Beeinflussen Filtermethoden die Liste differenziell exprimierter Gene?

Merk S, Dugas M
Institut für Medizinische Informatik und Biomathematik, Universität Münster, Deutschland
Sylvia.Merk@ukmuenster.de

Einleitung und Fragestellung

Mit Hilfe der Microarray-Technologie ist es möglich, die momentane simultane Genaktivität (Transkriptom) einer Zelle bzw. eines Gewebes zu messen[1-2]. Microarrays der neueren Generation enthalten Messpunkte für mehr als 50.000 ProbeSets, die eine enorme Datenmenge generieren. Es werden meist Whole Genome-Arrays verwendet, die inzwischen fast das gesamte menschliche Genom umfassen. Da in dem zu untersuchenden Gewebe allerdings nicht alle auf dem Chip enthaltenen Gene tatsächlich exprimiert werden, ist die resultierende Datenmenge wesentlich größer als erforderlich. Die Durchführung eines Filterschritts vor der eigentlichen Analyse führt zu einer Verkleinerung der Datenmenge und trägt somit auch dazu bei, den Einsatz der Computerressourcen sowie die Rechenzeit für nachfolgende Analysen zu reduzieren. Von Interesse ist hierbei die Frage, ob der Einsatz von Filtern die Liste differenziell exprimierter Gene beeinflusst.

Material und Methoden

Für die Analysen wurden drei frei verfügbare Datensätze mit jeweils zwei Analysegruppen verwendet. Die Auswertung der Daten erfolgte mit R 2.2.1 [3] sowie den Bioconductor-Packages affy 1.8.1, multtest 1.8.0 und genefilter 1.8.0 [4]. Nach Einlesen der CEL-Files und anschließender RMA-Normalisierung wurden mit Hilfe von drei unterschiedlichen Filterkriterien Auswertedatensätze erstellt und mittels Welch-t-Test Listen von differenziell exprimierten Genen generiert. Zur p-Wert-Korrektur für multiples Testen wurde die False-Discovery-Rate (<0.02 für Datensatz 1 und 2, <0.001 für Datensatz 3) nach Benjamini & Yekutieli bestimmt.

Beschreibung der Datensätze:

Datensatz 1 [5]: Kardiologie, Affymetrix GeneChip® HG-U133A, 22283 ProbeSets, 10 Vorhofflimmern vs. 20 Sinusrhythmus
Datensatz 2 [6]: Pneumologie, Affymetrix GeneChip® HG-U133A, 22283 ProbeSets, 34 Raucher vs. 23 Nichtraucher
Datensatz 3 [7]: Kardiologie, Affymetrix GeneChip® HG-U133Plus2, 54675 ProbeSets, 27 Idiopathische Kardiomyopathie vs. 32 Ischämische Kardiomyopathie

Beschreibung der Filter:

1. Absent/Present-Call: Es werden die Perfect Match und Mismatch-Informationen des Chips ausgenutzt und nach Vergleich mit einem Schwellenwert entschieden, ob das einzelne ProbeSet in dieser Probe als Present, Marginal oder Absent gewertet wird. Um der Tatsache Rechnung zu tragen, dass Gene nur in einer der zu untersuchenden Gruppe reguliert sein können, wird gefordert, dass mindestens die Hälfte der Probenzahl der kleineren Gruppe als Present oder Marginal gewertet sein muss.
2. Expressionswert: Es sollen diejenigen ProbeSets aussortiert werden, die einen geringen Expressionswert (<100) aufweisen. Um eine mögliche gruppenspezifische Regulation zu berücksichtigen, wird als Schwelle die Hälfte der Probenzahl der kleineren Gruppe verwendet.
3. Variationskoeffizient: Mit diesem Filter sollen ProbeSets mit niedriger Variabilität über sämtliche Proben ausgeschlossen werden. Der Variationskoeffizient wird für jedes ProbeSet bestimmt und anschließend absteigend sortiert. Es werden diejenigen 50% der ProbeSets selektiert, die den höchsten Korrelationskoeffizienten aufweisen.

Ergebnisse

Durch den Einsatz der verschiedenen Filter wurden die Datensätze deutlich reduziert (Tabelle 1). Der Anteil an differenziell exprimierten Genen erhöht sich durch den Einsatz von Filtern (Tabelle 2).

	DS 1	DS 2	DS 3
F1	58.6	46.1	48.0
F2	79.6	26.6	69.7
F3	50.0	50.0	50.0

Tab.1: Anteil (%) der nach dem Filtern verbliebenen ProbeSets bezogen auf die Gesamtzahl der ProbeSets

DS: Datensatz, F: Filter, KF: Kein Filter

	DS 1	DS 2	DS 3
KF	7.0	2.5	6.0
F1	10.0	5.6	12.1
F2	8.4	7.2	9.0
F3	13.5	5.4	13.2

Tab.2: Anteil (%) der nach dem Filtern differenziell exprimierten ProbeSets bezogen auf die Anzahl der verbliebenen ProbeSets

Der Vergleich aller vier aus den gefilterten Datensätzen resultierenden Genlisten ergab für den Datensatz 1 eine Übereinstimmung von 57%, für den Datensatz 2 eine Übereinstimmung von 48% und für den Datensatz 3 eine Übereinstimmung von 63%. Die Tabellen 3-5 enthalten eine Auflistung des Vergleichs der Genlisten von jeweils zwei Filtermethoden für jeden Datensatz.

DS 1	KF	F1	F2	F3
KF		88.0	92.6	78.8
F1	72.4		75.8	71.3
F2	87.4	87.0		79.1
F3	74.8	82.2	79.5	

Tab.3

DS 2	KF	F1	F2	F3
KF		78.6	74.7	79.1
F1	82.1		85.2	82.9
F2	58.1	63.4		57.7
F3	86.9	87.1	81.5	

Tab.4

DS 3	KF	F1	F2	F3
KF		76.3	91.3	72.4
F1	73.1		78.6	78.0
F2	87.3	85.2		80.4
F3	78.7	88.7	84.3	

Tab.5

Tab. 3-5: Paarweiser Vergleich der Filterverfahren. Übereinstimmung (%) der ProbeSet-Listen bezogen auf die Anzahl der ProbeSets beider beteiligter Listen.

Diskussion

Ziel der Arbeit war die Untersuchung, ob und inwiefern die Anwendung von Filterkriterien die Liste der als differenziell exprimiert identifizierten Gene verändert.

Ein Vergleich der durch mindestens einen Filter eliminierten ProbeSets mit der Liste der differenziell exprimierten ProbeSets des ungefilterten Datensatzes zeigt für alle drei Datensätze, dass ProbeSets entfernt werden, die in dem ungefilterten Datensatz als differenziell exprimiert eingestuft sind. Dass diese „falschen" ProbeSets überhaupt in der Liste der differenziell exprimierten ProbeSets erscheinen ist möglicherweise dadurch zu erklären, dass das Target auf dem Chip nicht spezifisch genug ist und durch unspezifische Transkriptanlagerung ein Fluoreszenzsignal erzeugt wird, welches dann von dem Absent/Present-Call-Algorithmus identifiziert wird. Aufgrund dieser Beobachtung scheint es bei dem Einsatz von Whole Genome Arrays unerlässlich zu sein, vor der Analyse geeignete Filterverfahren anzuwenden, um ProbeSets zu eliminieren, die in den zu untersuchenden Proben keine Expression aufweisen.

Für die Datensätze 1 und 3 zeigt sich, dass ein großer Anteil der aussortierten ProbeSets aufgrund des Absent/Present-Kriteriums als nicht exprimiert gewertet wird. Hingegen wird bei dem Datensatz 2 ein großer Anteil der ProbeSets aufgrund des Expressionswert-Filters ausgeschlossen. In Kombi-

nation mit der Tatsache, dass bei diesem Datensatz nach dem Filtern nur 26.6% der ProbeSets verbleiben, liegt der Schluss nahe, dass hier das Expressionsniveau insgesamt niedrig ist.
Das Kriterium, dass mindestens die Hälfte der Probenanzahl der kleineren Gruppe den Filter passieren muss, ist wenig stringent und führt trotzdem zu einer deutlichen Datenreduktion. So wird der Datensatz 1 durch die Anwendung des Absent/Present-Filters für eine Probenzahl von 5 (16% von 30) von 22283 ProbeSets um 42% auf 13067 ProbeSets reduziert.
Die Untersuchung verschiedener Filtermethoden hat gezeigt, dass z.B. durch die Berücksichtigung der Absent/Present-Informationen zum einen der Datensatz erheblich verkleinert werden kann und zum anderen ProbeSets eliminiert werden können, die in den zu untersuchenden Proben nicht exprimiert sind. Die Auswahl einer geeigneten Filtermethode hängt allerdings sowohl von der Fragestellung als auch von dem zu untersuchenden Datensatz ab.

Literatur

[1] Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999; 21(1 Suppl):20-4.
[2] Butte A. The use and analysis of microarray data. Nat Rev Drug Discov. 2002; 1:951-60.
[3] R: Development core team (2004). R: A language and enviroment for statistical computing. R foundation for statistical computing. Vienna, Austria. http://www.r-project.org
[4] Gentleman RC, Carey VJ, Bates DM et al. Open software development for computational biology and bioinformatics. Genome Biology 2004; 5:R80.
[5] http://www.ncbi.nlm.nih.gov/geo/ Accession number GSE2240
[6] http://www.ncbi.nlm.nih.gov/geo/ Accession number GSE994
[7] http://cardiogenomics.med.harvard.edu/groups/proj1/pages/download_home1.html