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

Yves Moreau, Stein Aerts, Bart De Moor, Bart De Strooper, Michal Dabrowski

Y. Moreau[1*], S. Aerts[1], B. De Moor[1], B. De Strooper[2], M. 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

[*] Corresponding author: moreau@esat.kuleuven.ac.be

Summary

The upcoming availability of public microarray repositories and of large compendia of gene expression opens up a new realm of possibilities for microarray data analysis. An essential challenge along this road (and still mostly an open problem) 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) efficient access to and exchange of microarray data, (2) validation and comparison of data from different platforms (cDNA and short and long oligos), and (3) integrated statistical analysis of multiple data sets.

In the last 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 especially efficient comparison and integration of data obtained in related biological systems provide biologists and geneticists with an enormous opportunity to address complex questions in an effective way.

Tellingly, larger microarray projects are gearing up towards the generation of large compendia of gene expression. Those 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 associated with a diverse series of human pathologies [4]. We envision 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 big areas of the transcriptome while smaller-scale projects will refine the mesh, starting from a careful analysis of publicly-available microarray (and sequence) data to design experiments that sharpen and validate primary hypotheses.

But what are the barriers to this bonanza of information and how do we open them up? 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 has been scattered around the web, often as supplementary data to a paper. Consequently it has been difficult for investigators to know where relevant data is available. Several databases have started to address this problem by making some published microarray data available for query after uniform processing and filtering – while providing links to the original publications for more detailed information. These databases have diverse purposes: (1) platform specific (such as the Stanford Microarray Database [5]), (2) organism specific (such as the yeast Microarray Global Viewer [6]), or (3) project specific (such as the Lifecycle database on *Drosophila* development [1], the NeuroDiff database on neuronal differentiation in mouse [7], or the HugeIndex database on normal expression in human tissues [8]).

Although web supplements and microarray databases provide access to many data sets, they have the drawbacks that (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 reanalyze the data and (2) they do not use a standard format for microarray data and experi-

ment description. 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 problems. The Minimum Information About a Microarray Experiment (MIAME) specification [9] is practically a checklist that guides the investigator in the annotation of microarray experiments. As numerous biological and experimental factors influence gene expression measurements (from lighting conditions in plant experiments to the exact histopathology of a tumor, from the difference in specificity of different reporter sequences for the same gene to the particularities of a single batch of slides or to the laser intensity at which a slide is scanned), this minimum information includes the experimental design, the array design, the details of the samples and any treatments, the hybridization conditions, the measurements, and the 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 laboratory practice, a local MIAME-supportive database will allow gradual recording of this information. Upon publication, the database can directly export the data to a public repository. For a compendium project (such as the Compendium of Arabidopsis Gene Expression that will contain about 4,000 full-genome *Arabidopsis* microarrays for the plant community; 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 most other databases are working towards MIAME support [13,14]. Some journals already require publication of MIAME-compliant data to one of the two current repositories: ArrayExpress or GEO (http://www.ncbi.nlm.nih.gov/geo/) [15].

Although at this early stage observance of the MIAME guidelines has yet to demonstrate actual improvements in comparability of microarray experiments, it is clear that without this information meaningful comparison and integration of data generated by different labs or on different platforms will be fatally impaired and that major errors or misunderstandings could go undetected for a long time. Even with this information, however, comparison will remain difficult because so many factors come into play and new flexible statistical procedures will be needed that make the most of all this information.

Comparison of microarray technologies and validation of microarray results

Microarray data can be obtained from arrays of cDNA clones [16], of short (25-mer) [17] or long (60-mer) [18] oligonucleotides, or of 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 of the different technologies.

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, while in other species the cDNA clones may not be full length or may 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 (see 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 (as compared to 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 while microarrays of cDNA clones, long oligos, or genomic DNA use two-channel measurements.

Absolute measurements vs. expression ratios

Upon careful design, the 25-mer oligochips of 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., of reporter sequence and length) at the cost of losing the important information [21] about the absolute levels of expression. Long oligonucleotide platforms (typically 60- to 80-mers) also use the competitive hybridization, because on this platform relative measurements were shown to result in higher precision than absolute ones [18].

Another key difference between absolute measurements and ratios is in the design of experiment [22-24], which aims at maximizing the statistical informativeness 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 measuring the same reference again and again. In many situations, more powerful designs are possible by putting both channels on an equal footing (e.g., dye swap, loop designs, or factorial designs) [22].

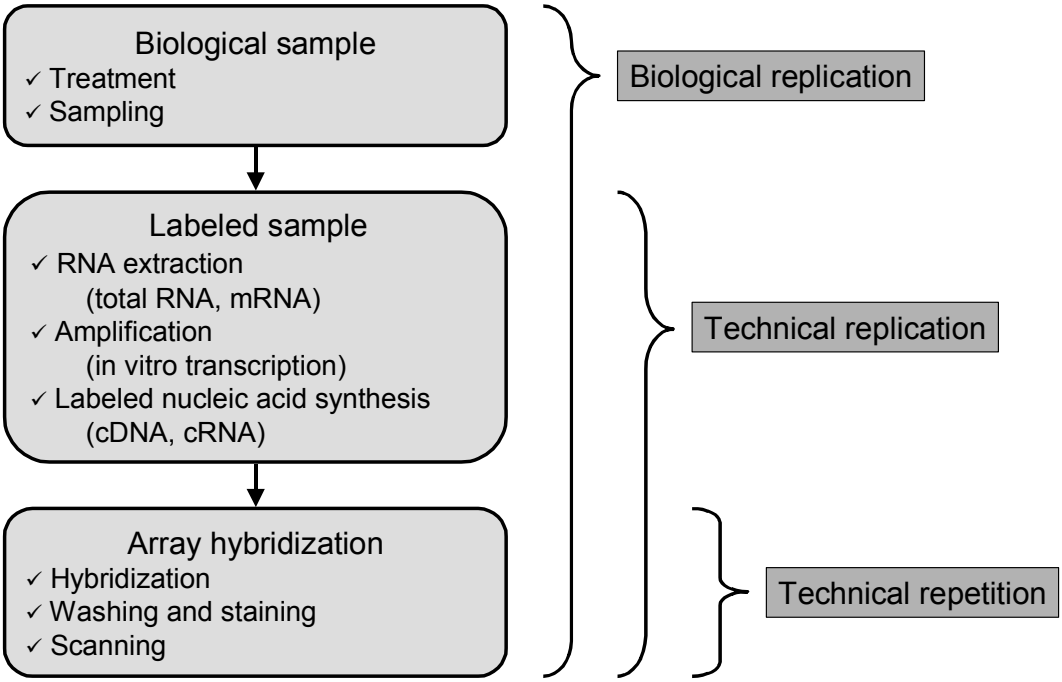

Figure 1. The different steps of a microarray experiment and the different types of replication. Given the many platforms, there are also many protocols for performing a microarray experiment. We can however distinguish three phases: (1) production of the biological sample, (2) RNA extraction and production of the sample of labeled nucleotides, and (3) 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.

Assessment of technology performance

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

Box 2. Parameters of microarray performance

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

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

Precision Reproducibility determines for statistical analysis the ability to detect the presence of a transcript or a difference in expression. In many experimental systems the biological variability of gene expression may be greater than the variability of measurements. In general, biologically replicated experiments (i.e., repeated measurements on mRNA samples from independent experiments; see Figure 1) are needed to filter out the biological variability. However here, for platform comparison, we review the reproducibility of the technically replicated measurements taken on the same mRNA samples (see Figure 1 also). 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 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. On 25-mer Affymetrix S98 yeast oligochips, the coefficient of variation of triplicate intensity measurements (calculated for the 86% genes with highest transcript abundance) ranged from 0.2 to 0.29 [26]. For a 60-mer array, the manufacturer (Box 5-a) reports for replicated self-on-self competitive hybridizations (of the same cDNA sample labeled with two fluorescent dyes) a median SD of $\log_{10}$ ratios of 0.018 with 94% of ratios below 1.5.

Reporter identity On 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 happened 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 5-b).

Specificity On cDNA microarrays, non-intended targets with sequence identity greater than 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 5-c,d) provides an additional level of assurance. To provide a similar level of specificity on a clone-based platform for *Arabidopsis*, the CATMA project designed gene-specific PCR amplicons from genomic DNA by choosing 150-500 bp regions of each transcript with less than 70% sequence identity to any other transcript [20], whenever possible.

Sensitivity On HG-U95v2 oligochips (25-mers), transcripts spiked into human RNA were detected with 90% accuracy at 1 pm and with 100% at 2 pm (Box 5-e), corresponding to about 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 pm (Box 5-a). In a study comparing the sensitivity of cDNA microarrays and Northern blots for 84 genes, the authors concluded that sensitivity of the two 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 high-to-medium level of expression, as determined by SAGE, while the 30% of genes with the lowest abundance by SAGE were never detected.

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

Validation

Although countless papers include validation of microarray results [35], this validation is most often subordinate to the study-specific biological conclusion and 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 cDNA microarray. They performed on both platforms triplicate measurements for samples from two conditions. For the 47 genes common to both platforms, they also performed quantitative reverse transcription real-time PCR (QRTPCR) and identified among those genes 17 genes with definitely 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 0 out of the 10 unchanged genes. By comparing the relative expression measured by QRTPCR against cDNA and oligonucleotide microarrays, the authors demonstrated that both platforms systematically underestimate high expression ratios.

Kothapali 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, 4 cDNA clones (24%) were not the ones claimed by the manufacturer, 8 genes were confirmed as differentially expressed by Northern blot (47%), and 5 were not confirmed (31%).

Zirlinger et al. [38] used *in situ* hybridization to verify oligochip results that had shown differential expression of 35 transcripts between distinct anatomical regions of mouse brain. They found that for approximately 60% of genes the results of *in situ* hybridization were consistent with the oligochip results, for 20% the results were inconsistent (7% regional pattern different from the oligochip results, 13% expression high in all the regions), and for 20% *in situ* hybridization did not produce any signal.

Cross-platform comparison

By contrast with 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 producing 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}=2,910) (Box 5-f). 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 an even higher correlation: r=0.97 (n_{genes}=4,598) [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}=2,895, $n_{samples}$=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 with both microarrays were concordant between themselves and with results of QRTPCR. Also, in a study on hippocampal neurons [7], a comparison of cDNA microarray data of differentiating hippocampal neurons in vitro against mouse 11K oligochip (25-mers) data for the differentiation of intact hippocampi in vivo [42] provided also a high average gene correlation between the log ratios from both platforms: r=0.646 (n_{genes}=475, $n_{samples}$=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 oligo (70-mer) platform and for a short oligo (25-mer) platform (U95Av2).

The key point is that the good agreement in [36] and [43] and between [7] and [42] was obtained after filtering or averaging out irreproducible profiles thanks to the use of replicates from different experiments, while no replicates were available to Kuo et al. [41] – seriously jeopardizing 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 thus amenable to integration and useful for results validation.

Meta-analysis of microarray data

What if, in the light of our previous argument, several studies addressing the same question are available to us? Can 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 is to determine 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 3). These tests score genes generally by reporting a *p* value that expresses the chance 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 4).

Box 3. Detecting differential expression

The most basic setup of a microarray experiment is to measure gene expression for two distinct groups of samples (for example, mice with treatment vs. control mice) and to ask which genes are expressed differently between the two groups. Other more advanced experiment designs are of course possible, but we leave this issue aside and refer to recent reviews on these topics [22,24]. The simplest approach to detecting differential expression is to consider a *t* statistic 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 [a]. We mention only a few possibilities, such as the nonparametric approach in Significance Analysis of Microarrays [b], Bayesian tests [c], or analysis of variance (ANOVA) [d,e]. 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 some 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 thus 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 have observations that suggest no differential expression (false negative).

References

a. Nadon, R. et al. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.* 18, 265-71

b. Tusher, V.G. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U S A* 98, 5116-21

c. 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, 509-19

d. Kerr, M.K. et al. (2000) Analysis of variance for gene expression microarray data. *J. Comput Biol.* 7, 819-37

e. Jin, W. et al. (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat. Genet.* 29, 389-95

Box 4. Controlling false positives

The genomewide character of microarrays has a nasty 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 about (20.000-100)*0.05=995 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 [a,b].

Although there is no easy way out of this conundrum, there are several approaches to improve the situation. The first approach is to require that the probability of at least one false positive among all genes tested (called the *familywise error rate*) be lower than some threshold. This approach leads to the Bonferroni correction [c] that consists in multiplying each *p* value by the number of genes tested to obtain a corrected *p* value. Unfortunately, because of large number of measurements and the noisy nature of microarray data, this may 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 Holm's correction [d] and the Westfall and Young min*P* and max*T* adjusted *p* values [e]. 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 I try to validate turn out to be true? The *false discovery rate* (FDR) addresses this question 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 [f].

References

a. Delaunay, F. et al. (2002) Circadian clock and microarrays: mammalian genome gets rhythm. *Trends Genet.* 18, 595-7

b. Cooper, S. (2002) Cell cycle analysis and microarrays. *Trends Genet.* 18, 289-90

c. Miller, R.G. (1966) *Simultaneous statistical inference*, McGraw Hill

d. Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 65-70

e. Westfall, P.H. et al. (1993) On adjusting P-values for multiplicity. *Biometrics* 49, 941-45

f. Benjamini, Y. et al. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc., Series B* 57, 289-300

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. Since p values from continuous statistics are (as a result of their 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 on the different data sets need not even be the same! For example, we could imagine combining a data set for tumors with good and bad responses to chemotherapy with a data set for the same type of tumors with good and bad prognosis (as we might be interested in identifying genes associated with both bad response and bad prognosis).

A first method [46] to test the significance of combined results is to take for one gene the minimum p value $p_{\min}$ observed over the k different data sets but test this minimum p value at a more stringent level than the single-study rejection threshold α (see Box 3):

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 in 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 an open question at this moment; 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: $(\mu_1-\mu_2)/SD$). 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 cancer (two cDNA microarray studies [51,52] and two oligochip studies [53,54]) to determine genes differentially expressed between benign prostate tissue and clinically localized prostate cancer. The procedure they propose is a variant of Fisher's method followed by a multiple testing correction through false discovery rate (FDR) adjustment (Box 4). While the individual studies called, at an FDR adjusted value of 0.1, respectively 758 [51], 665 [52], 0 [53], and 1194 [54] genes as overexpressed, the meta-analysis identified 50 genes as consistently overexpressed across the studies at the same FDR adjusted value. The method used by Rhodes *et al.* is however highly conservative because of a particular choice of null hypothesis and we do not recommend it. In our reanalysis of the data of the three reliable studies [50,51,54] using the classical version of Fisher's method, we found 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 upon 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 benefits bigger 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 constantly evolving platforms will coexist for a long time. However, sequence identity error in cDNA clones (at least in higher organisms) is worryingly high and sequence specificity is not optimal. So we can expect spotted cDNA arrays to be progressively replaced by spotted arrays of long oligos 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 (such as an equimolar mixtures of PCR products or oligos complementary to all array features [55,56] or external normalization spikes [57]) could greatly improve precision and accuracy – and may even allow recovering absolute measurements.

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

At the sociological 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 carried only

if good software tools that minimize it are developed and if the return on this effort becomes rapidly clear.

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 with, for each gene, a virtual expression profile summarizing all the available expression data [58,59]. Even automatic discovery alerts could 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 with meta-analysis techniques. Similarly, classification methods could benefit from similar treatments. In fact, because decent statistics lies at the basis of any serious data mining, an improved statistical treatment of microarray data across platform probably means that most data mining techniques applied to microarray data will eventually be able to deal with multiple data sets.

If we address properly these real difficulties and boldly pursue these exciting opportunities, we can hope that, in the next decade, exploring transcriptomes will become almost as natural as exploring genomes.

Box 5. Technical notes from microarray manufacturers (non peer reviewed)

a. 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/LiteratureResults.asp?iProdGroup=10&iProdLine=15&iModel=1245&iProdInfotype=68

b. 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

c. Brzoska P. Background Analysis and Cross Hybridization. 2001. Technical note: Publication number 5988-2363EN. http://www.chem.agilent.com/scripts/LiteratureResults.asp?iProdGroup=10&iProdLine=15&iModel=1188&iProdInfotype=68

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

e. 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

f. 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

Acknowledgments

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, 2270-5

2 Gasch, A.P. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241-57

3 Hughes, T.R. et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109-26

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, 94-6

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. Gene profiling of hippocampal neuronal culture. *J. Neurochem.* in press

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, 214-7
9 Brazma, A. et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365-71
10 Stoeckert, C.J., Jr. et al. (2002) Microarray databases: standards and ontologies. *Nat. Genet.* 32 Suppl, 469-73
11 Spellman, P.T. et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 3, RESEARCH0046
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, 143-58
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., Schöning, H., and Rahm, E., eds), pp. 482-502, Bonner Köllen Verlag, Bonn, Germany (http://www.btw2003.de/proceedings/paper/96.pdf)
15 Genetics-Editorial, N. (2002) Coming to terms with microarrays. *Nat. Genet.* 32 Suppl, 333-4
16 DeRisi, J.L. et al. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-6
17 Lipshutz, R.J. et al. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* 21, 20-4
18 Hughes, T.R. et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19, 342-7
19 Kim, H. et al. (2003) Gene expression analyses of Arabidopsis Chromosome 2 using a Genomic DNA Amplicon Microarray. *Genome Res.* 13, 327-40
20 Crowe, M.L. et al. (2003) CATMA: a complete Arabidopsis GST database. *Nucleic Acids Res.* 31, 156-8
21 Kuruvilla, F.G. et al. (2002) Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* 3, RESEARCH0011
22 Yang, Y.H. et al. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3, 579-88
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, 490-5
25 Yang, I.V. et al. (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* 3, research0062
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, 37001-8
27 Halgren, R.G. et al. (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic. Acids Res.* 29, 582-8
28 Wurmbach, E. et al. (2001) Gonadotropin-releasing hormone receptor-coupled gene network organization. *J. Biol. Chem.* 276, 47195-201
29 Knight, J. (2001) When the chips are down. *Nature* 410, 860-1
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, 4552-7
32 Taniguchi, M. et al. (2001) Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. *Genomics* 71, 34-9
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, 409-13
34 Datson, N.A. et al. (2001) Expression profile of 30,000 genes in rat hippocampus using SAGE. *Hippocampus* 11, 430-44
35 Chuaqui, R.F. et al. (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.* 32 Suppl, 509-14
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, 5270-5
39 Velculescu, V.E. et al. (1995) Serial analysis of gene expression. *Science* 270, 484-7
40 Ishii, M. et al. (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68, 136-43
41 Kuo, W.P. et al. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18, 405-12

42 Mody, M. et al. (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl. Acad. Sci. U S A* 98, 8862-7
43 Barczak, A. et al. (2003) Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* 13, 1775-85.
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, 4427-33
46 Tippet, L.H.C. (1931) *The methods of statistics*, Williams and Norgate
47 Wilkinson, B. (1951) A statistical consideration in psychologial research. *Psych. Bull.* 48, 156-8
48 Fisher, R.A. *Statistical methods for research worker*, Oliver and Boyd
49 Good, I.J. (1955) On the weighted combination of statistical tests. *J. Royal Stat. Soc., Series B* 17, 264-5
50 Normand, S.L. (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.* 18, 321-59
51 Luo, J. et al. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* 61, 4683-8
52 Dhanasekaran, S.M. et al. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822-6
53 Magee, J.A. et al. (2001) Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res.* 61, 5692-6
54 Welsh, J.B. et al. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* 61, 5974-8
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, 7554-9
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 Reports* 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, 219-23
59 Hubbard, T. et al. (2002) The Ensembl genome database project. *Nucleic. Acids Res.* 30, 38-41