Reliability and reproducibility issues in DNA microarray measurements

Sorin Draghici[1], Purvesh Khatri[1], Aron C. Eklund[2] and Zoltan Szallasi[3]

[1]Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA
[2]Laboratory of Functional Genomics, Brigham and Women's Hospital, 65 Landsdowne Street, Cambridge, MA 02139, USA
[3]Children's Hospital Informatics Program, Harvard Medical School, Boston, MA 02115, USA

DNA microarrays enable researchers to monitor the expression of thousands of genes simultaneously. However, the current technology has several limitations. Here we discuss problems related to the sensitivity, accuracy, specificity and reproducibility of microarray results. The existing data suggest that for relatively abundant transcripts the existence and direction (but not the magnitude) of expression changes can be reliably detected. However, accurate measurements of absolute expression levels and the reliable detection of low abundance genes are difficult to achieve. The main problems seem to be the sub-optimal design or choice of probes and some incorrect probe annotations. Well-designed data-analysis approaches can rectify some of these problems.

Introduction

Since its introduction in 1995 [1], DNA microarray technology has evolved rapidly. Although all DNA microarrays are based on hybridization of nucleic acid strands, the available technical choices differ widely between platforms [2]. An important distinction is the length of the probes. Microarrays can be categorized as either: (i) cDNA arrays, usually using probes constructed with PCR products of up to a few thousands base pairs; or (ii) oligonucleotide arrays, using either short (25–30mer) or long oligonucleotide (60–70mer) probes. The probes can be either contact-spotted, ink-jet deposited or directly synthesized on the substrate. Each of these approaches has its own requirements in terms of the amount of RNA needed, data acquisition, and transformation and normalization techniques [3]. These requirements together with the differences in types and composition of probes, deposition technologies, and labeling and hybridization protocols, frequently result in poor reproducibility of results from one platform to another. Here, we discuss these issues and the extent to which they can affect the outcome of a microarray experiment.

What is expected from microarrays?

Searching for determinants of a phenotype using gene expression levels requires suitable coverage of the genome coupled with reasonable reproducibility, accuracy (Box 1) and sensitivity in the technology employed. These limitations matter less if microarrays are used for screening because changes in gene expression can be verified independently. However, the stakes were raised when microarrays were suggested as a diagnostic tool in molecular disease classification [4,5] (Box 2), because regulatory agencies, such as the Food and Drug Adminis tration (FDA), require solid, empirically supported data about the accuracy, sensitivity, specificity, reproducibility and reliability of diagnostic techniques (http://www.fda.gov/cdrh/oivd/guidance/1210.pdf). As it will become apparent, the first decade of microarray technology produced rather limited data pertinent to these issues.

Basic considerations of microarray measurements

There have been few methodologies in molecular biology where expectations were raised to such a high level, with so little evidence for the actual capabilities of the technology. For the first six years during which microarrays were available commercially, probe sequence information was not available. End users had to trust the manufacturer that a given probe actually quantified a specific transcript. In reality, many cDNA microarrays used a substantial number of incorrect probes [6–9] and a surprisingly large portion of Affymetrix microarray probes (up to 30–40% depending on the actual chip) were not present in high-quality sequence databases such as Refseq [10,11]. To its credit, Affymetrix (http://www.affymetrix.com) subsequently released its probe sequences, but few competitors followed. This situation seems to be changing, in no small part because of a community-wide microarray-validation project initiated and led by FDA researchers: the MicroArray Quality Control Project (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/index.htm). Apart from probe design issues, the discrepancies between an intended probe sequence and the actual sequence synthesized or deposited on the microarray also deserve some attention. The synthesis of nucleotide chains performed on solid surfaces, such as the technology used by Affymetrix, Agilent (http://www.agilent.com), Combimatrix (http://www.combimatrix.com) and NimbleGen (http://www.nimblegen.com) is not 100% accurate. This means that microarray probes directly synthesized on substrates will contain a significant number of nucleotide chains that are different from the design

Corresponding authors: Draghici, S. (sorin@wayne.edu), Szallasi, Z. (zszallasi@chip.org).

Available online 27 December 2005

www.sciencedirect.com 0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2005.12.005

Box 1. Definitions

Accuracy: can be defined as the degree of conformity of the measured quantity to its actual (true) value. Usually, measurements are affected by a bias, which makes the mean depart from the actual value. Given a set of measurements, the accuracy of the instrument or technique is usually measured by comparing some measure of central tendency of the measurements (e.g. mean and median) to the actual value. An ideally accurate technique would have the mean exactly equal to the actual value.

Precision: (also called reproducibility or repeatability) is the degree to which repeated measurements of the same quantity will show the same or similar results. Usually, measurements are affected by an error that makes repeated measurements differ from each other. Given a set of measurements, the precision is usually measured by comparing some measure of dispersion (e.g. variance or standard deviation) with zero. An ideally precise technique would have all measurements exactly equal (zero variance).

Accuracy and precision are completely independent. A technique can be accurate but not precise (the mean of several measurements is close to the actual value but the individual measurements vary considerably), precise but not accurate (the individual measurements are close to each other but their mean is far from the actual value) neither or both. If a result is both accurate and precise, it is valid.

Specificity: in the context of DNA microarrays, refers to the ability of a probe to bind to a unique target sequence. A specific probe will provide a signal that is proportional to the amount of the target sequence only. A non-specific probe will provide a signal that is influenced by the presence of other molecules. The specificity of a probe can be diminished by cross-hybridization, a phenomenon in which sequences that are not strictly complementary according to the Watson–Crick rules bind to each other. Cross-hybridization is also called non-specific hybridization.

sequence, owing to base skipping [12]. Microarray platforms using probes that are purified by high-performance liquid chromatography (HPLC) and then deposited on the solid surface, such as the CodeLink Arrays from GE healthcare (http://www.gehealthcare.com), have the advantage of containing an almost homogeneous population of probes, which increases the specificity of hybridization [13]. Unfortunately, these probes need to be deposited after synthesis, a process often marred by inconsistent feature shape and size. However, perhaps the most important concern related to microarrays is that all current technologies are based on the fundamental assumption that most microarray probes produce specific signals under a single, rather permissive hybridization condition. As we will see, this is probably not true [14].

Sensitivity

The sensitivity threshold of microarray measurements defines the concentration range in which accurate measurements can be made. In an attempt to assess the dynamic range of microarrays, Holland measured the range of transcript abundance for 275 genes in yeast, using kinetically monitored reverse-transcribed PCR (kRT–PCR) [15], and compared the results with data from cDNA and oligonucleotide arrays. The results from cDNA and Affymetrix arrays were reasonably consistent

Box 2. Gene expression microarrays in clinical research and diagnostics

The concept of disease classifiers based on gene expression assumes that a particular disease state is associated with the expression levels of a given set of genes. In principle, microarrays can be used for either the identification of multi-genic disease classifiers or they could be directly applied to clinical samples as a diagnostic tool. The use of microarrays in clinical diagnostics will depend on several factors, including the number of genes in a given multi-genic classifier and whether accurate, robust and platform-independent quantification of the appropriate gene markers can be achieved by microarrays. As we outline, achieving the second criteria will require careful optimization of microarray technology.

Studies based on expression profiling of human cancer samples with various clinical outcomes usually produce gene-expression classifiers involving ten to 100 genes [4,5,57–59], although microarray-based classifiers with as few as two genes have also been published [60,61]. However, the multi-genic disease classifiers published so far do not seem to be consistent across studies used to predict the clinical outcome for the same type of cancer. For example, two groups attempted to develop a prognostic signature to predict survival in diffuse large B-cell lymphoma using different microarray platforms. The studies produced two completely different gene classifiers with 13 and 17 genes each, without a single overlapping gene between them [57,58]. Similar inconsistencies were found in studies that aimed to develop gene-expression classifiers to predict the likelihood of distant metastasis in breast cancer [4,59]. It seems, however, that these inconsistencies are due, in part, to the methods by which the classifiers are extracted from the microarray data. There is evidence that a multi-genic classifier based on prior biological knowledge and extracted from the literature can be predictive in a given microarray data set, but the same classifier could not be identified from the same microarray data owing to the associated computational difficulties. For example, it was shown that survival in diffuse large-B-cell lymphoma can be predicted based on the PCR-quantified expression levels of six genes that were selected from a larger set of genes that were previously known to be associated with disease outcome [62]. The same six-gene classifier was also predictive in microarray data sets, although it could not be derived from the same data sets [57,58].

If reliable disease classifiers will involve, for example, a few tens of genes, then well-established technologies for gene expression quantification, such as real time quantitative RT–PCR, are probably a better alternative to microarrays.

However, the diagnostic application of microarrays could be still considered in some cases. There is an implicit assumption that various cancer states are defined by the complex interaction of a non-trivial number of genes, therefore a large number of diagnostic marker genes are better suited to microarrays than alternative technologies that have a lower throughput. Therefore, if the reliable prediction of a clinical outcome requires >100 genes, optimized microarrays might be a viable option for clinical diagnostics. However, the identification of classifiers with more than ten genes suffers from the same problems observed in cancer-based gene-expression-profiling studies. The number of genes (usually thousands) that needs to be evaluated does not permit the reliable extraction of complex classifiers from the 200–300 clinical samples (or fewer) that most studies contain [63]. This is a well-known difficulty when the number of alternative hypotheses is too large relative to the number of samples [64]. Furthermore, the quantification of certain clinically relevant genes, such as *ERBB2*, by RT–PCR was shown to be difficult in clinical samples [65]. Microarray hybridization might not be affected by the mechanism interfering with PCR in these examples.

with those from kRT–PCR, down to the level of two copies per cell. However, the microarrays failed to produce meaningful measurements below that threshold. Czechowski *et al.* also obtained similar results when comparing RT–PCR profiling of >1400 *Arabidopsis* transcription factors with the 22K *Arabidopsis* Affymetrix array [16]. Although 83% of those genes could be reliably quantified by RT–PCR, Affymetrix GeneChips could detect <55% of the 1400 transcription factors, which are usually expressed at the lower end of the dynamic range of the transcriptome (most are present at <100 copies per cell). Finally, Kane *et al.* compared the sensitivity of 50mer oligonucleotide probes with that of PCR products [17]. These results showed that for rat liver RNA, microarrays had a minimum reproducible detection limit of approximately ten mRNA copies per cell.

It was expected that varying the probe length would provide various trade-offs among sensitivity, signal strength and specificity. Signal strength increases with probe length in a certain range. For example, 30mers provide twice the intensity of a 25mer probe [18]. A comparison of the CodeLink (30 nt) and Affymetrix platforms (25 nt) also suggested a tenfold greater sensitivity of the former platform [19]. Therefore, in theory, the sensitivity issue can be addressed by simply using longer probes. However, it is not a simple as this. A further increase in probe length produced only a limited enhancement, whereas the specificity of probes, as quantified by the relative intensity of perfect match versus single base pair mismatch probes, actually decreased [18].

In summary, the detection limit of current microarray technology seems to be between one and ten copies of mRNA per cell. This sensitivity threshold is probably lower for cell types with a more-limited concentration range of transcripts, such as yeast [15]. Although this sensitivity is impressive, it might still be insufficient to detect relevant changes in low abundance genes, such as transcription factors [15,16]. It remains to be seen whether novel technological developments, such as labeling with quantum dots [20], will further increase the sensitivity of microarray platforms.

Accuracy

Microarrays can be used to measure either absolute transcript concentrations or relative transcript concentrations (i.e. expression ratios). In principle, accurate absolute concentration measurements will also provide accurate measurements of expression ratios but the reverse does not necessarily hold true. Estimating ratios requires a less detailed understanding of how the signal intensity of a given microarray probe is related to the concentration of the measured transcript. As long as a probe binds to its target specifically and the produced signal intensity is proportional to the amount of transcripts bound (up to a multiplicative constant), the expression ratios will reflect the reality to a significant extent. However, estimating absolute concentrations, particularly at the current insufficient level of understanding of DNA and RNA hybridization, requires careful calibration with known concentrations of the transcripts.

Traditionally, two-channel, cDNA array data (e.g. using Cy3 and Cy5 dyes) are usually used to measure ratios*, whereas single channel, oligonucleotide array data (e.g. Affymetrix) are intended to represent absolute expression values. However, some important issues become apparent when examining the signal intensities produced by two different Affymetrix probes of the same probe set that are targeted against closely placed or overlapping sequences on a given transcript. These are likely to hybridize to the same labeled RNA fragment and still can produce signals varying by orders of magnitude (Figure 1). This suggests that the same transcript concentration can produce rather different probe signal intensities depending on the specific probes representing each gene, which, in turn, means that interpreting the measured signals as proportional to the absolute concentrations is not necessarily advisable.

Ratios can be measured with a greater accuracy [16], which is reflected by the fact that probes in a given Affymetrix probe set (i.e. probes designed to recognize the same gene transcript) not only produce significantly different intensity but also produce consistent ratio values across the same probe set when two RNA samples are compared with each other (Figure 2).

Assessing the accuracy of microarray measurements requires that true concentrations, or ratios, be available for many transcripts. True concentrations can be obtained by either spike-in or dilution experiments [24] or measuring transcript levels by independent means, for example, quantitative RT–PCR or northern blots. A few spike-in or dilution data sets were provided by Affymetrix (Affymetrix-Latin square data, http://www.affymetrix.com/support/technical/sampledata/datasets.affx) and its industrial partners (http://www.genelogic.com/newsroom/studies/index.cfm). A wide variety of Affymetrix DNA chip analysis methods were evaluated based on these data sets (http://affycomp.biostat.jhsph.edu). However, the variance of the average chip intensity among these spike-in data sets is much lower than those measured in most real-life data sets, casting doubts on the general applicability of these data for developing analytical tools for highly diverse, clinical gene-expression profiles. Furthermore, the limited number of spike-in genes, 42 at most, also makes it difficult to use this data set for the comprehensive evaluation of both the technology and the data analysis methods. Even assuming that the genes were selected in a completely unbiased manner (i.e. assuming that genes known to produce 'good' results were not favored and problematic transcripts were not ignored), this is still a limited number of genes. In the light of the strong dependency of the measurements on the specific target genes and the specific probe sequence selected, extrapolating the accuracy measured on 42 genes to another 10 000–30 000 genes seems more like an act of faith than a scientific inference.

A few spike-in data sets were also produced by some academic laboratories. Choe and co-workers produced Affymetrix GeneChip data for *Drosophila* RNA samples with ~1300 spiked-in genes against a fairly well-defined

* Usually but not always, see, for example, the loop designs proposed by M.K. Kerr *et al.* [21], G.A. Churchill [22] and S. Draghici [23].

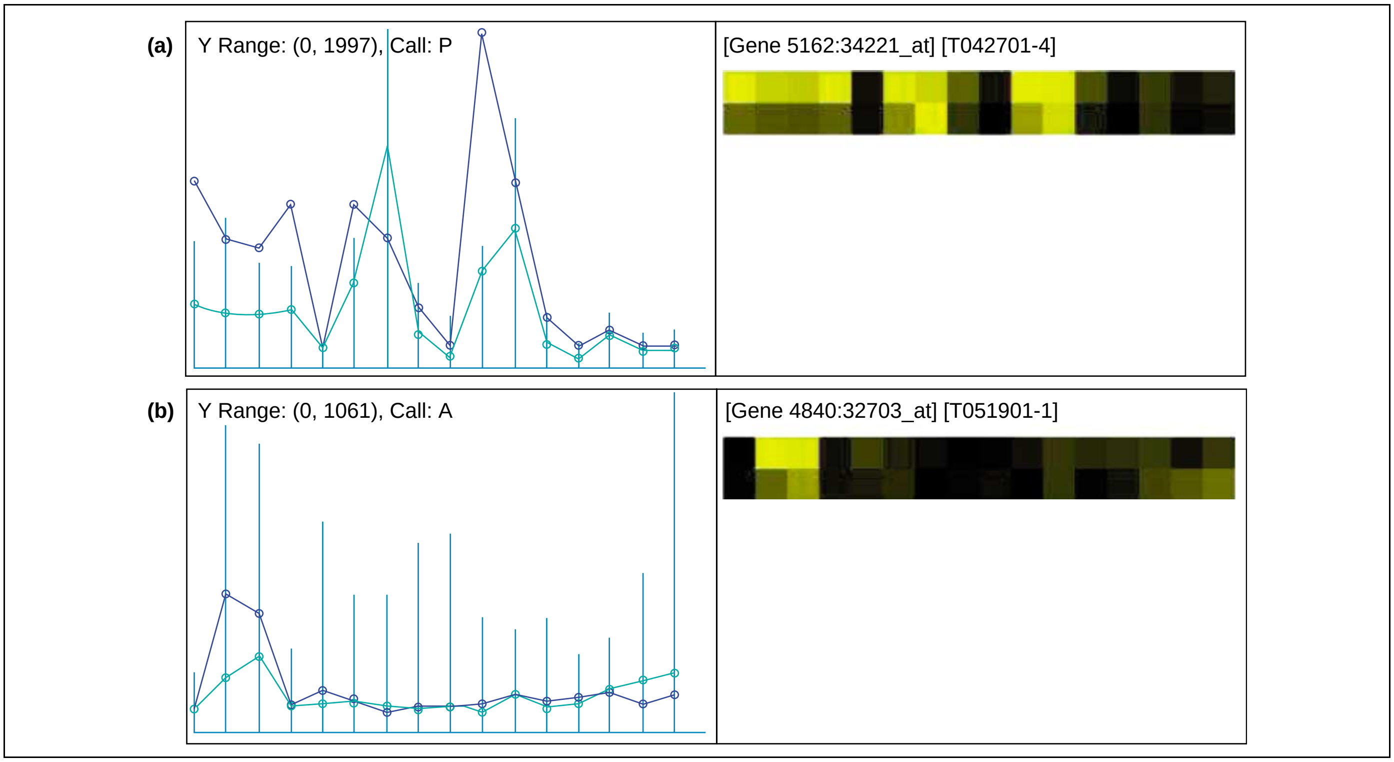

Figure 1. Different probes meant to represent the same transcript can yield widely different signals. The left panels show the perfect match (PM, in blue) and mismatch (MM, in green) values across the probes in the same probe set. **(a)** For the gene shown, probe 10 (from left to right) is close to saturation level, whereas probes 5, 9, 13, 15 and 16 are close to the background level. Programs such as MAS 5.0 call this gene present (P) and calculate the expression level based on the average difference between PM and MM probes. This illustrates how sensitive the actual numbers are to the choice of the specific probes. **(b)** Most probes corresponding to this gene are expressed at the background level; therefore, this gene is absent (A). However, probes 2 and 3 produce high levels of signal intensity. In some cases, the PM–MM difference for some of the probes of an absent gene can be so large that the average difference, often used to represent the expression level of the gene, can be higher for some absent genes than some of those declared present (reproduced with permission from Ref. [23]).

background of ∼2500 genes [25]. The agreement between observed and actual fold changes was significant ($R^2=0.86$), when the probe sets in the lowest quartile of signal intensity were filtered out. This filtering step seems to be the single most important preparatory step to ensure accurate and consistent microarray measurements. Their results also suggested that the detection of ∼70% of true positives can be achieved before reaching a 10% false-discovery rate. However, this seems to add further support to the idea that microarray measurements are not reliable for genes expressed at low levels. A similar, large-scale spike-in data set for human genes would also be most welcome for calibration purposes.

The alternative to spike-in-based validation requires the independent quantification of transcripts by RT–PCR or northern-blot analysis. Owing to the cost associated with independent verifications, most such studies measure expression levels by independent means only for a limited number of transcripts, typically <20 genes [26,27]. The transcripts selected for verification are usually widely studied genes with well agreed upon sequences [26]. For these genes, considering the good quality of the associated annotations, one would expect a reasonable level of accuracy. Indeed, widely used microarray platforms, such as the Affymetrix GeneChip, produce verifiable differential expression calls in ∼85–90% of the genes [27]. However, this applies only to the range of expression levels that is greater than the sensitivity threshold of the given microarray platform, effectively eliminating up to 40–50% of the transcripts present in the RNA samples from the analysis [28]. Furthermore, the failure to detect the correct fold changes for a highly relevant gene such as the epidermal growth

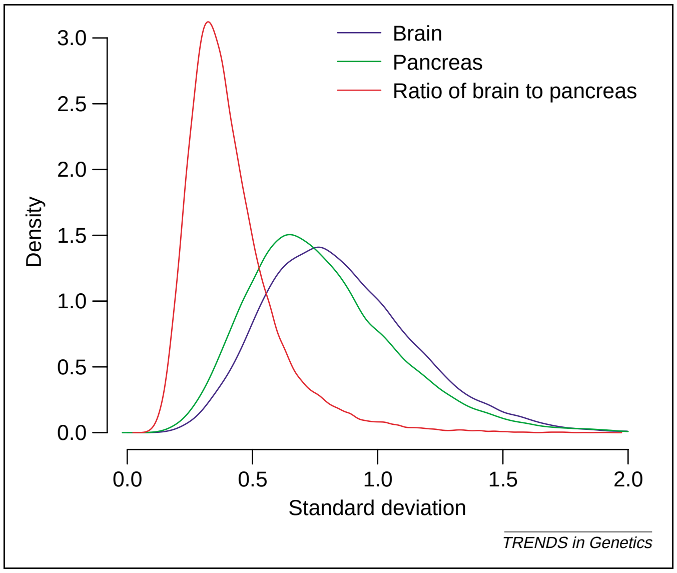

Figure 2. Probe intensity ratios are more consistent (have a smaller variance) than absolute probe intensities. Probe-level intensities from two Affymetrix HG-U133A arrays were $\log_2$ transformed and the standard deviation was calculated for the 11 probes in each probe set. The smoothed distribution of these standard deviations are plotted for brain (blue), pancreas (green) and the probe-level ratios of brain to pancreas (red). Most variances of the ratios are smaller than the variances of either tissue. Data are from the GNF expression atlas [66].

factor receptor (*EGFR*) [27], a gene often implicated in cancer diagnostics, should also encourage a pause for reflection for those interested in the diagnostic applications of microarrays.

A few studies produced independent measurements by RT–PCR for a more comprehensive set of transcripts from ~50 to 1400 [15,16,29]. The Microarray Quality Control Project is currently producing the first such comprehensive data set for human RNA, containing ~1000 genes. Based on these studies, the picture emerging for the most widely used microarray platforms such as cDNA microarrays and Affymetrix GeneChips seems to warrant the following conclusions:

- Above their sensitivity threshold, microarray measurements accurately reflect the existence and direction of expression changes in ~70–90% of the genes. However, the magnitude of such expression changes as reflected by microarray data tends to be different from the magnitude of the changes measured with other technologies such as RT–PCR.
- Microarrays (both single and dual channel) tend to measure ratios more accurately than absolute expression levels. For example, in the most comprehensive data set [16] that quantified 1400 genes by RT–PCR, Czechowski *et al.* found a poor correlation between normalized raw data produced by RT–PCR and normalized raw data produced by Affymetrix arrays in the same RNA sample. However, when RNA from shoots and roots of *Arabidopsis* were compared, a more promising result emerged. The ratios between these two RNA samples extracted from RT–PCR and array measurements produced a Pearson correlation of 0.73 for the most highly expressed set of 50 genes. However, one should note that a correlation of ~0.7 is not impressive for two platforms that are expected to measure the abundance of given transcripts in the same samples.
- The relatively good correlation between microarray-based and RT–PCR-based gene expression ratios does not necessarily mean that the microarray technology directly produces accurate estimates of gene expression ratios. In fact, it has been well known that microarray-based expression ratios are compressed [29] (i.e. the ratios of mRNA expression levels are consistently underestimated). It seems that the ratio compression is a significantly more consistent phenomenon for cDNA microarrays than for short oligonucleotide chips [29].

In conclusion, a handful of independent validation studies and spike-in data sets have enabled an empirical assessment of the accuracy of microarray technology. Although an accurate measurement of absolute transcript levels by microarrays is probably beyond the current capabilities of the technology, ratios can be estimated reasonably well, particularly when the significant level of ratio compression is taken into consideration and corrected for. However, this favorable assessment applies only to the measurement of transcripts that are expressed well above the sensitivity level of microarrays, rendering perhaps half of the transcriptome beyond the reach of microarrays.

Reproducibility

Reproducibility is the most readily assessable characteristic of any microarray platform. Unfortunately, a platform can have an excellent reproducibility without necessarily producing measurements that are accurate or consistent with other platforms. This is because reproducibility only requires that a given probe binds to the same number of labeled transcripts in repeated measurements of the same sample. Badly designed probes that cross-hybridize with several other transcripts, can easily provide highly reproducible and yet useless data. Therefore, reproducibility is a necessary but not sufficient requirement. In their appropriate sensitivity range, most microarray platforms produce highly reproducible measurements. Some oligonucleotide arrays (Affymetrix, Agilent and Codelink) [30,31] provide correlation coefficients of >0.9. For other platforms, such as cDNA microarrays or the Mergen platform (http://www.mergen-ltd.com/company.htm), the reported Pearson correlation coefficient between technical replicates can range between the disappointing level of 0.5 and the reassuring level of 0.95 [7,31,32]. It is, therefore, not surprising, as we discuss in the following section, that these platforms show poor correlation with commercial oligonucleotide-based platforms [31,33].

Cross-platform consistency

If microarray data were highly reproducible across various platforms and if they provided information about the absolute transcript levels, one could use appropriately normalized gene expression data without regard to the platform on which the data was obtained. This in turn would reduce the need to replicate experiments and would enable researchers to build universal gene-expression databases that would compile many different data sets from a variety of experimental conditions. This consideration is particularly relevant for clinical samples with limited amounts of mRNA.

Owing to the relative scarcity of comprehensive, large-scale, spike-in or independently measured gene expression data sets, cross-platform consistency has been used as a surrogate measure of microarray reliability. In this approach, aliquots from the same RNA sample, or RNA isolated from the same biological source are profiled on different microarray platforms. The consistency of these results is considered an indication of the reliability of all platforms compared. Lack of consistency can be caused by the inferior performance of at least one of the platforms, without a clear indication of the relative merit of each platform. Interpreting the cross-platform consistency as a proof of accuracy and reliability is not necessarily warranted because highly similar results across platforms could be simply caused by consistent cross-hybridization patterns without either platform measuring the true level of expression. Nevertheless, a high level of cross-platform consistency is desirable, because, if both platforms performed accurate measurements, cross-platform consistency would automatically follow. Cross-platform consistency is a necessary but insufficient requirement to validate the technology. Despite this limitation, cross-platform consistency studies produced several useful lessons for microarray users.

Cross-platform comparison of various microarray platforms depends on the availability of data sets based on same RNA aliquots profiled on different microarray platforms. Until recently there were only two such data that were widely available: the NCI60 cell-line panel profiled using cDNA microarray [34] and the Affymetrix platform [35]. These data sets were reanalyzed several times for cross-platform consistency with gradually improving results, highlighting the importance of probe sequence verification. One of the difficulties in the cross-platform comparison of microarray data is to ascertain that probes on the various platforms aimed at the same gene do in fact quantify the same mRNA transcript. The various strategies to match probes between different platforms can be constrained by the amount of information provided by the manufacturers of the given microarray. Before actual probe sequence information was released, probe matching could be based only on gene identifiers such as the Unigene ID [36]. This is known to produce a significant number of incorrect pairings [37,38]. Therefore, it is not surprising that in an early study, while comparing the two NCI60 data sets using this microarray probe matching strategy, Kuo *et al.* found an alarming level of inconsistencies (a Pearson correlation <0.34) [36]. While measuring the extent of within-array cross-hybridization, Kuo. *et al.* observed that the genes represented by cDNA probes with a greater number of cross-matches to other genes (defined as sequence similarity using BLAST) have less correlation with the oligonucleotide data. This suggested that cross-hybridization is a possible cause for the poor cross-platform consistency. They also found low correlations for genes with low intensity values on cDNA arrays and low average difference in the Affymetrix arrays (Pearson coefficient was 0.03 and Spearman coefficient was 0.02), indicating that the low-abundance transcripts were not measured reliably on either platform. As partial or complete probe sequence data have become available, more-accurate strategies could be implemented. Probes could be matched across the various platforms on the basis of whether they can be sequence-mapped to the same transcript. When microarray probe pairs across the platforms that obviously mapped to different transcripts, while still sharing, erroneously, the same UniGene ID, were filtered out, the mean correlation of gene expression between the two previously described NCI60 data sets increased to 0.6 [37]. Similar results were obtained in a study when RNA aliquots were profiled and compared across several Affymetrix platforms and the Agilent Human 1 cDNA microarray platform [38]. UniGene-matched probes that failed the direct sequence-mapping test showed significantly lower expression correlation across the two microarray platforms [38].

Finally, probe sequences can be used to ascertain that microarray probes on different platforms are targeted against the same region of a given transcript. This ensures that the two platforms are quantifying the same splice variants but also increases the chance of similar undesired cross-hybridization patterns. For the two NCI60 data sets, probes targeting the same region of the transcripts showed the greatest correlation (mean Pearson coefficient of ~ 0.7) [39]. This relatively high correlation was obtained only after filtering out, again, the genes yielding low intensity signals. The rigorous sequence-mapping strategy employed in this article also ensured that only those Affymetrix probes that could be verified by high-quality sequence databases were used in the final analysis. By the application of an appropriate common reference produced *in silico*, the two data sets (Affymetrix and cDNA microarray) could be pooled, and hierarchical cluster analysis produced meaningful results on the combined gene expression profiles [39]. The results of this study seem to answer positively the important question regarding whether microarray data from different laboratories and different platforms can be assembled into a coherent unique database. However, this study also shows that if such a 'universal database' is to ever be constructed, this should not be done by merely storing expression data reported by the various platforms in a common database, but rather revisiting the biochemical foundations of the technology and using such knowledge to interpret and filter the numerical data generated by the arrays.

In a much-cited article, Tan *et al.* produced gene expression profiles for technical and biological replicates of RNA samples on three different platforms: the oligonucleotide-based CodeLink arrays, Affymetrix GeneChips and cDNA arrays from Agilent [40]. Although, the intra-platform consistency was good for replicates within all platforms (~ 0.9), the Pearson correlation coefficient was more moderate across the various platforms. The correlation of matched gene measurements between oligonucleotide arrays (Affymetrix and Codelink) was the greatest (0.59), whereas the correlation between cDNA and oligonucleotide arrays was lower (0.48–0.50). It should be noted, however, that owing to the lack of comprehensive probe sequence information it was not possible to apply rigorous sequence-mapping criteria, and the correlation coefficients listed here were calculated without filtering out genes that had low expression levels. Without the application of these noise-reducing strategies, it is not surprising that the three platforms showed a rather disappointing level of concordance in their ability to predict gene expression changes. Approximately two hundred genes were predicted to be differentially expressed by at least one platform but only four genes were detected as differentially expressed by all three platforms. The importance of intensity filtering and careful probe sequence verification is further supported by several other studies [7,26]. As a general rule, any set of results in which probe-level sequence matching and low-level filtering has not been performed should invite caution. Although disappointing in terms of concordance between the specific genes reported as differentially regulated, this work also showed that meaningful biological conclusions can still be obtained by a higher level of analysis, in which the sets of differentially regulated genes are mapped on their associated biological processes and cellular locations, using gene ontology (GO) annotations. Although the specific differentially regulated genes were very different between the platforms, all platforms were consistent in terms of the biological processes involved. Fortunately, several tools and techniques now exist that can automate this type of analysis for numerous genes, thus helping researchers circumvent some of the limitations of the technology [41–45].

The performance of the platforms that are used less frequently was also assessed recently. In their analysis of mouse microarray platforms, Yauk *et al.* used the Mergen and Agilent oligonucleotide platforms in addition to the Affymetrix, Codelink and cDNA microarray platforms [33]. After intensity filtering for 'present' calls, ratios measured by the Affymetrix, Codelink and Agilent oligonucleotide arrays showed a more satisfactory correlation among those three technologies (>0.7), whereas the custom cDNA microarray and the Mergen platform showed a significantly lower correlation (correlation coefficient of ≤ 0.5) with the other platforms. The Agilent cDNA microarray platform was placed in between these two groups. The Toxicogenomics Research Consortium (http://www.niehs.nih.gov/dert/trc/home.htm) has run an even wider comparison of the various mouse microarrays [31]. In addition to the commercial oligonucleotide arrays from Affymetrix, Agilent and GE Healthcare (Codelink), they also analyzed spotted oligonucleotide arrays from Compugen and spotted cDNA microarrays from two different sources (The Institute for Genomics Research http://www.tigr.org/ and National Institute for Aging http://www.nia.nih.gov/). This project examined cross-platform reproducibility and the bias introduced when the same platform was used by different laboratories analyzing aliquots from the same RNA sample. For the 500 genes represented on all platforms the cross-platform consistency varied between 0.11 (Codelink versus spotted cDNA) and 0.76 (two different version of spotted cDNA microarrays.) When the same platform was used by two different laboratories, the Affymetrix platform produced by far the greatest correlation (0.91) across laboratories.

Sources of inaccuracy and inconsistencies in microarray measurements

As a reasonable approximation, signals produced by any given microarray probe can be considered as the composite of three signals: (i) specific signal produced by the originally targeted labeled transcript; (ii) cross-hybridization signal produced by transcripts that have a non-perfect but still significant sequence similarity with the probe; (iii) a non-specific, background signal that is present in the absence of any significant sequence similarity. On an ideal, high-specificity microarray platform the second and third components would be negligible relative to the targeted specific signal. However, even under such ideal conditions, microarray technology in its current state would face significant limitations for several reasons as discussed in the following paragraph.

First, the relationship between probe sequences, target concentration and probe intensity is rather poorly understood. A given microarray probe is designed as a perfect complementary strand to a given region of the transcript. Based on the Watson–Crick pairing, the probe will capture a certain number of the transcripts. This number is proportional to the concentration of the transcript, but the actual relationship between transcript concentration and the number of molecules bound to the probe, and thus the signal produced, also depends on the affinity of the probe, or free energy change values, under the given hybridization conditions. This affinity is determined to a large extent by the actual nucleotide sequence stretch participating in the binding. This sequence-affinity relationship is not sufficiently understood. Although the sequence dependence of DNA–DNA hybridization in solutions has been studied in detail [46], DNA–RNA hybridization has received significantly less attention. Remarkably, the results of Sugimoto *et al.* [47] suggest that the sequence dependence of DNA–RNA hybridization can still hold surprises. For example, for certain sequences the binding energy of a DNA–RNA duplex can be stronger for a single mismatch than for the corresponding perfectly complementary strands [47,48]. The kinetics of hybridization are further complicated by the incorporation of modified nucleotides into the target transcripts during the most widely used labeling protocols. Furthermore, the results obtained in solutions cannot be directly applied to the hybridization of microarray probes attached to surfaces [49]. Various researchers have tried to investigate the dependence of affinities on the microarray probe sequence [50–52] but no convincing model has emerged.

Second, splice variants constitute another dimension that can introduce difficulties in the microarray analysis. It is estimated that at least half of the human genes are alternatively spliced, and might have many potential splice variants [53]. A given short oligonucleotide probe is targeted at either a constitutive exon (present in all splice variants) or at an exon specific for certain splice variants. In the former case, the probe intensity will reflect the concentration of all splice variants that are present in the sample, therefore obscuring expression changes occurring in certain splice variants. In the latter case, the specific splice variant will be measured, but other splice variants of the same gene will be ignored. Covering the various types of exons on short oligonucleotide-based arrays is necessary to dissect the splice variant associated composite signals. cDNA microarrays usually have a unique long probe with which the abundance of several splice variants can be measured. This might explain some of the discrepancies often observed between cDNA and short oligonucleotide microarrays.

Third, folding of the target transcripts [54] and cross-hybridization [14] can also contribute to the variation between different probes targeting the same region of a given transcript (Figure 1). It has been shown previously that a large proportion of the microarray probes produce significant cross-hybridization signals [14,55] for both oligonucleotide and cDNA microarrays. Even a limited stretch of sequence complementarity might be sufficient to enable binding between two unrelated sequences. However, evaluating the overall impact of cross-hybridization on the accuracy of microarray measurements is not easy. For example, in Affymetrix arrays, the effect of a single cross-hybridizing probe can be down-weighted by the rest of the probe set (ten other probes in HU-133 chips). Furthermore, the impact of cross-hybridization strongly depends on the relative concentration and the relative affinities of the correct target and the cross-hybridizing target(s). The latter must be present in sufficient quantities to interfere with specific signals. Cross-hybridization, in conjunction with splice variants, is probably a prime candidate to explain the discrepancies in differential gene expression calls between various microarray platforms, although no systematic study has yet been undertaken. Removing and/or redesigning the microarray probes prone to cross-hybridization is a reasonable strategy to increase the hybridization specificity and hence, the accuracy of the microarray measurements. However, this requires a good understanding of cross-hybridization, towards which only limited progress has been made owing to the lack of appropriate experimental data.

In light of the complexity of microarray signals described, issues such as the compression of expression ratios can be reasonably explained. The presence of cross-hybrization signals on a given probe, for example, might prevent the detection of large changes in gene expression levels because a probe will always produce a certain level of 'false' signal, even if the true signal is much lower or perhaps undetectable. As our understanding of splice variants, specific and non-specific nucleic acid hybridization and other relevant issues deepens, we will design probes that will quantify transcripts in an increasingly optimal fashion. The quest for increasing microarray performance by regularly eliminating and redesigning probes can be easily tracked in, for example, the Affymetrix technology. Only a small proportion of probes are retained between successive generations of Affymetrix arrays, even for probe sets targeting the same gene [56]. Although noble in purpose, this constant probe redesign has the undesirable side effect that data sets obtained on different generations of arrays cannot be combined easily.

Conclusions and future directions

Microarrays are a popular research and screening tool for differentially expressed genes. Their ability to monitor the expression of thousands of genes simultaneously is unsurpassed. However, certain limitations of the current technology exist and have become more apparent during the past couple of years. In its appropriate sensitivity range, the existence and direction of gene expression changes can be reliably detected for the majority of genes. However, accurate measurements of absolute expression levels and the reliable detection of low abundance genes are currently beyond the reach of microarray technology. Therefore, the ability to detect changes in the expression of specific individual genes might be affected. Basic research in the areas of nucleic acid hybridization, and technological advances in detection methods and hybridization conditions will certainly increase the measurement capabilities of microarray technology.

Acknowledgements

S.D. and P.K. are supported in part by the following grants: NIH (NHGRI) 1R01HG003491-01A1, NSF DBI-0234806, NSF SOD-0438970, NIH (NCI) 1U01CA117478-01, NIH (NCRR) 1S10RR017857-01, NINDS 1R01NS045207-01, NIH (NCI) 1R21 CA100740-01, NIBIB 5R21EB000990-03, DOD DAMD 17-03-02-0035, MLSC MEDC-538 and MEDC GR-352. Z.S. was supported in part by NIH grants P50 CA089693-06 and 1PO1CA-092644-01.

References

1 Schena, M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
2 Hardiman, G. (2004) Microarray platforms–comparisons and contrasts. *Pharmacogenomics* 5, 487–502
3 Barrett, J.C. and Kawasaki, E.S. (2003) Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov. Today* 8, 134–141
4 Wang, Y. *et al*. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679
5 van de Vijver, M.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *New Engl. J. Med.* 347, 1999–2009
6 Halgren, R.G. *et al*. (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res.* 29, 582–588
7 Jarvinen, A.K. *et al*. (2004) Are data from different gene expression microarray platforms comparable? *Genomics* 83, 1164–1168
8 Taylor, E. *et al*. (2001) Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques* 31, 62–65
9 Watson, A. *et al*. (1998) Technology for microarray analysis of gene expression. *Curr. Opin. Biotechnol.* 9, 609–614
10 Harbig, J. *et al*. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.* 33, e31
11 Mecham, B.H. *et al*. (2004) Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* 18, 308–315
12 Forman, J.E. *et al.* (1998) Thermodynamics of duplex formation and mismatch discrimination on photolitographically synthesized oligonucleotide arrays. In *Molecular Modeling of Nucleic Acids* (Vol. 862) (Leontis, N.B. and SantaLucia, J.J., eds) pp.205–228, American Chemical Society
13 Ramakrishnan, R. *et al*. (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Res.* 30, e30
14 Zhang, J. *et al*. (2005) Detecting false expression signals in high-density oligonucleotide arrays by an *in silico* approach. *Genomics* 85, 297–308
15 Holland, M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.* 277, 14363–14366
16 Czechowski, T. *et al*. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* 38, 366–379
17 Kane, M.D. *et al*. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* 28, 4552–4557
18 Relogio, A. *et al*. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* 30, e51
19 Shippy, R. *et al.* (2004) Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* DOI:10.1186/1471-2164-5-61 (http://www.biomedcentral.com/1471-2164/5/61)
20 Liang, R.Q. *et al*. (2005) An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe. *Nucleic Acids Res.* 33, e17
21 Kerr, M.K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819–837
22 Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32(Suppl.), 490–495
23 Draghici, S. (2003) *Data Analysis Tools for DNA Microarrays*, Champan and Hall, CRC Press
24 Hubbell, E. *et al*. (2002) Robust estimators for expression analysis. *Bioinformatics* 18, 1585–1592
25 Choe, S.E. *et al*. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* 6, R16

26 Larkin, J.E. *et al*. (2005) Independence and reproducibility across microarray platforms. *Nat Methods* 2, 337–344

27 Gold, D. *et al.* (2004) A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays. *BMC Genomics* DOI:10.1186/1471-2164-5-2 (http://www.biomedcentral.com/1471-2164/5/2)

28 Kuznetsov, V.A. *et al*. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics* 161, 1321–1332

29 Yuen, T. *et al*. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* 30, e48

30 Bakay, M. *et al.* (2002) Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics* DOI:10.1186/1471-2105-3-4 (http://www.biomedcentral.com/1471-2105/3/4)

31 Bammler, T. *et al*. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2, 351–356

32 Jenssen, T.K. *et al*. (2002) Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res.* 30, 3235–3244

33 Yauk, C.L. *et al*. (2004) Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* 32, e124

34 Scherf, U. *et al*. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244

35 Staunton, J.E. *et al*. (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10787–10792

36 Kuo, W.P. *et al*. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18, 405–412

37 Lee, J.K. *et al*. (2003) Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol.* 4, R82

38 Mecham, B.H. *et al*. (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

39 Carter, S.L. *et al.* (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* DOI:10.1186/1471-2105-6-107 (http://www.biomedcentral.com/1471-2105/6/107)

40 Tan, P.K. *et al*. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31, 5676–5684

41 Draghici, S. *et al*. (2003) Global functional profiling of gene expression. *Genomics* 81, 98–104

42 Khatri, P. *et al.* (2005) Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res.* 33, W762–765

43 Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587–3595

44 Khatri, P. *et al.* (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.* 32, W449–456

45 Khatri, P. *et al*. (2002) Profiling gene expression using onto-express. *Genomics* 79, 266–270

46 SantaLucia, J., Jr. *et al*. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555–3562

47 Sugimoto, N. *et al*. (2000) Thermodynamics-structure relationship of single mismatches in RNA–DNA duplexes. *Biochemistry* 39, 11270–11281

48 Naef, F. and Magnasco, M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 68, 011906

49 Peterson, A.W. *et al*. (2002) Hybridization of mismatched or partially matched DNA at surfaces. *J. Am. Chem. Soc.* 124, 14601–14607

50 Mei, R. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11237–11242

51 Hekstra, D. *et al*. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.* 31, 1962–1968

52 Zhang, L. *et al*. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* 21, 818–821

53 Modrek, B. *et al*. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859

54 Mir, K.U. and Southern, E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.* 17, 788–792

55 Wu, C. *et al*. (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.* 33, e84

56 Nimgaonkar, A. *et al*. (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* DOI:10.1186/1471-2105-4-27 (http://www.biomedcentral.com/1471-2105/4/27)

57 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

58 Shipp, M.A. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8, 68–74

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

60 Gordon, G.J. *et al*. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* 62, 4963–4967

61 Ma, X.J. *et al*. (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 5, 607–616

62 Lossos, I.S. *et al*. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New Engl. J. Med.* 350, 1828–1837

63 Michiels, S. *et al*. (2005) Prediction of cancer outcome with microarrays. *Lancet* 365, 1685–1686

64 Ioannidis, J.P. (2005) Why most published research findings are false. *PLoS Med* 2, e124

65 Willey, J.C. *et al*. (2004) Standardized RT-PCR and the standardized expression measurement center. *Methods Mol. Biol.* 258, 13–41

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