Comparison and meta-analysis of microarray data: from the bench to the computer desk

Yves Moreau[1], Stein Aerts[1], Bart De Moor[1], Bart De Strooper[2] and Michal Dabrowski[2]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium
[2]Laboratory for Neuronal Cell Biology, Center for Human Genetics, Katholieke Universiteit Leuven and VIB (Flemish Interuniversity Institute for Biotechnology), Herestraat 49, 3000 Leuven, Belgium

The upcoming availability of public microarray repositories and of large compendia of gene expression information opens up a new realm of possibilities for microarray data analysis. An essential challenge is the efficient integration of microarray data generated by different research groups on different array platforms. This review focuses on the problems associated with this integration, which are: (1) the efficient access to and exchange of microarray data; (2) the validation and comparison of data from different platforms (cDNA and short and long oligonucleotides); and (3) the integrated statistical analysis of multiple data sets.

In the past few years, a myriad of microarray experiments has been produced, overwhelming the research community with a wealth of potentially valuable data. Efficient access to these data and, in particular, efficient comparison and integration of data obtained in related biological systems provide researchers with an opportunity to address complex questions in an effective way.

Tellingly, larger microarray projects are scaling up towards the generation of large compendia of gene expression. These will provide a comprehensive view of the transcriptome in different organisms at different stages of development [1] or under different environmental [2] or genetic [3] conditions and of the changes in gene expression that are associated with a diverse series of human pathologies [4]. We envisage a radical change in microarray studies – comparable to what happened in sequence analysis with the advent of the genome projects – where a division of labor takes place between a few large consortium-based projects on the one hand and the many smaller investigation-specific projects on the other hand. The 'compendium projects' will chart large areas of the transcriptome whereas smaller-scale projects will refine the details, starting from a careful analysis of publicly available microarray (and sequence) data to design experiments that validate and refine primary hypotheses.

But what are the barriers to this bonanza of information and how can they be overcome? In this review, we examine: (1) how microarray standards and repositories allow data exchange; (2) how a detailed understanding of the specifics of different platforms permits cross-center and cross-platform comparison and validation; and (3) how meta-analysis enables integrated analysis of multiple data sets.

Data access and exchange

Until now, most of the publicly available microarray data have been scattered around the internet, often as supplementary data to a published article. Consequently, it has been difficult for investigators to know where the relevant data are available. This problem has been addressed in several databases by making it possible to search for published microarray data that has undergone uniform processing and filtering and by providing links to the original publications for more detailed information. These databases have diverse purposes and are either: (1) platform specific (e.g. the Stanford Microarray Database; http://genome-www5.stanford.edu/MicroArray/SMD [5]); (2) organism specific (e.g. yeast Microarray Global Viewer; http://www.transcriptome.ens.fr/ymgv [6]); or (3) project specific {e.g. the Lifecycle database on *Drosophila* development, (http://genome.med.yale.edu/Lifecycle [1]), the NeuroDiff database on neuronal differentiation in mouse (http://www.east.kuleuvan.ac.be/neurdiff [7]), or the HugeIndex database on normal expression in human tissues (http://zlab.bu.edu/HugeSearch [8])}.

Although supplements and microarray databases on the internet provide access to many data sets, they have some drawbacks. (1) They lack direct access to the experimental information that is needed to judge the quality of the data, to repeat a study or to re-analyze the data. (2) A standard format for microarray data and experiment description is not used. These drawbacks make identifying, collecting and analyzing publicly available data sets a cumbersome and error-prone process.

Microarray standards and repositories

The Microarray Gene Expression Data (MGED) Society (http://www.mged.org) provides guidelines, formats and tools to overcome these two drawbacks. The Minimum Information About a Microarray Experiment (MIAME) specification [9] is a checklist that guides the investigator in the annotation of microarray experiments. Because

Corresponding author: Yves Moreau (moreau@esat.kuleuven.ac.be).

 0168-9525/$ - see front matter © 2003 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2003.08.006

numerous biological and experimental factors influence gene expression measurements (e.g. lighting conditions in plant experiments, the exact histopathology of a tumor, the difference in specificity of different reporter sequences for the same gene, the particularities of a single batch of slides or the laser intensity at which a slide is scanned), this MIAME specification includes the experimental design, array design, details of the samples and any treatments, hybridization conditions, measurements and normalization controls. Furthermore, the MGED ontology [10] provides a framework of microarray concepts for this annotation and the MicroArray Gene Expression Object Model (MAGE-OM) and Markup Language (MAGE-ML) conceptualize MIAME for data storage and exchange [11].

In practice, a local MIAME-supportive database will allow gradual recording of the information generated in the laboratory. Upon publication of a study, the database can directly export the data to a public repository. For a compendium project [e.g. the Compendium of *Arabidopsis* Gene Expression which will contain ~4000 full-genome *Arabidopsis* microarrays (http://www.psb.rug.ac.be/CAGE)], the data can be first transferred to a consortium database and later to a repository [10].

Currently, the only fully MIAME-supportive database is the ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress) [12], although other microarray databases are being developed so that they will eventually support MIAME [13,14]. Some journals already require publication of MIAME-compliant data to one of the two current repositories, ArrayExpress or Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) [15].

Although at this early stage observance of the MIAME guidelines has yet to demonstrate improvements in the comparability of microarray experiments, it is clear that without this information meaningful comparison and integration of data generated in different laboratories or on different platforms will be impaired and errors or misunderstandings could go undetected. Even if these data conforms to MIAME standards, however, comparison will remain difficult because many variables are involved and new flexible statistical procedures will be needed that make the most of this information.

Comparison of microarray technologies and validation of microarray results

Microarray data can be obtained from arrays containing cDNA clones [16], short (25-mer) [17] or long (60-mer) [18] oligonucleotides, or gene-specific PCR products amplified from genomic DNA [19,20]. These platforms differ in sequence content and measurement methodologies (Box 1) and thus produce qualitatively different data. If we are to integrate data from multiple sources, we must understand the specifics and the trade-offs in the different technologies.

Absolute measurements versus expression ratios

Carefully designed 25-mer oligochips from Affymetrix provide an absolute measurement of expression in an RNA sample (Box 1). By contrast, cDNA microarrays perform a two-color competitive hybridization (Box 1) that gives the ratio of transcript expression in two samples. Competitive hybridization results in the cancellation of multiple unwanted effects (e.g. reporter sequence and length) at the cost of losing important information [21] about the absolute levels of expression. Long oligonucleotide platforms (typically 60- to 80-mers) also use competitive hybridization because on this platform relative measurements were shown to result in higher precision than absolute measurements [18].

A key difference between absolute measurements and ratios is in the design of the experiment [22–24], which aims to maximize the statistical power of the experiment. For a series of two-channel hybridizations, the easiest setup is to compare all the test samples against the same reference. However, this setup wastes almost half the resources by re-measuring the same reference. In many situations, more powerful designs are possible that collect informative measurements from both channels (e.g. dye swap, loop designs or factorial designs) [22].

Assessment of technology performance

Several indicators (precision, reporter identity, specificity and sensitivity) (Box 2) capture the different aspects of platform performance.

Precision

Reproducibility determines for statistical analysis the ability to detect the presence of a transcript or a difference

Box 1. Sequence content and measurement principle

cDNA microarrays consist of cDNA clones spotted orderly at high density at defined positions on glass slides. In yeast, the full-length sequence of each cDNA is known, whereas in other species the cDNA clones might not be full length or might be only partially sequenced.

The oligonucleotide microarrays are currently produced in two formats. On the short (25-mer) oligonucleotide platform, each transcript is probed with a set of several reporters, arranged as pairs of perfect match–mismatch, which permits estimation of the specificity (Box 2) of the signal for each target. On the long oligonucleotide platforms, each transcript is probed with a 60-mer reporter, providing higher sensitivity (compared with the 25-mers) but no target-by-target estimation of specificity.

Two principles of measurement of expression are employed: (1) hybridization of a single labeled sample derived from the RNA sample, followed by one-channel detection, in which the intensity of the hybridization signal is used to determine the concentration of the target (absolute quantification); and (2) competitive hybridization of two labeled samples, each of them derived from one of the two compared RNA samples (usually named 'test' and 'reference') and labeled with a different fluorescent dye. The two labeled samples are mixed and hybridized to the same slide. After two-channel detection the ratio of fluorescence intensities from the two dyes measures the ratio of concentrations of the same target between the two samples. The short oligonucleotide platforms use single-channel measurements whereas microarrays of cDNA clones, long oligonucleotides or genomic DNA use two-channel measurements.

Box 2. Parameters of microarray performance

Precision
Precision describes how accurately the measurement (here a hybridization signal intensity or a ratio of two intensities) can be reproduced and is usually reported as a standard deviation or average replicate error. It can be determined by running replicated experiments on the same RNA sample.

Accuracy
Accuracy describes how close to a true value a measurement lies. It can be estimated in experiments where relevant RNA populations are spiked with several realistic targets of known concentrations, or from comparisons with validation experiments.

Specificity
Specificity is the proportion of the signal of a reporter that originates from the intended target. Imperfect specificity is for the main part caused by cross-hybridization from other transcripts.

Sensitivity
Sensitivity is the lowest target concentration at which an acceptable accuracy is obtained.

in expression. In many experimental systems the biological variability of gene expression can be greater than the variability of measurements. In general, biologically replicated experiments (i.e. repeated measurements on mRNA samples from independent experiments) (Fig. 1) are needed to filter out the biological variability. However, for platform comparison, we review the reproducibility of the technically replicated measurements taken on the same mRNA samples (Fig. 1). In a series of self-on-self hybridizations for cDNA microarrays [25], the standard deviation (SD) of replicated $\log_2$ ratios [filtered and dye-normalized using locally weighted scatterplot smoothing (LOWESS) fit] was 0.27, with 5.5% of genes outside the 2 SD limit of 1.46-fold change. A similar percentage of false positives for differential expression can be expected when comparing different cDNA samples. To avoid these false positives, replicates (using different arrays) are essential to filter out irreproducible measurements. The coefficient of variation of triplicate intensity measurements (calculated for the 86% genes with highest transcript abundance) ranged from 0.2 to 0.29 using 25-mer Affymetrix S98 yeast oligochips [26]. For a 60-mer array, the manufacturer (Box 3, point 1) reports a median SD of $\log_{10}$ ratios of 0.018 with 94% of ratios below 1.5 for replicated self-on-self competitive hybridizations (of the same cDNA sample labeled with two fluorescent dyes).

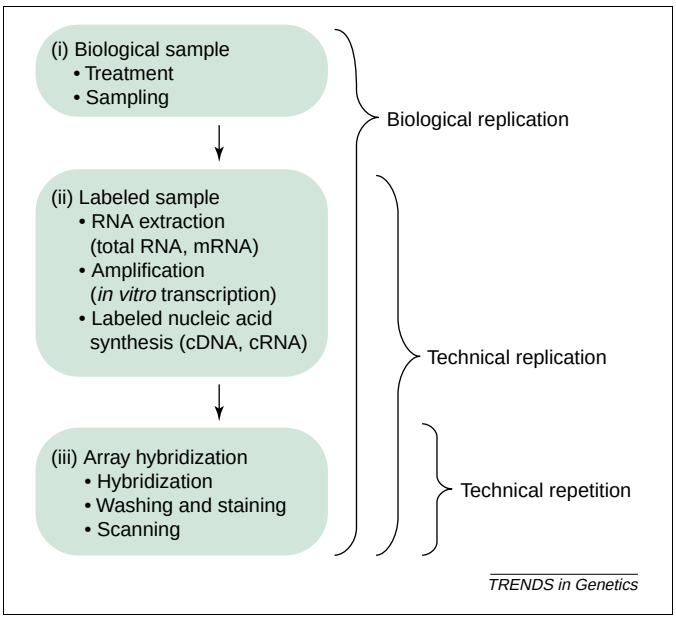


Fig. 1. The different steps of a microarray experiment and the different types of replication. Given the many platforms, there are also many protocols for performing a microarray experiment. We can however distinguish three phases: (i) production of the biological sample; (ii) RNA extraction and production of the sample of labeled nucleotides; and (iii) array hybridization. For the biological sample, by treatment we mean almost any attribute of a biological experiment – which can range from a specific choice of microbial strain under given growth conditions, to treating a mouse with a specific drug or to collecting a specific type of tumor from different patients. For the production of the labeled sample, many variants are possible, depending on the choice of cDNA or cRNA as the nucleic acids for hybridization and on the choice of labeling strategy (and possibly also the use of an amplification strategy). If a microarray experiment is replicated by producing a new biological sample, we talk about a biological replicate. If an experiment is replicated by producing a new sample of labeled nucleic acids from the same biological sample, we talk about a technical replicate. If the same labeled sample is hybridized to another array, we talk about a repetition. When performing a microarray experiment, biological replicates are crucial because conclusions drawn from an unreplicated microarray are applicable only to the observed individuals and not to the biological population it is intended for. When assessing the performance of a microarray platform, technical replicates are appropriate because biological variability is out of scope in this case. Technical repetitions are somewhat less appropriate for technology assessment because sample labeling is an integral part of the technology.

Reporter identity

For cDNA microarrays the correct identity of the reporter deposited on the slide cannot be taken for granted, given the incidence of errors in large clone collections, at least in mouse [27]. For example, a random sample of 119 clones, mostly from the mouse NIA 15K library, was shown to contain 91% correct clones [28]. Major errors in reporter design (identity) also occurred with mouse oligochips [29]. To prevent such errors in the future, Affymetrix has published its reporter sequences and a detailed description of its design pipeline (Box 3, point 2).

Specificity

For cDNA microarrays, non-intended targets with sequence identity >70% cross-hybridize to the spotted cDNA reporters [30], which makes it impossible to distinguish closely related gene family members. Oligonucleotide reporters can have high specificity for the intended targets [17,18,31] and the possibility of estimating the specificity for every probe set on 25-mer oligochips (including a mismatch or deletion control for each perfect match probe) (Box 3, points 3 and 4) provides an additional level of assurance. To provide a similar level of specificity on a clone-based platform for *Arabidopsis*, the Complete *Arabidopsis* Transcriptome Microarray (CATMA) project designed gene-specific PCR amplicons from genomic DNA by choosing 150–500 bp regions of each transcript with <70% sequence identity to any other transcript whenever possible [20].

Box 3. Technical notes from microarray manufacturers (non-peer reviewed)

(1) Fulmer-Smentek SB. Performance of Agilent Technologies 60 mer *in situ* synthetized oligonucleotide microarrays. 2001. Technical note: Publication number 5988–5063EN. http://www.chem.agilent.com/scripts/LiteraturePDF.asp?iWHID = 30998

(2) Array Design for the GeneChip Human Genome U133 Set. 2001. Technical note: Part No. 701133 rev 1. http://www.affymetrix.com/support/technical/technotesmain.affx

(3) Brzoska P. Background Analysis and Cross Hybridization. 2001. Technical note: Publication number 5988–2363EN. http://www.chem.agilent.com/scripts/LiteratureResults.asp?iProdGroup = 10andiProdLine = 15andiModel = 1188 andiProdInfotype = 68

(4) Statistical Algoritm Description Document. 2002. White paper: Part Number 701137 Rev 3. http://www.affymetrix.com/support/technical/whitepapers.affx

(5) New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays. 2001. Technical note: Part No. 701097 Rev 3. http://www.affymetrix.com/support/technical/technotesmain.affx.

(6) Performance and Validation of the GeneChip Human Genome U133 Set. 2002. Technical note: Part No. 701211 Rev 1. http://www.affymetrix.com/support/technical/technotesmain.affx

Sensitivity

For HG-U95v2 oligochips (25-mers) (available from Affymetrix) transcripts spiked into human RNA were detected with 90% accuracy at 1 picomole and with 100% at 2 picomole (Box 3, point 5), corresponding to ~1 transcript in 100 000. The 60-mer platform has a higher sensitivity of 1 transcript in 1 000 000 [18], with a dynamic range of 0.05–5.00 picomole (Box 3, point 5). In a study comparing the sensitivity of cDNA microarrays and northern blots for 84 genes, the authors concluded that the sensitivity of both methods was comparable [32]. Evans *et al.* [33] compared oligochips with SAGE on a sample from a complex tissue (hippocampus) [34]. The RG-U34A oligochip reproducibly detected 30% of transcripts with a high-to-medium level of expression, as determined by SAGE, whereas the 30% of genes with the lowest abundance detected by SAGE were never detected.

The above indicators show that, although grossly similar, the performance of different microarray technologies (cDNA, short oligonucleotides, long oligonucleotides and genomic amplicons) is far from identical. From the limited literature, it is difficult to predict which technology, if any, will prevail.

Validation

Although countless published articles include validation of microarray results [35], this validation is most often secondary to the study-specific biological conclusion and is thus biased (i.e. only a small, non-random sample of the changed genes is verified). We focus here on dedicated studies that permit assessment or comparison of platform accuracy (Box 2).

In a particularly careful study, Yuen *et al.* [36] assessed the accuracy of the U74A mouse oligochip (25-mers) and of their custom-made cDNA microarray. They performed triplicate measurements of samples from two conditions on both platforms. For the 47 genes common to both platforms, they performed quantitative reverse transcription real-time PCR (QRT-PCR) and identified 17 genes with changed expression and 10 genes with unchanged expression. On either platform, the difference in expression was confirmed for 16 out of the 17 changed and for none out of the 10 unchanged genes. By comparing the relative expression measured by QRT-PCR against cDNA and oligonucleotide microarrays, the authors demonstrated that both platforms systematically underestimate high expression ratios.

Kothapalli *et al.* [37], who used human cDNA microarrays and oligochips, concentrated on verifying the results of the cDNA platform. Of 17 clones classified as differentially expressed by cDNA microarrays, four cDNA clones (24%) did not correspond to those described by the manufacturer, eight genes were confirmed as differentially expressed by northern blot (47%) and five were not confirmed (31%).

Zirlinger *et al.* [38] used *in situ* hybridization to verify oligochip results that had shown differential expression of 35 transcripts in distinct anatomical regions of mouse brain. They found that for ~60% of genes the results of *in situ* hybridization were consistent with the oligochip results, for 20% of genes the results were inconsistent (7% of genes had a regional pattern that was different from those identified using the oligochips and 13% of genes had a high expression in all the regions examined) and for 20% of genes *in situ* hybridization failed to produce a signal.

Cross-platform comparison

In contrast to low-throughput techniques that allow only limited validation, cross-platform comparisons could be an efficient way to validate results for large numbers of genes (by protecting us from the idiosyncrasies of a particular platform). Such comparisons are also necessary for developing techniques to integrate multiple data sets.

Several comparisons between platforms that produce the same type of measurements have revealed good agreement. The log ratios of intensities from hybridizations of two labeled samples from human brain and kidney to two generations of 25-mer oligochips had high correlation $r = 0.89$ ($n_{genes} = 2910$) (Box 3, point 6). The log ratios from a competitive hybridization of two samples to a cDNA microarray and from a competitive hybridization of the same two samples to a 60-mer oligonucleotide microarray had a higher correlation: $r = 0.97$ ($n_{genes} = 4598$) [18]. Correlation between intensity measurements and tag counts resulting from SAGE [39] was also good: $r = 0.817$ ($n_{genes} = 224$) [40].

The situation is less clear when ratio measurements are compared with absolute intensity measurements. On the one hand, Kuo *et al.* [41] compared two published data sets from 56 human cancer cell lines for cDNA microarrays (ratios) and for HU6800 oligochips (intensities). The average gene correlation found in the study was worryingly low: $r = 0.278$ ($n_{genes} = 2895$, $n_{samples} = 56$). Kothapalli *et al.* [37] also remarked that 'a large variation of expression profiles from the two platforms was clearly evident'. On the other hand, the correlation coefficient between the log ratios measured with cDNA microarrays

and the log ratios of the intensities measured with 25-mer oligochips by Yuen *et al.* [36] was high: 0.793 ($n = 47$). Thus, in this study the results from both microarrays were concordant and were also consistent with results using QRT-PCR. Also, in a study of hippocampal neurons [7], a comparison between cDNA microarray data of differentiating hippocampal neurons *in vitro* and mouse 11K oligochip (25-mers) data for the differentiation of intact hippocampi *in vivo* [42] provided a high average gene correlation between the log ratios derived from both platforms: $r = 0.646$ ($n_{\text{genes}} = 475$, $n_{\text{samples}} = 5$) (even though the biological systems were not identical). Very recently, Barczak *et al.* [43] found strong correlations ($r = 0.8–0.9$) (using at least four replicate samples from K562 erythroleukemia cells from a single culture) between expression ratios for a long oligonucleotide (70-mer) platform and for a short oligonucleotide (25-mer) platform (U95Av2).

The key point is that the good agreement between platforms in the studies of Yuen *et al.* [36] and Barczak *et al.* [43] and between the studies of Dabroswki *et al.* [7] and Mody *et al.* [42] was obtained after filtering or averaging out non-reproducible profiles by using replicates from different experiments. However, no replicates were available to Kuo *et al.* [41], which seriously decreased the value of these important data. We thus conclude that, after appropriate filtering, ratio and intensity data from different platforms can be compared and are amenable to integration and are useful for the validation of results.

Meta-analysis of microarray data

What if, in the light of our previous argument, several studies addressing the same question are available to us? Could we analyze those data sets in an integrated fashion and extract more information than from a single data set? Before considering more advanced data analyses, let us look at the most basic question: which genes are differentially expressed between two groups of samples? Meta-analysis is a set of classical statistical techniques [44] to combine results from several studies. Recently, its applicability to microarray data was demonstrated for the first time [45]. Such meta-analysis is built on top of statistical tests for the detection of differential expression (Box 4). These tests score genes generally by reporting a P value that expresses the probability that the observed level of differential expression could have occurred by chance. However, because such procedures test thousands of genes (and thus generate many false positives), there is a need to adjust P values to control this effect (Box 5).

Once P values are available for each gene in each study, some simple methods (called omnibus procedures) [44] are available to test the statistical significance of P values combined from several tests. Because P values from continuous statistics are (by definition) uniformly distributed between 0 and 1, combining only P values frees us from any dependency on the statistical test or on the distribution of the data. The hypotheses tested in the different data sets need not even be the same. For example, we could imagine combining a data set for tumors with positive and negative responses to chemotherapy with a data set for the same type of tumors with good and bad prognosis.

One method [46] to test the significance of combined results is to take for one gene the minimum P value ($P_{\min}$) observed over the different data sets (k) but test this minimum P value at a higher stringency than the single-study rejection threshold α Box 4: reject 'no differential expression' if $P \min < 1 - (1 - \alpha)^{1/k}$. This method is sensitive to outliers, so a variant uses the nth smallest value as the test statistic [47]. Another method is Fisher's inverse chi-square method [48]. It consists of computing a combined statistic S from the different P values

$$S = -2 \log P_1 - \cdots - 2 \log P_k$$

and using this statistic for testing. It is also possible to extend Fisher's method by giving each data set a different weight [49], which will be important for microarray data where the quality of different data sets can be highly variable. How to determine good weights given the data of a microarray experiment remains undecided at present, but weights will probably summarize the discrimination power and the noise in the data.

Although omnibus procedures are versatile and easy to implement, they have the major drawback that, by working only with the P values, it is impossible to estimate the level of differential expression observed [effect size: $(\mu_1 - \mu_2)/\text{SD}$, where μ_1, μ_2 are the group means] Many procedures can tackle this question [44] and they often closely resemble the procedures for the detection of differential expression (Box 1) but they incorporate the study as an additional explanatory variable [50].

In the first application of meta-analysis to microarrays, Rhodes *et al.* [45] combined four data sets on prostate

Box 4. Detecting differential expression

The most basic setup of a microarray experiment is to measure gene expression for two distinct groups of samples (e.g. mice with treatment versus control mice) and to ask which genes are expressed differently between the two groups. Other more advanced experimental designs are of course possible [22,24]. The simplest approach to detecting differential expression is to consider a statistic t that expresses the difference between the observed average expression levels or ratios across the two groups divided by the estimated standard deviation over these groups. This approach can be extended in many ways as witnessed by the recent flurry of publications on the detection of differential expression [60]. We mention only a few possibilities, such as the nonparametric approach in significance analysis of microarrays [61], Bayesian tests [62] or analysis of variance (ANOVA) [63,64]. Most of these approaches then associate to each gene a P value that assesses the probability that the level of differential expression observed for this gene could have occurred by chance. If the P value is lower than a rejection threshold α (e.g. $P < 0.05 = \alpha$;) then the (null) hypothesis that the gene does not show any differential expression between the groups is rejected and the (alternative) hypothesis that there is differential expression is accepted. It is therefore possible that, by chance and because of experimental and biological noise, the observations for a gene that is truly not differentially expressed appear to indicate differential expression (such a gene is a false positive for our test). Conversely, a gene that is actually differentially expressed could give rise to observations that suggest no differential expression (a false negative).

Box 5. Controlling false positives

The genome-wide character of microarrays has a statistical drawback when trying to detect differential expression. If we use the classical statistical threshold of $\alpha = 0.05$ on a microarray experiment with 20 000 genes, 100 of which are truly differentially expressed, we can expect approximately $(20\,000-100) \times 0.05 = \sim 1000$ false positives. Thus, the true positives get buried under the false positives. This situation can create a lot of confusion – for an example on the detection of cell cycle genes by microarrays, see [65,66].

Although there is no easy way out of this conundrum, there are several approaches to improve the situation. The first approach is to ensure that the probability of at least one false positive among all genes tested (called the family-wise error rate) be lower than a threshold. This approach leads to the Bonferroni correction [67] that consists in multiplying each *P* value by the number of genes tested to obtain a corrected *P* value. Unfortunately, because of a large number of measurements and the noisy nature of microarray data, this might lead to the reverse situation where most truly differentially expressed genes get rejected because the statistical requirement becomes so stringent. Improvements to this procedure are available, such as the Holm correction [68] and the Westfall and Young min *P* and max *T* adjusted *P* values [69]. Another approach is more intuitive to the biologist. The validation of microarray data is strongly driven by economics. How many true hypotheses can I discover after validation? How many genes can I afford to validate? Which proportion of the genes that I try to validate turn out to be true? The false discovery rate (FDR) addresses these questions and is the expected ratio of the number of true positives over the number of true positives plus false positives and procedures are available to correct *P* values according to the FDR [70].

cancer (two cDNA microarray studies [51,52] and two oligochip studies [53,54]) to determine genes that are differentially expressed between benign prostate tissue and clinically localized prostate cancer. The procedure they proposed was a variant of Fisher's method followed by a multiple testing correction through false discovery rate (FDR) adjustment (Box 5). Although the individual studies (with an FDR adjusted value of 0.1) identified 758 [51], 665 [52], 0 [53] and 1194 [54] genes as being overexpressed, the meta-analysis identified 50 genes as being consistently overexpressed across the studies at the same FDR-adjusted value. The method used by Rhodes *et al.* is highly conservative because of the choice of null hypothesis and we would not recommend it. In our re-analysis of the data from the three reliable studies [50,51,54], which used the classical version of Fisher's method, we found that 233 out of the 2126 genes common to the three studies to be reliably overexpressed at the same FDR-adjusted value.

Microarray analysis in the era of repositories and compendia

A new era is dawning on microarray analysis with large public resources of microarray easily available for retrieval and integrated analysis across platforms. But what are the obstacles lying ahead? And can we expect more benefits than just the improved statistical efficiency offered by meta-analysis?

At the technological level, trade-offs in costs and available expertise probably mean that several platforms will coexist for at least several years. However, sequence identity error in cDNA clones (at least in higher organisms) is worryingly high and sequence specificity is not optimal. Therefore, we can expect spotted cDNA arrays to be progressively replaced by spotted arrays of long oligonucleotides or other methodologies that improve sequence identity and specificity [20]. For compendium projects on two-channel platforms, where the use of a common reference is standard practice, using a specific and calibrated reference (e.g. an equimolar mixture of PCR products or oligonucleotides complementary to all array features [55,56] or external normalization spikes [57]) could greatly improve precision and accuracy – and might even allow recovering absolute measurements.

At the methodological level, there is now enough evidence to suggest that replicates of microarray experiments are essential if the data are to be of any value [41]. It must become standard practice to require sufficient biological replication before lending any credence to results based on microarray data.

At the practical level, we should not underestimate the burden placed on investigators to keep the annotation and data of each experiment MIAME compliant. This burden will be lessened if good software tools are developed.

At the infrastructure level, we can expect many new powerful features (much beyond simple storage and query). For example, data alerts could be automatically generated when a new data set relevant to your research is deposited – just like MEDLINE can generate publication alerts based on keywords. Extensive gene-centric views of the transcriptome could be made available for each gene, with a virtual expression profile summarizing all the available expression data [58,59]. Even automatic discovery alerts might be possible, after semi-automated data collection, by repeatedly performing a standard analysis script as new data becomes available and dispatching each incremental discovery to the investigator – just like automatic daily BLASTing of a sequence of interest for homolog detection.

At the data analysis level, we limited ourselves to meta-analysis for the improvement of the detection of differential expression because this is the current state of affairs. But the underlying ideas are clearly more broadly applicable. For example, clustering of gene expression profiles across multiple data sets will probably be achieved through the integration of clustering techniques and meta-analysis techniques. Classification methods could benefit from similar treatments. In fact, because reliable statistics is the basis of serious data mining, an improved statistical treatment of microarray data across platforms probably means that most data mining techniques applied to microarray results will eventually be able to deal with multiple data sets.

If we address fully these real challenges and pursue these exciting opportunities, in the next decade exploring transcriptomes should become almost as natural as exploring genomes.

Acknowledgements

We thank Joke Allemeersch for re-executing the meta-analysis of the prostate cancer data. This work was supported by the VIB, the K.U. Leuven (research council, GOA Mefisto 666, IDO), the FWO-Vlaanderen,

IWT (STWW, GBOU), AWI (Bil.Int.Coll.), the EU, FP5 DIADEM and CAGE and the DWTC (IUAP V-22 and IUAP V-19). M.D. is a Marie Curie fellow (QLK6-CT-2000–52154). Y.M. is a postdoctoral fellow of the FWO-Vlaanderen.

References

1 Arbeitman, M.N. *et al.* (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270–2275
2 Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257
3 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
4 Ramaswamy, S. *et al.* (2003) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49–54
5 Gollub, J. *et al.* (2003) The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Res.* 31, 94–96
6 Marc, P. *et al.* (2001) yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res.* 29, E63–3
7 Dabrowski, M. *et al.* (2003) Gene profiling of hippocampal neuronal culture. *J. Neurochem.* 85, 1279–1288
8 Haverty, P.M. *et al.* (2002) HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res.* 30, 214–217
9 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371
10 Stoeckert, C.J. Jr *et al.* (2002) Microarray databases: standards and ontologies. *Nat. Genet.* 32 (Suppl.), 469–473
11 Spellman, P.T. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, 3RESEARCH0046
12 Brazma, A. *et al.* (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68–71
13 Gardiner-Garden, M. *et al.* (2001) A comparison of microarray databases. *Brief. Bioinform.* 2, 143–158
14 Do, H.H. *et al.* (2003) Comparative evaluation of microarray-based gene expression databases. In *GI-Edition Lecture Notes in Informatics P-26* (Weikung, G. *et al.*, eds), pp. 482–502, Bonner Köllen Verlag
15 Genetics-Editorial, N. (2002) Coming to terms with microarrays. *Nat. Genet.* 32 (Suppl.), 333–334
16 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
17 Lipshutz, R.J. *et al.* (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* 21, 20–24
18 Hughes, T.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19, 342–347
19 Kim, H. *et al.* (2003) Gene expression analyses of *Arabidopsis* chromosome 2 using a genomic DNA amplicon microarray. *Genome Res.* 13, 327–340
20 Crowe, M.L. *et al.* (2003) CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.* 31, 156–158
21 Kuruvilla, F.G. *et al.* (2002) Vector algebra in the analysis of genome-wide expression data. *Genome Biol.*, 3RESEARCH0011
22 Yang, Y.H. *et al.* (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3, 579–588
23 Kerr, M.K. *et al.* (2001) Experimental design for gene expression microarrays. *Biostatistics* 2, 183–201
24 Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32 (Suppl.), 490–495
25 Yang, I.V. *et al.* (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, 3RESEARCH0062
26 Piper, M.D. *et al.* (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 277, 37001–37008
27 Halgren, R.G. *et al.* (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res.* 29, 582–588
28 Wurmbach, E. *et al.* (2001) Gonadotropin-releasing hormone receptor-coupled gene network organization. *J. Biol. Chem.* 276, 47195–47201
29 Knight, J. (2001) When the chips are down. *Nature* 410, 860–861
30 Xu, W. *et al.* (2001) Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. *Gene* 272, 61–74
31 Kane, M.D. *et al.* (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* 28, 4552–4557
32 Taniguchi, M. *et al.* (2001) Quantitative assessment of DNA microarrays – comparison with northern blot analyses. *Genomics* 71, 34–39
33 Evans, S.J. *et al.* (2002) Evaluation of Affymetrix gene chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial analysis of gene expression. *Eur. J. Neurosci.* 16, 409–413
34 Datson, N.A. *et al.* (2001) Expression profile of 30,000 genes in rat hippocampus using SAGE. *Hippocampus* 11, 430–444
35 Chuaqui, R.F. *et al.* (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.* 32 (Suppl.), 509–514
36 Yuen, T. *et al.* (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* 30, e48
37 Kothapalli, R. *et al.* (2002) Microarray results: how accurate are they? *BMC Bioinformatics* 3, 22
38 Zirlinger, M. *et al.* (2001) Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5270–5275
39 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
40 Ishii, M. *et al.* (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68, 136–143
41 Kuo, W.P. *et al.* (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18, 405–412
42 Mody, M. *et al.* (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* 98, 8862–8867
43 Barczak, A. *et al.* (2003) Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* 13, 1775–1785
44 Hedges, L.V. *et al.* (1985) *Statistical Methods For Meta-Analysis*, Academic Press
45 Rhodes, D.R. *et al.* (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* 62, 4427–4433
46 Tippet, L.H.C. (1931) *The Methods Of Statistics*, Williams and Norgate
47 Wilkinson, B. (1951) A statistical consideration in psychologial research. *Psychol. Bull.* 48, 156–158
48 Fisher, R.A. (1925) *Statistical Methods For Research Worker*, Oliver and Boyd
49 Good, I.J. (1955) On the weighted combination of statistical tests. *J. R. Stat. Soc. Ser. B* 17, 264–265
50 Normand, S.L. (1999) Meta-analysis: formulating, evaluating, combining and r eporting. *Stat. Med.* 18, 321–359
51 Luo, J. *et al.* (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* 61, 4683–4688
52 Dhanasekaran, S.M. *et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822–826
53 Magee, J.A. *et al.* (2001) Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res.* 61, 5692–5696
54 Welsh, J.B. *et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* 61, 5974–5978
55 Dudley, A.M. *et al.* (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7554–7559
56 Sterrenburg, E. *et al.* (2002) A common reference for cDNA microarray hybridizations. *Nucleic Acids Res.* 30, e116
57 van de Peppel, J. *et al.* Monitoring global mRNA changes with externally controlled microarray experiments. *EMBO Rep.* (in press)
58 Diehn, M. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies and gene expression data. *Nucleic Acids Res.* 31, 219–223
59 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41

60 Nadon, R. *et al.* (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.* 18, 265–271
61 Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121
62 Baldi, P. *et al.* (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519
63 Kerr, M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819–837
64 Jin, W. *et al.* (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* 29, 389–395
65 Delaunay, F. *et al.* (2002) Circadian clock and microarrays: Mamm. Genome gets rhythm. *Trends Genet.* 18, 595–597
66 Cooper, S. (2002) Cell cycle analysis and microarrays. *Trends Genet.* 18, 289–290
67 Miller, R.G. (1966) *Simultaneous Statistical Inference*, McGraw Hill
68 Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 65–70
69 Westfall, P.H. *et al.* (1993) On adjusting P-values for multiplicity. *Biometrics* 49, 941–945
70 Benjamini, Y. *et al.* (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300

Could you name the most significant papers published in life sciences this month?

Updated daily, **Research Update** presents short, easy-to-read commentary on the latest hot papers, enabling you to keep abreast with advances across the life sciences.
Written by laboratory scientists with a keen understanding of their field, **Research Update** will clarify the significance and future impact of this research.

Our experienced in-house team is under the guidance of a panel of experts from across the life sciences who offer suggestions and advice to ensure that we have high calibre authors and have spotted the 'hot' papers.

Visit the **Research Update** daily at http://update.bmn.com and sign up for email alerts to make sure you don't miss a thing.

This is your chance to have your opinion read by the life science community, if you would like to contribute, contact us at research.update@elsevier.com