REVIEW

Microarray truths and consequences

R Šášik¹, C H Woelk¹ and J Corbeil¹,²
¹School of Medicine, University of California San Diego, La Jolla, California 92093-0679, USA
²Department of Medicine, Université Laval, Québec, Canada, G1K 7P4

(Requests for offprints should be addressed to J Corbeil; Email: jcorbeil@ucsd.edu)

Abstract

For many, analysis of a microarray experiment starts with a spreadsheet of expression levels. While great attention is duly paid to RNA extraction, preparation and hybridization, relatively little care is devoted to extraction of expression levels from the fluorescent image. By delegating this step to a click of the mouse the exact extraction process is masked and researchers may be unwittingly compromising their data. In this review, we describe the most common mistakes committed on the path from the image to the spreadsheet and their impact on data quality. Remedies are further proposed for most of the popular microarray platforms in use today.

Journal of Molecular Endocrinology (2004) 33, 1–9

Introduction

Array technology was in use as early as the 1980s (Augenlicht 1984, 1987) but did not come into prominence until the mid 1990s when cDNA microarrays emerged as an exciting new biomolecular tool capable of probing the entire transcriptome of the cell (DeRisi et al. 1996, Lockhart et al. 1996). Today, microarrays are widespread in genomic research and have a diverse range of applications in biology and medicine. A few recent applications include microbe identification (Wang et al. 2003), tumor classification (van’t Veer et al. 2002), evaluation of the host cell response to pathogens (van’t Wout et al. 2003) and analysis of the endocrine system (Jiang & Wang 2003, Cheng et al. 2004). Following commercialization of microarray technology, many researchers have abandoned the manufacturing of their own arrays. On the whole, the emphasis for the researcher has shifted away from manufacturing toward data analysis, which involves image acquisition and quantification. Image acquisition pertains to scanning the array and quantification refers to the conversion of images into numerical data, which are stored in a spreadsheet. This is where biologists start to get interested; however, we will backtrack a little and discuss the microarray platforms used to generate the raw data known as the image file.

Types of microarrays

The first widely used technology was the spotted cDNA microarray, which consists of numerous probes of PCR-amplified cDNA fragments deposited in a matrix pattern of spots on a treated glass surface, crosslinked and dried. The target for these probes is a solution of differentially fluorescently labeled cDNA derived from reverse-transcribed mRNA extracted from two cell populations (e.g. treatment and control). Hybridization of cDNA derived from two different sources on the same microarray is called competitive hybridization and such microarrays must be scanned at two wavelengths. Scanning initially produces two gray-scale (16-bit tiff) images corresponding to the different conditions, which can be merged and a pseudo-color scale applied for display purposes (Fig. 1a). Target DNA should not bind in the empty spaces surrounding the spots on
the microarray and in theory such areas can be used for background calculations since they should not fluoresce. Unfortunately this is an idealization and background ‘issue’ must be addressed, as discussed later. The main advantage of this technology is that it obviates the need for expensive sequencing of entire genomes since it allows the evaluation of gene expression in organisms for which little sequence data are available (Rathod et al. 2002). cDNA microarrays are mostly maintained in those laboratories with substantial expertise and experience or by those whose organism under study has not been developed into a commercial product. The next microarray technology to emerge involved in situ-synthesized oligonucleotide arrays using photolithographic technology (Affymetrix). Each gene target is probed by a number of distinct probes (11–20) collectively termed a probe set although some probes within a set are known to overlap in sequence. In theory, each probe consists of millions of single strands of DNA of exact length and sequence, confined to a small square area, that probe the same 25 bp segment of a target gene. In practice many of these DNA strands are not identical in sequence or length due to premature termination of synthesis during the manufacturing process. Since the probes are aligned edge to edge it is not possible to determine the level of background in the same manner as for spotted arrays. Biotinylated cRNA derived from a biological sample is hybridized onto the microarray, stained and scanned for fluorescence at a single wavelength (Fig. 1b). This is not a competitive hybridization method and in order to compare two samples two separate microarrays are required. Affymetrix microarrays have become the industry standard and every advance in the field is measured against this technology. Human arrays of various size are available as well as those for eight other organisms. Custom arrays are possible.

Spotted oligonucleotide arrays (Codelink, Amersham) have recently emerged in the field and are made with pre-synthesized 30 mer oligonucleotides piezoelectrically deposited on a gel-coated plastic film. This is also a single-dye technique so

Figure 1 Examples of microarray images: a) cDNA spotted microarray, b) high-density oligonucleotide array (Affymetrix), c) spotted oligonucleotide array (Codelink), and d) in situ-synthesized oligonucleotide array (Agilent).
sample preparation and scanning proceeds analogous to Affymetrix microarrays. The scanned image contains characteristic rings (Fig. 1c), and very little background. Amersham claims to have experimentally optimized the oligonucleotide probe length and sequence content for sensitivity and selectivity. A small percentage of spots on the array are labeled as defective prior to release as a result of the manufacturing process; the extent of this varies between batches. It is advantageous to know such limitations before hybridizations are made. This is particularly relevant if genes of interest to the research project are deleted. Arrays are available for human, rat and mouse, and custom arrays are offered.

Ink-jet technology (Agilent) is also in use for the in situ synthesis of oligonucleotide microarrays consisting of probes that are 60 bases long, which is more than twice the length of those used by Affymetrix. On the other hand, the Affymetrix platform uses several probes to sample a single gene, so there are different benefits to each technology. Agilent microarrays are similar to the original cDNA arrays because they use two-dye technology. It is clear from Fig. 1d that Agilent microarrays have improved overall spot uniformity but this is not critical beyond aesthetics. Ostensibly, the main advantage of these microarrays is improved selectivity due to increased probe length and sequence optimization. In practice, this advantage may be compromised by a relatively high background fluorescence. Human microarrays and five other model organisms are offered, as well as custom microarrays.

Which platform is the best for my experiment? A clear answer does not exist since a comprehensive comparative study has not been done. Cost is probably the most important factor when deciding on a platform. Another consideration is gene content of the microarray. It is important to ensure that the genes of the pathway or process of interest can all be probed by the microarray platform selected. In fact, if only a subset of the genome is of interest, then it may make more sense to build a custom array of those genes as opposed to pursuing a potentially more expensive whole genome approach.

It is worth noting that new microarray technologies are currently emerging. Fully customizable, photolithographically manufactured 60 mer arrays from NimbleGen have already made a following for themselves. Illumina markets bead oligonucleotide arrays comprised of thousands of microwells, each containing a single bead, each of which probes a target gene using >10^5 copies of a 50 mer oligonucleotide probe. Beads that carry the same probe are scattered randomly across the microarray, and each chip has a unique address file which encodes the identity of each bead. Finally, Applied Biosystems recently announced the release of promising new microarrays based on chemiluminescence rather than fluorescence. Regardless of the underlying technology, the raw data always comes in the same form – the scanned image.

**Data analysis**

Biologists should not have to worry about image acquisition and quantification in so far as these steps should be taken care of by a click of the mouse. This is certainly possible with current bioinformatics applications but the quality of the resulting data is often sacrificed for convenience. Take for instance the background issues mentioned earlier in the context of microarrays of the types shown in Fig. 1a, c and d. Most researchers believe that local background intensity should be subtracted from the signal. Commercial software typically estimates local background around a spot from image intensities in the annulus between the white and green circles in Fig. 2a. The problem is that the annulus has a small surface area and therefore represents a small sample of the fluctuating background. Estimating a mean from a small sample is prone to error especially when the fluctuations are large. This error is then propagated into data when the ‘background’ is subtracted from the ‘signal’. A systematic error may also result when the dye leaks from a spot, as in Fig. 1a. Errors in background estimation are the cause of the familiar splay of expression data in scatter plots at low intensity (e.g. Fig. 3d), and in extreme cases can lead to negative spot ‘intensities’, for which there is no good remedy.

A valid approach to background treatment for spotted microarrays is to first remove isolated extremal pixels – often an artifact of the scanner or photomultiplier – from the image by applying a pair of gray-scale thinning and thickening filters (Soille 1998). The image is then subtracted from its morphological opening which leads to the desired removal of background, whether uniform or
non-uniform, as exhibited by the transformation from Fig. 2a to Fig. 2b. Since the background has been removed and only scanner noise remains, the signal for the spot can be extracted from inside the white circle without further correction. This procedure is graphically illustrated in Fig. 2c, where the black line is the cross-section of intensity along the corresponding line in Fig. 2a, and where the blue line relates in the same way to Fig. 2b. This translates as a substantial reduction of data variance at small intensities for the scatter plots. All spotted microarray platforms benefit from this kind of image processing and Spot (Yang et al. 2001) is one example of commercial software that uses morphological opening.

Background intensity, as mentioned earlier, must be estimated differently with high-density oligonucleotide arrays (Fig. 1b) because, in contrast to spotted microarrays, no space exists between the probes (Naef et al. 2001). The design of these arrays incorporates mismatch (MM) probes that differ from perfect match (PM) probes at one base. Intensity from these MM probes is considered by the manufacturer to be background suitable for subtraction from PM probes. The expression level of a gene is then estimated as the mean of these differences across all the probes in the probe set of that gene. Averaging over the probes implicitly and incorrectly assumes that all probes have equal sensitivity to the target. Furthermore, as shown by the definitive analysis of Naef and coworkers (2003), the MM probes in many cases act as surrogates capable of binding targets of the PM probes leading to signal removal from the experiment. Despite these facts, the widely used commercial software – Affymetrix Microarray Suite 5.0 (AMS) – still treats the data in this way. Fortunately, the collection of probes into probe sets permits the application of appropriate statistical methods as indicated by Li and Wong (2001) and further improved by Šášik et al. (2002) and Irizarry et al. (2003).

It is worthwhile briefly explaining the appropriate method of plotting expression data, in this case derived from Affymetrix oligonucleotide arrays. The data for Fig. 3a, b and c were produced using an academic software package Corgon (Šášik et al. 2002, available for free download at http://corgon.ucsd.edu/~sasik). The raw data for these plots were obtained from a Latin-square study using the U95A GeneChip made available by Affymetrix at http://www.affymetrix.com/support/datasets.affx. This experiment contains 59 microarrays hybridized with the same RNA solution except for 14 transcripts spiked-in at various concentrations according to the Latin-square design. For our purposes, two microarrays (X and I) were chosen from this experiment for plots in Fig. 3. Microarray I was hybridized with twice as much RNA than X for 11 transcripts, which are not all visible in Fig. 3a because of spot overlap but can be clearly distinguished in Fig. 3b and c. Microarray X had a high concentration of one particular transcript.
while \( Y \) had none; this transcript is visible as the red dot near the \( X \)-axis of Fig. 3a and b but is out of range in Fig. 3c. Finally, of the original 14 transcripts, there were two at concentrations too low to be detected (only genes whose value for presence was less than 0·1 in \( X \) or \( Y \) were plotted).

Figure 3a depicts the data plotted on a linear scale with the identity line in green, and blue lines denoting \( \pm 2 \)-fold change. This is a valid representation of the data but is difficult to interpret because the scale makes it impossible to resolve data points near the origin, as illustrated by Figure 3a, three views of the same data: a) linear, b) logarithmic, c) A-M plot. \( X \) and \( Y \) are expression levels measured by identically hybridized Affymetrix microarrays, except for 14 spiked-in transcripts. Twelve of these were detected and labeled red. The green line is the identity; the solid blue lines are lines of \( \pm 2 \)-fold change, the dashed blue lines are lines of \( \pm 1·25 \)-fold change. Processing was done using the academic package Corgon. d) The same raw data processed with Affymetrix Microarray Suite 5·0.
the spiked-in transcripts. The same data are plotted on a logarithmic scale in Fig. 3b where the ± 2-fold lines become parallel and the main features of the data become visible. However, because of geometric constraints, there is still a false sense of concordance of the data with the identity line. The optimal plot of this data is presented in Fig. 3c and shows that clockwise rotation facilitates amplification away from the identity line for better visualization. This is called the A–M plot as introduced by Yang et al. (2002). Closer inspection of Fig. 3c reveals that all fold changes appear attenuated since all the red dots do not fall along the blue 2-fold change line as they should. There is also a trend for apparent fold change to become progressively smaller with increasing expression level, as evident by the tailing off of red dots with progression along the A-axis. These observations were also made by Naef et al. (2003) and probably reflect the properties of this technology that lead to gradual saturation of signal at high expression levels. It should also be noted that by using a significance cut-off of ± 1·25-fold most of the spikes are detected with very few false positives (black dots). A fold change of ± 1·25 probably reflects the practical detection level of the Affymetrix technology when there are no replicates. Smaller fold changes could of course be detected by increasing the number of experimental replicates. The biological significance of such small fold changes remains to be seen.

Figure 3d shows the same raw data analyzed with AMS 5·0 software. Only nine out of 14 spike-ins were detected as ‘present’ with default settings in the software but even more disturbing is the fact that there is no cut-off significance level that would isolate the genes that changed significantly from genes that did not change at all. By inspection of the black dots alone it appears that perhaps a 2-fold cut-off (blue lines) in Fig. 3d is informative and it is understandable why this value has perpetuated in the literature. However, it is clear that only one of the 2-fold spiked-in transcripts (red dots) would be detected by using this cut-off. It should be stressed that all commercial platforms, given the proper processing, can detect biologically significant fold changes that are lower than 2-fold. For example, processing multiple replicates of Affymetrix microarray data with Corgon (Šásik et al. 2002) versus AMS 5·0 led to a reduction in the type-I (false-positive) error rate from 29·0% to 4·4% using a fold change cut-off of 1·25. Finally, replicate hybridizations will of course greatly alleviate problems in detecting genes that are significantly expressed, which begs the following question:

**How many replicates do I need?**

This question is closely tied to experimental design and the nature of the experiment. Often the number of replicates performed is dictated by cost rather than by statistical demands. However, reasonable guidelines should be followed in order to capitalize on available resources. Biological variability is usually greater than technical variability; therefore the hybridization should be repeated with independently harvested RNA rather than repeat hybridizations with the same RNA. Replicating experiments is good practice even if the experiments are carried out with well-characterized standard cell lines because very small variations in cell culture can alter phenotype. For example, if some of the cells in culture enter into cell death due to excessive splitting of the cell line then the gene expression monitored may not reflect the process that the researcher is attempting to evaluate. A single replicate of a microarray experiment is justified as a ‘first pass’ when attempting to generate initial hypotheses for further testing. For more powerful biological inferences, at least three replicates per condition are needed for sensible statistical analysis. Although we think that microarrays alone are quantitative enough to carry the burden of proof, the scientific community demands that to ensure biological validity different methods of quantifying gene expression (quantitative RT-PCR, Northern blotting) should be employed to corroborate the induction of key factors and to confirm previously generated hypotheses.

Pooling of RNA samples is possible in microarray experiments under judiciously assessed conditions in order to lessen unwanted biological variability between RNA donors while keeping the microarray expenses low. This goes without saying for experiments with small organisms like Caenorhabditis elegans or Dictyostelium discoideum, where thousands of individuals are harvested at one time and pooling is par for the course. However, it is critical when measuring the gene expression of any time-dependent phenomenon that synchronization of cells or organisms occurs before pooling is done.
For example, pooling is acceptable in mouse developmental studies when sampling is required at different stages of development. In this case mice are synchronized by the developmental stage such that RNA can be pooled from several different animals at the same stage for a single microarray hybridization. Pooling should be avoided when donors can not be synchronized. In this case, it is better to perform several separate experiments to obtain independent time courses and then look for common salient features.

In another context, when using a two-dye platform to analyze a time course of treatment versus control, pooling is sometimes applicable to the reference mixture used as the control. Mixing the reference sample from all the treatments over the time course ensures that all the transcripts are present in the control and division by zero is avoided when ratios of fluorescent intensities are taken.

In laboratories working extensively with the same kind of tissue and the same kind of microarray it might make sense to measure the biological variability beforehand by repeated hybridizations, and calculate the variability of expression for every gene. The actual experiment can then be performed only once enabling a well-defined value to be assigned to every change of gene expression in the experiment. An example of this is in Fig. 4 where we plot the standard deviations of \( \log_2 \) expression levels of all genes detected in the Latin-square experiment. We used a subset of 57 samples from this set of 59, as two appeared to be of inferior quality. The distribution of expression levels for virtually all genes was nearly Gaussian (data not shown) allowing the use of the empirical s.d. to calculate significance of any measured change. However, Fig. 4 shows that there is no simple dependence of the s.d. on mean expression level. This might be surprising at first, but we need to realize that although each gene is probed by a fixed set of probes, some of them perform poorly, so that a highly expressed gene may only be effectively probed by a few probes whereas a poorly expressed gene may be probed by many. The former gene will have more variance than the latter.

In some cases, such as cancer studies, identification of differentially expressed genes between two classes is required. A sorted list of t-statistics will indicate highly expressed genes but an estimate of the false-positive error rate, or more preferably, the false discovery rate is needed in order to derive a cut-off such that significantly differentially expressed genes can be identified. The false-positive error rate is assessed using \( p \)-values, the probability that a truly insignificant gene will look as, or more, significant than the observed gene. The false discovery rate is measured using the \( q \)-value, which is the probability that the observed gene is truly insignificant. Tools are readily available for calculation of \( q \)-values from expression data including SAM (Tusher et al. 2001) and QVALUE (Storey & Tibshirani 2003). There is some evidence that the false discovery theory, while keeping strong control over the type-I (false-positive) error rate, maintains only weak control over type-II errors (\( \sim 50\% \) false-negative error rate). A machine learning algorithm Focus (Cole et al. 2003) can be used if this is of concern. It identifies significantly expressed genes without the reliance on conventional statistics and is also useful when the number of replicates is small.

### Normalization

The process of applying corrections to the data to account for any bias between microarrays...
generated by systematic non-biological causes is referred to as normalization. Non-biological differences with respect to two-dye experiments pertain to variable incorporation of dyes into cDNA and their different fluorescent properties; however, sources of bias in terms of single-dye techniques are less well defined. Perhaps subtle differences in manufacturing or quality control between individual microarrays exist and contribute to biases. One successful technique to remove such bias but to leave the fine structure of the data intact is to use the loess smoother (Cleveland 1979, Dudoit et al. 2000), which captures the gross trend in the data for subsequent subtraction. This technique should always be employed with two-dye arrays and should be used whenever generating expression scatterplots except in cases where gross global shifts in expression are possible (i.e. cell death).

The current perception is that the loess smoother is only practical for normalizing two-dimensional data. For example, the loess is easily applied to microarray experiments using two-dye platforms because each array in a time course need only be normalized intrinsically before comparison with others in the experiment. However, for single-dye platforms the loess must be applied across all \( n > 2 \) arrays in the experiment at once. The presumed difficulty arises because of \( n(n-1)/2 \) possible pairings and hence the same number of loess regressions to perform, in addition to multiple iterations since one pass does not harmonize all of the data. The following outlines how loess can be used to normalize all arrays simultaneously in a single pass. The procedure is simple, but the explanation requires a little math: let us denote \( \mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{in}) \) as the expression vector of gene \( i \). Assuming that most genes do not change appreciably throughout the experiment, the \( (n\)-dimensional) scatterplot defined by points \( \mathbf{x} \) will be a cloud surrounding the diagonal whose direction is defined by the unit vector \( \mathbf{u} \equiv 1/\sqrt{n} (1, \ldots, 1) \). Motivated by this, we decompose the expression vector as \( \mathbf{x}_i \equiv A_i \mathbf{u} + \mathbf{m}_i \), where \( A_i \) is the length of the orthogonal projection of \( \mathbf{x}_i \) onto the diagonal, and \( \mathbf{m}_i \) is the residual deviation from the diagonal. The residuals are still \( n\)-dimensional vectors, but what matters is that \( A_i \) is a real number (scalar) and the loess procedure can proceed as defined by Cleveland (1979). The \( A \) coordinate is the generalization of the abscissa, and \( m \) are the generalized ordinates in the \( A-M \) plot for \( n \geq 2 \).

Conclusion

The number of microarray-based publications is rapidly growing and consequently the volume of accumulated microarray data. How much of that data are good quality is debatable but researchers should first make sure that their own data are processed with the best possible tools. This means that all stages of data acquisition must be optimized, including image acquisition and quantification, data quality analysis and normalization. We have shown that it is not always sensible to do what is expedient. Free academic software and comprehensive handbooks are available and we would urge the users to explore such possibilities (Lorkowski & Cullen 2003, Parmigiani et al. 2003). By way of final advice, do not entrust your data to an application that requires a few clicks of the mouse and presents a simple spreadsheet of results without prior scrutiny of the raw images and scatter plots. Perhaps there is a subtle bias in one of the microarrays and all highly modulated genes are false positives? It is good practice to have a bioinformatician or a biostatistician review your data especially since microarray results often represent a starting point for further studies.

Acknowledgements

We acknowledge the support of all our collaborators and the members of the Genomics Core Laboratory.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases (AI46237 and AI47703), a National Institute of Health AIDS Training Grant (AI07384), the Center for AIDS Research Genomics Core Laboratory (AI36214), the Universitywide AIDS Research Program (IS99-S.D.213), and the San Diego Veterans Medical Research Foundation. J C holds a Canada Research Chair in Medical Genomics and is funded by the Canadian Institute for Health Research. There is no conflict of interest that would prejudice the impartiality of this research.
References


Received in final form 16 April 2004
Accepted 6 May 2004