0). True positives (true CNVs) are better separated from true negatives by cn.FARMS as the smaller variance of true negatives, which is indicated by dark blue density at the bottom. The red arrows, e.g. at positions 65 or 85 Mb in the upper cn.FARMS panel, indicate verified CNVs which were detected by one method, in this case cn.FARMS, but not by both others. cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip.

Table 2. Area under the PRCs on HapMap SNP 6.0 arrays for cn.FARMS, CRMA_v2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad *et al.* (38)

Method	Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of			
		3	4	5	7		3	4	5	7		3	4	5	7		3	4	5	7
cn.FARMS	**1**	0.20	0.23	0.23	0.25	**7**	0.16	0.19	0.19	0.22	**13**	0.22	0.26	0.26	0.26	**19**	0.17	0.21	0.23	0.26
CRMA_v2		0.16	0.19	0.20	0.23		0.10	0.12	0.13	0.16		0.10	0.14	0.18	0.23		0.11	0.13	0.15	0.19
dChip		0.14	0.18	0.19	0.22		0.09	0.11	0.12	0.15		0.08	0.11	0.16	0.20		0.09	0.11	0.13	0.16
cn.FARMS	**2**	0.19	0.21	0.22	0.25	**8**	0.27	0.28	0.29	0.27	**14**	0.06	0.06	0.06	0.06	**20**	0.24	0.26	0.25	0.31
CRMA_v2		0.12	0.14	0.16	0.21		0.18	0.20	0.22	0.26		0.06	0.06	0.06	0.07		0.12	0.18	0.19	0.25
dChip		0.10	0.13	0.15	0.20		0.13	0.16	0.19	0.21		0.05	0.06	0.06	0.06		0.11	0.14	0.15	0.21
cn.FARMS	**3**	0.27	0.31	0.34	0.39	**9**	0.14	0.13	0.12	0.12	**15**	0.11	0.12	0.14	0.16	**21**	0.05	0.11	0.12	0.19
CRMA_v2		0.16	0.20	0.23	0.29		0.09	0.08	0.10	0.09		0.07	0.09	0.11	0.14		0.04	0.05	0.10	0.10
dChip		0.13	0.16	0.20	0.25		0.06	0.07	0.07	0.07		0.06	0.08	0.09	0.12		0.03	0.04	0.06	0.07
cn.FARMS	**4**	0.25	0.30	0.31	0.34	**10**	0.14	0.17	0.20	0.23	**16**	0.21	0.31	0.36	0.35	**22**	0.54	0.61	0.64	0.70
CRMA_v2		0.16	0.20	0.22	0.26		0.09	0.12	0.15	0.19		0.14	0.21	0.25	0.33		0.41	0.50	0.54	0.62
dChip		0.12	0.16	0.19	0.21		0.08	0.11	0.15	0.18		0.12	0.19	0.23	0.32		0.37	0.44	0.49	0.58
cn.FARMS	**5**	0.19	0.22	0.22	0.24	**11**	0.24	0.25	0.25	0.27	**17**	0.17	0.22	0.22	0.26	**X**	0.99	0.99	0.99	0.99
CRMA_v2		0.11	0.15	0.17	0.21		0.14	0.16	0.19	0.24		0.16	0.22	0.22	0.26		0.99	0.99	0.99	0.99
dChip		0.09	0.11	0.13	0.16		0.08	0.11	0.15	0.19		0.14	0.20	0.21	0.26		0.99	0.99	0.99	0.99
cn.FARMS	**6**	0.23	0.25	0.25	0.27	**12**	0.26	0.32	0.33	0.38	**18**	0.11	0.14	0.15	0.16	**all**	0.20	0.22	0.24	0.26
CRMA_v2		0.16	0.20	0.22	0.25		0.16	0.22	0.25	0.30		0.05	0.06	0.10	0.11		0.13	0.16	0.18	0.21
dChip		0.13	0.16	0.20	0.23		0.12	0.17	0.21	0.26		0.03	0.04	0.05	0.08		0.11	0.14	0.16	0.19

A larger value means that the method has lower FDR averaged over different given recall values. 'Chr' gives the chromosome; 'Area under the PRC for combined loci of' reports the area under the PRCs for different number of combined loci. Note, that large windows can increase the FDR again because CNV regions are overestimated. cn.FARMS clearly outperforms the other methods.

Computational complexity

Finally, we give the computation time for cn.FARMS, dChip and CRMA_v2. The required computation time can be an important factor for choosing an appropriate method because for many samples and large arrays (e.g. Affymetrix SNP 6.0 comprises 6.6 million probes), CNV analysis can take some hours. Table 3 shows the computational times for the compared methods. cn.FARMS requires less time than other methods. cn.FARMS's low computational load is due to the fact that FARMS's update rules both for single and multi-loci modeling are based on an EM algorithm which converges in only a few iterations.

DISCUSSION

Variation across samples versus variation across the chromosome

cn.FARMS identifies regions in the genome that have variable copy numbers across samples. If a CNV is found, it is straightforward to select the samples which caused the variation. In a next step (not considered here), integer copy numbers will be assigned by segmentation methods which find deviations along the chromosome. Thus, segmentation methods serve as a second filter which are able to sort out wrongly detected CNVs stemming from few high variable (noisy) or outlier samples. High variable samples inject locally variation across samples which may be detected by cn.FARMS as a CNV. However, if segmentation methods scan along a chromosome of a high variable sample, the local variation may be considered as being in the range of copy number two. Concluding, cn.FARMS finds variations across samples and segmentation finds variations across the chromosome—only locations having variations in both directions are finally considered as CNV regions.

Affymetrix Mapping250K_NSP to SNP 6.0 arrays

Affymetrix Mapping250K and 500K arrays contain only SNP probes which are allele A or allele B, strand or antistrand, prefect match or mismatch, shifted or not. In contrast to these arrays, Affymetrix SNP 6.0 arrays have, besides single CNV probes, for each SNP and allele three identical probes on one strand. One may think that single locus modeling is superfluous for SNP 6.0 arrays, but we observed that for SNP loci it still improves the results. Though the probes are identical, their fixed array location leads to consistent intensity differences which are captured by single locus modeling.

cn.FARMS for other platforms

Of course, cn.FARMS is not limited to the Affymetrix platform and can be applied to other platforms like Illumina bead arrays or Agilent arrays. The concept remains the same: do genomically adjacent measurements agree on copy numbers? If they contain variation, then the more they agree to each other, the more confident cn.FARMS is in its copy number estimates.

Combining array types and platforms

cn.FARMS can integrate a mixture of arrays or a mixture of platforms if normalization is done carefully to make

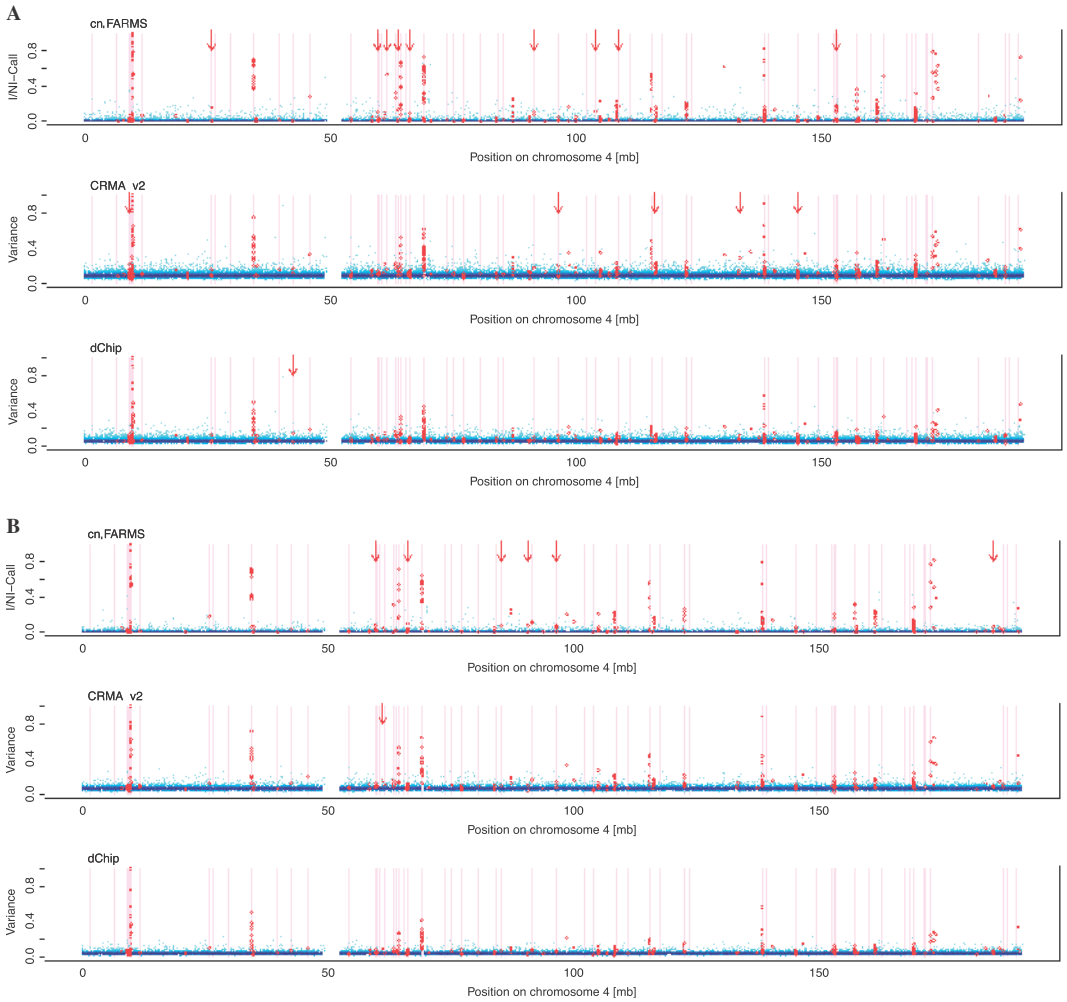

Figure 6. (**A**) CNV calling plots across chromosome 4 for 3 loci regions (each point in the plot summarizes 3 loci). The *y*-axis gives the I/NI call estimated by cn.FARMS and for both CRMA_v2 and dChip it gives the variance. Calling values are scaled such that the maximum is one. Local calling densities are encode by blue color shades. True CNVs [reported in Conrad *et al.* (38)] are marked as light rose bars and calls at these loci by red circles. A perfect calling method would call all true CNVs (red circles at 1) and does not call others (dark blue background at 0). cn.FARMS separates called true positives (true CNVs) from true negatives better than other methods which can be seen at less variance in true negatives indicated by dark blue density at the bottom. The red arrows, e.g. at positions 65 or 85 Mb in the upper cn.FARMS panel, indicate verified CNVs which were detected by one method, in this case cn.FARMS, but not by both others. cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip. (**B**) The same plot for 5 loci (each point in the plot summarizes 5 loci). The FDR is further reduced, as can be seen by the lower variance of non-call values at the bottom. Again, cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip.

single arrays comparable. We created metaprobe sets for the Affymetrix 500K where the metaprobes for one set are from both the 250K_NSP and the 250K_STY array. Metaprobe sets can in principle consist of metaprobes from different platforms like from Affymetrix and Illumina. The combination of meta-probes across array types or platforms has the advantage that it increases the resolution and coverage, but, on the other hand, it may introduce between array type or between platform variations. It may even be possible to combine array metaprobes with metaprobes obtained from next-generation sequencing (NGS). To provide these NGS metaprobes, we currently work on adapting the idea of cn.FARMS to NGS data by a mixture of Poissons model.

Table 3. Computational time (in seconds) to process 60 Mapping250K_NSP arrays, respectively SNP 6.0. cn.FARMS, is faster than CRMA and dChip

time (s)	cn.FARMS	CRMA_v2	dChip
60 250K	**1055**	3363	2493
60 SNP 6.0	**3657**	11 850	6210

CONCLUSION

We introduced a novel method for detecting CNVs called 'cn.FARMS' which controls the FDR. In experiments, cn.FARMS outperformed its competitors both with respect to FDR and sensitivity, i.e. has fewer false positives while detecting more true CNVs. The reduced FDR increases the discovery power of studies and avoids that researchers are misguided by spurious correlations between CNVs and diseases.

ACKNOWLEDGEMENTS

The authors thank Dr Ulrich Bodenhofer for helpful discussion and comments.

FUNDING

Funding for open access charge: Funds from the Institute of Bioinformatics, Johannes Kepler University Linz.

Conflict of interest statement. None declared.

REFERENCES

1. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
2. Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet.*, **39**, S30–S36.
3. Jakobsson,M., Scholz,S.W., Scheet,P., Gibbs,J.R., VanLiere,J.M., Fung,H.-C., Szpiech,Z.A., Degnan,J.H., Wang,K., Guerreiro,R. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.
4. Fanciulli,M., Norsworthy,P.J., Petretto,E., Dong,R., Harper,L., Kamesh,L., Heward,J.M., Gough,S.C., deSmith,A., Blakemore,A.I. *et al.* (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.
5. Gonzalez,E., Kulkarni,H., Bolivar,H., Mangano,A., Sanchez,R., Catano,G., Nibbs,R.J., Freedman,B.I., Quinones,M.P., Bamshad,M.J. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.
6. Wellcome-Trust-Case-Control-Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.
7. Scott,L.J., Mohlke,K.L., Bonnycastle,L.L., Willer,C.J., Li,Y., Duren,W.L., Erdos,M.R., Stringham,H.M., Chines,P.S., Jackson,A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
8. Zeggini,E., Weedon,M.N., Lindgren,C.M., Frayling,T.M., Elliott,K., Lango,H., Timpson,N., Perry,J., Rayner,N., Freathy,R. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.
9. Frayling,T.M. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.*, **8**, 657–662.
10. Estivill,X. and Armengol,L. (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.*, **3**, e190.
11. Jacobs,S., Thompson,E.R., Nannya,Y., Yamamoto,G., Pillai,R., Ogawa,S., Bailey,D.K. and Campbell,I.G. (2007) Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer Res.*, **67**, 2544–2551.
12. Tuefferd,M., Bondt,A.D., Wyngaert,I.V.D., Talloen,W., Verbeke,T., Carvalho,B., Clevert,D.-A., Alifano,M., Raghavan,N., Amaratunga,D. *et al.* (2008) Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays. *Genes Chromosomes Cancer*, **47**, 957–964.
13. Baross,A., Delaney,A., Li,I.H., Nayar,T., Flibotte,S., Qian,H., Chan,S., Asano,J., Ally,A., Cao,M. *et al.* (2007) Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics*, **8**, 368.
14. Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
15. Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C. *et al.* (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.*, **65**, 6071–6079.
16. Van de Wiel,M.A., Picard,F., van Wieringen,W.N. and Ylstra,B. (2010) Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief. Bioinformatics*, **12**, 10–21.
17. Affymetrix. (2007) CNAT 4.0: Copy number and loss of heterozygosity estimation algorithms for the GeneChip human mapping 10/50/100/250/500K array set. *Technical report*. Affymetrix Inc.
18. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
19. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
20. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, 1–8.
21. Lin,M., Wei,L.-J., Sellers,W.R., Lieberfarb,M., Wong,W.H. and Li,C. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.
22. Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
23. Bengtsson,H., Irizarry,R., Carvalho,B. and Speed,T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.
24. Bengtsson,H., Wirapati,P. and Speed,T.P. (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**, 2149–2156.
25. Dellinger,A.E., Saw,S.-M., Goh,L.K., Seielstad,M., Young,T.L. and Li,Y.-J. (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.*, **38**, e105.
26. Hupe,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA region. *Bioinformatics*, **20**, 3413–3422.

27. Broët,P. and Richardson,S. (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**, 911–918.
28. Andersson,R., Bruder,C.E.G., Piotrowski,A., Menzel,U., Nord,H., Sandgren,J., Hvidsten,T.R., deStähl,T.D., Dumanski,J.P. and Komorowski,J. (2008) A segmental maximum a posteriori approach to genome-wide copy number profiling. *Bioinformatics*, **24**, 751–758.
29. Wang,K., Li,M., Hadley,D., Liu,R., Glessner,J., Grant,S., Hakonarson,H. and Bucan,M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.
30. Scharpf,R.B., Parmigiani,G., Pevsner,J. and Ruczinski,I. (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Stat.*, **2**, 687–713.
31. Hochreiter,S., Clevert,D.-A. and Obermayer,K. (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.
32. Talloen,W., Clevert,D.-A., Hochreiter,S., Amaratunga,D., Bijnens,L., Kass,S. and Göhlmann,H.W.H. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.
33. Talloen,W., Hochreiter,S., Bijnens,L., Kasim,A., Shkedy,Z. and Amaratunga,D. (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl Acad. Sci. USA*, **107**, 173–174.
34. Olshausen,B.A. and Field,D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609.
35. Lewicki,M.S. and Sejnowski,T.J. (2000) Learning overcomplete representations. *Neural Comput.*, **12**, 337–365.
36. Girolami,M. (2001) A variational method for learning sparse and overcomplete representations. *Neural Comput.*, **13**, 2517–2532.
37. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–22.
38. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
39. Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.
40. McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., deBakker,P.I.W., Maller,J.B., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
41. Bourgon,R., Gentleman,R. and Huber,W. (2010) Independent filtering increases detection power for high-throughput experiment. *Proc. Natl Acad. Sci. USA*, **107**, 9546–9551.

BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 12 2010, pages 1520–1527
doi:10.1093/bioinformatics/btq227

Gene expression

Advance Access publication April 23, 2010

FABIA: factor analysis for bicluster acquisition

Sepp Hochreiter[1,*], Ulrich Bodenhofer[1], Martin Heusel[1], Andreas Mayr[1], Andreas Mitterecker[1], Adetayo Kasim[2], Tatsiana Khamiakova[2], Suzy Van Sanden[2], Dan Lin[2], Willem Talloen[3], Luc Bijnens[3], Hinrich W. H. Göhlmann[3], Ziv Shkedy[2] and Djork-Arné Clevert[1,4]

[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria, [2]Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, [3]Johnson & Johnson Pharmaceutical Research & Development, Division of Janssen Pharmaceutica, Beerse, Belgium and [4]Department of Nephrology and Internal Intensive Care, Charité, Berlin, Germany

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Biclustering of transcriptomic data groups genes and samples simultaneously. It is emerging as a standard tool for extracting knowledge from gene expression measurements. We propose a novel generative approach for biclustering called 'FABIA: Factor Analysis for Bicluster Acquisition'. FABIA is based on a multiplicative model, which accounts for linear dependencies between gene expression and conditions, and also captures heavy-tailed distributions as observed in real-world transcriptomic data. The generative framework allows to utilize well-founded model selection methods and to apply Bayesian techniques.

Results: On 100 simulated datasets with known true, artificially implanted biclusters, FABIA clearly outperformed all 11 competitors. On these datasets, FABIA was able to separate spurious biclusters from true biclusters by ranking biclusters according to their information content. FABIA was tested on three microarray datasets with known subclusters, where it was two times the best and once the second best method among the compared biclustering approaches.

Availability: FABIA is available as an R package on Bioconductor (http://www.bioconductor.org). All datasets, results and software are available at http://www.bioinf.jku.at/software/fabia/fabia.html

Contact: hochreit@bioinf.jku.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 23, 2009; revised on March 26, 2010; accepted on April 20, 2010

1 INTRODUCTION

Recent technologies such as the Affymetrix array plates and next-generation sequencing open up new possibilities for high-throughput expression profiling. These technologies in turn require advanced analysis tools to extract knowledge from the huge amount of data. If the experimental conditions are known, supervised techniques such as support vector machines are suitable to extract the dependencies between conditions and gene expression or to identify condition-indicative genes. However, conditions may not be known or biologists and medical researchers are interested in dependencies within or across conditions. For instance, it could be possible to refine pathways across conditions or to identify new subgroups within one condition. For these tasks, unsupervised methods such as clustering are required, which are usually insufficient, because samples may only be similar on a subset of genes and vice versa. In drug design, for example, researchers want to reveal how compounds affect gene expression; the effects of compounds, however, may be similar only on a subgroup of genes. Under such circumstances, *biclustering* is the proper unsupervised analysis technique.

A *bicluster* in a transcriptomic dataset is a pair of a gene set and a sample set for which the genes are similar to each other on the samples and vice versa. If multiple pathways are active in a sample, it belongs to different biclusters. If a gene participates in different pathways for different conditions, it belongs to different biclusters, too. Thus, biclusters can overlap.

A survey of biclustering approaches has been given by Madeira and Oliveira (2004). In principle, there exist four categories of biclustering methods: (1) variance minimization methods, (2) two-way clustering methods, (3) motif and pattern recognition methods and (4) probabilistic and generative approaches. Transcriptomic data are usually supplied as a matrix, where each gene corresponds to one row and each sample to one column; the matrix entries themselves are the expression levels.

(1) *Variance minimization methods*: define clusters as blocks in the matrix with minimal deviation of their elements. This definition has been already considered by Hartigan (1972) and extended by Tibshirani *et al.* (1999). The δ-cluster methods search for blocks of elements having a deviation ('variance') below δ. One example are δ-ks clusters (Califano *et al.*, 2000), where the maximum and the minimum of each row need to differ less than δ on the selected columns. A second example are δ-pClusters (Wang *et al.*, 2002), which are defined as 2×2 submatrices with pairwise edge differences less than δ. A third example are the Cheng and Church (2000) δ-biclusters having a mean squared error below δ after fitting an additive model with a constant, a row and a column effect. FLexible Overlapped biClustering (FLOC; Yang *et al.*, 2005) extend Cheng–Church δ-biclusters by dealing with missing values via an occupancy threshold θ and by using both l_1 and l_2 norms.

(2) *Two-way clustering methods* apply conventional clustering to the columns and rows and (iteratively) combine the results. Coupled Two-Way Clustering (CTWC; Getz *et al.*, 2000) iteratively

*To whom correspondence should be addressed.

performs standard clustering of the rows (columns) using previously constructed columns (rows) clusters as features. Also Interrelated Two-Way Clustering (ITWC; Tang *et al.*, 2001) using k-means and Double Conjugated Clustering (DCC; Busygin *et al.*, 2002) using self-organizing maps combine column and row clustering.

(3) *Motif and pattern recognition methods* define a bicluster as samples sharing a common pattern or motif. To simplify this task, some methods discretize the data in a first step, such as xMOTIF (Murali and Kasif, 2003) or Bimax (Prelic *et al.*, 2006), which even binarizes the data and searches for blocks with an enrichment of ones. Order-Preserving SubMatrices (OPSM; Ben-Dor *et al.*, 2003) searches for blocks having the same order of values in their columns. Using partial models, only the column order on subsets must be preserved. Spectral clustering (SPEC; Kluger *et al.*, 2003) performs a singular value decomposition of the data matrix after normalization. SPEC extracts columns (samples) with the same conserved gene expression pattern using the fact that they are linearly dependent and span a subspace associated with a certain singular value. The Iterative Signature Algorithm (ISA; Ihmels *et al.*, 2004) selects samples that have a given gene signature and then uses these samples to define a new sample signature. This sample signature, in turn, is used to select genes and to define a new gene signature. For each bicluster to be extracted, this process is initialized by a randomly selected binary gene signature and repeated iteratively. A related approach uses a Hough transform for identifying groups of linearly dependent genes and samples (Gan *et al.*, 2008). Contiguous column coherent (CCC biclustering; Madeira and Oliveira, 2009; Madeira *et al.*, 2010) is a method for gene expression time series, which finds patterns in contiguous columns.

(4) *Probabilistic and generative methods* use model-based techniques to define biclusters. Statistical-Algorithmic Method for Bicluster Analysis (SAMBA; Tanay *et al.*, 2002) uses a bi-partitioned graph, where both conditions and genes are nodes. An edge from a gene to a condition means that the gene responds to the condition. With a probabilistic objective, subgraphs are found that have a significantly higher connectivity than the overall graph. In another approach, Sheng *et al.* (2003) use Gibbs sampling to estimate the parameters of a simple frequency model for the expression pattern of a bicluster. However, the data must first be discretized and then only one bicluster with constant column values at each step can be extracted. Probabilistic Relational Models (PRMs; Getoor *et al.*, 2002) and their extension ProBic (Van den Bulcke, 2009) are fully generative models that combine probabilistic modeling and relational logic. Another generative approach is cMonkey (Reiss *et al.*, 2006), which models biclusters by Markov chain processes. Both PRMs and cMonkey are able to integrate non-transcriptomic data sources.

In the plaid model family (Lazzeroni and Owen, 2002), the i-th bicluster is extracted by row and column indicator variables ρ_{ki} and κ_{ij}. The values of each bicluster are explained by a general additive model $\theta_{kij} = \mu_i + \alpha_{ki} + \beta_{ij}$. Parameters are estimated by a least square fit. Gu and Liu (2008) generalized the plaid models to fully generative models called Bayesian BiClustering model (BBC). To avoid the high percentage of overlap in the plaid models, BBC constrains the overlapping of biclusters to only one dimension. Further it allows different error variances per bicluster. Caldas and Kaski (2008) also extended the plaid model to a fully generative model using a Bayesian framework and found that the plaid model is equivalent to the PRM model for specific parameters.

The latter models (Caldas and Kaski, 2008; Gu and Liu, 2008) are generative models which have the advantage that (i) they select models using well-understood model selection techniques such as maximum likelihood, (ii) hyperparameter selection methods (e.g. to determine the number of biclusters) can rely on the Bayesian framework, (iii) signal-to-noise ratios can be computed, (iv) they can be compared with each other via the likelihood or posterior, (v) tests such as likelihood ratio test are possible and (vi) they produce a global model to explain all data. These models are additive and assume that all effects are Gaussian to utilize Gibbs sampling for parameter estimation. However, after prefiltering, real microarray datasets are not Gaussian distributed and have heavy tails (Hardin and Wilson, 2009), even after log transformation. This can be seen in Supplementary Figures S8, S9 and S19 for gene expression datasets. In this article, we propose a *generative multiplicative model tailored to the special characteristics of gene expression data*.

This article is organized as follows. Section 2 introduces the multiplicative bicluster model class. Section 3 describes the model selection (training) algorithm for the new model class. Section 4 highlights how biclusters can be ranked according to the information they contained about the data. Section 5 describes how to extract bicluster members from our new models. Finally, Section 6 provides an experimental evaluation of the new method.

2 THE FABIA MODEL

We propose a multiplicative model class for analyzing gene expression datasets for several reasons. First, a multiplicative model allows to model heavy tailed data, as observed in gene expression. Second, it can relate the strength of gene expression patterns to characteristics of the induced condition such as elapsed time or concentration of compounds. After log transformation, exponential dynamics such as decay (mRNA or compound) or saturation can also be modeled. Note that supervised multiplicative models, e.g. support vector machines, were successfully applied to log-transformed gene expression datasets. Further, artificial multiplicative effects are introduced during data preprocessing, for example, if expression values are standardized, then variations stemming from noise scale the signal.

We assume that the gene expression dataset is preprocessed and filtered for genes that contain a signal (e.g. informative call or signal strength). The resulting data is given as a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times l}$, where every row corresponds to a gene and every column corresponds to a sample; the value x_{kj} corresponds to the expression level of the k-th gene in the j-th sample. The matrix $\boldsymbol{X}$ is the input to biclustering methods.

We define a *bicluster* as a pair of a row (gene) set and a column (sample) set for which the rows are similar to each other on the columns and vice versa. In a multiplicative model, two vectors are similar if one is a multiple of the other, that is, the angle between them is zero or, as realization of random variables, their correlation coefficient is (minus) one. It is clear that such a linear dependency on subsets of rows and columns can be represented as an outer product $\boldsymbol{\lambda}\, z^T$ of two vectors $\boldsymbol{\lambda}$ and z. The vector $\boldsymbol{\lambda}$ corresponds to a *prototype column vector* that contains zeros for genes not participating in the bicluster, whereas z is a vector of *factors* with which the prototype column vector is scaled for each sample; clearly z contains zeros for samples not participating in the bicluster. Vectors containing many zeros or values close to zero are called *sparse vectors*. Figure 1 visualizes this representation by sparse vectors schematically.

The overall model for p biclusters and additive noise is

$$\boldsymbol{X} = \sum_{i=1}^{p} \boldsymbol{\lambda}_i \, z_i^T + \boldsymbol{\Upsilon} = \boldsymbol{\Lambda}\, \boldsymbol{Z} + \boldsymbol{\Upsilon}\,, \tag{1}$$

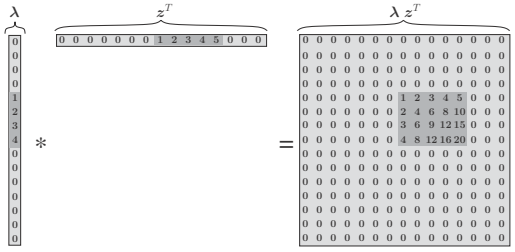

Fig. 1. The outer product $\boldsymbol{\lambda}\, z^T$ of two sparse vectors results in a matrix with a bicluster. Note that the non-zero entries in the vectors are adjacent to each other for visualization purposes only.

where $\boldsymbol{\Upsilon} \in \mathbb{R}^{n \times l}$ is additive noise; $\boldsymbol{\lambda}_i \in \mathbb{R}^n$ and $z_i \in \mathbb{R}^l$ are the sparse prototype vector and the sparse vector of factors of the i-th bicluster, respectively. The second formulation above holds if $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times p}$ is the sparse prototype matrix containing the prototype vectors $\boldsymbol{\lambda}_i$ as columns and $\boldsymbol{Z} \in \mathbb{R}^{p \times l}$ is the sparse factor matrix containing the transposed factors z_i^T as rows. Note that Equation (1) formulates biclustering as sparse matrix factorization.

According to Equation (1), the j-th sample $\boldsymbol{x}_j$, i.e. the j-th column of $\boldsymbol{X}$, is

$$\boldsymbol{x}_j = \sum_{i=1}^{p} \boldsymbol{\lambda}_i \, z_{ij} + \boldsymbol{\epsilon}_j = \boldsymbol{\Lambda} \, \tilde{z}_j + \boldsymbol{\epsilon}_j \,, \tag{2}$$

where $\boldsymbol{\epsilon}_j$ is the j-th column of the noise matrix $\boldsymbol{\Upsilon}$ and $\tilde{z}_j = (z_{1j}, \ldots, z_{pj})^T$ denotes the j-th column of the matrix $\boldsymbol{Z}$. Recall that $z_i^T = (z_{i1}, \ldots, z_{il})$ is the vector of values that constitutes the i-th bicluster (one value per sample), while $\tilde{z}_j$ is the vector of values that contribute to the j-th sample (one value per bicluster).

The formulation in Equation (2) facilitates a generative interpretation by a factor analysis model with p factors (Everitt, 1984)

$$\boldsymbol{x} = \sum_{i=1}^{p} \boldsymbol{\lambda}_i \, \tilde{z}_i + \boldsymbol{\epsilon} = \boldsymbol{\Lambda} \, \tilde{z} + \boldsymbol{\epsilon} \,, \tag{3}$$

where $\boldsymbol{x}$ is the observation, $\boldsymbol{\Lambda}$ is the loading matrix, $\tilde{z}_i$ is the value of the i-th factor, $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_p)^T$ is the vector of factors and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the additive noise. Standard factor analysis assumes: the noise is independent of $\tilde{z}$, $\tilde{z}$ is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$-distributed and $\boldsymbol{\epsilon}$ is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$-distributed (the covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ is diagonal—expressing independent Gaussian noise). The parameter $\boldsymbol{\Lambda}$ explains the dependent (common) and $\boldsymbol{\Psi}$ the independent variance in the observations $\boldsymbol{x}$. Additive noise in gene expression is normally distributed (Hochreiter *et al.*, 2006).

That the covariance matrix for $\tilde{z}$ is the unit matrix means that the biclusters should not be correlated. This assumption ensures that one true bicluster in the data will not be divided into dependent small model biclusters—thereby ensuring maximal model biclusters. Note, however, that this assumption still allows for overlapping biclusters.

Standard factor analysis does not consider sparse factors and sparse loadings that are essential in our formulation to represent biclusters. Sparseness is obtained by a component-wise independent *Laplace* distribution (Hyvärinen and Oja, 1999), which is now used as a prior on the factors $\tilde{z}$ instead of the Gaussian:

$$p(\tilde{z}) = \left(\tfrac{1}{\sqrt{2}}\right)^p \prod_{i=1}^{p} e^{-\sqrt{2}\,|\tilde{z}_i|}$$

Sparse loadings $\boldsymbol{\lambda}_i$ and, therefore sparse $\boldsymbol{\Lambda}$, are achieved by two alternative strategies. In the first model, called FABIA, we assume a component-wise independent *Laplace* prior for the loadings (like for the factors):

$$p(\boldsymbol{\lambda}_i) = \left(\tfrac{1}{\sqrt{2}}\right)^n \prod_{k=1}^{n} e^{-\sqrt{2}\,|\lambda_{ki}|} \tag{4}$$

The FABIA model contains the product of Laplacian variables which is distributed proportionally to the 0th order modified Bessel function of the second kind (Bithas *et al.*, 2007). For large values, this Bessel function is a negative exponential function of the square root of the random variable. Therefore, the tails of the distribution are heavier than those of the Laplace distribution. The Gaussian noise, however, reduces the heaviness of the tails such that the heaviness is between Gaussian and Bessel function tails—about as heavy as the tails of the Laplacian distribution. These *heavy tails* are exactly the desired model characteristics.

The second model, called FABIAS, uses a prior distribution for the loadings that is non-zero only in regions where the loadings are sparse. Following (Hoyer, 2004), we define sparseness as

$$\mathrm{sp}(\boldsymbol{\lambda}_i) = \frac{\sqrt{n} - \sum_{k=1}^{n} |\lambda_{ki}| \;/\; \sum_{k=1}^{n} \lambda_{ki}^2}{\sqrt{n} - 1}$$

leading to the prior with parameter spL

$$p(\boldsymbol{\lambda}_i) = \begin{cases} c & \text{for} \quad \mathrm{sp}(\boldsymbol{\lambda}_i) \leq \mathrm{spL} \\ 0 & \text{for} \quad \mathrm{sp}(\boldsymbol{\lambda}_i) > \mathrm{spL} \end{cases} . \tag{5}$$

Relation to Independent Component Analysis (ICA): our models are closely related to ICA (Hyvärinen, 1999). ICA searches for a matrix factorization, where the components of $\tilde{z}$ in model Equation (3) without noise $\boldsymbol{\epsilon}$ should be mutually independent. The matrix decomposition for ICA is

$$\boldsymbol{X} = \boldsymbol{\Lambda}_{\mathrm{ICA}} \, \boldsymbol{Z}_{\mathrm{ICA}}, \;\; \text{where} \;\; \boldsymbol{Z}_{\mathrm{ICA}} \, \boldsymbol{Z}_{\mathrm{ICA}}^T = \boldsymbol{I} \,.$$

ICA results in sparse $\boldsymbol{Z}_{\mathrm{ICA}}$, whereas $\boldsymbol{\Lambda}_{\mathrm{ICA}}$ is not sparse as in our models.

3 MODEL SELECTION

To identify the biclusters, we have to select the model parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ that explain the data best. Maximum likelihood is the most common approach for selecting a generative model. Unfortunately, in our case, the likelihood is analytically intractable. The reason is that we aim at generating sparse values, for which we use Laplacian priors (in contrast to the commonly used Gaussian priors). The resulting integral defining the likelihood cannot be computed analytically. In such situations, variational approaches can be applied, where a lower bound of the likelihood is maximized instead of the likelihood itself.

Expectation maximization (EM; Dempster *et al.*, 1977) is the most popular method for maximizing the likelihood. The EM algorithm has been extended to variational EM (Girolami, 2001; Palmer *et al.*, 2006). We follow this approach. However, we also assume a prior on the loadings in order to make the loadings sparse as well. Therefore, we use variational EM for maximizing the posterior—in line with our previous approaches (Hochreiter *et al.*, 2006; Talloen *et al.*, 2007).

3.1 Variational approach for sparse factors

As mentioned above, the likelihood

$$p(\boldsymbol{x} \,|\, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \int p(\boldsymbol{x} \,|\, \tilde{z}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) \; p(\tilde{z}) \; \mathrm{d}\tilde{z}$$

cannot be computed analytically for a Laplacian prior $p(\tilde{z})$. Girolami (2001) introduces a model family that is parameterized by $\boldsymbol{\xi}$, where the maximum over models in this family is the true likelihood:

$$\underset{\boldsymbol{\xi}}{\arg\max} \; p(\boldsymbol{x} | \boldsymbol{\xi}) = p(\boldsymbol{x}) \,.$$

The variational EM algorithm does not only maximize the lower bound on the likelihood with respect to the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, but also with respect to the variational parameter $\boldsymbol{\xi}$.

In the following, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ denote the parameter estimates in the current iteration. According to Girolami (2001) and Palmer *et al.* (2006), we obtain

the following variational E-step:

$$\mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j) = \left(\boldsymbol{\Lambda}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda} + \boldsymbol{\Xi}_j^{-1}\right)^{-1}\boldsymbol{\Lambda}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{x}_j \quad \text{and}$$

$$\mathrm{E}(\tilde{z}_j\,\tilde{z}_j^T\,|\,\boldsymbol{x}_j) = \left(\boldsymbol{\Lambda}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda} + \boldsymbol{\Xi}_j^{-1}\right)^{-1} + \mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j)\,\mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j)^T\,,$$

where $\boldsymbol{\Xi}_j$ stands for $\mathrm{diag}(\boldsymbol{\xi}_j)$. The update for $\boldsymbol{\xi}_j$ is

$$\boldsymbol{\xi}_j = \mathrm{diag}\left(\sqrt{\mathrm{E}(\tilde{z}_j\,\tilde{z}_j^T\,|\,\boldsymbol{x}_j)}\right).$$

3.2 New update rules for sparse loadings

The M-step for FABIA (Laplace prior on loadings) is

$$\boldsymbol{\Lambda}^{\mathrm{new}} = \frac{\frac{1}{l}\sum_{j=1}^{l}\boldsymbol{x}_j\,\mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j)^T - \frac{\alpha}{l}\,\boldsymbol{\Psi}\,\mathrm{sign}(\boldsymbol{\Lambda})}{\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}(\tilde{z}_j\,\tilde{z}_j^T\,|\,\boldsymbol{x}_j)} \tag{6}$$

$$\mathrm{diag}\left(\boldsymbol{\Psi}^{\mathrm{new}}\right) = \boldsymbol{\Psi}^{\mathrm{EM}} + \mathrm{diag}\left(\frac{\alpha}{l}\,\boldsymbol{\Psi}\,\mathrm{sign}(\boldsymbol{\Lambda})(\boldsymbol{\Lambda}^{\mathrm{new}})^T\right), \quad \text{where}$$

$$\boldsymbol{\Psi}^{\mathrm{EM}} = \mathrm{diag}\left(\frac{1}{l}\sum_{j=1}^{l}\boldsymbol{x}_j\boldsymbol{x}_j^T - \boldsymbol{\Lambda}^{\mathrm{new}}\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j)\,\boldsymbol{x}_j^T\right).$$

The M-step for FABIAS updates $\mathrm{diag}(\boldsymbol{\Psi}^{\mathrm{new}}) = \boldsymbol{\Psi}^{\mathrm{EM}}$ and $\boldsymbol{\Lambda}$ according to the standard EM. However, we must take into account that the prior on $\boldsymbol{\lambda}_i$ has restricted support. This is ensured by a projection of $\boldsymbol{\lambda}_i$ according to Hoyer (2004). The projection is a convex quadratic problem, which minimizes the Euclidean distance to the original vector subject to $\|\boldsymbol{\lambda}_i\| = 1$ and $\mathrm{sp}(\boldsymbol{\lambda}_i) = \mathrm{spL}$, see Equation (5). The final update is

$$\boldsymbol{\Lambda}^{\mathrm{new}} = \mathrm{proj}\left(\frac{\frac{1}{l}\sum_{j=1}^{l}\boldsymbol{x}_j\,\mathrm{E}(\tilde{z}_j\,|\,\boldsymbol{x}_j)^T}{\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}(\tilde{z}_j\,\tilde{z}_j^T\,|\,\boldsymbol{x}_j)}, \mathrm{spL}\right).$$

For $n > p$, the algorithm has a complexity of $O(lp^2n)$ per iteration, i.e. it is linear in n and l.

3.3 Extremely sparse priors

Some microarray data are extremely sparse. For example, we observed a kurtosis larger than 30 for Affymetrix SNP 6 arrays [see copy number variation (CNV) data on FABIA homepage]. We want to generalize our model class to deal with such sparse datasets and define extremely sparse priors both on the factors and the loadings utilizing the following (pseudo) distributions:

Generalized Gaussians:	$p(z) \propto \exp\left(-\,\lvert z\rvert^{\beta}\right)$ *for* $0 < \beta \leq 1$
Jeffrey's prior:	$p(z) \propto \exp\left(-\,\ln\lvert z\rvert\right) = 1/\lvert z\rvert$
Improper prior:	$p(z) \propto \exp\left(\lvert z\rvert^{-\beta}\right)$ *for* $0 < \beta$

The latter may only exist on an interval $[\epsilon, a]$ with sufficiently small ϵ.

For updating the *loadings* in the M-step, we need the derivatives of the negative log-priors, which can be expressed proportionally to $|z|^{-\mathrm{spl}}$ for a specific exponent spl, where $\mathrm{spl} = 0$ ($\beta = 1$) corresponds to the Laplace prior and $\mathrm{spl} > 0$ to sparser priors. The M-step for the loadings is finally as in Equation (6), where $\mathrm{sign}(\boldsymbol{\Lambda})$ is replaced by $|\boldsymbol{\Lambda}|^{-\mathrm{spl}}\,\mathrm{sign}(\boldsymbol{\Lambda})$ with element-wise operations (absolute value, sign, exponentiation and multiplication).

For the *factors*, we represent the priors by a convex variational form. According to Palmer *et al.* (2006), this is possible if $g(z) = -\ln p(\sqrt{z})$ is increasing and concave for $z > 0$. Our priors fulfill this, because first-order derivatives are positive and second-order derivatives are negative. Then the update for the variational parameter $\boldsymbol{\xi}_j$ is

$$\boldsymbol{\xi}_j \propto \mathrm{diag}\left(\mathrm{E}(\tilde{z}_j\,\tilde{z}_j^T\,|\,\boldsymbol{x}_j)^{\mathrm{spz}}\right)$$

where spz is the exponent of $|z|$ in the first derivative of $g(z)$; $\mathrm{spz} = 1/2$ ($\beta = 1$) represents the Laplace prior and $\mathrm{spz} > 1/2$ leads to sparser priors.

3.4 Data preprocessing and initialization

The data should be centered to zero mean, zero median or zero mode (Supplementary Material). If the correlation of weak signals is of interest too, we recommend to normalize the data.

The iterative model selection procedure requires initialization of the parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\xi}_j$. We initialize the variational parameter vectors $\boldsymbol{\xi}_j$ by ones, $\boldsymbol{\Lambda}$ randomly and $\boldsymbol{\Psi} = \mathrm{diag}(\max(\delta, \mathrm{covar}(\boldsymbol{x}) - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T))$.

4 INFORMATION CONTENT OF BICLUSTERS

A highly desired property for biclustering algorithms is the ability to rank the extracted biclusters analogously to principal component which are ranked according to the data variance they explain. We rank biclusters according to the information they contain about the data. The information content of $\tilde{z}_j$ for the j-th observation $\boldsymbol{x}_j$ is the mutual information between $\tilde{z}_j$ and $\boldsymbol{x}_j$ as

$$\mathrm{I}(\boldsymbol{x}_j; \tilde{z}_j) = \mathrm{H}(\tilde{z}_j) - \mathrm{H}(\tilde{z}_j\,|\,\boldsymbol{x}_j) = \tfrac{1}{2}\ln\left|\boldsymbol{I}_p + \boldsymbol{\Xi}_j\,\boldsymbol{\Lambda}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda}\right|,$$

where H is the entropy. The independence of $\boldsymbol{x}_j$ and $\tilde{z}_j$ across j gives

$$\mathrm{I}(\boldsymbol{X}; \mathbf{Z}) = \tfrac{1}{2}\sum_{j=1}^{l}\ln\left|\boldsymbol{I}_p + \boldsymbol{\Xi}_j\,\boldsymbol{\Lambda}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda}\right|.$$

To assess the information content of one factor, we consider the case that factor $\tilde{z}_i$ is removed from the final model and, consequently, the explained covariance $\xi_{ji}\,\boldsymbol{\lambda}_i\,\boldsymbol{\lambda}_i^T$ must be considered as noise:

$$\boldsymbol{x}_j\,|\,(\tilde{z}_j \setminus z_{ij}) \sim \mathcal{N}\left(\boldsymbol{\Lambda}\,\tilde{z}_j|_{z_{ij}=0}\,,\ \boldsymbol{\Psi} + \xi_{ij}\,\boldsymbol{\lambda}_i\,\boldsymbol{\lambda}_i^T\right)$$

The information of z_{ij} given the other factors is

$$\begin{aligned}\mathrm{I}\left(\boldsymbol{x}_j; z_{ij}\,|\,(\tilde{z}_j \setminus z_{ij})\right) &= \mathrm{H}(z_{ij}\,|\,(\tilde{z}_j \setminus z_{ij})) - \mathrm{H}(z_{ij}\,|\,(\tilde{z}_j \setminus z_{ij}), \boldsymbol{x}_j) \\ &= \tfrac{1}{2}\ln\left(1 + \xi_{ij}\,\boldsymbol{\lambda}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}_i\right).\end{aligned}$$

Again independence across j gives

$$\mathrm{I}\left(\boldsymbol{X}; z_i^T\,|\,(\boldsymbol{Z} \setminus z_i^T)\right) = \tfrac{1}{2}\sum_{j=1}^{l}\ln\left(1 + \xi_{ij}\,\boldsymbol{\lambda}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}_i\right).$$

This information content gives that part of the information in $\boldsymbol{x}$ that z_i^T conveys across all examples. Note that the information content grows with the number of non-zero $\boldsymbol{\lambda}_i$'s (size of the bicluster).

5 EXTRACTING MEMBERS OF BICLUSTERS

After model selection and ranking of bicluster, the i-th bicluster has *soft gene memberships* given by the absolute values of $\boldsymbol{\lambda}_i$ and *soft sample memberships* given by the absolute values of z_i^T. Soft clustering has the advantage that gradual memberships are able to account for ambiguities that occur in gene expression datasets (where hard memberships can be obscured by noise). However, some applications require *hard 'yes/no' memberships*. We determine the members of the i-th bicluster by selecting absolute values λ_{ki} and z_{ij} above thresholds thresL and thresZ, respectively.

First, the second moment of each factor is normalized to 1 resulting in a factor matrix $\hat{\boldsymbol{Z}}$ [in accordance with $\mathrm{E}(\tilde{z}\tilde{z}^T) = \boldsymbol{I}$]. Consequently, $\boldsymbol{\Lambda}$ is rescaled to $\hat{\boldsymbol{\Lambda}}$ such that $\boldsymbol{\Lambda}\boldsymbol{Z} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{Z}}$. Now the threshold thresZ can be chosen to determine which percentage of samples will on average belong to a bicluster. For a Laplace prior, this percentage can be computed by $\frac{1}{2}\exp(-\sqrt{2}/\mathrm{thresZ})$.

We extract one bicluster for each factor $\hat{z}_i$. In gene expression, a gene pattern is either absent or present, but not negatively present. Therefore, the i-th bicluster is either determined by the positive or negative values of $\hat{z}_{ij}$. Which of these two possibilities is chosen is decided by whether the sum over $|\hat{z}_{ij}| > \mathrm{thresZ}$ is larger for the positive or negative $\hat{z}_{ij}$.

We may not normalize $\hat{\boldsymbol{\Lambda}}$ for extracting loadings, since the factors have been normalized already. We suggest to estimate the average contribution of

$\hat{\lambda}_{ki}\hat{z}_{ij}$ first. Therefore, we compute the standard deviation of $\hat{\mathbf{\Lambda}}\hat{\mathbf{Z}}$ by

$$\text{sdLZ} = \sqrt{\frac{1}{p\,l\,n}\sum_{(i,j,k)=(1,1,1)}^{(p,l,n)}\left(\hat{\lambda}_{ki}\,\hat{z}_{ij}\right)^2}\,.$$

Now we choose thresL = sdLZ/thresZ that corresponds to extracting those loadings which have an above-average contribution.

6 EXPERIMENTS

6.1 Evaluating biclustering results

Before comparing biclustering methods, we have to consider how to evaluate the performance of biclustering methods. If the true biclusters are known, the performance of a biclustering method should be evaluated by the consensus between the set of extracted biclusters and the set of true biclusters.

Previous consensus measures such as the one in Gu and Liu (2008) do not take overlapping biclusters into account. Other consensus measures do not consider the numbers of biclusters in both sets (e.g. Prelic *et al.*, 2006, Li *et al.*, 2009). Thus, the set of true biclusters would be in consensus with very large sets of random biclusters. We introduce a novel *consensus score* for two sets of biclusters which avoids the drawbacks mentioned above as follows:

(1) compute similarities between all pairs of biclusters, where one is from the first set and the other from the second set;

(2) assign the biclusters of one set to biclusters of the other set by maximizing the assignment by the Munkres algorithm (Munkres, 1957); and

(3) divide the sum of similarities of the assigned biclusters by the number of biclusters of the larger set.

Step (3) penalizes different numbers of biclusters as emphasized above.

We use the Jaccard index for computing the similarity of two biclusters. It measures the relative proportion of overlap of two biclusters as the quotient of the number of matrix elements contained in the intersection of the biclusters and the number of matrix elements contained in the union of the biclusters.

The highest consensus is 1 and only obtained for identical sets of biclusters. Further note that the consensus score defined above can be applied analogously to comparing standard clustering results.

6.2 Compared methods

We compare the following 13 biclustering methods:

(1) FABIA: our new method with sparse prior Equation (4).

(2) FABIAS: our new method with sparseness projection Equation (5).

(3) MFSC: matrix factorization with sparseness constraints (Hoyer, 2004).

(4) plaid: plaid model (Lazzeroni and Owen, 2002).

(5) ISA: Ihmels *et al.* (2004).

(6) OPSM: Ben-Dor *et al.* (2003).

(7) SAMBA: Tanay *et al.* (2002).

(8) xMOTIF: conserved motifs (Murali and Kasif, 2003).

(9) Bimax: divide-and-conquer algorithm (Prelic *et al.*, 2006).

(10) CC: Cheng–Church δ-biclusters (Cheng and Church, 2000).

(11) plaid_t: improved plaid model (Turner *et al.*, 2003)

(12) FLOC: a generalization of Cheng–Church δ-biclusters (Yang *et al.*, 2005).

(13) spec: spectral biclustering (Kluger *et al.*, 2003).

We used the following software: for (1)–(3) our R package 'fabia', for (4) the software http://www-stat.stanford.edu/∼owen/plaid/, for (5) the R package 'isa2', for (6) the software BicAT (Barkow *et al.*, 2006), for (7) the software EXPANDER (Shamir *et al.*, 2005), for (8)–(13) the R package 'biclust' (Kaiser and Leisch, 2008).

In all experiments, rows (genes) were standardized to mean 0 and variance 1. For a fair comparison, the parameters of the methods were optimized on auxiliary toy datasets. If more than one setting was close to the optimum, all near optimal parameter settings were tested. In the following, these variants are denoted as *method_variant* (e.g. plaid_ss). A complete list of all settings and variants is available in the Supplementary Material.

Among the compared methods, not only FABIA and FABIAS but also ISA, OPSM and SPEC are geared to identifying biclusters based on a multiplicative model. Additionally, we included MFSC, although it is not a biclustering method in the strict sense, but it is a standard method for multiplicative factorization and hence provides a baseline for our comparison.

6.3 Simulated datasets with known biclusters

Benchmark datasets published in Prelic *et al.* (2006) and Li *et al.* (2009) are small (50 to 100 genes), have low noise, equally sized biclusters, and only simultaneous row and column overlaps. FABIA performed very well on these datasets (see Supplementary, S6.3.1 and S6.3.2). However, we use more realistic simulated datasets that match the characteristics of gene expression data better, especially in terms of the heavy tails. This can be seen in the Supplementary Material by comparing the densities and moments of our simulated datasets (Supplementary Fig. S7) with real gene expression data (Supplementary Figs S8, S9 and S19).

We assumed $n = 1000$ genes and $l = 100$ samples and implanted $p = 10$ multiplicative biclusters with the model given by Equation (1).

The $\boldsymbol{\lambda}_i$'s are generated by (i) randomly choosing the number N_i^{λ} of genes in bicluster i from $\{10,\ldots,210\}$, (ii) choosing N_i^{λ} genes randomly from $\{1,\ldots,1000\}$, (iii) setting $\boldsymbol{\lambda}_i$ components not in bicluster i to $\mathcal{N}(0, 0.2^2)$ random values and (iv) setting $\boldsymbol{\lambda}_i$ components that are in bicluster i to $\mathcal{N}(\pm 3, 1)$ random values, where the sign is chosen randomly for each gene.

The $\boldsymbol{z}_i$'s are generated by (i) randomly choosing the number N_i^{z} of samples in bicluster i from $\{5,\ldots,25\}$, (ii) choosing N_i^{z} samples randomly from $\{1,\ldots,100\}$, (iii) setting $\boldsymbol{z}_i$ components not in bicluster i to $\mathcal{N}(0, 0.2^2)$ random values and (iv) setting $\boldsymbol{z}_i$ components that are in bicluster i to $\mathcal{N}(2, 1)$ random values.

Finally, we draw the $\boldsymbol{\Upsilon}$ entries (additive noise on all entries) according to $\mathcal{N}(0, 3^2)$ and compute the data $\boldsymbol{X}$ according to Equation (1). Using these settings, noisy biclusters of random sizes between 10×5 and 210×25 (genes×samples) are generated.

With this procedure, we created 100 independent datasets. Table 1 shows the biclustering results for these datasets. The methods are evaluated by the average consensus score of the extracted biclusters

Table 1. Results on the 100 simulated datasets

Method	Score	Method	Score
FABIA	*0.478* (1e-2)	SAMBA	0.006 (5e-5)
FABIAS	**0.564** (3e-3)	xMOTIF	0.002 (6e-5)
MFSC	0.057 (2e-3)	Bimax	0.004 (2e-4)
plaid_ss	0.045 (9e-4)	CC	0.001 (7e-6)
plaid_ms	0.072 (4e-4)	plaid_t_ab	0.046 (5e-3)
plaid_ms_5	0.083 (6e-4)	plaid_t_a	0.037 (4e-3)
ISA_1	0.333 (5e-2)	FLOC	0.006 (3e-5)
ISA_2	0.299 (6e-2)	spec_1	0.032 (5e-4)
ISA_3	0.188 (4e-2)	spec_2	0.011 (5e-4)
OPSM	0.012 (1e-4)		

The numbers denote average consensus scores with the true biclusters as defined in Section 6.1 (standard deviations in parentheses). The best results are highlighted in bold and the second best in italics ('better' means significantly better according to both a paired *t*-test and a McNemar test of correct elements in biclusters).

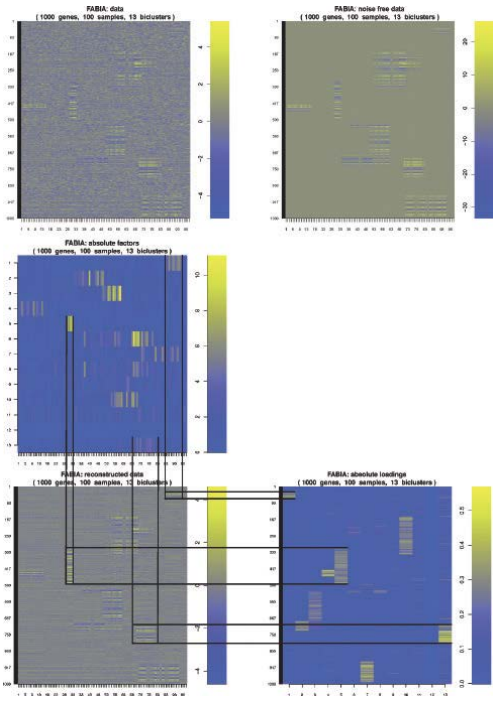

Fig. 2. An example of FABIA model selection. The data have 10 true biclusters. We have trained the model with 13 biclusters. Only for visualization purposes, the biclusters are generated as contiguous blocks. Top: data (left) and noise-free data (right). Middle: factors $\mathbf{Z}$. Bottom: data reconstructed by the FABIA model as $\mathbf{\Lambda}\,\mathbf{Z}$ (left) and loadings $\mathbf{\Lambda}$ (right). The lines indicate three biclusters and connect each bicluster in the reconstructed data with its corresponding factors (middle) and loadings (bottom right).

and the true biclusters as defined in Section 6.1. Our new methods FABIA and FABIAS outperform all other methods considerably.

Figure 2 illustrates a FABIA result on a simulated dataset, where, in contrast to our 100 benchmark datasets, the biclusters have been created as contiguous blocks for visualization purposes.

We observed the following characteristics of the methods, also confirming earlier findings of Gu and Liu (2008): SAMBA and OPSM excluded many relevant biclusters; SAMBA, Bimax, xMOTIF, CC and FLOC found many small random biclusters (overfitting). spec produces a partition of the samples for each gene set. The plaid models and ISA extract large overlapping clusters.

Ranking by information content: to verify that the information content is useful for ranking the extracted biclusters, we performed a two-sided Spearman rank correlation test comparing (i) the information content and (ii) the Jaccard similarity to the assigned true bicluster. We obtained *P*-values of 1.7×10^{-5} for FABIA and 6.1×10^{-3} for FABIAS, which shows that true biclusters can indeed be identified by their information content.

Data based on an additive model: we also generated data according to an additive model structure in order to analyze how well FABIA and FABIAS perform on data not satisfying the multiplicative model assumptions. We generated 100 datasets with the above settings, but using the general additive model from Section 1, category (4). Both FABIA and FABIAS outperform all other methods, followed by plaid_ms_5. Specifically, for three different signal levels, FABIAS gave average consensus scores of 0.15–0.27–0.55, FABIA 0.10–0.20–0.48 and plaid_ms_5 0.10–0.14–0.22 (detailed results, also for all other methods, are reported in the Supplementary Material). One would assume plaid methods to perform better than FABIA and FABIAS. We explain the superiority of our methods on datasets that do not even match the data generation model as follows: (i) they construct biclusters simultaneously, thereby, taking overlaps into account; (ii) the decorrelation of factors minimizes redundancy of biclusters; (iii) the low complexity of the model ensures low parameter interdependencies, which facilitates model selection.

6.4 Gene expression datasets

We consider three gene expression datasets that have been provided by the Broad Institute and were previously analyzed by Hoshida *et al.* (2007). They first clustered the samples using additional datasets and then confirmed the clusters by gene set enrichment analysis. Our goal was to study how well biclustering methods are able to re-identify these clusters without any additional information.

(A) The *'breast cancer' dataset* (van't Veer *et al.*, 2002) was aimed at a predictive gene signature for the outcome of a breast cancer therapy. We removed the outlier array S54 that leads to a dataset with 97 samples and 1213 genes. After standardization, skewness was 0.45 and excess kurtosis 0.93. In Hoshida *et al.* (2007), three biologically meaningful subclasses were found that should be re-identified.

(B) The *'multiple tissue types' dataset* (Su *et al.*, 2002) are gene expression profiles from human cancer samples from diverse tissues and cell lines. The dataset contains 102 samples with 5565 genes. After standardization, skewness was 0.15 and excess kurtosis 1.3. Biclustering should be able to re-identify the tissue types.

(C) The *'diffuse large-B-cell lymphoma (DLBCL)' dataset* (Rosenwald *et al.*, 2002) was aimed at predicting the survival

Table 2. Results on the breast cancer, multiple tissue samples, DLBCL datasets measured by the consensus score from Section 6.1

	Breast cancer				Multiple tissues				DLBCL			
Method	Score	#bc	#g	#s	Score	#bc	#g	#s	Score	#bc	#g	#s
FABIA	**0.52**	3	92	31	0.53	5	356	29	**0.37**	2	59	62
FABIAS	**0.52**	3	144	32	0.44	5	435	30	*0.35*	2	104	60
MFSC	0.17	5	87	24	0.31	5	431	24	0.18	5	50	42
plaid_ss	*0.39*	5	500	38	0.56	5	1903	35	0.30	5	339	72
plaid_ms	*0.39*	5	175	38	0.50	5	571	42	0.28	5	143	63
plaid_ms_5	0.29	5	56	29	0.23	5	71	26	0.21	5	68	47
plaid_a_ss	*0.37*	5	796	35	**0.65**	5	3711	31	0.28	5	389	68
plaid_a_ms	0.34	5	194	35	*0.58*	5	583	34	0.27	5	95	61
plaid_a_ms_5	0.16	5	5	26	0.20	5	11	25	0.18	5	4	68
ISA_1	0.03	25	55	4	0.05	29	230	6	0.01	56	26	8
ISA_2	0.25	2	466	42	0.37	3	1904	28	0.22	1	267	74
ISA_3	0.22	1	742	33	0.35	3	2856	28	0.18	2	385	58
OPSM	0.04	12	172	8	0.04	19	643	12	0.03	6	162	4
SAMBA	0.02	38	37	7	0.03	59	53	8	0.02	38	19	15
SAMBA_01	0.01	79	33	8	0.01	128	53	9	0.01	70	18	14
xMOTIF	0.07	5	61	6	0.11	5	628	6	0.05	5	9	9
Bimax	0.01	1	1213	97	0.10	4	35	5	0.07	5	73	5
CC	0.11	5	12	12	nc	nc	nc	nc	0.05	5	10	10
plaid_t_ab	0.24	2	40	23	0.38	5	255	22	0.17	1	3	44
plaid_t_a	0.23	2	24	20	0.39	5	274	24	0.11	3	6	24
spec_1	0.12	13	198	28	0.37	5	395	20	0.05	28	133	32
spec_2	0.07	14	77	22	0.21	1	117	39	0.08	8	82	44
FLOC	0.04	5	343	5	nc	nc	nc	nc	0.03	5	167	5

An 'nc' entry means that the method did not converge for this dataset. The best results are in bold and the second best in italics (again 'better' means significantly better according to a paired *t*-test). The columns '#bc', '#g' and '#s' provide the numbers of biclusters, their average numbers of genes and their average numbers of samples, respectively.

after chemotherapy. It contains 180 samples and 661 genes, and after standardization the skewness was -0.05 and excess kurtosis 0.35. The three classes found by Hoshida *et al.* (2007) should be re-identified.

The biclustering results are summarized in Table 2. For the methods assuming a fixed number of biclusters, we chose five biclusters—slightly higher than the number of known clusters to avoid biases toward prior knowledge about the number of actual clusters. The performance was assessed by comparing known classes of samples in the datasets with the sample sets identified by biclustering as defined in Section 6.1, in this case on sample clusters instead of biclusters. For the multiple tissue dataset, plaid performs best and our methods FABIA and FABIAS are second best. For breast cancer and DLBCL datasets, our new methods FABIA and FABIAS detected the clusters most accurately. Further, note that FABIA and FABIAS have considerably fewer genes in their bicluster than the next-best methods.

For the biological interpretation of the FABIA results, we applied gene ontology (GO), Kyoto encyclopedia of genes and genomes (KEGG) pathway and protein interaction network analysis. We provide a summary of these analysis results, details of which can be found in the Supplementary Material.

Breast cancer: GO and KEGG agree that genes in bicluster 1 are related to the cell cycle (KEGG *P*-value: 9.7×10^{-8}; GO *P*-value: 2.8×10^{-9}), especially to M-phase (GO *P*-value: 2.5×10^{-15}). Proteins which drive this bicluster are the cell division control protein CDC2 and the mitosis-related KIF proteins. Genes in bicluster 2 are related to immune response (GO *P*-value: 1.4×10^{-26}) and cytokine–cytokine receptor interaction (KEGG *P*-value $<10^{-10}$), involving cytokine-related proteins such as CCR5, CCL4 and CSF2RB. Note that cytokines are important regulators and mobilizers of the immune response. Bicluster 3 is too small to allow for a reliable biological interpretation.

DLBCL: the most significant GO terms and KEGG pathways found for bicluster 1 are related to the ribosome (GO *P*-value: 2.2×10^{-6}; KEGG *P*-value: 1.3×10^{-8}) and to B-cell receptor signaling (KEGG *P*-value: 9.6×10^{-8}). The latter fits especially well to the kind of cells the data stem from. The most significant GO terms and KEGG pathways for bicluster 2 are immune system-related (GO *P*-value: 3.2×10^{-6}; KEGG *P*-value: 5.7×10^{-8}).

Multiple tissues: this dataset is very heterogeneous and the samples differ in many biological processes; hence, it is difficult to provide a comprehensible biological interpretation.

6.5 Drug design

In a drug design project, Affymetrix GeneChip HT HG-U133+ PM array plates with 96 samples (12 × 8) per plate were used to analyze the effect of different compounds on gene expression. The compounds were selected to be active on a cancer cell line and were tested in groups of three replicates.

Raw expression data were summarized with FARMS (Hochreiter *et al.*, 2006) and informative genes are selected by I/NI calls (Talloen *et al.*, 2007). The preprocessed data matrix was 1413×95 (one array was missing) with skewness of -0.39 and excess kurtosis larger than 3.0 (i.e. heavier tails than Laplace). We tested FABIA on this dataset. Biclusters were extracted with thres$Z=1.5$ to obtain an average of 5–6 samples in a bicluster (note that, for the Laplacian prior, $\frac{1}{2}\exp(-\sqrt{2}\,1.5)\approx0.06$).

FABIA found four biclusters. The first bicluster consisted of two replicate sets (6 arrays), the second consisted of five replicate sets with one replicate missing (14 arrays). The third bicluster consisted of three replicate sets and an additional array (10 arrays). The fourth bicluster consisted of arrays located at the last column of the plate—corresponding to border arrays which dry out. In the meantime, this problem has been fixed by Affymetrix. That replicates are clustered together shows that our biclustering approach works correctly.

The bicluster with highest information content (two sets of replicates) extracted genes related to mitosis (GO analysis gave a *P*-value $<10^{-13}$). Regulation of mitosis genes is biologically plausible, as inhibiting cell division would be consistent with an active compound that does not kill the cell. The compounds of this bicluster are now under investigation by Johnson & Johnson Pharmaceutical Research & Development.

7 CONCLUSION

We have introduced a novel biclustering method that is a generative multiplicative model. It assumes realistic non-Gaussian signal distributions with heavy tails. The generative model allows to rank biclusters according to their information content. Model selection is performed by maximum *a posteriori* via an EM algorithm based on a variational approach.

On 100 simulated datasets with known true biclusters, FABIA clearly outperformed all 11 competing methods. On three gene expression datasets with previously verified subclusters, it was once

the second best and twice the best performing method. The biological relevance of the FABIA biclusters has been demonstrated by GO and KEGG analyses. Finally, FABIA has been successfully applied to drug design to find compounds with similar effects on gene expression.

Funding: Janssen Pharmaceutica N.V. and Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT project 80536).

Conflict of Interest: none declared.

REFERENCES

Barkow,S. *et al.* (2006) BicAT: a biclustering analysis toolbox. *Bioinformatics*, **22**, 1282–1283.

Ben-Dor,A. *et al.* (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, **10**, 373–384.

Bithas,P.S. *et al.* (2007) Distributions involving correlated generalized gamma variables. In *Proceedings of the International Conference on Applied Stochastic Models and Data Analysis*, vol. 12, Chania.

Busygin,S. *et al.* (2002) Double conjugated clustering applied to leukemia microarray data. In *Proceedings of the 2nd SIAM International Conference on Data Mining/Workshop on Clustering High Dimensional Data*, Arlington, VA, USA.

Caldas,J. and Kaski,S. (2008) Bayesian biclustering with the plaid model. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, vol. XVIII, Cancún, Mexico, pp. 291–296.

Califano,A. *et al.* (2000) Analysis of gene expression microarays for phenotype classification. In *Proceedings of the International Conference on Computational Molecular Biology*, ACM, Tokyo, Japan, pp. 75–85.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 8, ACM, Tokyo, Japan, pp. 93–103.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B Met.*, **39**, 1–22.

Everitt,B.S. (1984) *An Introduction to Latent Variable Models*. Chapman and Hall, London.

Gan,X. *et al.* (2008) Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics*, **9**, 209.

Getoor,L. *et al.* (2002) Learning probabilistic models of link structure. *J. Mach. Learn. Res.*, **3**, 679–707.

Getz,G. *et al.* (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Girolami,M. (2001) A variational method for learning sparse and overcomplete representations. *Neural Comput.*, **13**, 2517–2532.

Gu,J. and Liu,J.S. (2008) Bayesian biclustering of gene expression data. *BMC Genomics*, **9** (Suppl. 1), S4.

Hardn,J. and Wilson,J. (2009) A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, **10**, 446–450.

Hartigan,J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, **67**, 123–129.

Hochreiter,S. *et al.* (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Hoshida,Y. *et al.* (2007) Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS ONE*, **2**, e1195.

Hoyer,P.O. (2004) Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Hyvärinen,A. (1999) Survey on independent component analysis. *Neural Comput. Surv.*, **2**, 94–128.

Hyvärinen,A. and Oja,E. (1999) A fast fixed-point algorithm for independent component analysis. *Neural Comput.*, **9**, 1483–1492.

Ihmels,J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Kaiser,S. and Leisch,F. (2008) A toolbox for bicluster analysis in R. In Brito,P. (ed.) *Compstat 2008 – Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, pp. 201–208.

Kluger,Y. *et al.* (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**, 703–716.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, 61–86.

Li,G. *et al.* (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.*, **37**, e101.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE ACM Trans. Comput. Biol.*, **1**, 24–45.

Madeira,S.C. and Oliveira,A.L. (2009) A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithm Mol. Biol.*, **4**, 8.

Madeira,S.C. *et al.* (2010) Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE ACM Trans. Comput. Biol.*, **7**, 153–165.

Munkres,J. (1957) Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.*, **5**, 32–38.

Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. In *Pacific Symposium on Biocomputing*, Lihue, Hawaii, USA, pp. 77–88.

Palmer,J. *et al.* (2006) Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems 18*, The MIT Press, Vancouver, BC, Canada, pp. 1059–1066.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Reiss,D.J. *et al.* (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **2**, 280–302.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New Engl. J. Med.*, **346**, 1937–1947.

Shamir,R. *et al.* (2005) EXPANDER – an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.

Sheng,Q. *et al.* (2003) Biclustering micrarray data by Gibbs sampling. *Bioinformatics*, **19** (Suppl. 2), ii196–ii205.

Su,A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.

Talloen,W. *et al.* (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl. 1), S136–S144.

Tang,C. *et al.* (2001) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, IEEE Computer Society, Bethesda, MD, USA, pp. 41–48.

Tibshirani,R. *et al.* (1999) Clustering methods for the analysis of DNA microarray data. *Technical report*, Department of Health Research and Policy, Department of Genetics and Department of Biochemestry, Stanford University.

Turner,H. *et al.* (2003) Improved biclustering of microarray data demonstrated through systematic performance tests. *Comput. Stat. Data Anal.*, **48**, 235–254.

Van den Bulcke,T. (2009) *Robust Algorithms for Inferring Regulatory Networks Based on Gene Expression Measurements and Biological Prior Information*. PhD Thesis, Katholieke Universiteit Leuven, Lirias number: 236073.

van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wang,H. *et al.* (2002) Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, pp. 394–405.

Yang,J. *et al.* (2005) An improved biclustering method for analyzing gene expression profiles. *Int. J. Artif. Intell. T.*, **14**, 771–790.

GENES, CHROMOSOMES & CANCER 47:957–964 (2008)

Genome-Wide Copy Number Alterations Detection in Fresh Frozen and Matched FFPE Samples Using SNP 6.0 Arrays

Marianne Tuefferd,[1*†] An De Bondt,[2†] Ilse Van Den Wyngaert,[2] Willem Talloen,[2] Tobias Verbeke,[3] Benilton Carvalho,[4] Djork-Arne Clevert,[5,6] Marco Alifano,[1,7] Nandini Raghavan,[8] Dhammika Amaratunga,[8] Hinrich Göhlmann,[2‡] Philippe Broët,[1‡] and Sophie Camilleri-Broët[1,9‡]

[1]JE2492 Department, Faculté de Médecine Paris-Sud, IFR69, Villejuif, France
[2]Functional Genomics Department, Johnson & Johnson Pharmaceutical Research & Development, a Division of Janssen Pharmaceutica, Beerse, Belgium
[3]Business & Decision, Life Science Department, Benelux Division, Brussels, Belgium
[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[5]Institute of Bioinformatics, Johannes Kepler Universität Linz, Austria
[6]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany
[7]Department of Thoracic Surgery, Hôtel-Dieu Hospital, Paris V University, Paris, France
[8]Non-Clinical Biostatistics Department, Johnson & Johnson Pharmaceutical Research & Development, Raritan, NJ 08869, USA
[9]Faculté Paris-Descartes, Assistance Publique - Hôpitaux de Paris (AP-HP), Paris, France

SNP arrays offer the opportunity to get a genome-wide view on copy number alterations and are increasingly used in oncology. DNA from formalin-fixed paraffin-embedded material (FFPE) is partially degraded which limits the application of those technologies for retrospective studies. We present the use of Affymetrix GeneChip SNP6.0 for identification of copy number alterations in fresh frozen (FF) and matched FFPE samples. Fifteen pairs of adenocarcinomas with both frozen and FFPE embedded material were analyzed. We present an optimization of the sample preparation and show the importance of correcting the measured intensities for fragment length and GC-content when using FFPE samples. The absence of GC content correction results in a chromosome specific "wave pattern" which may lead to the misclassification of genomic regions as being altered. The highest concordance between FFPE and matched FF were found in samples with the highest call rates. Nineteen of the 23 high level amplifications (83%) seen using FF samples were also detected in the corresponding FFPE material. For limiting the rate of "false positive" alterations, we have chosen a conservative False Discovery Rate (FDR). We observed better results using SNP probes than CNV probes for copy number analysis of FFPE material. This is the first report on the detection of copy number alterations in FFPE samples using Affymetrix GeneChip SNP6.0.

INTRODUCTION

Copy number alterations (CNA) occur in most solid tumors (Weir et al., 2004; Kops et al., 2005; Albertson, 2006). Those chromosomal aberrations confer selective advantages to clonal expansion, being essential for tumorigenesis (Rajagopalan and Lengauer, 2004) and characterize clinical phenotype or cancer histological subtype (Thomas et al., 2006). Moreover, this genomic information can also be used for predicting response to targeted therapy, e.g., *HER2* amplification in breast cancer (Yaziji et al., 2004).

In recent years, high-resolution array-based comparative genomic hybridization (aCGH) has replaced conventional metaphase CGH, becoming the standard protocol for identifying subchromosomal regions that are over/under represented in the genome (Pinkel D et al., 1998; Pinkel and Albertson, 2005). The strategy of aCGH technique is to cohybridize genomic DNA from a tumor sample (labeled with one fluorochrome) with genomic DNA from a reference sample (non tumoral and labeled with a different fluorochrome) to the aCGH probes. These probes correspond to genomic clones (such as BAC, PAC, or cosmid clones) or nonoverlapping oligonucleotides of different length that are spotted or directly synthesized onto

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

[†,‡]These authors contributed equally to this work.

*Correspondence to: Tuefferd Marianne, JE2492, Faculté de Médecine Paris-Sud, IFR69, Villejuif, France, Europe.
E-mail: tuefferd@vjf.inserm.fr

Received 13 March 2008; Accepted 17 June 2008
DOI 10.1002/gcc.20599
Published online 28 July 2008 in Wiley InterScience (www.interscience.wiley.com).

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2897–2902
doi:10.1093/bioinformatics/btm478

Gene expression

I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data

Willem Talloen[1,*,†], Djork-Arné Clevert[2,3,†], Sepp Hochreiter[2], Dhammika Amaratunga[4], Luc Bijnens[1], Stefan Kass[1] and Hinrich W.H. Göhlmann[1]

[1]Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica n.v., Beerse, Belgium, [2]Institute of Bioinformatics, Johannes Kepler Universität Linz 4040 Linz, Austria, [3]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany and [4]Johnson & Johnson Pharmaceutical Research & Development, Raritan, USA

Received on April 13, 2007; revised on September 7, 2007; accepted on September 18, 2007

Advance Access publication October 5, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: DNA microarray technology typically generates many measurements of which only a relatively small subset is informative for the interpretation of the experiment. To avoid false positive results, it is therefore critical to select the informative genes from the large noisy data before the actual analysis. Most currently available filtering techniques are supervised and therefore suffer from a potential risk of overfitting. The unsupervised filtering techniques, on the other hand, are either not very efficient or too stringent as they may mix up signal with noise. We propose to use the multiple probes measuring the same target mRNA as repeated measures to quantify the signal-to-noise ratio of that specific probe set. A Bayesian factor analysis with specifically chosen prior settings, which models this probe level information, is providing an objective feature filtering technique, named informative/non-informative calls (I/NI calls).

Results: Based on 30 real-life data sets (including various human, rat, mice and Arabidopsis studies) and a spiked-in data set, it is shown that I/NI calls is highly effective, with exclusion rates ranging from 70% to 99%. Consequently, it offers a critical solution to the curse of high-dimensionality in the analysis of microarray data.

Availability: This filtering approach is publicly available as a function implemented in the R package FARMS (www.bioinf.jku.at/software/farms/farms.html).

Contact: wtalloen@prdbe.jnj.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

High-density oligonucleotide microarrays, and in particular Affymetrix GeneChip arrays (Lockhart *et al.*, 1996), are now fruitfully being used in many areas of biomedical research. The wealth of information generated by this DNA microarray technology is key to its power and success, but also constitutes its major weakness. The large number of gene expression comparisons between experimental groups, combined with the commonly present noisy genes showing irrelevant variation, leads to false positives in the identification of truly differentially expressed genes (Dudoit *et al.*, 2003) and increases the risk of overfitting in classification methods (Bellman, 1961). Ideally, the high-dimensionality of microarray data should be reduced before the actual analysis by excluding all the non-informative genes. This need for suitable data reduction approaches resulted in the development of many feature selection methods to separate signal from noise, i.e. the informative from the non-informative genes. Most selection algorithms are supervised like the various methods implemented within classification algorithms (Vapnik, 2000), and the ranking of genes on fold changes or test-statistics. As supervised feature selection approaches often suffer from overfitting (Varshavsky *et al.*, 2006) and selection bias (Ambroise and McLachlan, 2002), unsupervised feature filtering techniques started to emerge, like ranking of features on variation (Herrero *et al.*, 2003), principal components (Hastie *et al.*, 2000) or SVD-entropy (Varshavsky *et al.*, 2006). But still, these filtering techniques are based on assumptions that are not necessarily universally valid, and therefore still can distort the subsequent statistical analyses. This is unfortunate, as unsupervised filtering increases the significance level of the final result after multiple testing correction (Dudoit *et al.*, 2003) because genes are excluded without looking at the label.

Making use of domain knowledge, when available, is key in feature selection (Guyon and Elisseeff, 2003). Affymetrix microarray chips consist of probes that are designed to interrogate how much of the transcript sequence complementary to its DNA sequence is present in a sample (Lockhart *et al.*, 1996). They also provide the opportunity to assess whether or not genes were detected in every array. This is because each target transcript is probed by a pair of oligonucleotides; a perfect match (PM) measuring the target mRNA concentration, and a mismatch (MM) for background measurement (Affymetrix, 2002). The difference between PM and MM is used to determine whether the transcript was

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

detected (present) or not (absent). This lies at the base of one of the most objective filtering techniques, namely absent/present calls (A/P calls, Liu *et al.*, 2002). Although this method is not very efficient in filtering (McClintick and Edenberg, 2006), it is complementary to all feature filtering techniques mentioned above and therefore one of the most commonly applied. Another feature of Affymetrix microarray chips is that each target transcript is represented by 11–20 different probe pairs. The intensities of these probes are typically summarized for each probe set to provide one expression level for the respective target transcript (Wu and Irizarry, 2004). Summarization prevents the use of the information provided by the probes on the noise level of the probe set. In this article, we use this information in a rigorous way to assess whether the probe set will be informative for subsequent analyses or not.

This article introduces the concept and applicability of informative/non-informative calls (I/NI calls). I/NI-calls is—like A/P calls—more objective than, and completely complementary to, the existing filtering techniques. We demonstrate that I/NI-calls is a very stringent gene filtering tool using a spiked-in data set and 30 real-life data sets, and illustrate the consequences of I/NI calls on tests for differential expression.

2 METHODS

2.1 The model

I/NI calls expands upon the algorithm used in factor analysis for robust microarray summarization (FARMS) (Hochreiter *et al.*, 2006). FARMS has been developed for summarization, but its excellent application properties for gene filtering remained so far undiscovered. The core of the algorithm is a factor analysis—a multivariate technique to detect a common structure in the data of multiple probes that measure the same target. The assumption is that the probe intensity measurements of the perfect matches x depend on the true mRNA concentration z via:

$$x = \lambda z + \varepsilon \tag{1}$$

with λ being the loadings for the factor analysis (Hochreiter *et al.*, 2006). In Equation (1), a $N(0, 1)$-distributed z models the common factor in the data x, while the $N(0, \psi)$-distributed ε models the independent noise in each probe of each array. In essence, model (1) is explaining the observed covariance structure of the data x by representing the data as being $N(0, \lambda\lambda^T + \psi)$-distributed with an individual noise variance ψ and signal variance $\lambda\lambda^T$. Based on the model assumption, the variance of factor z given the data x, var$(z|x)$, can be computed through:

$$\mathrm{var}(z|x) = (1 + \lambda^T \Psi^{-1} \lambda)^{-1} \tag{2}$$

This value, ranging from 0 to 1, provides a measure of how much variation in the probe set data x is explained by the factor z. The more variation in x is dominated by the signal, the more variation of z is already explained by x, so that var$(z|x)$ comes closer to 0. Var$(z|x)$ can be directly translated to a signal-to-noise ratio. A var$(z|x)$ of 0.5 indicates that z and ε contribute in equal parts to the total variation, corresponding with a signal-to-noise-ratio of 1. Values smaller than 0.5 indicate that there is more signal than noise and these probe sets are therefore selected for further analysis.

The estimation of the parameters of the factor analysis model is done by a Bayesian approach (3), with a prior for λ from a normal distribution with mean μ_λ and variation σ_λ (4).

$$p(\lambda,\psi|\{x\}) \propto p(\{x\}|\lambda,\psi)p(\lambda,\psi) \tag{3}$$

$$\lambda \sim N(\mu_\lambda,\sigma_\lambda) \tag{4}$$

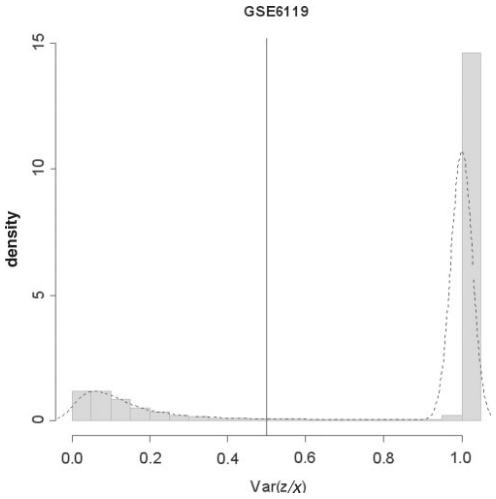

Fig. 1. Histogram of var$(z|x)$ for the real-life data set GSE6119 (see Supplementary Material 1 for the other data sets).

Setting μ_λ to zero makes loadings of λ equal to zero more likely. This implies that non-informative genes are more likely to be observed. Note that $\lambda = 0$ leads to var$(z|x) = 1$. This means that z is not determined by an observation x when it is only explained by noise. Var$(z|x)$ consequently shows a clear bimodal distribution with a very distinct mode for the non-informative and informative probe sets (see Fig. 1 and Supplementary Material 1). This data-driven bimodal distribution facilitates the use of 0.5 as an objective threshold for var $(z|x)$ to classify genes as informative or non-informative. Indeed, Fig. 1 (and Supplementary Material 1) show that the results are very robust against the choice of threshold value, as cut-offs between 0.3 and 0.9 would result in very similar conclusions.

Var$(z|x)$ is actually a multivariate measurement of the correlation between the components of x. According to model (1), the observations x are distributed according to a normal distribution with zero mean and covariance $\lambda\lambda^T + \psi$. So if the data covariance is mainly explained by λ then $x \approx \lambda z$, meaning that the noise is neglected. Then the components of x are $x_j \approx \lambda_j\ z \approx \lambda_j/\lambda_i\ x_i$, meaning that probes x_i and x_j are highly correlated. Conversely, a correlation between probes x_i and x_j is equivalent to a positive entry at position ij in the covariance matrix of x. Now, as ψ is diagonal, this entry can only be explained by $\lambda\lambda^T$. This means that highly correlated probes lead to high values of λ and low values of ψ According to (2), large λ and small ψ result in values of var$(z|x)$ near zero. Hence, a strong correlation among probes results in a var$(z|x)$ of 0, and—as can be proven analogously—a weak correlation results in a var$(z|x)$ of 1.

As FARMS is—like GCRMA—a multi-array summarization technique, it depends on the number of arrays being preprocessed. We show that I/NI filtering is useful when experiments have at least six arrays (see Supplementary Material 2).

2.2 Used data sets

We made use of the spike-in data set from the Affycomp website (Irizarry *et al.*, 2006) and 30 real-life data sets obtained from Gene

Table 1. The real-life datasets used for the assessment of I/NI calls

Accession number	Chip	Total	I/NI calls	A/P calls
E-MEXP-101	hgu133a	22 283	1726	12 898
E-MEXP-120	hgu133a	22 283	5027	13 850
E-MEXP-121	hgu133a	22 283	5105	16 574
E-MEXP-714	hgu133a	22 283	1242	13 711
E-MEXP-72	hgu133a	22 283	4385	13 801
Spike-in U133	hgu133a	22 300	113	12 869
E-MEXP-882	hgu133plus2	54 675	16 022	41 355
E-TABM-127	hgu133plus2	54 675	4962	41 022
E-TABM-34	hgu133plus2	54 675	12 810	35 162
E-TABM-84	hgu133plus2	54 675	6781	38 258
GSE3744	hgu133plus2	54 675	10 673	42 625
E-MEXP-834	Mouse430_2	45 101	8067	26 382
E-MEXP-835	Mouse430_2	45 101	5247	26 891
E-MEXP-839	Mouse430_2	45 101	8107	28 485
E-MEXP-842	Mouse430_2	45 101	1756	27 945
E-TABM-102	Mouse430_2	45 101	8858	29 934
E-MEXP-856	Mouse430A_2	22 690	5014	16 569
GSE2867	Mouse430A_2	22 690	3027	16 412
GSE2882	Mouse430A_2	22 690	4080	15 035
GSE3858	Mouse430A_2	22 690	2801	14 379
GSE4065	Mouse430A_2	22 690	984	12 181
E-MEXP-553	Rat230_2	31 099	3255	19 261
E-MEXP-920	Rat230_2	31 099	954	22 725
E-MEXP-948	Rat230_2	31 099	4080	19 378
GSE5606	Rat230_2	31 099	2723	20 626
GSE6119	Rat230_2	31 099	7449	22 030
GSE1491	ATH1-121501	22 810	3138	17 855
GSE3326	ATH1-121501	22 810	8186	17 827
GSE3350	ATH1-121501	22 810	5716	16 646
GSE3416	ATH1-121501	22 810	4635	15 159
GSE431	ATH1-121501	22 810	3593	15 653

The Accession number from either GEO or ArrayExpress is mentioned, together with the used chip type and the number of probe sets (total number on the array, and number of probe sets filtered using I/NI calls and A/P calls).

Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/). The 30 publicly available data sets were selected to cover six of the most commonly used Affymetrix gene chips, namely human genome chips (HGU133plus2 and HGU133A), mice genome chips (Mouse430_2 and Mouse430A_2), rat genome chips (Rat230_2) and Arabidopsis genome chips (ATH1-121501). See Table 1 for GEO and ArrayExpress accession numbers, a brief description of the data and the actual numbers of filtered probe sets.

3 RESULTS

3.1 Calling a probe set informative or non-informative

As the different probes of a probe set are designed to measure the same target transcript, most of them should be correlated if there is meaningful variation in the concentration of this target transcript across the arrays in the experiment. We call a probe set informative when many of its probes reflect the same increase or decrease in mRNA concentration across arrays. No common probe pattern across arrays indicates that the variation in probe expression values among arrays did not exceed the noise within a probe set, and suggests therefore the exclusion of this probe set. We call such a gene non-informative, as opposed to an undetectable gene, which is a gene that was called absent in all arrays using A/P calls (Liu *et al.*, 2002). The scatterplots in Figure 2 illustrate how 11 probes of a probe set are correlated for a non-informative (red) and an informative (green) probe set. In an informative probe set, the variation in mRNA concentration across arrays is apparent in all its probes, making these probes highly correlated. A non-informative probe set, on the other hand, has typically no consistent probe behavior. Here, increased expression values in certain arrays do not coincide in any of the joint probes. Empirical and simulated data show that probe sets with an intermediate behavior between these two clear examples are called informative as soon as at least half of their probes are correlated (see Supplementary Material 3).

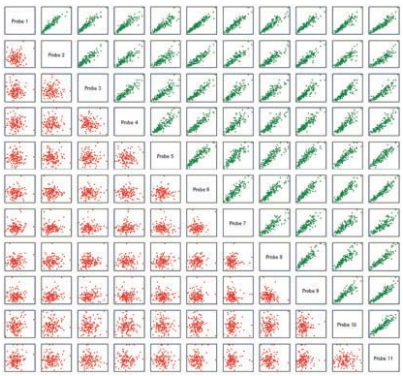

Fig. 2. Probe level patterns for an informative and a non-informative probe set. This scatterplot matrix shows all pair-wise correlations among the 11 probes of the same probe set across arrays for (1) an informative probe set (colored in green in the upper right panel) and for (2) a non-informative probe set (colored in red in the bottom left panel). Each dot represents an array.

3.2 Exclusion rates of I/NI calls

For the spike-in data set (Irizarry *et al.*, 2006) and 30 real-life data sets, on average 84 (±1.5)% of all probe sets could be excluded using I/NI calls, while A/P calls excluded only 33 (±1)%. This significant difference in filtering efficiency (paired *t*-test, $t_{30}=37$, $P<0.0001$) was apparent in all the different Affymetrix chips under study (Fig. 3). Such high exclusion rates generated by I/NI calls are expected when using high-content genome arrays where most probe sets are irrelevant for the interpretation of the experiment. In the spike-in data set, I/NI filtering excluded 99.5% of the probe sets. The remaining 0.5% included all spiked-in probe sets. In addition to this confirmed absence of false negatives, we have never observed—in all biological data sets examined so far—the exclusion of a gene that was proven to be biologically meaningful. On the contrary, instead of being too stringent,

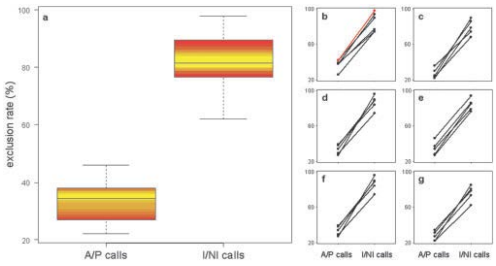

Fig. 3. Graphical comparison of exclusion rates between informative/non-informative (I/NI) calls and absent/present (A/P) calls. (**a**) Boxplots showing the distribution of the exclusion rates of both filtering techniques. The color gradient reflects the distribution within the interquartile range, going from yellow (=50%) to red (=25% and 75%). On the right, the exclusion rates of both filtering techniques are connected for each data set for (**b**) the hgu133a chip with the spiked-in data colored red, (**c**) the hgu133plus2 chip, (**d**) the Mouse430_2 chip, (**e**) the Mouse430A_2 chip, (**f**) the Rat230_2 chip and (**g**) the ATH1-121501 chip. See Table 1 for a description of the used data sets, which are obtained from GEO (www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/).

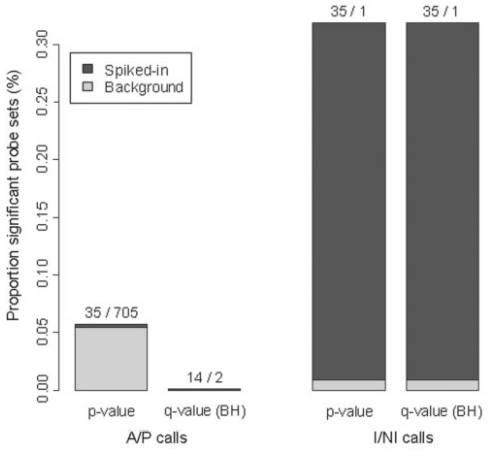

Fig. 4. Effect of gene filtering on tests for differential expression. Two differently spiked-in arrays, each done in triplicate (Experiments 5 and 6; Irizarry et al., 2006) were tested for differential expression with a *t*-test after filtering using both A/P and I/NI calls, using GCRMA summarized data. The proportion of significant probe sets ($\alpha = 0.05$) is given for the two filtering techniques before and after multiple testing correction with an FDR of 10% (Benjamini and Hochberg, 1995).

filtering using I/NI calls is too conservative as it still selects probe sets with low variation like a number of background probe sets in the spike-in data set. However, as I/NI calls in the current setting already results in exceptionally strong filtering, we suggest its use in this slightly conservative setting to prevent the exclusion of potentially interesting genes.

3.3 Impact on performance of statistical tests

Applying I/NI call selection prior to tests for differential expression has two major implications. First—apart from multiple testing correction—the list of significant genes shortens as some probe sets that would otherwise have been called significant have now been excluded. To illustrate this, we used two groups of three arrays (triplicates) that were spiked in at different concentrations [Experiments 5 and 6 of the spiked-in data set (Irizarry *et al.*, 2006)]. We tested for differential expression between these two groups with a *t*-test after filtering using both A/P and I/NI calls, using GCRMA summarized data as an independent comparison platform. After A/P filtering, the so-called significantly differentially expressed genes ($n = 740$) contained many false positives, i.e. probe sets that were not spiked-in (Fig. 4), while the list of significant genes is much shorter ($n = 36$) due to a much smaller number of false positives (Fig. 4). In various data sets, I/NI calls indeed weeded genes out that were statistically significant but had a biological function that seemed irrelevant in the respective experimental framework. Hence, filtering based on I/NI calls makes gene lists more interpretable as it seems to help excluding false positives.

A second implication of I/NI filtering on tests for differential expression is that multiple testing becomes less problematic, because the number of tests dramatically decreases. Of the 35 spiked-in probe sets that were initially called significant, only 14 remained after FDR correction (Benjamini and Hochberg, 1995) with A/P filtering while the number of significantly called genes remained unaffected after I/NI calls.

To illustrate the biological relevance of I/NI calls, we compared the genes called informative in two of the studied data sets with the conclusions of their respective papers. Nishiruma *et al.* (2003), describing public data set GSE431, provide a list containing the 684 significant probe sets, sorted by their fold change. Of the top ranked 50 genes of this list (i.e. the 50 significant genes with the highest fold change), 49 were called informative (98%), indicating that I/NI calls indeed filters the relevant genes. This high proportion of informative genes decreased gradually to 75% when including more genes with smaller fold changes (see Supplementary Material 4). This is in line with the expectation that significant genes with smaller fold-changes are more likely to be false positives, and suggests therefore that I/NI calls is capable to identify these false positives. In another paper, Glyn-Jones *et al.* (2007; data set GSE5606) compared animals with and without a treatment that induces diabetes. They conclude that the genes that were differentially expressed between the treatments were often related to proteins in the mitochondria and to genes regulating fatty acid metabolism (see Supplementary Material 4). A pathway analysis of the genes called informative using I/NI calls resulted in highly significantly affected pathways like 'Mitochondrial long chain fatty acid beta-oxidation' ($P = 1E-14$) and 'Mitochondrial unsaturated fatty acid beta-oxidation' ($P = 3E-12$). In contrast, an identical pathway analysis using non-informative genes resulted in much less

significant pathways that seemed to be irrelevant in the context of the article. Clearly, I/NI-calls made the analysis more focused on the relevant expression changes.

3.4 Properties of probe sets excluded by I/NI and by A/P calls

The expression values of the probe sets excluded by I/NI and A/P calls have different distributional properties (Fig. 5 and Supplementary Material 5 for all the 31 data sets under study). Most probe sets excluded with A/P calls have average expression values below 5 and variances of 0.1 or lower (Fig. 5b). This is because filtering based on A/P calls selects for probe sets that were called at least once present, making it dependent on the average expression value (the lower, the more likely absent) and on variation across arrays (the higher, the more likely at least one array is called present). This is however not a general pattern, as some probe sets with low average expression values and low variances are still filtered (Fig. 5a). Probe sets excluded by I/NI calls are—like A/P calls—also less variable (Fig. 5d), but can have either low or high expression values. The low-expressed probe sets excluded by I/NI calls are mostly probe sets where the technical noise was as high as the variation across arrays. The highly expressed, but excluded, probe sets code for transcripts with equally high concentrations in all arrays. These probe sets are present—and therefore selected by A/P calls—but not variable across arrays. Hence, as microarray experiments in principle try to discover differences between conditions, these genes are mostly regarded as being non-informative. Besides, e.g. house-keeping genes, these probe sets also include genes expressed at saturation levels (>13, see Fig. 5c). Figure 5c also indicates that lower expressed genes need to be more variable to be called informative by I/NI calls. This is because background noise increases with decreasing average expression levels. The signal, i.e. the true variation in gene expression across arrays, therefore needs to increase as well in order to call these genes informative. This is an objective approach similar to current common practice where people rather subjectively rely less on differentially expressed genes at lower intensity values. As these have indeed a higher potential of being false positives due to background noise, microarray users often ignore them when their fold change is rather low. Another common practice in microarray analysis is to select the most variable genes after deleting the always-absent ones. This approach not only involves arbitrary threshold choices like for instance the number of variable genes, but it also hampers the detection of truly differentially expressed genes at relatively small fold changes when they coincide with other, quite noisy—and therefore variable—genes. Hence, I/NI calls are providing a better alternative as they serve the same purpose and are based on the same reasoning as filtering on variance or on coefficient of variation, but have three main improvements: First, they do not use a general measure of probe set variation, but disentangle biological variation from variation due to technical noise, and use the mutual proportion between them as a kind of selection criterion. Second, they avoid the need of several decision steps (A/P calls, filtering on minimum variation and so forth), but incorporate all information into a single analysis. And third, no arbitrary threshold choices or assumptions have to be taken.

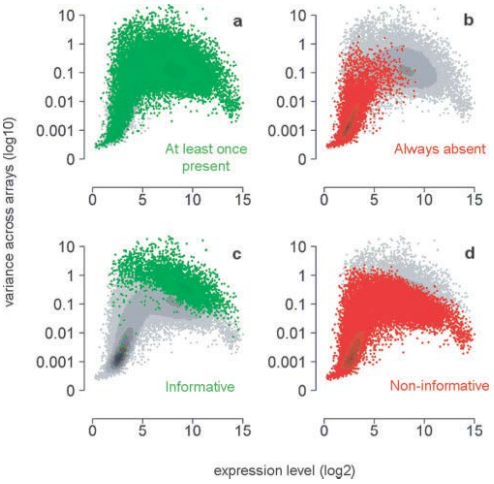

Fig. 5. Distributional properties (variance and mean) of GCRMA summarized genes selected by A/P calls and I/NI calls. Variance and mean are calculated per probe set across arrays using GCRMA summarization, and are plotted against each other. All probe set values are plotted in gray in the background and are superimposed by probe sets called at least once present (colored green in **a**), always absent (colored red in **b**), informative (colored green in **c**) and non-informative (colored red in **d**).

4 CONCLUSIONS

By incorporating probe level information to assess the noisy nature of probe sets, I/NI calls provide a highly powerful and objective tool for gene filtering. Consequently, I/NI calls offer a key solution to the main problem in the analysis of high-dimensional microarray data, being the high recurrence of false positive results because of multiple testing and overfitting. We therefore suggest that I/NI calls be used more routinely in combination with summarization techniques like FARMS (Hochreiter *et al*., 2006) or GCRMA (Wu *et al*., 2004).

ACKNOWLEDGEMENTS

We are grateful to Jeroen Aerssens and An De Bondt for reviewing and discussing earlier versions of this manuscript. We also want to explicitly acknowledge the current possibility to make easy use of publicly available microarray data sets. This is thanks to invaluable efforts of the people maintaining ArrayExpress and GEO, and in the first place thanks to the scientists who are willing to submit their data to the public domain.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2002) Statistical Algorithms Description Document. Available from www.affymetrix.com

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Bellman,R.E. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.*, **57**, 289–300.

Dudoit,S. *et al.* (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Glyn-Jones,S. *et al.* (2007) Transcriptomic analysis of the cardiac left ventricle in a rodent model of diabetic cardiomyopathy: molecular snapshot of a severe myocardial disease. *Physiol. Genomics*, **28**, 284–293.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *JMLR*, **3**, 1157–1182.

Hastie,T. *et al.* (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, RESEARCH0003.

Herrero,J. *et al.* (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.

Hochreiter,S. *et al.* (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Irizarry,R.A. *et al.* (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.

Liu,W.M. *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

McClintick,J.N. and Edenberg,H.J. (2006) Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, **7**, 49.

Nishimura,M.T. *et al.* (2003) Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science*, **301**, 969–972.

Vapnik,V.N. (2000) *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Varshavsky,R. *et al.* (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**, e507–e513.

Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658.

Wu,Z. *et al.* (2004) A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 8 2006, pages 943–949
doi:10.1093/bioinformatics/btl033

Gene expression

A new summarization method for affymetrix probe level data

Sepp Hochreiter*[1,2], Djork-Arné Clevert[1] and Klaus Obermayer[1]

[1]Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany and [2]Institute of Bioinformatics, Johannes Kepler Universität Linz, 4040 Linz, Austria

Received on October 7, 2005; revised on December 13, 2005; accepted on January 30, 2006
Advance Access publication February 10, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: We propose a new model-based technique for summarizing high-density oligonucleotide array data at probe level for Affymetrix GeneChips. The new summarization method is based on a factor analysis model for which a Bayesian maximum a posteriori method optimizes the model parameters under the assumption of Gaussian measurement noise. Thereafter, the RNA concentration is estimated from the model. In contrast to previous methods our new method called 'Factor Analysis for Robust Microarray Summarization (FARMS)' supplies both *P*-values indicating interesting information and signal intensity values.

Results: We compare FARMS on Affymetrix's spike-in and Gene Logic's dilution data to established algorithms like Affymetrix Microarray Suite (MAS) 5.0, Model Based Expression Index (MBEI), Robust Multi-array Average (RMA). Further, we compared FARMS with 43 other methods via the 'Affycomp II' competition. The experimental results show that FARMS with default parameters outperforms previous methods if both sensitivity and specificity are simultaneously considered by the area under the receiver operating curve (AUC). We measured two quantities through the AUC: correctly detected expression changes versus wrongly detected (fold change) and correctly detected significantly different expressed genes in two sets of arrays versus wrongly detected (*P*-value). Furthermore FARMS is computationally less expensive then RMA, MAS and MBEI.

Availability: The FARMS **R** package is available from http://www.bioinf.jku.at/software/farms/farms.html

Contact: hochreit@bioinf.jku.at

Supplementary information: http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps

1 INTRODUCTION

The microarray technique is currently one of the most successful experimental tools in microbiological research. It extracts a gene expression profile from a tissue sample and, therefore, supplies the expression state of tens of thousands of genes. Microarray experiments can be used to infer metabolic pathways, to characterize protein–protein interactions or to extract target genes for developing therapies for various diseases (e.g. cancer). One of the leading microarray chip technologies (GeneChips) has been developed by Affymetrix and is considered here.

A GeneChip contains probe sets of 10–20 probe pairs representing unique genes. Each probe pair consists of two oligonucleotides of length 25, namely the perfect match (PM) and the mismatch (MM) probe. The perfect match probe is the exact complement of a 25 bp subsequence in the target gene. It is supposed to bind a labeled RNA (hybridization) which is obtained from the gene's mRNA in the tissue sample. The mismatch is identical to the perfect match except that one base is changed at the center position of the oligonucleotide leading to lower affinity to the gene's labeled RNA. Mismatches are supposed to detect non-specific hybridization.

The data recorded with the microarray technique are characterized by high levels of noise induced by the preparation, hybridization and measurement processes. Noise originates from chip fabrication tolerances, tolerances in the efficiency of RNA extraction and reverse transcription, background intensity fluctuations, non-uniform target labeling, temperature fluctuations, pipette errors, hybridization efficiency and scanning deviations. Also biological effects may disturb the target signal in the data, e.g. tissue samples from the same experimental condition may not show equal levels of RNA.

In order to analyze and evaluate GeneChip data from an experiment with multiple arrays, the data preprocessing at probe-level is a crucial step. An expression summary value is calculated using a four-step procedure. (1) 'Background correction', which removes the unspecific background intensities of the scanner images; (2) 'normalization', which reduces the undesired non-biological differences between chips and normalizes the signal intensity of the arrays; (3) 'PM correction', which removes non-specific signal contributions such as unspecific binding or cross-hybridization from the PM probes and (4) 'summarization', which combines the multiple preprocessed probe intensities to a single expression value. Errors introduced in one of these steps may corrupt further processing, e.g. spurious correlation with target conditions may appear especially for few tissue samples (arrays) and large number genes. For new chip generations with more genes on a chip the probability of detecting random correlations increases and summarization techniques will become even more important. The probable number of random correlations is the number of genes multiplied by the probability of a random correlation for independent measurement noise. Recently the new generation of HGU_133+2 GeneChips has been introduced by Affymetrix which provides the coverage of the entire human genome on a single array. Here one chip contains more than 54 000 probe sets and 1 300 000 distinct oligonucleotides.

In this paper we focus on new techniques for summarization. The summarization method which comes with an Affymetrix scanner is the Affymetrix Microarray Suite 5.0 [MAS 5.0, Aff, (2001); Hubbell *et al*., 2002]. The two best known approaches to improve MAS 5.0 are the Model Based Expression Index [MBEI, Li and

*To whom correspondence should be adderessed.

Wong (2001)] and the Robust Multi-array Average [RMA, Irizarry *et al*. (2003a, b); Bolstad *et al*. (2003)]. The Affymetrix Microarray Suite 5.0 (http://www.affymetrix.com/support/technical/manuals.affx) provides a 'present call' for each gene to indicate whether the measurement is likely to contain signal rather than noise but disregards information available at the summarization step. In addition the relevance of a gene in a certain experimental setting is usually determined by how strongly it is expressed at the one or the other condition. This, however, may not be the best way to evaluate the chip data, because even if a signal is present and strong but Gaussian distributed, its 'information content' may be low [see Friedman and Tukey, 1974; Friedman and Stuetzle, 1981; Huber, 1985] and it may not be useful to distinguish between conditions. Here we propose a summarization method which supplies noise corrected measurement values and improved present calls for genes as well as quantitative measures for the 'relevance' of a gene in a given context. Benchmark results using datasets from the open challenge 'Affycomp II' http://affycomp.biostat.jhsph.edu, Cope *et al*., 2004 and the 'golden spike-in' dataset from Choe *et al*. (2005) show that FARMS performs better than state-of-the-art methods like MAS 5.0, MBEI and RMA.

2 FACTOR ANALYSIS FOR ROBUST MICROARRAY SUMMARIZATION (FARMS)

2.1 The model

2.1.1 The basic model Our approach to the summarization problem is based on a linear model with Gaussian noise. Denote the actually observed and to zero mean normalized log-PMs by $\boldsymbol{x}$ and the normalized log-RNA concentration in the hybridization mixture by z. Then we assume that the log-observations $\boldsymbol{x}$ depend on the true log-concentration z via

$$\boldsymbol{x} = \lambda z + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{x}, \lambda \in \mathbb{R}^n \tag{1}$$

and

$$z \sim \mathcal{N}(0,1), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}). \tag{2}$$

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multidimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ [$\mathcal{N}(0,1)$ is the one-dimensional standard Gaussian]. z is usually called a 'factor'. $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ is the diagonal noise covariance matrix while $\boldsymbol{\epsilon}$ and z are statistically independent. According to the model, the observation vector $\boldsymbol{x}$ is Gaussian distributed as shown in the following equation:

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \lambda\lambda^T + \boldsymbol{\Psi}). \tag{3}$$

Consequently, the PMs are log-normal distributed. The λ_j are the shape-parameters of the log-normal distribution for each PM_j. To introduce individual shape-parameter for the PMs is justified by the findings in Li and Wong (2001), where the authors found that probes of the same probe-set may have different response to the same RNA amount. In Li and Wong (2001) these probe-effects were consistent over various arrays which implies specific binding characteristics of the probes. However, λ_j subsumes also signal contributions via signal strength σ as seen in text before Equation (9), where we set $\lambda_j = \sigma + \tau_j$. Large signal leads to large σ which scales up the shape-parameter which in turn results in a more heavy tail and allows for higher PM values carrying a signal. In the following we will motivate our model assumptions and then describe how to use factor analysis to infer the 'summarized' values z from the multiple observations $\boldsymbol{x}$ for each array and gene.

2.1.2 Using PM values only and the assumption of Gaussian noise In this section we want to justify the model assumption, that the vector $\boldsymbol{x}$ is Gaussian distributed. In Naef *et al*. (2002) replicate experiments on different arrays were made and the PM values as well as the PM − MM values were analyzed. The authors found that the PM values ('PM') have lower noise at low intensity than PM minus MM ('PM − MM') whereas for intermediate and high intensities the noise levels for PM and PM − MM were similar. Therefore we will use in our model only PM measurements.

Naef *et al*. (2002) also found that the distribution p_{diff} of the difference $\log(\mathrm{PM}_x) - \log(\mathrm{PM}_y)$ (x and y denote arrays of replicate measurements) is Gaussian, where the width depends on the intensity of the probe. Let p_{pm} be the distribution of log(PM). If p_{diff} is Gaussian and the distribution p_{pm} symmetric around a mean value μ, then p_{pm} is a Gaussian. This can be derived by setting w.l.o.g. $\mu = 0$ (note that the difference of the log-PMs is considered) and

$$\begin{aligned} p_{\mathrm{diff}}(a) &= \int_{-\infty}^{\infty} p_{\mathrm{pm}}(b)\, p_{\mathrm{pm}}(a+b)db \\ &= \int_{-\infty}^{\infty} p_{\mathrm{pm}}(b')\, p_{\mathrm{pm}}(a-b')d(b'), \end{aligned} \tag{4}$$

where $b' = -b$ and where we used $p_{\mathrm{pm}}(-b') = p_{\mathrm{pm}}(b')$. Fourier transformation of both sides yields

$$\mathcal{F}(p_{\mathrm{diff}})(a) = (\mathcal{F}(p_{\mathrm{pm}})(a))^2. \tag{5}$$

Because the Fourier transformation of a Gaussian is a Gaussian and the square root of a Gaussian is also Gaussian, the above statement holds.

Freudenberg *et al*. (2004) also found log-transformed data are normally distributed using a probe-wise Shapiro–Wilk test. Using the Affymetrix HGU133A latin square dataset (cf. Section 3), we confirmed that the log-transformed perfect matches are closer to a Gaussian distribution than the original perfect matches (Fig. 1). In conclusion, the assumption of a Gaussian distribution for the $\log(\mathrm{PM}_x)$ values seems to be justified.

2.1.3 The factor model assumptions In this section we motivate our linear ansatz λz from Equation (1), where z is interpreted as the logarithm of the true amount of mRNA in the tissue sample. Consider one gene, N arrays i—one for each tissue sample—and n perfect matches PM_{ij}, $1 \leq j \leq n$, on each array i. For each array we have a true (ideal) signal s_i indicating the logarithm of the amount of mRNA from this gene which is present in the tissue sample. Let z_i be the signal s_i normalized to mean zero and variance 1, that is

$$s_i = z_i\, \sigma + \mu, \sigma > 0. \tag{6}$$

Now we assume that for each PM_{ij} the signal deviates by τ_j and γ_j from the true values σ and μ giving

$$S_{ij} = z_i(\sigma + \tau_j) + \mu + \gamma_j, \tag{7}$$

where we assume that both the τ_j and the γ_j a distributed with zero mean. The value $\sigma + \tau_j$ determines the variance of the j-th measurement $\mathrm{PM}_{\cdot j}$ and $\mu + \gamma_j$ its mean, i.e. we assume that each oligonucleotide corresponding to PM_j has its own characteristics (e.g. hybridization efficiency or crosstalk). Adding the measurement

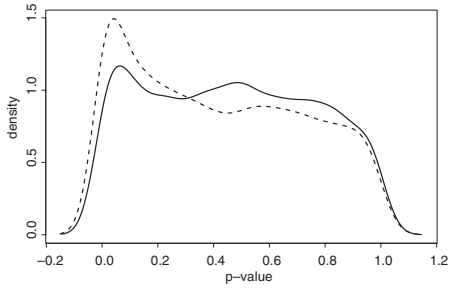

Fig. 1. Estimated density of *P*-values from the Shapiro–Wilk test for normality using 10 000 randomly selected PM intensities and 42 arrays from Affymetrix HGU133A latin square data. The continuous and dashed lines indicate the result for the $\log_2$-transformed and the original PMs, respectively. The deviation from a uniform distribution of the *P*-values indicates the deviation from Gaussian distributions. The $\log_2$-transformed PMs are closer to a Gaussian.

noise $\boldsymbol{\epsilon}$ to S_{ij} gives

$$\log(\mathrm{PM}_{ij}) = S_{ij} + \boldsymbol{\epsilon}_{ij} = z_i(\sigma + \tau_j) + \mu + \gamma_j + \boldsymbol{\epsilon}_{ij}, \tag{8}$$

where $\boldsymbol{\epsilon}_{ij}$ is a zero mean Gaussian (non-zero mean is accounted for by γ_j). The values τ_j, γ_j and the standard deviation of the $\boldsymbol{\epsilon}_{ij}$ may depend on the gene's signal intensities for the arrays. This takes the findings in Chudin *et al*. (2001); Naef *et al*. (2002); and Tu *et al*. (2002) into account, that the variance of the noise depends on the signal strength. Therefore, estimated values are only valid for the measurements under considerations, i.e. the actual signal strength.

If we set $\lambda_j = \sigma + \tau_j$ and normalize the observation $\boldsymbol{x}$ to zero mean by subtracting

$$\begin{aligned}\tfrac{1}{N}\sum_{i=1}^{N}\log(\mathrm{PM}_{ij}) &= (\sigma+\tau_j)\tfrac{1}{N}\left(\sum_{i=1}^{N} z_i\right)\\ &\quad + \mu + \gamma_j + \tfrac{1}{N}\left(\sum_{i=1}^{N}\boldsymbol{\epsilon}_{ij}\right)\\ &\approx \mu + \gamma_j = \mu_j,\end{aligned} \tag{9}$$

where the approximation is due to the zero mean assumptions then we arrive at Equation (1), the basic model. According to the model assumptions, $z \sim \mathcal{N}(0,1)$ [Equation (2)], our approach is best suited for genes with strong Gaussian distributed signal or for genes with low signal intensities (small σ), because the Gaussian noise is superimposed on the weak signal. The Gaussian signal assumption is justified for the majority of genes which are independent of the conditions, however it is not justified for the genes conveying a non-Gaussian signal. It will turn out that the model-based approach also provides good results for non-Gaussian distributions of z, because the non-Gaussianity of z has only a minor impact on the model likelihood as we will see at the end of Subsection 2.2.2.

2.2 Estimation of model parameters and signal

We now describe how to estimate the true signal strengths based on the data model of Section 2.1. The procedure consists of three steps:

(1) normalization of the observations to zero mean [cf. Equation (9)]
(2) the maximum a posteriori factor analysis to estimate model parameters λ_j in order to calculate σ and
(3) recovering the true signals s_i [Equation (6)] from z_i,

which we will describe in the following text.

2.2.1 Normalization of the observations In order to fulfill model assumptions, the log-PM values are normalized to zero mean by subtracting $\mu_j = \mu + \gamma_j$ which is estimated using Equation (9).

2.2.2 Maximum a posteriori factor analysis The Bayesian posterior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\})$ of the model parameters $(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ given the dataset $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is proportional to the product of the observation's likelihood $p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})$ of data $\{\boldsymbol{x}\}$ given the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}$ multiplied by the prior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ (e.g. DeGroot, 1970):

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \propto p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\ p(\boldsymbol{\lambda}, \boldsymbol{\Psi}). \tag{10}$$

For the prior we assume that $p(\boldsymbol{\lambda}, \boldsymbol{\Psi}) = p(\boldsymbol{\lambda})$, i.e. that the prior for the factor loadings $\boldsymbol{\lambda}$ is independent from the prior for $\boldsymbol{\Psi}$ and that the latter is uninformative (i.e. flat). The prior for $\boldsymbol{\lambda}$ is $p(\boldsymbol{\lambda}) = \prod_{j=1}^{n} p(\lambda_j)$ and for $p(\lambda_j)$ we choose the rectified Gaussian distribution $\mathcal{N}_{\mathrm{rect}}(\mu_\lambda, \sigma_\lambda)$ (see Hinton and Ghahramani, 1997) given by

$$\lambda_j = \max\{y_j, 0\} \ \text{ with } \ y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda). \tag{11}$$

σ_λ is chosen proportional to the mean of the variance $\mathrm{Var}(x_{*j})$ of the observations to allow the factor to explain the data variance, that is

$$\sigma_\lambda^2 = \rho \frac{1}{n}\sum_{j=1}^{n}\mathrm{Var}(x_{*j}). \tag{12}$$

The prior reflects the facts that

(1) the observed variance in the data is often low which makes high values of λ_j unlikely,
(2) a chip typically contains many more genes with constant signal ($\lambda_j \sim 0$) than genes with variable signal (large value of λ_j),
(3) negative values of λ_j are not plausible, because that would mean that increasing mRNA concentrations lead to smaller signal intensities.

The two hyperparameters ρ and μ_λ allow quantifying different aspects of potential prior knowledge. For example, μ_λ near zero assumes that most genes do not contain a signal and introduces a bias for λ-values near zero (items 1 and 2 from above).

The second factor of the posterior is the likelihood which is according to Equation (3)

$$p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \prod_{i=1}^{N}\mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi})\ (\boldsymbol{x}_i), \tag{13}$$

where $\mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi})\ (\boldsymbol{x}_i)$ is the distribution's density evaluated at $\boldsymbol{x}_i$.

Following Rubin and Thayer (1982), we estimate the parameters of the factor analysis model with the expectation-maximization (EM) algorithm of Dempster *et al*. (1977) modified to maximize the Bayesian posterior, Equation (10), of the model parameters given the data. The EM procedure estimates $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}$ and the posterior values for z for every $\boldsymbol{x}$. Analogous to the EM algorithm for

maximum likelihood, the EM algorithm maximizes a lower bound of the log-posterior

$$-\frac{1}{2}\boldsymbol{\sigma}_{\lambda}^{-2}(\boldsymbol{\lambda}-\mu_{\lambda}\mathbf{1})^{T}(\boldsymbol{\lambda}-\mu_{\lambda}\mathbf{1}) + \frac{nN}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{N} E_{z_i|x_i}((\boldsymbol{x}_i - \boldsymbol{\lambda}z_i)^{T}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\lambda}z_i)), \quad (14)$$

where $\boldsymbol{x}$ is already normalized to mean zero and

$$z_i \,|\, \boldsymbol{x}_i \sim \mathcal{N}(\mu_{z_i|x_i}, \sigma^2_{z_i|x_i}),$$
$$\mu_{z_i|x_i} = (\boldsymbol{x}_i)^{T}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}+\boldsymbol{\Psi})^{-1}\boldsymbol{\lambda} \quad \text{and} \quad (15)$$
$$\sigma^2_{z_i|x_i} = 1 - \boldsymbol{\lambda}^{T}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}+\boldsymbol{\Psi})^{-1}\boldsymbol{\lambda}.$$

A detailed derivation of both the lower bound and the complete EM algorithm can be found in the supplementary information (http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps).

Note that the maximum a posteriori factor analysis is also able to extract non-Gaussian signals. The likelihood covariance matrix is $\boldsymbol{\lambda}\boldsymbol{\lambda}^{T} + \boldsymbol{\Psi}$, therefore increasing the diagonal elements of $\boldsymbol{\Psi}$ would lead to a larger decrease of the likelihood than increasing one eigenvalue via $\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}$ (note that scaling a non-Gaussian to variance one increases $\boldsymbol{\lambda}$). Reason for the larger decrease of the likelihood in the first case is the cumulative effect of increasing n eigenvalues of the covariance matrix. Therefore, explaining data variance by a non-Gaussian factor has higher likelihood than explaining it by n measurement noise corrections.

2.2.3 Estimation of the true signals Finally we need to recover the 'true' signal s_i from the estimated values z_i, i.e. we need to estimate σ and μ in Equations (6) and (8). For each perfect match we have

$$\sigma = \lambda_j - \tau_j \text{ and } \mu = \mu_j - \gamma_j. \quad (16)$$

We determine σ and μ with the least squares fit, which is unbiased because we assumed in Subsection 2.1.3 that both τ_j and γ_j are drawn from a distribution with zero mean:

$$\sigma = \text{argmin}_{\tilde{\sigma}} \sum_{i=1}^{n}(\lambda_j - \tilde{\sigma})^2 = \frac{1}{n}\sum_{j=1}^{n}\lambda_j, \quad (17)$$

$$\mu = \text{argmin}_{\tilde{\mu}} \sum_{j=1}^{n}(\mu_j - \tilde{\mu})^2 = \frac{1}{n}\sum_{j=1}^{n}\mu_j. \quad (18)$$

The 'true' signal is then computed as

$$s_i = \sigma\, z_i\, f + \mu, \quad (19)$$

where f is a factor which compensates for the reduction of variance during preprocessing and factor analysis (some of the data variance is explained by the noise). The value of f is empirically determined on toy data for different normalization procedures: 2.0 for quantile normalization and 1.5 for cyclic loess (see Section 3.2 for the normalization procedures). Note that the factor f does not influence the AUC-values which we used to evaluate the different methods in Section 3.

We call the new summarization procedure which has been described 'Factor Analysis for Robust Microarray Summarization' (FARMS).

2.3 Extraction of the relevant genes

Using factor analysis we estimated the 'true' signals s_i. Their actual strengths, i.e. the value of σ, can be taken as a measure of the potential relevance of a gene in a given experimental setting: high value of σ indicates more relevant genes. A complementary and in several cases even better criterion, however, can be derived via the factor z and its distribution across arrays. Following the idea of projection pursuit of Friedman and Tukey (1974); Friedman and Stuetzle (1981); Huber (1985) interesting or 'relevant' variables are often not Gaussian distributed. This assumption is especially true for most microarray experiment designs, where genes are of interest if their expression levels are correlated with different experimental conditions. Often two conditions must be distinguished, thus genes which show a bimodal rather than a Gaussian distribution are of interest because they may be correlated with the conditions. But also for a larger number of conditions one would expect that non-Gaussianity is a good indicator for relevance. A quantitive measure can be obtained by a test of Gaussianity for the estimated variables z through the Shapiro–Wilk test (more robust in the case of a small sample size than the Kolmogorov–Smirnov test). FARMS is especially suited for this test because it assumes a Gaussian signal, thus violating this assumption indicates a strong signal. Genes can be ranked according to their σ-values or according to their non-Gaussianity, and the top candidates can then be investigated further.

3 EXPERIMENTS AND RESULTS

3.1 Datasets

For the following benchmarks we use four well-known evaluation datasets denoted by (A), (B), (C) and (D) which were produced by controlled experiments with known target expression values or known mutual relations. The first three datasets are from the open challenge 'Affycomp II' (http://affycomp.biostat.jhsph.edu/, Cope *et al*., 2004) whereas the fourth dataset is known as the 'golden spike-in' dataset from Choe *et al*. (2005).

Dataset A. This dataset is the original assessment dataset in Cope *et al*. (2004). It consists of two sub datasets with the Affymetrix human HGU95A array: the spike-in experiments and the dilution experiments.

For the first, spike-in dataset A1, the concentration of RNA for 14 genes, the so-called spike-in genes, was artificially controlled by adding RNA with predefined concentrations to the hybridization mixture. The 'latin square design' contained 20 experiments with different RNA concentrations of the 14 spike-in genes chosen from {0.0, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0, 256.0, 512.0, 1024.0} pM. For each experiment two replicate arrays were prepared except one with only two replicates. The datasets consist of 59 arrays stored in 'CEL' files. A 'CEL' file gives the 75 percentile pixel intensity of each spot, i.e. each gene in the array image.

The second, dilution dataset A2 from GeneLogic uses two tissue samples, human liver (HL) and human central nervous system (CNS), from which the RNAs were hybridized to the 75 HGU95A_v2 arrays. The dataset is based on changing dilutions (concentrations) and combinations of RNA taken from the two different tissues. Arrays are hybridized to a mixture of HL and CNS where the amount of RNA taken from each source is one from the six values {1.25, 2.5, 5.0, 7.5, 10.0, 20.0} µg. Each dilution

experiment is replicated five times and each replicate was evaluated on a different scanner.

Dataset B. This dataset is the first part of the new assessment from http://affycomp.biostat.jhsph.edu/. It is identical to dataset A1 but separately listed because of the separate Affycomp evaluation results.

Dataset C. This dataset is the second part of the new assessment from http://affycomp.biostat.jhsph.edu/. It is based on a 'latin square' experimental design which consists of 42 HGU133A arrays, with 42 spike-in genes with RNA concentrations from {0.0, 0.0125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0, 256.0, 512.0} pM. Here three spike-in genes of the same concentration were combined in order to create three replicates for each experiment.

Dataset D. Recently Choe *et al*. (2005) supplied a dataset consisting of six Affymetrix DrosGenome1 chips. This dataset mimics a common used microarray experimental setting, where two samples, i.e. a treatment and a control sample are compared in order to identify differentially expressed genes. The array can detect 3860 known individual RNA samples together with 2551 RNA samples as controls (and background) where the latter have the same concentration in all experiments. A total of 1309 RNAs samples mimic the differentially expressed genes, these RNAs were split into 8 subsets of about 80 to 180 RNAs. Each subset differs by one predefined relative concentration change from {1.2, 1.5, 1.7, 2.0, 2.5, 3.0, 3.5, 4.0} between the spike-in and control sample. Finally, the spike-in and control sample were hybridized in triplicates.

3.2 Benchmark details

We compare our method, FARMS, to the three best known summarization methods MAS5, MBEI and RMA as well with the 43 methods which participated at the challenge 'Affycomp II' (as of October 7, 2005). Microarray Suite (MAS) 5.0 is a non-parametric algorithm implemented by Affymetrix (Aff, 2001; Hubbell *et al*., 2002). The Model Based Expression Index [MBEI, Li and Wong (2001)] is like the Robust Multi-array Average [RMA, Irizarry *et al*. (2003a, b); Bolstad *et al*. (2003)] a model-based approach (software packages are available at http://www.dchip.org or www.bioconductor.org).

FARMS does not use background correction and uses either quantile normalization (Bolstad *et al*., 2003) or cyclic loess (Yang *et al*., 2002; Dudoit *et al*., 2002). FARMS uses quantile normalization as default normalization procedure because it is computational efficient. It does not apply PM corrections and uses PMs only. For all experiments with FARMS we set $\rho = 1/8$, $\mu_\lambda = 0$ and $f = 2.0$ for quantile normalization and $f = 1.5$ for cyclic loess. The maximal cycles for factor analysis were fixed to 100 and factor analysis was terminated if the λ-update vector has length smaller than 0.00001.

RMA can be improved through advanced background correction leading to a method called GCRMA (Wu *et al*., 2004, Available at http://ideas.repec.org/p/bep/jhubio/1001.html). GCRMA has lower performance on datasets A–C with respect to the AUC-values than FARMS as can be seen in the supplementary information but is superior to RMA. For our FARMS method we did not use background correction, however in future studies we want to investigate whether background correction can improve our FARMS method especially whether the GCRMA background corrections is suitable.

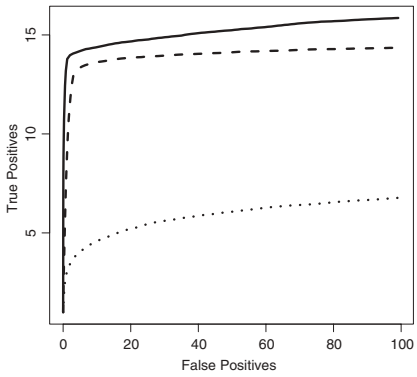

Fig. 2. ROC curves for all fold changes in dataset A1. ROC curve for FARMS with quantile normalization (solid line) is always above the ROC curve for RMA (dashed line) and MAS 5.0 (dotted line), therefore FARMS is better than RMA and MAS 5.0 for all false positive rates.

3.3 Results

For the evaluation of datasets A, B and C, we participated at the 'Affycomp II' challenge (http://affycomp.biostat.jhsph.edu/, Cope *et al*., 2004). For the complete challenge results see Tables 1–3 in the supplementary information (http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps).

3.3.1 AUC fold changes We think that from all challenge results the area under the curve (AUC) criterion is best suited to measure the quality of a summarization method. The AUC criterion is the area under the receiver operating characteristics (ROC) curve which plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) and serves a quality measure for classification methods. The AUC criterion can be applied here by defining gene classes: for a pair of arrays class 1 genes are the genes for which expression value differences exceed a certain relative factor (fold change). Now the output of a summarization method can be interpreted as classification by computing the class membership of genes based on the predicted expression values. We prefer the AUC criterion over other measures provided by 'Affycomp II' evaluation because it is independent of scaling of the results (log-expression values) and trades sensitivity against specificity. Other quality measures from the 'Affycomp II' evaluation focus either on sensitivity or specificity and are often not scaling independent. The AUC is computed for different fold changes, i.e. for different thresholds for being in class 1. Figures 2–4 show the fold change ROC curves for A1, C and D, respectively. Table 1 gives the corresponding AUC for datasets A–D. Note, that dataset D is especially suited to generate precise ROC curves because of the large number of defined RNAs. Except for dataset A, FARMS has the best AUC performance of the 43 competitors of the 'Affycomp II' challenge (the challenge method which has higher AUC values than FARMS in dataset A has lower AUC values for datasets B and C).

FARMS with quantile normalization is best for datasets A–B, whereas FARMS with cyclic loess is best for dataset D. However, both FARMS methods show higher performance than all its

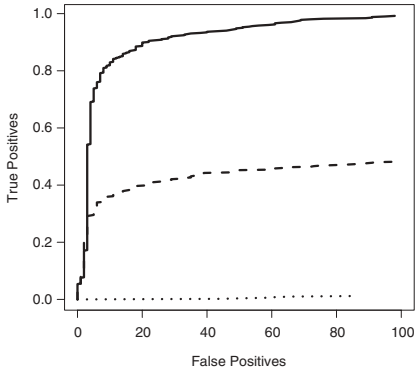

Fig. 3. ROC curve for fold changes of low intensity genes in dataset C. ROC curve for FARMS with quantile normalization (solid line) is always above the ROC curve for RMA (dashed line) and MAS 5.0 (dotted line) therefore FARMS is better than RMA and MAS 5.0 for all false positive rates.

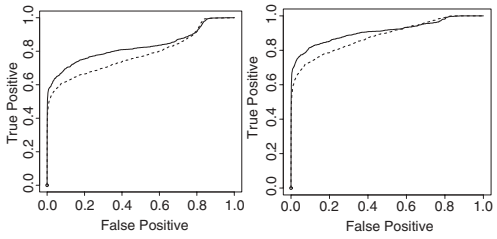

Fig. 4. Dataset D ROC curves for fold changes ⩾1.2 (left) and ⩾1.7 (right) for RMA (dashed line) versus FARMS (solid line). In both cases FARMS performs better than RMA as its ROC curve is above RMA's.

competitors (RMA, MAS and MBEI). FARMS shows a large improvement over RMA for small signal changes: for dataset A and fold change 2 the AUC value is 0.54 for RMA, 0.84 for FARMS (quantile normalization) and 0.78 for FARMS (cyclic loess), for dataset B the low intensity AUC is 0.51 for RMA, 0.89 for FARMS (quantile) and 0.80 FARMS (cyclic loess), and for dataset C the low intensity AUC is 0.57 for RMA, 0.94 for FARMS (quantile) and 0.91 for FARMS (cyclic loess). The AUC for random guessing is 0.5 and the maximal AUC is 1.0, therefore the improvement of FARMS over RMA is considerable.

3.3.2 AUC P-values Above AUCs for fold changes assess the quality of summarization methods with respect to the identification of differentially expressed genes in a pair of arrays. Here we want to go one step further and determine the quality of summarization methods with respect to the identification of significant differentially expressed genes in two conditions. To perform a significance test to the expression values in two conditions is a common experimental setting in biology and in medicine, therefore we evaluate the quality of different summarization methods by wrongly detected significant differences and missed differences.

Table 1. AUC results for fold changes for datasets A–D

AUC	FARMS q	FARMS l	RMA	MAS 5.0	MBEI	1	2	mean
FC Dataset A								
all	0.89	0.85	0.82	0.36	0.67	**0.91**	0.86	0.71
=2	0.84	0.78	0.54	0.07	0.17	**0.91**	0.69	0.42
1 Dataset B								
Low	**0.89**	0.80	0.51	0.07	0.21	0.74	0.68	0.44
Med	0.97	0.95	0.91	0.00	0.43	**0.98**	0.97	0.65
High	**0.97**	0.94	0.64	0.00	0.16	0.95	0.94	0.48
Mean	**0.91**	0.84	0.60	0.05	0.26	0.79	0.75	0.49
1 Dataset C								
Low	**0.94**	0.91	0.57	0.09	—	0.76	0.61	0.48
Med	**0.99**	**0.99**	0.91	0.00	—	0.95	0.95	0.64
High	**1.00**	**1.00**	0.96	0.00	—	0.99	0.99	0.61
Mean	**0.95**	0.93	0.65	0.06	—	0.81	0.66	0.44
FC Dataset D								
⩾1.2	0.72	**0.74**	0.70	0.52	0.49	—	—	—
⩾1.7	0.90	**0.91**	0.88	0.64	0.59	—	—	—

We compare FARMS with RMA, MAS 5.0 and MBEI and for dataset A–C also with 43 competitors from the *affycomp* Bioconductor Project benchmark where the best ('1'), the second best ('2') and the mean results ('mean') are given (as of October 7, 2005). FARMS results are reported for quantile normalization ('q') and for cyclic loess ('l'). The table reports AUC values for different fold changes ('FC', datasets A and D), i.e. detection of different concentrations changes, as well as different signal intensities ('I', datasets B and C). The best result is marked bold.

Analogous to the AUC for fold changes we define an AUC for *P*-values. Class 1 genes are the genes which have by design different expression values in the two conditions. A summarization method classifies a gene as being differently expressed in the two conditions if the *P*-value of a test is below a given threshold (we set it to 0.05). This allows us to compute the ROC curve.

A significance test, a modified *t*-test, for differentially expressed genes for microarray experiments with two conditions was suggested by Tusher *et al*. (2001). In the modified *t*-test a small positive constant ('fudge-constant') is added to the denominator to prevent genes with small variance from being selected as significant. According to Cui and Churchill (2003) we set the 'fudge-constant' to the 90th percentile of the standard deviation of all genes.

Datasets B and C encompass 19 and 14 experimental conditions, respectively, with 3 replicates for each condition. This leads to 171 and 91 experimental condition pairs (only unique variations), respectively, with 6 arrays (3 for each condition) for each experimental setting. The above-mentioned modified *t*-test is applied to these 171 (dataset B) and 91 (dataset C) experimental settings. The average AUC of all ROC curves for *P*-values is given in Table 2. For dataset B the average AUC for RMA is larger than for FARMS but the difference is not significant as confirmed by Wilcoxon-rank-sum test ($P = 0.19$). For dataset C FARMS shows significantly by ($P = 0.00027$) better results than RMA. Most reliable are the results on dataset D, where the number of defined RNAs is large. However, for dataset D there is only one experiment so that the Wilcoxon-rank-sum test cannot be applied, but the large number of spike-in genes allows to perform another test, the conservative McNemar test. It confirmed that FARMS performed significantly by better ($P = 0.000002$) than its competitors.

Table 2. AUC results for *P*-values for datasets B–D

	FARMS q	FARMS l	RMA	MAS 5.0	MBEI
Dataset B					
AUC [$e = 171$]	**0.955**	**0.955**	0.948	0.772	0.670
Dataset C					
AUC [$e = 91$]	**0.975**	**0.974**	**0.981**	0.892	0.875
Dataset D					
AUC [$e = 1$]	**0.802**	**0.823**	0.767	0.286	0.397

We compare FARMS with RMA, MAS 5.0 and MBEI on *e* experiments (one experiment consists of 6 arrays—3 arrays for each condition). Results which are significantly better than others are marked bold, where mutual differences between bold results are not significant.

Table 3. Computational time in (s) for dataset A2

	FARMS	RMA	MAS 5.0	MBEI
Computational time	**246**	472	1323	957

3.3.3 Computational time: The computational time of FARMS (quantile normalization), RMA, MAS 5.0 and MBEI is listed in Table 3. FARMS is the fastest method.

In conclusion FARMS performs better than all competitors with respect to the AUC criterion for fold changes as well as for *P*-values and was the fastest method.

4 CONCLUSION

We have presented a new method called FARMS for summarization of gene expression data obtained from Affymetrix chips. The new method outperforms known methods both with respect to sensitivity and specificity, i.e. detects more signals while being more robust against measurement noise. Further it is faster than the competitors.

ACKNOWLEDGEMENTS

The authors express their gratitude for the funding by the Anna-Geissler- and the Monika-Kutzner-Stiftung.

Conflict of Interest: none declared.

REFERENCES

Microarray Suite User Guide. (2001) Affymetrix, version 5 edition.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Choe,S.E. *et al*. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol*., **6**, R16.1–R16.16.

Chudin,E. *et al*. (2001) Assessment of the relationship between signal transcript concentration for Affymetrix GeneChip arrays. *Genome Biol*., **3**, research0005.1–0005.10.

Cope,L.M. *et al*. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Cui,X. and Churchill,G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.1–210.10.

DeGroot,M.H. (1970) *Optimal Statistical Decisions*. McGraw-Hill, NY.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–22.

Dudoit,S. *et al*. (2002) Statistical methods for identifying genes with differential expression in replicate cDNA microarray experiments. *Stat. Sin*., **12**, 111–139.

Freudenberg,J. *et al*. (2004) Comparison of preprocessing procedures for oligonucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. *Meth. Inform. Med*., **43**, 434–438.

Friedman,J.H. and Stuetzle,W. (1981) Projection pursuit regression. *J. Am. Stat. Assoc*., **76**, 817–823.

Friedman,J.H. and Tukey,J.W. (1974) A Projection Pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput*., **23**, 881–890.

Hinton,G.E. and Ghahramani,Z. (1997) Generative models for discovering sparse distributed representations. *Pilos. Trans. R. Soc. B*, **352**, 1177–1190.

Hubbell,E. *et al*. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.

Huber,P.J. (1985) Projection pursuit. *Ann. Stat*., **13**, 435–525.

Irizarry,R.A. *et al*. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, 1–8.

Irizarry,R.A. *et al*. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Naef,F. *et al*. (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol*., **3**, research0018.1–0018.11.

Rubin,D. and Thayer,D. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.

Tu,Y. *et al*. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl Acad. Sci. USA*, **99**, 14031–14036.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wu,Z., Irizarry,R., Gentleman,R., Murillo,F.M. and Spencer,F. (2004) A model based background adjustment for oligonucleotide expression arrays. Johns Hopkins University Dept. of Biostatistics Working Paper Series 1001, Berkeley Electronic Press.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.

Lebenslauf

Mein Lebenslauf ist aus Datenschutzgründen in der elektronischen Version meiner Arbeit nicht mitveröffentlicht.

Publikationsliste

ORIGINALARBEITEN

la.FARMS: Laplace FARMS for Detecting Rare Events in Microarray Data, A. Mitterecker, A. Mayr, G. Klambauer, W. Talloen, H. W. H. Göhlmann, D.-A. Clevert, and S. Hochreiter [**in Vorbereitung**]

Neurodevelopmental dysfunctions in the cerebellum cause autism spectrum disorders, S. Hochreiter, and D.-A. Clevert [**in Vorbereitung**]

cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in Next Generation Sequencing Data with A Low False Discovery Rate, G. Klambauer, K. Schwarzbauer, A. Mayr, A. Mitterecker, D.-A. Clevert, and S. Hochreiter **Nucleic Acids Research**, (**Accepted with minor changes**) [**ISI-JIF 7.84**]

cn.FARMS: a probabilistic latent variable model to detect copy number variations with a low false discovery rate, D.-A. Clevert, A. Mitterecker, A. Mayr, M. Tuefferd, A. De Bondt, W. Talloen, H. W. H. Göhlmann, and S. Hochreiter, **Nucleic Acids Research**, 2011, 39(12): e79, [**ISI-JIF 7.84**]

FABIA: Factor Analysis for Bicluster Acquisition, S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, L. Bijnens, H. W. H. Göhlmann, Z. Shkedy, and D.-A. Clevert, **Bioinformatics**, 2010, 26(12):1520-7, [**ISI-JIF 4.93**]

Informative or Noninformative Calls for Gene Expression: A Latent Variable Approach, Kasim A., Miss D. L., Van Sanden S., Clevert D.-A., Bijnens L., Göhlmann H., Amaratunga D., Hochreiter S., Shkedy Z., and Talloen W., **Statistical Applications in Genetics and Molecular Biology**, 2010, Vol. 9: Iss. 1, 4, [**ISI-JIF 2.25**]

Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays, Tuefferd M., De Bondt A., Van Den Wyngaert I., Talloen W., Verbeke T., Carvalho B., Clevert D.-A., Alifano M.,Raghavan N., Amaratunga D., Göhlmann H., Broët P., Camilleri-Broët S., **Genes, Chromosomes & Cancer**, 2008, 47(11):957-64. [**ISI-JIF 4.53**]

I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data, W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, H.W.H. Göhlmann, **Bioinformatics**, 2007, 23:2897-2902 [**ISI-JIF 5.04**]

A New Summarization Method for Affymetrix Probe Level Data, S. Hochreiter, D.-A. Clevert, and K. Obermayer, **Bioinformatics**. 2006 Apr 15;22(8):943-9, [**ISI-JIF 6.02**]

Color Doppler, power Doppler and B-flow ultrasound in the assessment of ICA stenosis: Comparison with 64-MD-CT angiography. Clevert DA, Johnson T., Jung EM, Clevert D.-A., Flach PM, Strautz TI, Ritter G, Gallegos MT, Kubale R., Becker C., Reiser M., **Eur. Radiol**. 2006 Nov 22; [**ISI-JIF 3.173**]

Imaging of aortic abnormalities with contrast-enhanced ultrasound. A pictorial comparison with CT. Clevert DA, Stickel M., Johnson T., Glaser C., Clevert D.-A., Steitz HO, Kopp R., Jauch KW, Reiser M., **Eur. Radiol**. 2007 [**ISI-JIF 2.437**]

High-grade stenoses of the internal carotid artery: Comparison of high-resolution contrast enhanced 3D MRA, duplex sonography and power Doppler imaging, Clevert DA, Johnson T., Michaely H., Jung EM, Flach PM, Strautz TI, Clevert D.-A., Reiser M., Schoenberg SO., **Eur J Radiol**. 2006 [**SI-JIF 1.332**]

Contrast-enhanced ultrasound versus MS-CT in blunt abdominal trauma, Clevert DA, Weckbach S., Minaifar N., Clevert D.-A., Stickel M., Reiser M., **Clin Hemorheol Microcirc.** 2008; 39(1-4):155-69. [**ISI-JIF 1.037** (**2007**)]

Color duplex ultrasound and contrast-enhanced ultrasound in comparison to MS-CT in the detection of endoleak following endovascular aneurysm repair., Clevert DA, Minaifar N., Weckbach S., Kopp R., Meimarakis G., Clevert D.-A., Reiser M., **Clin Hemorheol Microcirc.** 2008; 39(1-4):121-32. [**ISI-JIF 1.037 (2007)**]

Modern ultrasound diagnostics of deep vein thrombosis in lung embolism of unknown origin, Clevert DA, Jung EM, Pfister K., Stock K., Schulte-Altedorneburg G., Fink C., Clevert D.-A., Reiser M., **Radiologe**. 2007 Aug;47(8):673-684 [**ISI-JIF 0.72**]

Chip card assisted safety of communication in German public health, Clevert D.-A., Schober-Halstenberg H.-J. & Frei U. **Stud Health Technol Inform**: 77(2) 1096-7, 2000

BUCHBEITRÄGE

Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R [Chapter] Order Restricted Clustering, A. Kasim, S. Van Sanden, Z. Shkedy, S. Hochreiter, D.-A. Clevert, W. Talloen and D. Lin, **Springer**, 2011

Handbook of Research on Systems Biology Applications in Medicine, [Chapter] The Affymetrix GeneChip Microarray Platform, D.-A. Clevert, Axel Rasche, **IGI Publishing Group**, ISBN [978-1-60566-076-9], 2008

KONFERENZBEITRÄGE

ICHG/ASHG 2011 - 12th International Congress of Human Genetics & 61st MeetingThe American Society of Human Genetics:
„An analytical pipeline for detecting copy number variations with a low false discovery rate in microarray data“
„Copy number aberrations affecting adhesion genes involved in the development of the cerebellar vermis are associated with autism spectrum disorders“
„Accurate Detection of Copy Number Variations in Next Generation Sequencing Data by a Latent Variable Model“

ISMB/ECCB 2011 - 19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 9th European Conference on Computational Biology (ECCB):
„Identifying Copy Number Variations based on Next Generation Sequencing Data by a Mixture of Poisson Model“
„Detection of Nonlinear Effects in Gene Expression Pathways“
„FABIA: Factor Analysis for Bicluster Aquisition“
„ cn.FARMS: a probabilistic latent variable model to detect copy number variations with a low false discovery rate “
„Detecting rare copy number variations (CNVs) with sparse coding“

CAMDA 2011 - 11th Critical Assessment of Massive Data Analysis:
„Controlling the false discovery rate at detection of biological aberrations in -omic data“

HGV 2011 - 12th International Meeting on Human Genome Variation and Complex Genome Analysis:
„A low false discovery rate at detection of copy-number aberrations in microarray data “
„Copy Number Aberrations Affecting the Developing Cerebellar Vermis are Associated with Autism Spectrum Disorders“
„cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in Next Generation Sequencing Data“

IB 2011 - International Symposium on Integrative Bioinformatics 2011:
„Fabia: a biclustering method for simultaneous analysis of miRNA and mRNA data“

AGD 2010 - 25th Annual AGD Meeting 2010:
„FABIA: Factor Analysis for Bicluster Aquisition“

FGED 2010 - 13th International Meeting of the Functional Genomics Data Society:
„cn.FARMS - a probabilistic model to detect DNA copy numbers“

ISMB 2010 - 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB):
„Detection of Nonlinear Effects in Gene Expression Pathways“
„FABIA: Factor Analysis for Bicluster Aquisition“
„cn.FARMS - a probabilistic model to detect DNA copy numbers“
„Detecting rare copy number variations (CNVs) with sparse coding“

HGV 2009 - 11th International Meeting on Human Genome Variation and Complex Genome Analysis:
„I/NI-calls: a novel latent variable model for unsupervised feature selection“

ISMB/ECCB 2009 - 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB):
„Sparse Factor Analysis for Detecting Copy Number Variations (CNVs) “
„Construction of Metagenes by Conditional Factor Analysis“
„I/NI-calls: a novel latent variable model for unsupervised feature selection“

MGED11 -11th International Meeting of MGED:
„I/NI-calls: a novel latent variable model for unsupervised feature selection“
„cn.FARMS - a probabilistic model to detect DNA copy numbers“

Dagstuhl Seminar 09081 - Similarity-based learning on structure:
FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe level

AGD 2008, 23th Annual AGD Meeting 2008:
„Detecting DNA copy numbers with probabilistic latent variable models“

NCS 2008 - 2nd Non-Clinical Statistics Conference**:**
„FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe level“

ISMB 2008 - 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB):
„FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe set level“
„I/NI-calls: a novel latent variable model for unsupervised feature selection“
„cn.FARMS - a probabilistic model to detect DNA copy numbers“

AGD 2007, 22th Annual AGD Meeting 2007:
„FARMS - informative normalization“

GCB 2007 - 12th German Conference on Bioinformatics 2007
„I/NI-calls: a novel unsupervised feature selection criterion“

PMCB 2007 - Probabilistic Modelling in Computational Biology 2007
„I/NI-calls: a novel unsupervised feature selection criterion“

ISMB/ECCB 2007 - 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB):
„FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe set level“
„I/NI-calls: a novel unsupervised feature selection criterion“

DIA 2007 - 19th EuroMeeting Drug Information Association, Wien, 2007
„Gene Selection for Disease Identification Using Microarrays“

GMDS 2000 - 45. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, Hannover , 2000
„Chipkartengestützte Sicherheit bei der Kommunikation im deutschen Gesundheitswesen“

Selbstständigkeitserklärung

„Ich, Djork-Arné Clevert, erkläre, dass ich die vorgelegte Dissertation mit dem Thema:

Entwicklung und Vergleich biostatistischer Methoden zur Auswertung von Microarray Experimenten

selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe.“

Datum Unterschrift

Danksagung

Frau Prof. Dr. med. Petra Reinke danke ich herzlich für ihre Unterstützung bei der Umsetzung der Promotionsarbeit.

Mein ganz besonderer Dank gilt Herrn Prof. Dr. rer. nat. Sepp Hochreiter für sein Engagement, die anregenden Diskussionen, für seine außerordentliche Betreuung und die konstruktive Unterstützung bei allen Fragen.

Für die zahlreichen Hinweise und Anregungen sowie aufmunternde Gespräche bedanke ich mich bei Martin Heusel.

Des Weiteren bedanke ich mich bei allen Mitgliedern der Arbeitsgruppe für die angenehme Arbeitsatmosphäre und uneingeschränkte Hilfsbereitschaft.

Herrn Dr. rer. nat. Hinrich Göhlmann, Herrn Dr. rer. nat. Willem Talloen und Frau Dr. rer. nat. Marianne Tuefferd danke ich für ihre Unterstützung und die langjährige Kooperation.

Meinen Eltern möchte ich für ihre immerwährende Unterstützung und ihr geduldiges Verständnis herzlich danken.

Mein größter Dank geht an meine Ehefrau, Liane Clevert, für die moralische Unterstützung und den großartigen Rückhalt. Sie stand mir während meiner gesamten Promotionszeit immer zur Seite.