

Issues for Consideration in the Analysis of Microarray Data in Behavioural Studies

GORDON A. BARR^{1,2,3} & PUHONG GAO¹

¹Department of Developmental Psychobiology, New York State Psychiatry Institute, ²Department of Psychiatry, Columbia University Medical Center, and ³Department of Psychology, Hunter College and the Graduate School, City University of New York, USA

Abstract

Microarrays are one of several technologies that allow for measurement the expression of thousands of genes simultaneously. This technological advance provides a challenge for the analysis of these data. In this review we discuss these analytical issues from the initial quality control to normalization, differential expression, clustering and finally functional pathway analysis. We focus on Affymetrix data but many of the issues are the same for other array platforms.

Introduction

In the past decade, methods have been developed to assess simultaneously the expression of very large numbers of genes. Array technology (including both cDNA and oligonucleotide arrays) is one such method. Arrays can assess the expression of multiple genes displayed as multiple probes on a fixed matrix following a reaction in which labelled DNAs hybridize to the array. Alternate methods to measure RNA levels, such as serial analysis of gene expression (SAGE) or differential display may provide more comprehensive gene coverage, but are technically demanding for studies including large number of samples. Thus, there is an explosion of microarray studies. Now large-scale gene and protein analysis can augment studies that assay single gene/protein one or two at a time. The amount of data that can be generated in 48 hours exceeds that being gathered over decades by serial gene approaches. With the aid of sophisticated computer-based analysis algorithms, it is possible to analyse those hundreds of thousands of distinct probes in a single ‘chip’. Thus, microarrays represent a powerful tool for the study of gene expression on a genomic scale.

Although the number and type of genes on the chip are predetermined, there are no a priori assumptions necessary to define which genes are differentially regulated by the experimental conditions. Microarray studies do not require a hypothesis about which genes change. This is both an advantage and a challenge, and the ability to assess expression of tens of thousands of genes simultaneously has helped usher in a new era in science. In contrast to strict hypothesis testing approaches, the search for changes that are not specified in advance has been termed ‘Discovery Science’ by Huda Akil (Society for Neuroscience Newsletter, 2003). Established

criteria are used to indicate changes in expression. This is the ‘discovery’ aspect of discovery science. Combined with careful experimental design and follow-up studies that carry out test hypotheses, we now have the capacity to ‘discover’ novel changes in the cell under a variety of conditions. These results can provide a temporal snapshot of cellular function that was not possible 10 years ago. This approach is not without problems, and these are still being discussed and resolved. Particularly difficult are the analytical issues following a successful microarray experiment. In a sense, the technical difficulties have moved from the bench to the computer. Fortunately, there are a number of open source analysis tools that can be applied (Dudoit et al., 2003)

Microarray methods were applied initially to disease states such as tumour classification but increasingly are applied to behavioural and pharmacological problems such as addiction and mental illness. Therefore, it is important that biologists and psychologists who study behaviour understand the technology and the types of issues involved in the analysis of data from those technologies. What follows is a brief description of some of the issues that are important in the analysis of these types of data.

Types of arrays

There exist alternative microarray platforms. Two major types are cDNA and oligonucleotide designs. Although some arrays use radioactive signals, fluorescence detection is far more common. Both cDNA and oligonucleotide platforms have advantages and disadvantages. A list of these can be found at the NINDS/NIMH Microarray Consortium website (<http://arrayconsortium.cnmcresearch.org/NINDS/jsp/arrayPlat->

Correspondence to: Gordon A. Barr PhD, Department of Developmental Psychobiology, Unit 40, NYSPI, 1051 Riverside Drive, New York, New York 10032, USA. Tel: 212 543 5694; E-mail: gab5@columbia.edu <<http://maxweber.hunter.cuny.edu/~gbarr/index.htm>>

forms.jsp). Briefly, cDNA probes are sensitive because they are typically much longer than oligonucleotide probes. That sensitivity comes with a loss of specificity due to cross-hybridization. Because of variability of the cDNA clone, experimental groups must be compared to controls for each spotted gene and thus two dyes are needed. The result is a ratio of expression for each spot. Because the dyes used produce different signals, either common reference designs (a single control for all experimental groups) or dye swap experiments are needed. These issues have been described clearly and in depth by Churchill et al. (2002). There are also issues of bias due to spot location and reliability of printing of spots. cDNA probes are the method of choice for individual laboratories to create, but are also commercially available. That allows for significant flexibility in customizing the chips for unique purposes. However, because the type of chip may differ from laboratory to laboratory and vendor to vendor, inter-laboratory and inter-experiment comparisons are difficult.

Oligonucleotide arrays are sold by several commercial vendors. Three major sources are Affymetrix, Agilent and Amersham. The companies differ in their strategy for spotting the chips. Those details are beyond the scope of this review. Each produce arrays that are highly specific and the Agilent arrays are quite sensitive due to increased binding regions of their oligonucleotides (60 mers). The Agilent platform is highly customizable and the genes can be user defined. Both Agilent and Affymetrix platforms are highly reproducible, which allows comparison of results across groups or time. There is a large store of informatics for the Affymetrix chips and a very large community of researchers who use them and share analytical tools. Some of the advantages of the single channel platforms are: (1) increasing numbers of investigators use these chips, indicating that the platforms will remain viable and that methods of analysis will continue to be addressed; (2) because of the large number of users, there is a large and growing set of analytical tools available, especially for Affymetrix chips, much of it freeware; and (3) all three provide rat gene arrays. For behavioural research in general, and addiction research in particular, there is a wealth of behavioural data using the rat, and use of rat arrays does not require replicating that data in a different species (e.g. mouse). This may change as more rat arrays become available and/or as more behavioural studies use the mouse.

Experimental design

As in all experiments, design is critical. It may be more important in microarray analysis because the number of replicates is small compared to the amount of data generated. The most sophisticated statistical methods will not 'fix' a poorly designed experiment. Thus attention to proper experimental design *before* collecting tissue is essential. One issue is the number of replicates required. On microarray listserv discussion groups there are at times questions about why, with two biological replicates, there are no statistically significant differences in gene expression between groups and what statistics can be used to fix that problem. The problem is not the analysis but the experiment. The number of required replicates, of course, depends on the size of the effect and the variability of the data. It is a classic problem of statistical

power (Cohen, 1988) but one that cannot readily be answered for microarray studies. There have been some estimates that suggest six to eight biological replicates are the minimum needed (Pavlidis et al., 2003). Pooling tissue for multiple animals may help with power and some studies have detailed the conditions under which pooling is appropriate (Kendzioriski et al., 2003; Peng et al., 2003). This has been discussed in depth for two channel systems where design issues are more complicated (Churchill, 2002). We prefer experimental designs that reverse predictions, even when we do not know what those predictions might be. Thus, a design might assess the effects of chronic morphine on gene expression followed by a second experiment with a drug treatment that reverses the behavioural phenotype of tolerance or dependence (e.g. NMDA blockers, NOS inhibitors). Thus it would be predicted that genes involved in the phenotype would move in the same direction as the behavioural outcome.

Many of the general analytical issues detailed below are applicable to all platforms, although details will certainly differ. Further, analytical tools for analysis of microarray data are rapidly maturing. Our experience is solely with Affymetrix chips and thus we limit our discussion to that experience. Affymetrix uses 25-mer oligonucleotide probes. To improve sensitivity, each gene is represented by 11–20 of these probes in pairs (probe set). One of the pairs is the 'perfect match' which is the sequence of interest; the other is the 'mismatch' which has a single middle base switched to control, in theory, for non-specific binding (see below). Expression values are the robust average of the signal from those 11–20 probe pairs.

Analytical tools for gene expression analysis

Overview

There are at least seven components to the data analytical process. Here we define the issues of concern; each will be discussed in detail subsequently.

First, quality control of the chip and the raw data is critical. Spotting chips is not trivial and there are varieties of technical difficulties that can compromise arrays. There are problems with background, misalignment and with contamination from dust, water spots, etc. If the data from the hybridization on the chip are valid, the second issue is how to correct for background (noise) and combine the probe level data into gene expression numbers such that there is a single expression value for each gene. Two sources of noise are optical and non-specific hybridization to oligonucleotide or cDNA probes, with an increasing recognition that the nucleotide sequence composition of the probes is an important variable to consider (Gautier et al., 2004; Naef et al., 2003). Thirdly, because there are a number of variables that change that result in chip-to-chip variability, including but not limited to efficiency of mRNA isolation and hybridization, chip variability, differences in laser intensity and so forth, chips must be normalized to each other (Zien et al., 2001; Irizarry et al., 2003). Normalization schemes fall into two general classes. First are 'globalization' methods that include, among others, use of housekeeping genes, or adjustment of each array by some measure of central tendency. More sophisticated models provide non-linear methods (e.g. dChip, loess, quantile normalization, variance stabilizing methods) to account for

the relationship between expression levels and chip bias. Fourth are the procedures for determining whether or not a particular gene has changed because of the experimental manipulation, and whether or not that change replicates. What complicates these analyses is that, unlike traditional experimental designs with many replicates and few variables, expression arrays have few replicates and thousands of variables (genes). Many methods are based on well-characterized Bayesian models, with recently developed methods for correcting for Type I errors [e.g. false discovery rate (FDR; Benjamini and Hochberg, 1995)]. Note, however, that changes in single genes are probably far less critical than are changes in functional classes. Thus, the fifth and more important issue is how to detect clusters of gene expression that provide evidence of higher levels of organization not discovered by individual gene expression analysis. There are a number of clustering methods, supervised and unsupervised, that are available. Sixthly, methods that relate gene changes to functional pathways and biological processes, including network analyses, are perhaps the newest and the most difficult of these analytical problems. A number are available (e.g. GenMapp; KEGG; Cytoscape; Class Scoring). Finally, there is the need to validate the data independently using other methods, including measurements of gene message [e.g. quantitative real-time polymerase chain reaction (PCR); Northern analysis; solution hybridization] and of gene products (e.g. Western analysis; immunocytochemistry). These methods provide both confirmation of message change, its expression as protein and anatomical resolution. There is the requirement for storing data for public access. MIAME (Minimum Information About Microarray Experiments) is the standard format for these data (Brazma et al., 2001) and exists in a number of public access databases [e.g. ArrayExpress; GeneTraffic; Gene Expression Omnibus (GEO)]. All discussions of these issues presupposes strong experimental design with sufficient biological replicates.

First, a few caveats are needed. Because the methodology is new there are yet no agreed-upon protocols for analysis of these types of data. Although standard statistics can be used, the specific criteria for determining expression, analysis type and clustering method are under discussion by a number of groups. The field is developing rapidly; analytical tools that are increasingly powerful become available almost monthly. Our biases are to avoid arbitrary methods of analysis that might work *post hoc* for a particular data set but not for other data sets and thus are not reproducible, and to use methods well grounded in statistical theory.

Quality control

The quality of the chip and the raw data are critical. Spotting and reading chips is not trivial and there are varieties of technical problems that can compromise arrays. All analysed chips must meet specific criteria for background, specificity and so forth. To minimize problems with background, misalignment and contamination from dust, scratches, water spots, etc. we inspect each chip (the DAT image) to insure proper alignment and manually mask any such contamination using software provided by Affymetrix (MAS 5.0). Note that these are small imperfections and do not render the chip unusable; rather the affected probes are

masked and the expression value calculated from the remaining probes for that gene. This has become even less of a problem because Affymetrix now distributes probes for the same gene over the entire chip rather than grouping them together. Thus, any imperfection is likely not to affect expression of a particular gene, but rather just a few probes from many genes. In our experience about 50% of chips require some minor masking and to date in hundreds arrays that we have analysed only one has been misaligned. Plots of residuals is another method that can help identify outliers. Li and Hung Wong (2001) provide a function in their dChip software to assess the number of outliers for both probes and arrays that can be useful for examining anomalous data.

Affymetrix provides a tool to filter out genes whose expression is low ('absent/marginal/present call'). There are a number of other methods used alone or in combination with these calls to reduce the number of genes for analysis. There are some questions as to the statistical validity of these methods. For example, the Affymetrix decision process includes the mismatch probe; but as the mismatch signal is not random, the presence/absence call is unreliable. Absence of expression in some samples and presence in others may be important and excluding those data may serve to remove valuable information. Even if filtering methods were based on sound statistical bases, early filtering such as this has the danger of tossing genes whose expression may be low, but whose function may be important. This is especially a risk when changes in genes are viewed as part of changes in functional paths where a number of small changes result in important functional differences, as is typically true in brain, where changes are likely to be small. Thus we recommend extreme caution in filtering at all, and especially in early stages of analyses.

Probe level analysis

Background correction. There are two sources of noise in the signal that represents amounts of RNA expression. The first is optical noise. This is linear over a range of RNA concentrations. Thus, it is accounted for easily (Gautier et al., 2004). More problematic is non-specific hybridization of RNA to the probes. This problem exists for all designs of microarrays. The design of Affymetrix chips includes the MM probes, which were included to correct for non-specific binding. Unfortunately, signals from the MM probes are systematic and linear with increases in true signal; the mismatch data (MM) shows hybridization that is both systematic and of unknown source (see, for example, Fig. 1 from Naef et al., 2003). The literature has demonstrated clearly that use of the perfect match data (PM) alone reduces noise. Thus, those data should not be used in analyses until that systematic hybridization is better understood. Indeed, most recent methods ignore the MM data (Huber et al., 2003; Naef et al., 2003; Gautier et al., 2004). There are a number of physical and stochastic models that have been developed to account for specific and non-specific binding, for example the greater binding affinities of G & C than A & T nucleotides. This is an area of analysis that is of increasing interest and there are likely to be improvements in these methods.

Normalization. If the data from the hybridization on the chip are valid and that the gene expression values are accurate, the question of how to normalize the data is critical. Normalization corrects for differences in assays from chip to chip (Bolstad et al., 2003). There are almost as many normalization procedures as there are recipes for chocolate chip cookies. These include global methods such as normalizing the median (or mean) expression levels of experimental arrays to those of control arrays, or the use of spiked controls or housekeeping genes. Some of these are developed by the individual investigator following some rules that seem to make sense for that data set. These therefore may not hold for all data sets, and limit comparability between experiments. Other methods include more complicated statistical regression methods that are developed to account for the non-linearity of the expression values. In general, there are two issues. First, all normalization methods make certain biological assumptions that are probably not true. The consequences of violation of these assumptions vary with the assumption but are more limiting for global methods of normalization (e.g. by central tendency) than for non-linear methods. Secondly, variation between chips and probes is typically non-linear and to account for this variability non-linear methods are needed [e.g. different for different levels of expression (Irizarry et al., 2003; Bolstad et al., 2003)]. There are a number of current and statistically viable means, including loess and lowess, dChip, variance stabilizing methods (VSN) and quantile normalization. dChip is a statistically robust procedure, but has the disadvantage of requiring the investigator to choose an array against which all other arrays are normalized. In our experience the results differ depending on which criteria the array is chosen. Quantile and VSN methods have slight advantages over other methods in accuracy of identifying these genes and have the advantage of normalizing all chips to each other without the requirement of picking a standard against which to normalize. The VSN method is more computationally intense but also less aggressive than quantile normalization. Quantile normalization attempts to equate distributions among chips. VSN methods equate variance across different intensities of expression and, therefore, genes with high intensities have an equal chance of being ranked as highly as do genes with lower signal. The VSN normalization concedes some small loss of sensitivity for gains in precision. We have tried both quantile and VSN methods, comparing them to each other and to PCR data; there are no major differences and the correlation between these two methods is high. These issues have been discussed in depth previously (Zien et al., 2001; Irizarry et al., 2003). VSN and quantile normalization methods, and many other analytical tools, are implemented as options in Bioconductor, an available R program package for the analysis of microarray data.

The different methods developed to determine expression levels combine different background correction and normalization schemes. There exists a ‘competition’ to evaluate these strategies using samples with known levels of signal. These levels are from the Latin Square spike in the data set from Affymetrix and the dilution series from GeneLogic. These valuable contributions by these two companies allow at least

preliminary efforts to compare probe level analysis results to ‘true’ values. Comparisons of contributed analysis methods in this competition are available at <http://affycomp.biostat.jhsph.edu> (Cope et al., 2004). As valuable as these comparisons are, they are limited because they do not tell us which method allows the best results for assessing differences in gene expression levels from different experimental groups, or for finding functional changes in gene classes. Currently, GCRMA is implemented in Bioconductor. It provides perhaps the best available method for correcting both optical and non-specific binding noise.

Once background noise is accounted for and data are normalized, probe values need to be combined into expression values for individual genes. This is typically conducted by some robust averaging method (e.g. median polish). Data are typically transformed to log₂ values to reduce the magnitude of outliers and to stabilize variances at high levels of expression for subsequent analysis.

In all cases, the data should be visualized. There are a number of graphic methods that can be used. One useful plot for any two arrays is the MA plot that graphs variance versus intensity [$M = \text{minus} (\log \text{signal A} - \log \text{signal B})$; $A = \text{add} (\log \text{signal A} + \log \text{signal B})$; (Dudoit et al., 2003)]. The resultant plot should show equal variability over the entire range of intensities. Outliers can be identified and can be genes of interest.

Which genes differ?

Inferential statistics. A usual question is whether or not expression of a particular gene has changed because of the experimental manipulation. There are increasing questions as to the value of changes in single genes rather than gene families or functional groups. None the less, there are a number of commonly used algorithms for determining statistical differences among groups. These include simple empirical determinations of fold-change (DeRisi et al., 1997) that resulted in part in the ubiquitous use of a range of fold-changes in expression levels deemed significant, and the proprietary statistical methods of Affymetrix. Two problems with the use of fold-change are that the magnitude of change is unstable with low values of the denominator and that fold-change measures do not include any measure of variability. What complicates all types of inferential analyses is that, unlike traditional experimental designs with many replicates and few variables, expression arrays have few replicates and thousands of variables (genes). This creates problems in limiting Type I error and controlling familywise error rates (FWER), the error rate for a family of tests (Dudoit et al., 2003). Further, the distribution of microarray data is not normal, or distributed in t - or F -distributions. To account for that, several alternative methods have been developed to determine the distribution of the data empirically.

One approach is to permute the expression levels and compare the actual data distribution to the permuted distribution. A probability can then be associated with those actual data. The most popular of the permutation-based methods is Statistical Analysis of Microarrays (SAM) as implemented in Bioconductor, an open source bioinformatics software package (<http://www.bioconductor.org/>), as an Excel plugin (<http://www-stat.stanford.edu/~tibs/SAM>) or through

the Institute for Genomic Research (TIGR) (<http://www.tigr.org/software/tm4/mev.html>) (Holm, 1979; Tusher et al., 2001). One issue for all permutation methods is the number of biological replicates. Few replicates produce few unique permutations and the resulting distribution is coarse and probably not robust. Given adequate replicates, false positive rates must be controlled while identifying genes that actually change. The strongest methods are based on well-characterized Bayesian models with appropriate control for the Type I errors associated with multiple testing. The classic methods such as Bonferroni and Holm (Holm, 1979), stepdown Hochberg (Hochberg, 1988) and step-up corrections for FWER are too conservative. More recently developed methods include the calculation of the false discovery rate (FDR) and the Westfall–Young methods. The latter, although providing the more stringent FWER control, is also conservative and Westfall himself recommends the FDR methods for gene discovery work (Westfall et al., 2001). The FDR is the expected proportion of incorrect rejections of the null hypothesis among all rejections. This is conceptually different from the Type I error rate. There are various forms of the FDR method. The methods developed by Benjamini and colleagues and by Storey provide strong control over the FDR under a variety of assumptions (Benjamini et al., 1995; Benjamini and Yekutieli, 2001). Please note that because the FDR and alpha level are conceptually distinct, the generally accepted $\alpha = 0.05$ level does not imply a similar requirement for the FDR rate. Thus we can be less strict (e.g. 10%) if we wish to identify more candidate genes. Storey has discussed this in length and has provided a comparable statistic, the q -value as a metric of the FDR rate. This is the estimated proportion of ‘significant’ genes that are false positives. He also provides methods to calculate q - from p -values (Storey and Tibshirani, 2003).

A second approach is to improve power by pooling error estimates. Although the number of replicates for each gene may be small, there are a large number of such experiments (genes) conducted under the same conditions. The information from all genes, considering them single experiments under the same condition, is used to estimate the population random error. Variance estimates for each gene are pooled to obtain a single standard deviation estimate that is then used in the statistical of choice (Z -test, ANOVA). If variances differ over different intensity levels, as is typical, this method is not valid. Alternatively, local pooled error methods combine variances for genes of similar expression levels. When the random error differs for different expression levels local pooled error methods are necessary (Jain, 1998; Nadon and Shoemaker, 2002). Both methods substantially increase power and are implemented in commercial (ArrayStat; S-Plus) and non-commercial software (<http://hesweb1.med.virginia.edu/bioinformatics>). Correction methods for multiple testing as described above can be applied as described above.

As stated before, it is important to plot the results. Volcano plots, which compare $-\log$ of significance value to \log of fold-change, can be helpful in selecting specific genes to be studied further (Jin et al., 2001). These plots, implemented in Bioconductor and other sources, can also be calculated and plotted without special software. They provide information about the variability and magnitude of significant changes in expression.

Group and functional changes

Clustering. Because changes of gene expression patterns are probably more important than changes in individual genes, data mining through clustering is an important tool. Cluster methods fall into two general categories, supervised and unsupervised. Supervised methods require some a priori decisions, often from a training set of data. There are many supervised methods including support vector machines, BRB Array tool, Genes@Work (Califano et al., 2000; Klein et al., 2001; Pavlidis et al., 2004). A good starting point for information is <http://genopole.toulouse.inra.fr/bioinfo/microarray/index.php?page=logiciels&domaine=Microarray,+Classification&lang=en>, but there are many other resources available.

Unsupervised classification algorithms attempt to find patterns, trends or clusters in a data set, without prior knowledge of those patterns. An early and commonly used method is hierarchical clustering, for example Cluster (Eisen et al., 1998), an iterative procedure that selects at each iteration the two most similar vectors and merges them. Similarity is defined as some distance measure (Euclidean, correlation, etc.). The output of the algorithm is a binary tree in which similar expression vectors are close together. A cluster structure is then imposed upon this tree a posteriori (e.g. TreeView), which displays and browses the hierarchical results. Hierarchical clustering strategies have been applied to a fairly diverse set of experimental conditions to cluster genes with correlated expression patterns (Eisen et al., 1998). The drawbacks to classical hierarchical clustering are well documented, because it is a ‘greedy’ algorithm that makes early low-level decisions about the cluster structure. The validity of decisions sometimes results in unstable or arbitrary clusters (Bittner et al., 1999; Herrera et al., 2001). Furthermore, it forces all genes into a cluster structure, even though some might not fit into those clusters.

Self-organizing maps (SOMs) are a second unsupervised learning algorithm. The application of SOMs to expression data was pioneered in the Gene Cluster software developed at MIT and the Whitehead Institute (Tamayo et al., 1999). The SOM learning algorithm is an iterative procedure in which the vectors in the data set gravitate stochastically towards nodes in a pre-defined network of categories. Thus, the SOM algorithm is ‘slightly’ supervised in that the number of categories and their relative topology are specified in advance. This is also a ‘greedy’ algorithm and the resultant clusters are highly dependent on the initial nodes chosen. This often results in clusters that are unstable and dependent on the number of clusters specified. SOMs have similarly been used successfully for gene clustering and tissue clustering (Tamayo et al., 1999).

We use ArrayMiner, a commercial product, which provides a non-hierarchical method for clustering. It applies a Gaussian function to the clusters allowing the clusters to take into account cluster variability instead of distance (Faulkenaur and Marchand, 2002). There are three advantages to this method. First, the method is able to dissociate clusters that are overlapping in space. Secondly, the cluster structure is very consistent across different numbers of clusters. Thus, the experimenter can choose different levels of detail by increasing the numbers of clusters without losing the essential cluster structure. Thirdly, this method allows for genes to be

excluded from clusters (called ‘unclassified’); as mentioned above, traditional hierarchical and SOM clustering algorithms force all genes into clusters. There is still the necessity to define the numbers of clusters in advance, providing the opportunity to ‘try out’ a large number of alternatives until the desired results are found.

Functional paths. Methods to relate gene changes to functional pathways and biological processes are perhaps the newest and the most difficult of these analytical problems. They are also the most important. There are multiple approaches to this task. We describe only three different ways here. First, each gene that is determined to be statistically significant can be defined by function through searches in various databases (e.g. GenBank, Unigene, KEGG). One then can cull the literature through Medline searches to obtain a more detailed functional analysis. This labour-intensive method obviously works for only very limited sets of genes. Secondly, there are software approaches. These include AMIGO, GenMapp and its associated GeneFinder program to develop functional maps for a set of results. These programs search GeneOntology, the organizational database for gene function. There were originally few rat ‘maps’, the lack of which presents problems for most models of addiction. Recent efforts have provided greater depth to these functions in the rat. There are alternative strategies. First, we can start to build maps. Secondly, there are homologous genes for the mouse, and the more complete mouse maps can be used. However, most probably the rat genomic information, which is expanding rapidly, will be adequate in a short time. This is important considering the wealth of behavioural and physiological data on the effects of drugs in this species.

Thirdly, there are several methods to produce statistically significant functional groups based on the probability that certain genes in specified classes occur by chance. These include EASE DAVID, MAPPfinder and the Class Scoring method, among others (Dahlquist et al., 2002; Pavlidis et al., 2002; Dennis et al., 2003; Hosack et al., 2003). We use the latter method to provide a level of statistical confidence to functional groupings based on Gene Ontology classifications. Thus a collection of genes in the same family that are altered in a coherent manner, but non-significantly on a gene by gene basis, can be identified as an altered functional group. This is probably far more important than identifying individual genes as significantly regulated.

Validation of chip data. There is the need to validate data independently using other methods, including measurements of gene message (e.g. solution hybridization; qRT-PCR; Northern analysis, *in situ* hybridization) and of gene products (e.g. Western analysis; immunocytochemistry). We have used RNA samples for qRT-PCR from the same tissue samples as used for the microarray experiments. We then calculated expression levels of the treated animals relative to the controls for both the microarray data and the qRT-PCR data and averaged the data for the four replicates. Comparison of the expression ratios for the two methods shows that (1) the direction and magnitude of expression are quite similar in the two methods; and (2) the magnitude of the expression ratio is higher for the qRT-PCR method than the array method. This

has been reported before and is at least in part a function of normalizing the microarray data, which tends to reduce fold-change. When we compared average scores between array data and qRT-PCR data for 20 genes, we found a good correlation between the two assay methods ($r=0.82$), accounting for about 67% of the variance. These data provide independent confirmation of the validity of the microarray data. Assessment of protein is equally important as there are many mechanisms by which changes in gene expression are not translated into comparable changes in protein levels. Methods such as *in situ* hybridization and immunohistochemistry can not only confirm changes in gene and protein expression, albeit with lower sensitivity, but also provide anatomical resolution. That information is important, given that most microarray experiments lack anatomical and cellular resolution.

Public access. There is the requirement for storing data for public access and many journals now require that availability before publication (Brazma et al., 2001; Stoeckert et al., 2002). The current standard is to make data available in MIAME (minimum information about microarray experiments) form (publicly available databases include GeneTraffic, ArrayExpress, or GEO). MIAME requires information about the experimental design, relationship among samples and arrays and hybridization details, among other details. The idea is to be able to replicate exactly the conditions of any microarray experiment.

Conclusions

Microarrays offer a powerful tool to assess gene expression and as whole genomes are put on chips, this tool provides a window into gene function for the whole organism. They truly represent ‘discovery science’. The method is not without problems, not the least of which are poor design of experiments and analysis of data. Further, the jury is still out on the best platform, and probably different platforms will provide complementary information. As typically conducted, microarray experiments use heterogeneous tissue for hybridization and labelling. In brain, that includes not only anatomical diverse structures, but different cell types as well. These can have distinct gene expression signatures (Zirlinger and Anderson, 2003). Other methods, such as laser capture microscopy, will be needed for these types of studies. As paraphrased from a quote by the Manager of Technical Services, New England Nuclear, microarrays will not make your life easier; quite the contrary. But they will make your science interesting.

References

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc, Series B: Methodol* 57:289–300.
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165–1188.
- Bittner M, Meltzer P, Trent J (1999) Data analysis and integration of steps and arrows. *Nat Genet* 22:213–225.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.

- Brazma A (2001) Minimum information about a microarray experiment (MAIME)—towards standards for microarray data. *Nat Genet* 29:365–371.
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 29:365–371.
- Califano A, Stolovitsky G, Tu Y (2000) Analysis of gene expression microarrays for phenotype classification. *Bioinformatics* 8:75–85.
- Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 32:490–495.
- Cohen J (1988) *Statistical power analysis for the behavioural sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.
- Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP, Gautier L, Cope L, Bolstad BM, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20:323–331.
- Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 31:19–20.
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao G, HCL, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 4:3.
- DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–686.
- Dudoit S, Shaffer JP, Boldrick JC (2003a) Multiple hypothesis testing in microarray experiments. *Stat Sci* 18:71–103.
- Dudoit S, Gentleman RC, Quackenbush J (2003b) Open source software for the analysis of microarray data. *Biotechniques Suppl*:45–51.
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868.
- Faulkenaur E, Marchand A (2002) Clustering microarray data with evolutionary algorithms. In: *Evolutionary computation in bioinformatics* (Corn DW, Fogel GB, eds), Morgan Kaufmann.
- Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–315.
- Herrera JF, Valencia A, Dopazo J (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 17:126–136.
- Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802.
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70.
- Hosack DA, Dennis Jr. G, Sherman BT, Lane HC, Lempicki RA (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol* 4:60.
- Huber W, von Heydebreck A, Sultmann H (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol* 2:1–22.
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
- Jain RK (1998) The next frontier of molecular medicine: delivery of therapeutics. *Nat Med* 4:655–657.
- Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat Genet* 29:389–395.
- Joachims T (1999) Making large-scale SVM learning practical. In: *Advances in kernel methods - support vector learning* (Schölkopf B, Burges C, Smola A, eds), pp. Boston: MIT Press.
- Kendzierski CM, Zhang Y, Lan H, Attie AD (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics* 4:465–477.
- Klein U, Tu Y, Stolovitsky G, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, Nen A, Califano A, Dalla-Favera R (2001) Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med* 194:1625–1638.
- Li C, Hung Wong W (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2:RESEARCH0032. Epub 2001 3 August.
- Nadon R, Shoemaker J (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet* 18:265–271.
- Naef F, Magnasco MO, Hekstra D, Taussig AR, Magnasco M, Succi ND, Lim DA, Patil N (2003a) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys* 68:011906. Epub 2003 Jul 16.
- Naef F, Succi ND, Magnasco M, Lim DA, Patil N (2003b) A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics* 19:178–184.
- Pavlidis P, Lewis DP, Noble WS (2002) Exploring gene expression data with class scores. *Proc Pacific Symp Biocomputing* 474–485.
- Pavlidis P, Li Q, Noble WS (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics* 19:1620–1627.
- Pavlidis P, Wapinski I, Noble WS (2004) Support vector machine classification on the web. *Bioinformatics* 20:586–587.
- Peng X, Wood CL, Blalock EM, Chen KC, Landfield PW, Stromberg AJ (2003) Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* 4:26.
- Saeed A, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J (2002) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34:374–378.
- Stoeckert Jr CJ, Causton HC, Ball CA (2002) Microarray databases standards and ontologies. *Nature Genetics* 32:469–473.
- Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci USA* 100:9440–9445.
- Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96:2907–2912.
- Tusher VG, Tibshirani R, Chu C (2001) Significance analysis of microarrays applied to ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
- Westfall P, Zaykin DV, Young SS (2001) Multiple tests for genetic effects in association studies. In: *Biostatistical methods* (Looney SW, ed.), pp. 143–168. Totowa, NJ: Humana Press Inc.
- Zien A, Aigner T, Zimmer R, Lengauer T (2001) Centralization: a new method for the normalization of gene expression data. *Bioinformatics* 1:1–9.
- Zirlinger M, Anderson D (2003) Molecular dissection of the amygdala and its relevance to autism. *Genes Brain Behav* 2:282–294.