Acta Biochimica Polonica

Vol. 56 No. 2/2009, 271–277

on-line at: www.actabp.pl

Regular paper

Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays

Roman Jaksik[1], Joanna Polańska[1], Robert Herok[2] and Joanna Rzeszowska-Wolny[1,2]✉

[1]*System Engineering Group, Institute of Automation, Silesian University of Technology, Gliwice, Poland;*
[2]*Department of Experimental and Clinical Radiobiology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice, Poland*

Received: 09 February, 2009; revised: 27 April, 2009; accepted: 08 May, 2009
available on-line: 12 May, 2009

Microarray methods have become a basic tool in studies of global gene expression and changes in transcript levels. Affymetrix microarrays from the HGU133 series contain multiple probe-sets complementary to the same gene (4742 genes are represented by more than one probe-set in a microarray HGU133A). Individual probe-sets annotated to the same gene often show different hybridization signals and even opposite trends, which may result from some of them matching transcripts of more than one gene and from the existence of different splice-variant transcripts. Existing methods that redefine probe-sets and develop custom probe-set definitions use mathematical tools such as Matlab or the R statistical environment with the Bioconductor package (Gentleman *et al.*, 2004, *Genome Biol.* 5: 280) and thus are directed to researchers with a good knowledge of bioinformatics. We propose here a new approach based on the principle that a probe-set which hybridizes to more than one transcript can be recognized because it produces a signal significantly different from others assigned to the particular gene, allowing it to be detected as an outlier in the group and eliminated from subsequent analyses. A simple freeware application has been developed (available at www.bioinformatics.aei.polsl.pl) that detects and removes outlying probe-sets and calculates average signal values for individual genes using the latest annotation database provided by Affymetrix. We illustrate this procedure using microarray data from our experiments aiming to study changes of transcription profile induced by ionizing radiation in human cells.

Keywords: transcript profiles, Affymetrix microarrays, multiple probe-sets, outlier detection, NucleoDix computer program

INTRODUCTION

DNA microarray technology allows measurement of the abundance of thousands of specific transcripts in an RNA sample (Lockhart *et al.*, 1996; Ramsay, 1998; Stoughton, 2005; Perez-Iratxeta *et al.*, 2005). The Affymetrix technology is the most widely used in human transcript profiling, and microarrays from the HGU133 series have the particular characteristic that each probe-set consists of eleven Perfect Match (PM) 25-mer oligonucleotide probes that can hybridize with the respective gene, together with an additional set of eleven Mismatch probes (MM) containing a single mismatch at the 13th position, which serve as specificity controls by comparison with the corresponding Perfect Match probes. In the HGU133A microarrays used in the present study there are 247 965 PM and MM probes altogether grouped into 22 283 probe-sets that represent only about 13 thousand annotated genes. A PM probe-set and its corresponding MM set are adjacent to each other, but different pairs assigned to a given gene are located in different regions of the microarray to prevent errors caused by effects of increased luminescence intensity in certain regions (Affymetrix, 2004). In spite of the careful microarray design, differences in signal strength from different probes assigned to one probe-set were observed (Li & Wong, 2001) and some discrepancies in the original probe-set/gene assignments in these microarrays

✉Corresponding author: Joanna Rzeszowska-Wolny, Department of Experimental and Clinical Radiobiology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology (Gliwice Branch), Wybrzeże Armii Krajowej 15, 44-101 Gliwice, Poland; tel.: (48) 32 278 9677; fax: (48) 32 231 3512; e-mail: jwolny@io.gliwice.pl
Abbreviations: MM, mismatch; PM, perfect match.

have been revealed by improved genome sequence annotations (Chalifa-Caspi *et al.*, 2004) and show that many probes match transcripts from more than one gene or even do not match any transcribed sequence (Mecham *et al.*, 2004; Gautier *et al.*, 2004; Harbig *et al.*, 2005; Stalteri & Harrison, 2007; Yu *et al.*, 2007). Hybridization signals read from different probe-sets annotated to the same gene are often different and may even show opposite trends in the direction of change of transcript level. Attempts have been made to resolve these problems, most of which propose procedures for redefining probe-sets and for developing custom probe-set definitions for Affymetrix gene chips that greatly improve the reliability of the results (Dai *et al.*, 2005; Lu *et al.*, 2007; Ferrari *et al.*, 2007). Choosing one probe-set as representative or treatment of each probe-set as individual gene was also proposed (Jordan *et al.*, 2005; Elbez *et al.*, 2006; Bourquin *et al.*, 2006; Liao & Zhang, 2006; Li *et al.*, 2008). Most of the methods are based only on gene sequence and annotation information available in public databases and do not take into account possible inconsistencies in the experimental data.

Here we propose a different approach which is based on the principle that a probe-set which hybridizes to more than one transcript can be recognized because it produces a hybridization signal significantly different from those of other probe-sets assigned to the particular gene, allowing it to be detected as an outlier in the group and eliminated from subsequent analyses. We describe this method using examples of transcript profile changes induced by exposure of human cells to ionizing radiation assessed with Affymetrix HG-U133A microarrays and based on single probe-sets only or on all probe-sets annotated to a particular gene, and provide a link to a free computer program that allows such calculations for large numbers of genes.

MATERIALS AND METHODS

Cells and irradiation. K562 (human lymphoblastoid) and Me45 (human melanoma) cells were grown in suspension in DMEM (Sigma-Aldrich, St. Louis, MO, USA) with 10% fetal bovine serum (ICN, Irvine, CA, USA) and used at a density of 10^5 cells/ml. Cultures were irradiated 24 h after a change of medium using an X-ray dose of 4 Gy at 1 Gy/min from a Clinac 600 GMV (Varian, Palo Alto, CA, USA) at room temperature, and were resuspended in fresh medium after irradiation. Control untreated cells and irradiated cells were processed in parallel in similar conditions and collected at different times of incubation at 37°C.

Microarray assays of transcript levels and normalization of microarray data. Total RNA was isolated from about 3×10^6 cells with RNeasy Mini Kits (Qiagen, Valencia, CA, USA) including a digestion step with RNase-free DNase I, and its quantity and integrity were checked spectrophotometrically and by electrophoresis in 1% agarose gels. Materials and methods for microarrays were from Affymetrix (Santa Clara, CA, USA); double-stranded cDNA prepared with the GeneChip Expression 3′-Amplification One-Cycle cDNA Synthesis Kit was cleaned using the Sample Cleanup Module and biotinylated cRNA was synthesized with GeneChip Expression 3′-Amplification Reagents for IVT Labeling, cleaned on RNA Sample Cleanup columns, and fragmented at 94°C for 35 min in Fragmentation Buffer. The biotinylated cRNA was hybridized first to a control Test3 microarray to evaluate its quality and then to a Human Genome U133A array. Chips were stained with streptavidin-phycoerythrin conjugate and scanned in a Gene Array G2500A scanner (Agilent, Santa Clara, CA, USA). Signals from replicate experiments were normalized by Robust Multiarray Analysis (RMA) (Irizarry *et al.*, 2003; Bolstad, 2007).

Statistical methods. The multiple hybridization signal values obtained for a particular gene were tested for outliers by the Dixon test (Dixon, 1953), which applies to small series of 3–30 data points and is based on a ratio that describes the difference between minimal, maximal and adjacent values for two given extreme values in the group analyzed, allowing a conclusion to be made whether the minimal or maximal value is an outlier. A computer application which enables fast identification and elimination of outliers using the Dixon test, and calculation of average hybridization values for large groups of genes on the basis of the latest Affymetrix probe-set annotation database, is freely available at www.bioinformatics.aei.polsl.pl.

RESULTS AND DISCUSSION

Affymetrix microarray assays give different results from different probe-sets for the same gene

RNA was isolated at different times after exposure of K562 or Me45 cells to 4 Gy of X-radiation and the levels of transcripts from different genes were assessed by hybridization to Affymetrix HG-U133A microarrays which include multiple probe-sets assigned to the same gene. Table 1 shows the numbers of annotated genes which are characterized by one, two, or more probe-sets on this microarray (excluding 1239 sets with unspecified gene annotations), calculated using the annotation files obtained

Table 1. Probe-sets assigned to a single gene on HGU133A microarrays.

Number of probe-sets/gene	1	2	3	4	5	6	7	8	9	10+
Number of genes	8310	2780	1230	464	162	64	14	11	7	10

on 16 December 2008 from the Affymetrix website (Liu *et al.*, 2003).

We found that in many cases the probe-sets assigned to a single gene produced different hybridization signal values and sometimes opposite trends. Table 2 presents examples of results obtained with such probe-sets for transcripts in control cells and in cells at 1, 12, and 24 h after X-irradiation.

In Table 2 the probe-sets assigned to a particular gene by the Affymetrix database are grouped; the first two columns show the gene symbol and probe-set numbers. For these genes, signals obtained from different probe-sets varied widely and in some cases by an order of magnitude. To resolve this problem and to obtain a single hybridization signal value for each gene, we removed from the subsequent analysis those probe-sets that could be recognized as outliers in the group assigned to one gene.

Signal averaging combined with the Dixon test for outliers

The Dixon method for removing outliers was chosen for this purpose because it was designed for

Table 2. Microarray hybridization signals from multiple probe-sets.

Gene symbol	Probe-set	K562 cells				Me45 cells			
		Control	After irradiation			Control	After irradiation		
			1 h	12 h	24 h		1 h	12 h	24 h
PTBP1	***212016_s_at***	***329.7***	***299.1***	***252.1***	***410.2***	***175.4***	***129.4***	***58.7***	257.8
	212015_x_at	1059.6	1294.4	1127.0	1094.6	546.9	554.1	398.0	804.2
	211271_x_at	1215.3	1265.8	1204.0	1111.2	623.2	564.7	429.7	839.7
	202189_x_at	1455.0	1472.7	1390.5	1174.6	737.3	724.7	873.1	955.5
	211270_x_at	1589.2	1552.7	1492.0	1354.9	766.7	831.7	1034.4	1007.2
	216306_x_at	1606.3	1759.9	1581.4	1586.8	826.5	777.2	643.4	1174.5
BAT2D1	214052_x_at	27.7	25.7	22.1	23.9	25.7	31.9	17.9	24.2
	211947_s_at	54.8	101.0	88.6	97.0	62.3	55.3	37.1	51.2
	214055_x_at	85.4	113.7	71.4	64.8	45.7	46.9	25.1	38.7
	211944_at	105.3	110.8	71.9	93.5	58.2	49.8	26.4	53.2
	211948_x_at	180.1	217.9	159.5	138.1	110.2	106.2	64.5	90.9
	211946_s_at	***403.3***	***477.2***	***433.8***	***386.4***	***244.1***	***279.5***	***323.1***	***217.5***
HUWE1	***207783_x_at***	***2869.9***	***2718.8***	***2698.1***	***2986.0***	***3391.7***	***3502.1***	***3235.8***	***3349.9***
	208598_s_at	376.4	325.7	392.1	505.8	516.3	464.9	408.1	536.7
	208599_at	24.9	21.0	22.1	23.1	23.9	21.5	20.5	21.0
	214673_s_at	29.5	28.4	29.2	29.3	26.8	22.6	22.4	21.6
ATP5C1	*214132_at*	*29.5*	*24.7*	*31.6*	*34.1*	*33.9*	*41.9*	*59.4*	*36.7*
	208870_x_at	1781.7	1680.9	1499.0	2326.2	1615.7	1972.1	1709.9	1639.5
	213366_x_at	1840.0	1936.5	1773.3	2236.5	1707.7	1982.7	1985.6	1623.9
	205711_x_at	1906.4	1883.1	1622.8	2475.6	1683.5	2025.0	1824.3	1674.4
RPL38	202028_s_at	182.3	209.0	199.6	147.9	233.8	308.3	325.1	283.4
	221943_x_at	288.3	312.9	289.8	199.7	416.3	471.8	375.6	443.0
	202029_x_at	***4534.8***	***4320.7***	***4672.5***	***4040.6***	***5499.5***	***5144.2***	***6413.6***	***5089.9***

Table 3. Average hybridization results after detection and removal of outliers.

Gene symbol	K562 cells				Me45 cells			
	Control	After irradiation			Control	After irradiation		
		1 h	12 h	24 h		1 h	12 h	24 h
PTBP1	1385.1	1469.1	1359.0	1264.4	1669.0	1993.3	1839.9	1645.9
RPLP0	8173.1	7742.5	7995.9	8738.7	60.4	58.0	34.2	51.6
BAT2D1	90.6	113.8	82.7	83.5	700.2	690.5	572.9	956.2
HUWE1	143.6	125.0	147.8	186.1	189.0	169.6	150.3	193.1
ATP5C1	1842.7	1833.5	1631.7	2346.1	325.1	390.1	350.4	363.2
RPL38	235.3	261.0	244.7	173.8	8165.3	8320.1	8169.3	7979.9

small groups of values, from 3 to 30 (Dixon, 1953). In Table 2, rows in bold italic type show probe-sets which were classified as outliers by this test at a significance level of $\alpha \leq 0.05$. The first example (top row) shows probe-sets classified as outliers because of their significantly lower signal value than the others. The *PTBP1* gene has four different transcript variants and four probe-sets that match these transcripts exactly (10–11/11 probes hybridize with a transcript of this gene according to its sequence in the Reference Sequence database). The location of the sequence which hybridizes with the fifth probe-set is very similar, whereas the probe-set which gives the outlying hybridization value is complementary to a sequence about 1000 bp closer to the 5′ end of the gene. The lower level of transcripts matching probe-set number 212016_s_at could therefore result from degradation of transcripts starting from the 5′ end.

For the *BAT2D1* gene, one probe-set which hybridize with sequences located about 3000–4000 bp closer to the 3′ end than the others and showing hybridization signals significantly higher than the others, were recognized as outlier, and removed from the subsequent analysis. The sequences of probe-set 214052_x_at do not match sequences in the *BAT2D1* gene at all; however, this set showed a signal close to the noise level measured for others and thus was not recognized as an outlier by the Dixon test. The gene *HUWE1* is represented by four probe-sets, with set 207783_x_at classified as an outlier because it showed a much higher signal compared to the others representing that gene; the main reason may be that 6 of 11 probes belonging to that set hybridize with the gene *TPT1* instead of *HUWE1*. The average expression value of the other probe-sets specific for *TPT1* was 3000, which may have caused the strongly increased value of 207783_x_at. One of the probe-sets assigned to the gene *ATP5C1* was classified as an outlier in all experiments; the transcript level detected by this set was close to the noise level while other sets showed signal values about 60 times higher. Probes of the removed set (214132_at) did not match the transcript sequence, while those in the other sets showed 10–11/11 matches. Gene *RPL38* illustrates a weak point in our approach: only one of the three probe-sets for this gene showed 10/11 sequence matches to the transcript, but it was classified as an outlier since its value was significantly different from the other two although those did not match the gene sequence at all (0/11 matches).

The hybridization signals measured by probe-sets for the same gene can be not only significantly different in some or all experiments, but can also show opposite trends (up- or down-regulation) as seen in the example of the *PTBP1* gene (Table 2), where the probe-set classified as an outlier shows a reduced transcript level immediately after irradiation while the others show an increased level. Such differences in expression can result not only from inaccurate measurements or normalization methods, but also from a different location of the sequence hybridizing with the probe-set; for example, transcripts could be broken as a consequence of genotoxic effects and therefore hybridize differently with different probe-sets. RNAs that do not have a marked poly-A tail are not converted to cRNA and therefore are not detected, although they could potentially hybridize with a probe-set. Different expression values can also be a result of inaccurate gene sequencing at the time when the probes were designed (Heydebreck *et al.*, 2004), but also of hybridization with different members of a family of similar genes or with various splice variants of a single gene, which increase the diversity of over 60% of human genes (Ladd & Cooper, 2002) and very often can lead to totally different expression levels of individual mature transcripts (Buck *et al.*, 1992; Lim *et al.*, 2006).

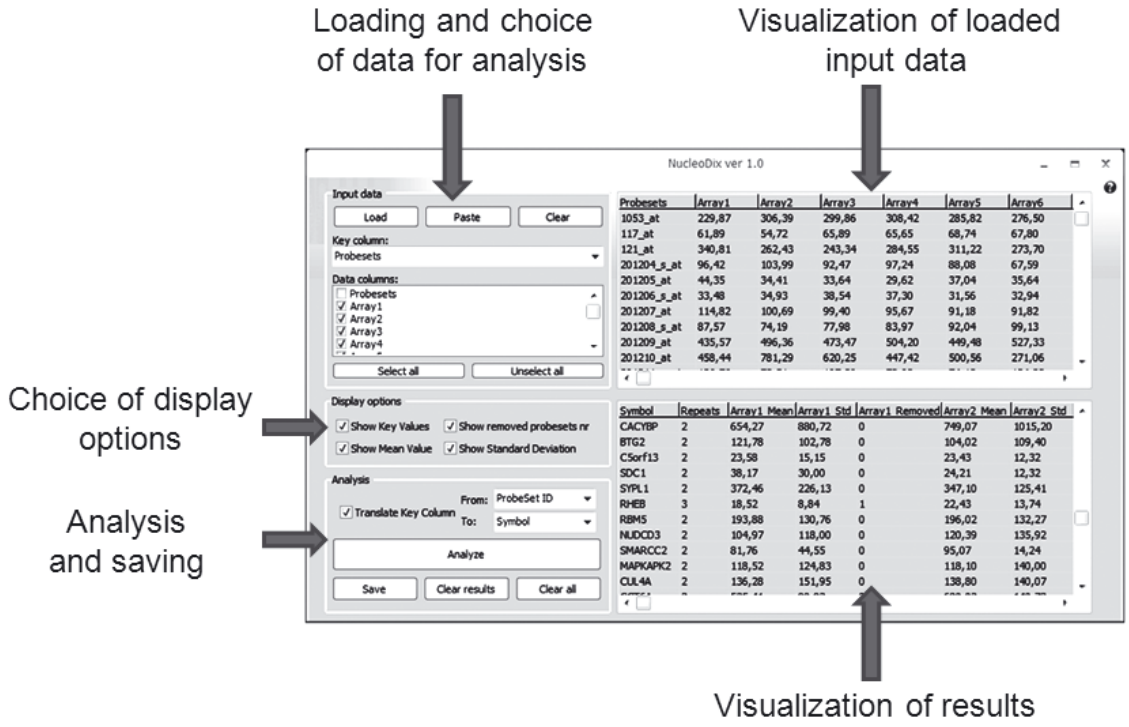

Figure 1. The graphical interface of computer application NucleoDix.

The differences in read-outs from individual probe-sets make it impossible directly to assess from microarray data the changes in gene expression caused, for example, by ionizing radiation. The method described here, in which outlying values are detected and removed from the analysis, enables calculation of the average signal assigned to each gene and reasonable comparisons of transcript levels between samples, since it removes any probe-sets which show a significantly different fold-change of expression of a gene. Table 3 presents the data for the genes in Table 2 after recalculation by this method. On the average, 400 of 22215 probe-sets were removed by this operation from the data set from each microarray.

Data analysis software

We developed a user-friendly bioinformatics tool to perform the Dixon test, as well as signal averaging for probe-sets for the same gene. Access to the data-processing algorithms is provided by a simple graphical user interface (Fig. 1).

This program detects outlying signals using the Dixon test and removes the data for the corresponding probe-set, followed by signal averaging. The interface usage can be divided into three steps: data import, display options, and analysis. In the first step the user can import the data from many microarray experiments in the form of text files or by pasting the data table. Imported data appear in the upper Table (see Fig. 1). The option is provided to select a labelled column (probe-sets, gene symbols or other gene identifiers) to be analyzed ("Key" column) and columns for calculations from the displayed list (only the selected columns will be analyzed in the particular run). The "Display option" box enables the choice of the result data that will be displayed in one row with each gene identifier, the mean expression value ("Mean value") calculated after excluding the values from the outlying probe-sets, the standard deviation ("Standard deviation"), the number of outlying probe-sets ("Removed probe-sets") and, in the last column of the results, the names of all probe-sets which were annotated to each gene ("Key values"). Coming to the analysis step, the user should choose the input and output symbol type (i.e., wheather the key column labels should be translated from probe-sets to symbols or other identification numbers before looking for values annotated to the same gene). The "Analyze" button starts the calculations and the results appear in the lower Table (see Fig. 1) and can be saved as a text file by using the "Save" button. The columns of the lower Table present the results of the analysis and contain the information chosen in the "Display options" box and also, in the second column, the information on the number of probe-sets annotated to the particular gene ("Repeats"). The program was designed especially for analyzing data

from HGU133A microarrays, but it can also be used for other types of microarrays or other types of experiments as long as the labels column is provided with the experimental data.

CONCLUDING REMARKS

The analysis of microarray results raises many questions, mainly because even a single experiment provides a huge amount of information that needs to be stored, processed and analyzed. The Affymetrix GeneChip is a very popular microarray platform for characterizing transcription profiles and has been widely used in functional genomics and in clinics, for example for classification of cancers (Györffy & Schäfer, 2008; Karlsson *et al.,* 2008). The results of an experiment depend highly on the quality of the probe-set annotations, which should be as specific for a single gene as possible. The first Affymetrix microarrays were designed some years ago, and since that time the number of human gene sequences available in databases has increased enormously so that a large amount of the information used at the time these probe-sets were designed is outdated and some on the HGU133A microarray can hybridize to more than one gene or are not properly annotated to genes (Zhang *et al.,* 2005; Lu & Zhang, 2006; Okoniewski *et al.,* 2006). Comparisons of probes representing the same gene on different microarray platforms or on different generations of the same platform also show discrepancies (Kuo *et al.,* 2002; Kothapalli *et al.,* 2002; Carter *et al.,* 2005; Hwang *et al.,* 2004; Elo *et al.,* 2005). Many attempts have been made to match the probe-set definitions with the new information in public databases (Chalifa-Caspi *et al.,* 2004; Hwang *et al.,* 2004; Mecham *et al.,* 2004).

Our approach is based on averaging the signal from probe-sets representing the same gene based on the most up-to-date probe annotation data from Affymetrix. In theory, these signals should not be averaged since we cannot determine if they hybridize with the same efficiency (i.e., have equal weight) and therefore we first carry out a Dixon test for outliers and remove those probe-sets which give a signal significantly different from the others for the same gene, and only then the remaining values are averaged. This strategy can be justified in a biological sense by considering that the diversity among genes is much higher than the diversity between probe-sets related to the same gene after removal of outliers. The disadvantage of this method is that the results are questionable when there are only two probe-sets representing a single gene so that we cannot carry out the Dixon test, or when there are many values which vary over a large range. The main positive aspects are that no genes are removed from the analysis, and that we obtain a single expression value for each of them that can be used for further comparisons. This approach is not as strongly dependent on gene sequence as most others are; none of the probe-sets are removed if we cannot determine their relation to a transcript because the current knowledge about the individual gene sequences is not so precise as we would like it to be. The identification of differentially expressed genes is still the major goal of microarray-based expression studies, and a combination of modern bioinformatics technologies with the most up-to-date genomic information can significantly improve the outcome of microarray analyses even using data from experiments that were performed a few years ago.

Acknowledgements

This research was supported by the Ministry of Science and Higher Education, grant PBZ-MNiI-2/1/2005.

REFERENCES

Affymetrix (2004) *GeneChip Expression Analysis — Technical Manual*. pp 185.

Bolstad B (2008) RMAExpress. http://rmaexpress.bmbolstad.com/

Bourquin J, Subramanian A, Langebrake C, Reinhardt D, Bernard O, Ballerini P, Baruchel A, Cave H, Dastugue N, Hasle H, Kaspers G, Lessard M, Michaux L, Vyas P, Wering E, Zwaan C, Golub T, Orkinar S (2006) Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proc Natl Acad Sci USA* **103**: 3339–3344.

Buck K, Vanek M, Groner B, Ball RK (1992) Multiple forms of prolactin receptor messenger ribonucleic acid are specifically expressed and regulated in murine tissues and the mammary cell line HC11. *Endocrinology* **130**: 1108–1114.

Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z (2005) Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* **6**: 107.

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

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**: e175.

Dixon WJ (1953) Processing data for outliers. *Biometrics* **9**: 74–89.

Elbez Y, Farkash-Amar S, Simon I (2006) An analysis of intra array repeat: the good, the bad and the noninformative. *BMC Genomics* **7**: 136.

Elo LL, Lahti L, Skottman H, Kylaniemi M, Lahesmaa R, Aittokallio T (2005) Integrating probe-level expression

changes across generations of Affymetrix arrays. *Nucleic Acids Res* **33**: e193.

Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S (2007) Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics* **8**: 446–452.

Gautier L, Møller M, Friis-Hansen L, Knudsen S (2004) Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics* **5**: 111.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Györffy B, Schäfer R (2008) Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. *Breast Cancer Res Treat* Epub ahead of print.

Harbig J, Sprinkle R, Enkemann SA (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res* **33**: e31.

Heydebreck A, Huber W, Gentleman R (2004) Differential Expression with the Bioconductor Project. *Bioconductor Project Working Papers* Working Paper 7.

Hwang KB, Kong SW, Greenberg SA, Park PJ (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics* **5**: 159.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264.

Jordan I, Marino-Ramirez L, Koonin E (2005) Evolutionary significance of gene expression divergence. *Gene* **345**: 119–126.

Karlsson E, Delle U, Danielsson A, Olsson B, Abel F, Karlsson P, Helou K (2008) Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer. *BMC Cancer* **8**: 254.

Kothapalli R, Yoder SJ, Mane S, Loughran TPJ (2002) Microarray results: how accurate are they? *BMC Bioinformatics* **3**: 22.

Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**: 405–412.

Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3**: reviews0008.

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

Li H, Zhu D, Cook M (2008) A statistical framework for consolidating "sibling" probe sets for Affymetrix GeneChip data. *BMC Genomics* **9**: 188.

Liao B, Zhang J (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**: 530–540.

Lim SJ, Jung HH, Cho YA (2006) Postnatal development of myosin heavy chain isoforms in rat extraocular muscles. *Mol Vis* **12**: 243–250.

Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* **31**: 82–86.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**: 1675–1680.

Lu X, Zhang X (2006) The effect of GeneChip gene definitions on the microarray study of cancers. *BioEssays* **28**: 739–746.

Lu J, Lee JC, Salit ML, Cam MC (2007) Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics* **8**: 108.

Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* **32**: e74.

Okoniewski MJ, Miller CJ (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7**: 276.

Perez-Iratxeta C, Palidwor G, Porter CJ, Sanche NA, Huska MR, Suomela BP, Muro EM, Krzyzanowski PM, Hughes E, Campbell PA, Rudnicki MA, Andrade MA (2005) Study of stem cell function using microarray experiments. *FEBS Lett* **579**: 1795–1801.

Ramsay G (1998) DNA chips: state-of-the art. *Nat Biotechnol* **16**: 40–44.

Stalteri MA, Harrison AP (2007) Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* **8**: 13–27.

Stoughton RB (2005) Applications of DNA microarrays in biology. *Annu Rev Biochem* **74**: 53–82.

Yu H, Wang F, Tu K, Xie L, Li YY, Li YX (2007) Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics* **8**: 194.

Zhang J, Finney RP, Clifford RJ, Derr LK, Buetow KH (2005) Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85**: 297–308.