Fundamentals of cDNA microarray data analysis

Yuk Fai Leung and Duccio Cavalieri

Bauer Center For Genomics Research, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA

Microarray technology is a powerful approach for genomics research. The multi-step, data-intensive nature of this technology has created an unprecedented informatics and analytical challenge. It is important to understand the crucial steps that can affect the outcome of the analysis. In this review, we provide an overview of the contemporary trend on various main analysis steps in the microarray data analysis process, which includes experimental design, data standardization, image acquisition and analysis, normalization, statistical significance inference, exploratory data analysis, class prediction and pathway analysis, as well as various considerations relevant to their implementation.

The development of microarray technology has been phenomenal in the past few years. It has become a standard tool in many genomics research laboratories. The reason for this popularity is that microarrays have revolutionized the approach to biological research. Instead of working on a gene-by-gene basis, scientists can now study tens of thousands of genes at once. Unfortunately, they are often daunted and confused by the complexity of data analyses. Although it is advisable to collaborate with statisticians and mathematicians on performing a proper data analysis, it is crucial to understand the fundamentals of data analysis. In this review, we explain these fundamentals step-by-step (Figure 1; Table 1). Instead of discussing any particular analysis software, we focus primarily on the rationale behind the analysis processes and the key factors that affect the quality of the result. For a compilation of current microarray analysis software see a recent article [1] and author's website (http://ihome.cuhk.edu.hk/~b400559/arraysoft.html; permanent link: http://genomicshome.com). We also focus on the use of the two-dye cDNA microarray data analysis, although most of our discussions are also applicable to the single-dye oligonucleotide platform (i.e. Affymetrix) (Box 1). We hope that by appreciating the fundamentals novices will become successful at microarray data analysis.

Experimental design and implementation

> 'If the experimental design is wisely chosen, a great deal of information is readily extractable, and no elaborate analysis might be necessary. In fact, in many happy situations all the important conclusions are evident from visual examination of the data'. [2]

'Well begun is half done', is an aphorism that is especially true of for microarray experiments. Good design is very important at the beginning of a microarray experiment. A typical microarray usually consists of tens of thousands of elements. On the one hand, it provides a comprehensive coverage that almost always promises some new discoveries. On the other hand, analyzing the vast amount of data being generated can be daunting to scientists. It is therefore, more important now than ever, to design a microarray project carefully to generate high-quality data and to maximize the efficiency of data analysis.

Good microarray experimental design should comprise at least four elements: (i) a clearly defined biological question and/or hypothesis; (ii) treatment, perturbation and observation of the biological materials, as well as the microarray experimental protocols, should be as little affected by systematic and experimental errors as possible; (iii) a simple, sensible and statistically sound microarray experimental arrangement that will give the maximal amount of information given the cost structure and complexity of the study [3–5]; and (iv) compliance with the standard of microarray information collection, which will be further discussed in the next section.

Standardization of information generated by microarray experimentation

The adoption of international standards have long been seen as vital in science because of the confusion generated through the use of various units. We have been experiencing a similar issue in the microarray field. The same increase or decrease in gene expression observed by two different laboratories might actually be different, especially when they are using different experimental protocols and data-analysis methods. Without a standard, it is almost impossible to judge the validity of a result just by inspecting the expression changes or even the raw data [6]. In view of this problem, the Microarray Gene Expression Data (MGED) Society (http://www.mged.org), an international initiative to develop standards for microarray data, has recently proposed a standard Minimum Information About a Microarray Experiment (MIAME) (http://www.mged.org/Workgroups/MIAME/miame.html) [7]. The research community has embraced it and many major journals now require compliance with MIAME for any new submission [8]. It is therefore advisable to ensure that the experimental design, implementation and data analysis comply with the MIAME standard.

Corresponding author: Yuk Fai Leung (yfleung@cgr.harvard.edu).

http://tigs.trends.com 0168-9525/$ - see front matter doi:10.1016/j.tig.2003.09.015

Glossary

Adaptive circle segmentation: a segmentation process in which the diameter of the circle being applied to the spot is calculated case by case in order to address the variation of spot diameter. The pixels that fall within the circle are regarded as foreground.
Background estimation: the background fluorescence signal usually originates from non-specific hybridization of the labeled samples or auto-fluorescence of the glass slide. This unwanted background signal needs to be estimated and removed from foreground signal during image analysis.
Background intensity subtraction: the calculation of fluorescence signal from the background pixels of a spot identified during the segmentation process. Usually the median of the pixel intensities is used.
Dye-swapping experiment: two hybridizations of the sample pair of samples in which the labeling dye of the two samples is reversed in one of hybridizations. Averaging the two expression ratios would give one a good estimate of the true ratio.
Fixed circle segmentation: a segmentation process in which a circle with a constant diameter is applied to all spots on the image. The pixels that fall within the circle are regarded as foreground.
Intensity extraction: the process that calculates the foreground (signal) and background intensities from the pixels after the segmentation process.
Local background estimation: a commonly used background estimation method in which the immediate background pixels surrounding the spot, as identified by the segmentation process, are used for estimating the background signal.
Segmentation: a computational process which differentiates the pixels within a spot-containing region into foreground (true signal) and background.
Spot intensity extraction: the calculation of fluorescence signal from the foreground pixels of a spot identified during the segmentation process. Usually the mean of the pixel intensities is used.
Spot recognition or gridding: a computational process which locates each spot on the microarray image.

MIAME represents the minimal information to be recorded that enables faithful experimental replication, the verification of the validity of the reported result, and the facilitation of the comparison among similar experiments. Besides, the information should be structured with controlled vocabularies and ontology to assist in developing database and automated data analysis. Currently, the minimal information includes the six parts: (i) experimental design; (ii) array design; (iii) samples; (iv) hybridizations; (v) measurements; and (vi) normalization controls. A detailed description of each part and a convenient checklist are available on the MIAME website (http://www.mged.org/Workgroups/MIAME/miame_checklist.html).

Image acquisition and analysis

After performing all biological and hybridization experiments, the first step of data analysis is scanning the slide and extracting the raw intensity data from the images. There are four basic steps in image acquisition and analysis: (i) scanning; (ii) SPOT RECOGNITION OR GRIDDING (see Glossary); (iii) SEGMENTATION; and (iv) INTENSITY EXTRACTION and ratio calculation.

Image acquisition is a very important step in data analysis. Once an image has been scanned, all data, high or poor-quality, are essentially fixed. A poor-quality image requires further manipulations, which will lead to a decrease in the power of analysis. There are two prerequisites for obtaining a high-quality image. First, all steps in array construction, RNA extraction, labeling, and array hybridization have to be performed to the highest possible standards. These endeavors ensure that all images would be least affected by contamination (e.g. dust or dirt), and have consistent spots with high signal-to-noise ratios. Second, the choice of scanning parameters is also important. We discuss the settings for the Axon scanner, but the general principle is applicable to other platforms. A low laser power (30%) should be used whenever possible to prevent photo-bleaching. The photomultiplier tube (PMT) gain settings are adjusted during the scanning process to balance the overall intensities between the two channels (i.e. cy3 and cy5) as much as possible. This balance can be evaluated in several ways: (i) visual inspection of the scanning image. The non-differentially expressed spots should appear

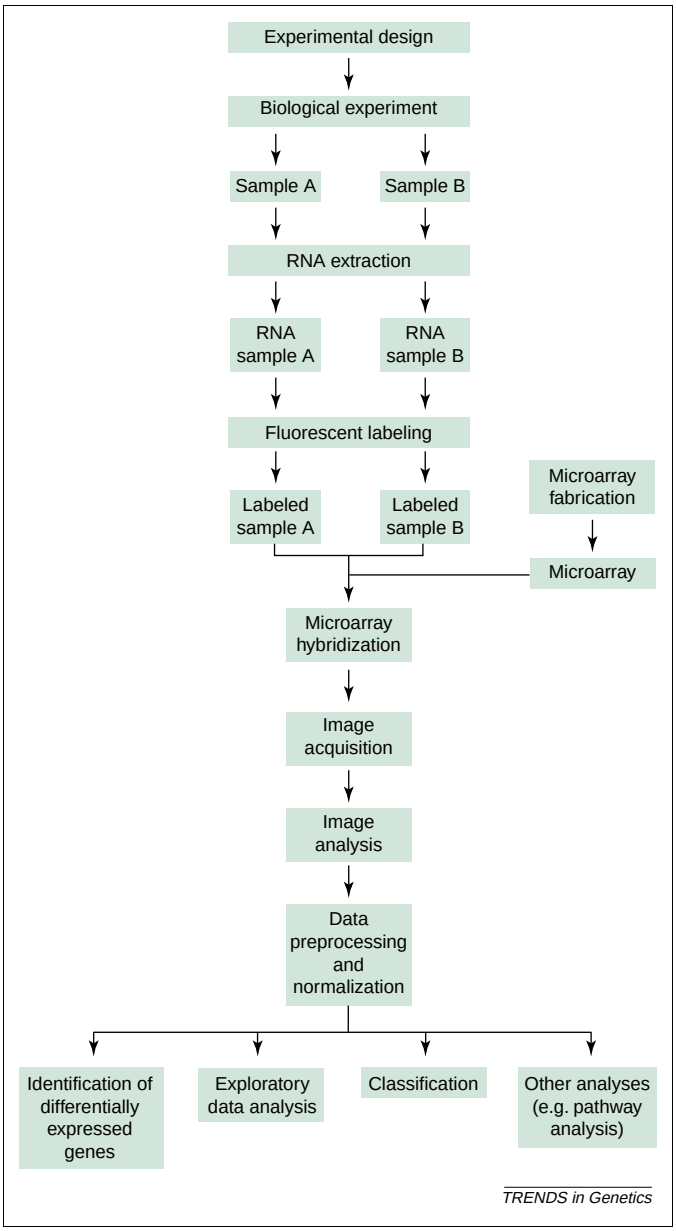


Figure 1. Flow of a typical microarray experiment. A typical microarray experiment begins with good experimental design. After carrying out the biological experiment, the samples, either tissues from patient or animal model, or cells from *in vitro* cultures, are collected. Their RNAs are then extracted and labeled with different fluorescent dyes, and co-hybridized to a microarray. The hybridized microarray is scanned to acquire the fluorescent images. Image analysis is performed to obtain the raw signal data for every spot. Poor quality data are filtered out and the remaining high quality data are normalized. Finally depending on the aim of the study, one can infer statistical significance of differential expression, perform various exploratory data analyses, classify samples according to their disease subtypes and carry out pathway analysis. Note that data from all the steps should be collected according to certain standards, minimum information about a microarray experiment (e.g. MIAME), and archived properly.

Table 1. Summary of microarray analysis steps[a]

Analysis step	Caveats
Experimental design and implementation	Define the biological question and hypothesis clearly
	Design the microarray experimental scheme carefully; include biological replication in experimental design
	Avoid experimental errors
Data collection and archival	Compliance with microarray information collection standards (e.g. MIAME)
Image acquisition	Avoid photo-bleaching
	Try to balance the overall intensities between the two dyes
	Scan image at appropriate resolution
Image analysis	Inspect the gridding result manually; adjust the mask and flag poor-quality spots if necessary
	Choose and apply an appropriate segmentation algorithm
	Apply quality measures to aid decision of spot quality
Data pre-processing	Remove poor-quality spots
	Remove spots with intensity lower the background plus two standard deviations.
	Log-transform the intensity ratios
Data normalization	Use diagnostic plots to evaluate the data
	Consider using LOWESS and its variants for normalization
Identifying differentially expressed genes	Do not use fixed threshold (i.e. two-fold increase or decrease) to infer significance
	Calculate a statistic based on replicate array data for ranking genes
	Select a cut-off value for rejecting the null-hypothesis that a gene is not differentially expressed; remember to adjust for multiple hypothesis testing
Exploratory data analysis	Use different analysis tools with different setting to 'explore' the data
	Validate the result by follow-up experiments
Class prediction and classification	Do not over-train the classifier; try to balance the accuracy and generalizability
Pathway analysis	Try to understand the microarray data in a pathway perspective and not genes in isolation

[a]Abbreviations: LOWESS, locally weighed scatterplot smoothing; MIAME, minimum information about a microarray experiment.

yellow (i.e. ratio equals to 1) on a balanced image (Figure 2a). In many cases, most of the spots on the array are non-differentially expressed; (ii) examining the extent of overlap between the pixel distribution histograms of both channels (Figure 2b); and (iii) computation of the global normalization factor for all the spots contained in the two channels, for example the sum of signals in one channel divided by the sum of signals in the other one. A well-balanced image should have a factor close to 1.

The choice of a suitable scanning resolution depends on the array specification. A rule of thumb is that the resolution setting should be at least 10% of the spot diameter. At the same time, the number of spots with saturated pixels should be kept to a minimum (e.g. <3–5 spots in a whole yeast genome array with 6240 elements) to maximize the dynamic range usage of the scanner. Excessive scanning of a slide should be avoided to prevent photo-bleaching. Images of high-quality can be acquired routinely when all these factors are taken into consideration (Figure 2a).

Spot recognition or gridding is not a difficult problem for most contemporary image analysis software, although it is often necessary to adjust the grid for some spots manually afterwards. In fact, many scientists prefer to visually inspect the images for adjusting the grid and flagging low quality spots instead of totally relying on software recognition. Segmentation is a process used to differentiate the foreground pixels (i.e. the true signal) in a spot grid from the background pixels. This is a tricky computational problem because the spot morphology in a poor-quality image can vary substantially and the background can be high. Furthermore, the image can contain other

Box 1. Different microarray technologies

In general, there are two types of microarray platforms depending on the method of nucleic acid deposition on the chip surface: robotically spotted [52] or *in situ* synthesis by photolithography, a technology that is commonly used in computer chips fabrication [53]. The latter is commercially available from Affymetrix™. Historically the robotically spotted microarrays were referred to as cDNA microarrays because the nucleic acids being spotted were PCR products amplified from cDNA libraries. And the photolithographically synthesized arrays were commonly called oligonucleotides arrays or oligoarrays because shorter oligonucleotides (~25mers) were placed on the arrays and each gene is represented by multiple oligos. It is inaccurate to use the type of probes on arrays to differentiate different platforms because researchers now also prepare oligoarrays by robotically spotting oligonucleotides (~50 to 70mers) on the slide.

Nonetheless, there is still a fundamental difference in the experimental setup between the robotically spotted arrays and photolithographically synthesized ones. In the robotically spotted array experiments, the two samples under comparison are labeled with two different fluorescent dyes and co-hybridized to the same array. This is essentially a comparative hybridization experiment. The ratio between the two dyes indicates the relative abundance of a gene in these two samples. In the photolithographically synthesized array experiments, the two samples under comparison are labeled with the same dye and individually hybridized to different arrays.

Although most downstream analyses like exploratory analysis are similar for the two-microarray platforms, the differences in sample labeling and hybridization have created different requirements in upstream data pre-processing. In particular, because the samples are individually hybridized to different arrays in the case of photolithographically synthesized array experiments, there are specific concerns on features selection [54,55], background adjustment [56], the relationship between signal intensity and transcript abundance [56,57], probe-specific biases [58] and normalization across different arrays [55,56]. This review is focused on the data analysis of the spotted cDNA microarrays, the most accessible microarray platform for general biologists.

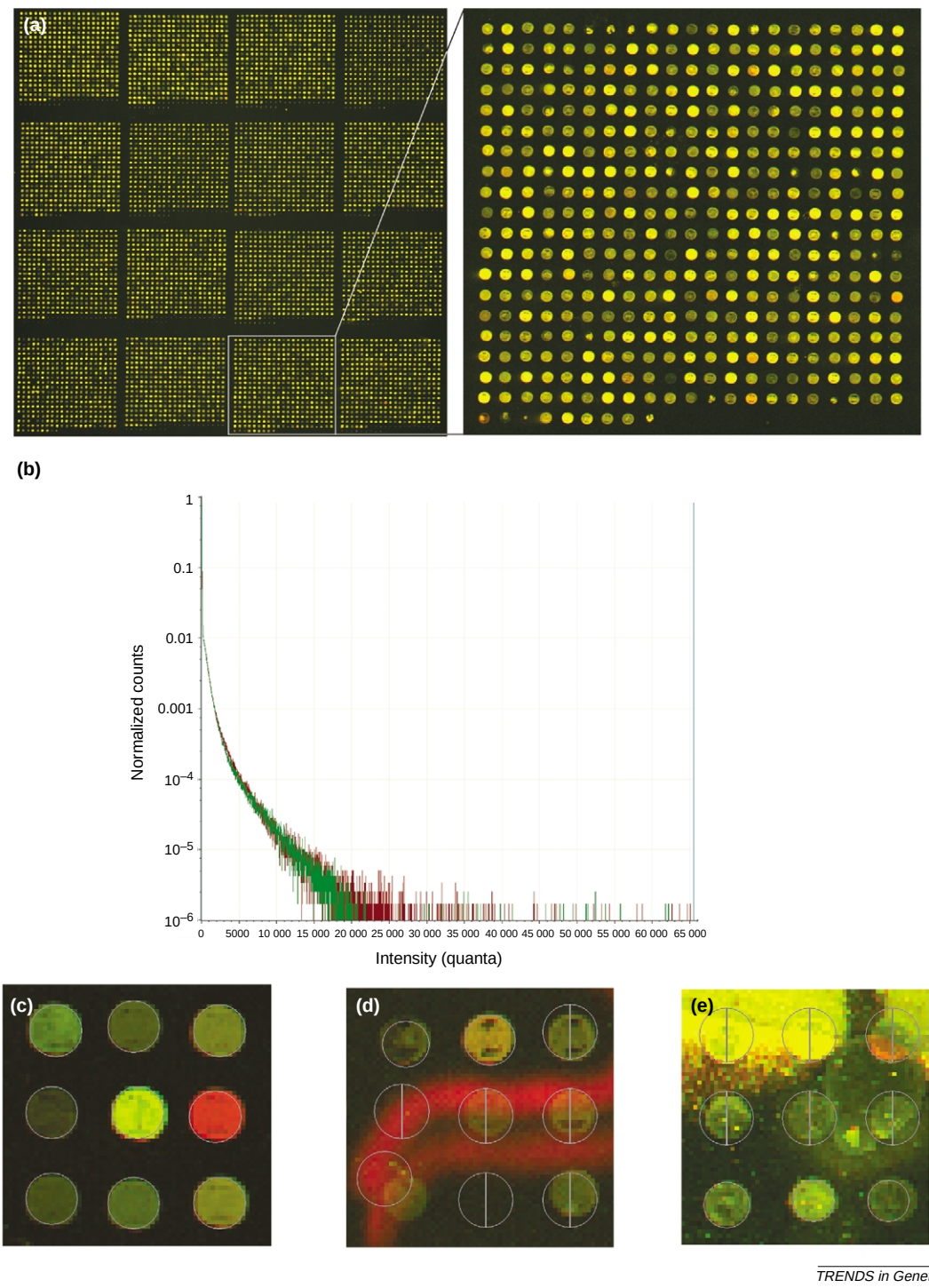


Figure 2. A typical microarray image, pixel distribution histogram for image acquisition, and the effect of image quality on spot recognition and segmentation. **(a)** In this microarray experiment yeast cells treated with a chemical that induced a subtle expression change was compared with the untreated cells by hybridization to a microarray with a complete set of yeast open reading frames (ORFs). **(b)** Pixel histogram for image acquisition. The histograms of the two channels should overlap as much as possible. (c–e) Effect of image quality on spot recognition and segmentation. **(c)** A high-quality image. **(d)** Image with dust contamination. **(e)** Image with high background. (More poor-quality images and how to trouble shoot are available at http://stress-genomics.org/stress.fls/expression/array_tech/trouble_shooting/troubles_index.htm.

imperfections. This can make a proper segmentation difficult. There are several algorithms for segmentation, including FIXED CIRCLE SEGMENTATION, ADAPTIVE CIRCLE SEGMENTATION, adaptive shape segmentation and histogram segmentation. There are also several algorithms for BACKGROUND ESTIMATION, for example constant background, LOCAL BACKGROUND and morphological opening. These algorithms are implemented in different image analysis software [9]. The adaptive circle segmentation and local background estimation algorithms work efficiently for us, but the choice of appropriate algorithms obviously depends on the quality of the raw images. For example, the adaptive circle segmentation that estimates the diameter separately for each spot, works best when all

the spots are circular. Figures 2c–e show the recognition and adaptive circle segmentation results of spots with different background contaminations. When the image quality is high, the algorithm can predict the size of the spots and segment their signal accurately (Figure 2c). If there is dust contamination (Figure 2d) or a high background signal in the image (Figure 2e), the algorithm will not only reject those poor-quality spots, but might also recognize the contamination as a spot (Figure 2d). In this case, both the true signal and background signals will be erroneously estimated. Because it is much more robust for various algorithms to perform segmentation and background estimation processes on a high-quality image than on a low-quality one, it is crucial to produce a high-quality microarray and collect a high-quality image from it in the first place.

Recently there has been an interesting experimental segmentation method reported in which the DNA spots on the microarray were counterstained by 4′, 6′ -diamidino-2-phenylindole (DAPI) and the counterstained image used to assist in the segmentation process [10]. This new experimental approach has apparently resolved many limitations of the algorithmic approach and potentially facilitated the development of a fully automated image analysis system.

After the segmentation process, the pixel intensities within the foreground and background masks (i.e. the areas in the image defined as foreground and background by the software, respectively) are averaged separately to give the foreground and background intensities, respectively. Median or other intensity extraction methods can be used when there are extreme values in the spots that skew the distribution of pixel intensities. Subtracting the BACKGROUND INTENSITY from the foreground intensity in each channel gives the SPOT INTENSITY for calculating the expression ratio between the two channels.

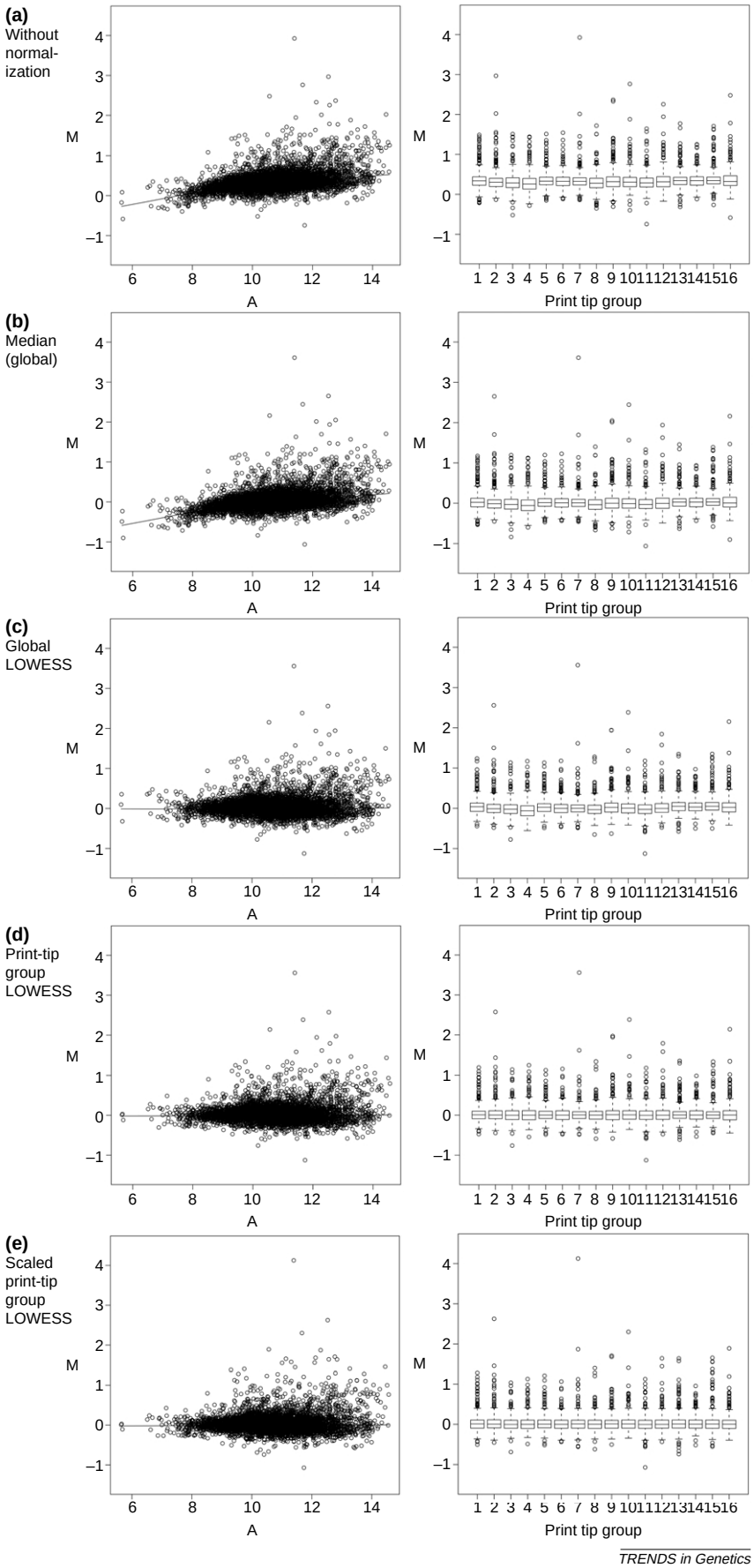
A rapidly developing area that assists in image analysis is the measurement of quality. Some software apply criteria such as diameter, spot area, circularity and replicate uniformity to judge whether a spot is of sufficiently good quality for downstream analysis. The underlying assumption of these criteria is usually a perfect spot, which can be too idealized. A working definition of a good spot is therefore necessary. There is also a need to relate these measures to more common statistical concepts in order that they can be useful for a routine image analysis [9]. A combination of the empirical counterstain segmentation method discussed above [10] and theoretical quality measures can be a practical solution. The DNA counterstain provides information about actual spot morphology and DNA distribution in the spots, which helps to formulate an improved basis for applying different theoretical measures to evaluate the spot quality.

Data pre-processing and normalization

The data extracted from image analysis need to be pre-processed to exclude poor-quality spots and normalized to remove many systematic errors as possible before downstream analysis. Any spot with intensity lower than the background plus two standard deviations should be excluded. The intensity ratios should also be log-transformed so that upregulated and downregulated values are of the same scale and comparable [11].

The process of normalization aims to removing systematic errors by balancing the fluorescence intensities of the two labeling dyes. The dye bias can come from various sources including differences in dye labeling efficiencies, heat and light sensitivities, as well as scanner settings for scanning two channels. Some commonly used methods for calculating normalization factor include: (i) global normalization that uses all genes on the array (Figure 3b); (ii) housekeeping genes normalization that uses constantly expressed housekeeping/invariant genes; and (iii) internal controls normalization that uses known amount of exogenous control genes added during hybridization (http://www.dnachip.org/mged/normalization.html) [11]. Unfortunately these normalization methods are inadequate because dye bias can depend on spot intensity and spatial location on the array. Housekeeping genes are not as constantly expressed as was previously assumed [12]. As a result, using housekeeping genes normalization might introduce another potential source of error. Dye-swapping experiments are seen as a plausible solution to reduce the dye bias problem, but can be impractical because of the limited supply of certain precious samples.

Recently there have been suggestions for using a non-linear normalization method on the basis of gene intensity and spatial information [4,11], which is believed to be superior to the other methods. Figure 3 provides a comparison of various normalization methods, using the data extracted from Figure 2a. All data analyses and graph plotting were performed using statistical microarray analysis (SMA) package (http://stat-zww.berkeley.edu/users/terry/zarray/Software/smacode.html) running in R statistical environment (http://www.r-project.org/). The plots show Log_2 of the expression ratio versus average spot intensity. Ideally the center of the distribution of log-ratios should be zero, the log-ratios should be independent of spot intensity, and the fitted line should be parallel to the intensity axis. In our example, the global locally weighted scatterplot smoothing (LOWESS) normalization is a good choice because it provides a good balance on the three factors mentioned above (Figure 3c). The fluorescent images (Figure 2a) do not suffer from serious spatial effects, as indicated by a very similar log expression ratio distribution among all the print-tips in the bloxplot for the global LOWESS normalization (Figure 3c). However, when there is a significant difference in the distribution of log-ratios among the print-tips in the bloxplot, suggesting a possible spatial effect, print-tip group LOWESS (Figure 3d) or scaled print-tip group LOWESS normalization (Figure 3e) should be considered. Apart from within-a single array, the distribution of gene expression ratios from replicate experiments might have different distribution of log ratios due to the difference in experimental conditions. Therefore scaling adjustment is often necessary to standardize the distribution of log-ratios across replicate experiments to prevent any particular experiment becoming dominant and affecting downstream statistical analysis.


(a)
Without normal-ization
(b)
Median (global)
(c)
Global LOWESS
(d)
Print-tip group LOWESS
(e)
Scaled print-tip group LOWESS
M
A
Print tip group
TRENDS in Genetics

Data analysis

The next stage of analysis is to apply various statistical and data mining techniques to study the data. There are several typical approaches that are discussed in the following sections.

Significance inference – identifying significantly differentially expressed genes

Traditionally, differentially expressed genes are inferred by a fixed threshold cut off method (i.e. a two-fold increase or decrease), but this is statistically inefficient, the main reason being that there are numerous systemic and biological variations that occur during a microarray experiment. Although some of the systemic variations such as dye bias can be effectively removed by normalization, random biological variations such as sample-to-sample and physiological variations are more difficult to handle [13,14] (for a comprehensive review of various statistical issues, variations and errors of microarray experiment see Ref. [15]). Because of these underlying variations, merely using a fixed threshold to infer significance might increase the proportion of false positives or false negatives. A better framework of significance inference includes calculation of a statistic based on replicate array data for ranking genes according to their possibilities of differential expression and selection of a cut-off value for rejecting the null-hypothesis that the gene is not differentially expressed.

Replication of a microarray experiment is essential to obtain the variation in the gene expression for statistics calculation. It has been suggested that every microarray experiment should be performed in triplicate to increase data reliability [16]. There are two types of replication: biological and technical. Biological replication refers to the analysis of multiple independent biological samples (e.g. one tissue type obtained from different patients with the same disease, or individual samples of a particular cell line under the same treatment), whereas technical replication refers to the repetition of microarray experiment using the same extracted RNA samples. Biological replication is particularly important for expression profiling of disease tissues, because there might be variability of expression among the same tissue type or tissue heterogeneity. Any particular tissue might not be representative of the whole disease sample group. Technical replication provides a precise measurement of gene expression for a particular sample and eliminates many technical variations introduced during the experiment. Unfortunately, merely obtaining a precise expression measurement of a tissue by technical replication will not resolve the problem of biological variation. Therefore it is usually preferable to have biological replication rather than technical replication if there are not enough tissues or resources to perform several microarray experiments, provided the experiment procedures are carried out carefully [4,5]. Statistical methods such as Student's *t*-test and its variants [17,18], ANOVA [19,20], Bayesian method [17,20,21], or Mann–Whitney test [22], can be used to rank the genes from replicated data.

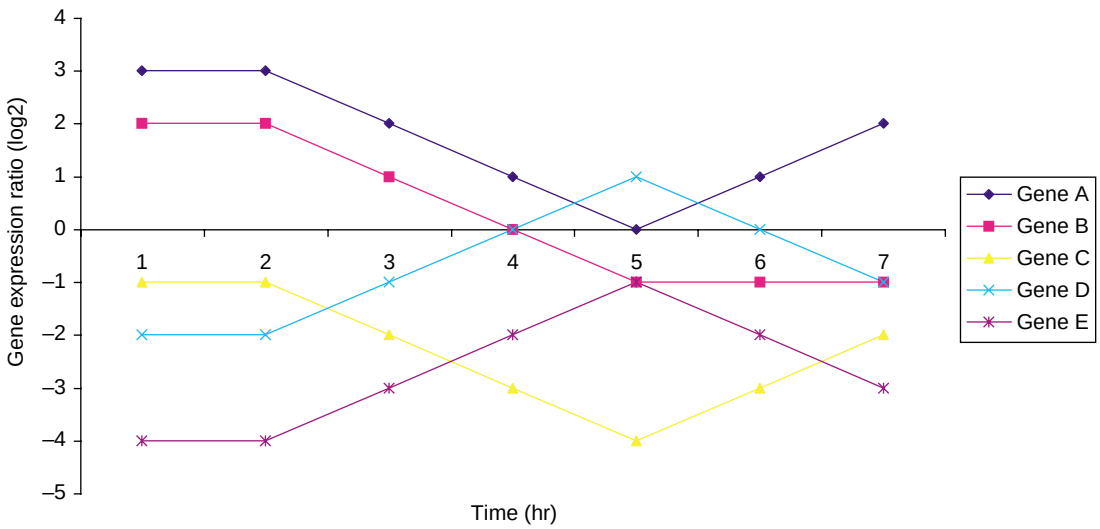
Setting a cut-off for differential expression is tricky, because one has to balance the false positives (Type I error) and the false negatives (Type II error). Furthermore, performing statistical tests for tens of thousands of genes creates a multiple hypothesis-testing problem. For example, in an experiment with a 10 000-gene array in which the significance level α is set at 0.05, $10\,000 \times 0.05 = 500$ genes would be inferred as significant even though none is differentially expressed. Therefore using a *p*-value of 0.05 is likely to exaggerate Type I errors. The multiple hypothesis testing problem is conventionally tackled by conservative approaches that control the family-wise error rate (FWER), the probability of having at least one false positive among all testing hypotheses [23]. A classical example is the Bonferroni correction. However, controlling the FWER can be too stringent and limits the power to identify significantly differentially expressed genes. In fact, differential expression is usually confirmed by RT-PCR, northern blots or *in situ* hybridization [24]. It is often acceptable to have few false positives if the majority of true positives are chosen. Therefore it might be more practical to control the false discovery rate (FDR) [25], the expected proportion of false positives among the number of rejected hypotheses. A program, statistical analysis of microarray (SAM), has been developed to utilize this FDR concept as a tool to assist in determining a cut-off after performing adjusted *t*-tests (http://www-stat.stanford.edu/~tibs/SAM/index.html) [18].

Exploratory data analysis – understanding the (dis)similarities of the gene expression levels among all samples

Also known as unsupervised data analysis, exploratory data analysis does not require the incorporation of any prior knowledge in the process. It is essentially a grouping technique that aims tot find genes with similar behaviors (i.e. expression profiles). Some commonly used examples include principal component analysis (PCA) [26] or singular value decomposition (SVD) [27] for dimensionality reduction, as well as hierarchical clustering [28], K-means clustering [29] and self organizing maps (SOMs) [30] for clustering. There are already several excellent reviews on various unsupervised analyses and their applications in microarray data mining [31–33], therefore we do not discuss their details here.

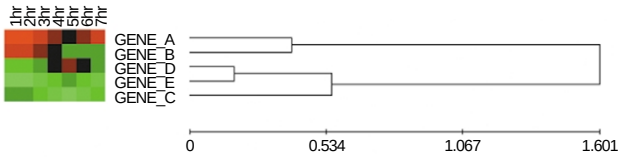
There is perhaps no unsupervised data analysis that can suit all situations. Different analyses or even different parameters of the same analysis can reveal unique aspects of the data. This idea is illustrated in Figure 4, in which five genes from a hypothetical time series data are clustered using various distance or similarity measures and unweighted pair group method with arithmetic mean (UPGMA) algorithm. Each distance or similarity measure

Figure 3. A comparison of various normalization methods. The raw data was extracted from Figure 2a. Any spot with intensity lower than the background plus two standard deviations or of poor-quality was excluded from further analysis. From top to bottom: Log_2 ratios (M) versus average intensities (A) plot and boxplot of the data without normalization **(a)** and with four different kinds of normalization methods: **(b)** median, **(c)** global locally weighted scatterplot smoothing (LOWESS), **(d)** print-tip group LOWESS, **(e)** scaled print-tip group LOWESS.



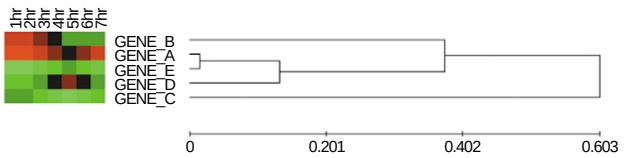
Distance similarity measure

(a) Correlation coefficient without centering

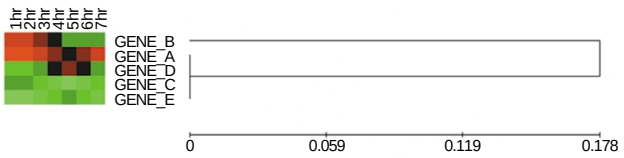


(b) Correlation coefficient with centering

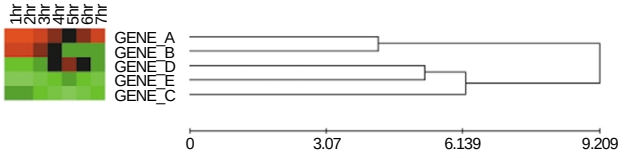
(c) Absolute correlation coefficient without centering



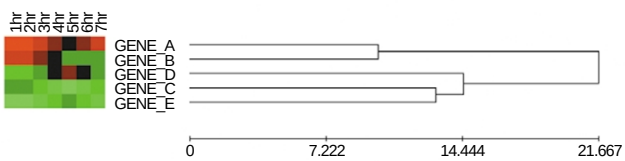
(d) Absolute correlation coefficient with centering



(e) Euclidean distance



(f) Manhatten distance



TRENDS in Genetics

can assign the genes to different clusters. For example, Euclidean and Manhattan distances are sensitive to absolute expression levels, and are able to reveal those genes that have similar expression levels in the cluster. Two main clusters are identified in the data, one for gene *A* and *B* and the other cluster for gene *C*, *D* and *E* (Figure 4e,f). *A* and *B* are clustered with each because their overall expression ratios more similar when compared with *C*, *D* and *E*, and vice versa. The similarity between their expression profiles suggests the genes in the two clusters might be co-regulated. However, if the researchers conclude the analysis at this stage, they are likely to miss some other interesting relationship among the genes. A slightly different picture is revealed by using correlation coefficient with centering, a similarity measure that is sensitive to the expression profile shape, regardless of the expression levels (Figure 4b). Gene *A*, *B* and *C* are grouped in the same cluster whereas *D* and *E* are in another. Intriguingly, *A* and *C*, gene *D* and *E* are correlated with each other perfectly using this distance measure. An inspection of the expression profile offers a hint. Although *A* and *C* differ largely in expression level, the shape of their expression profiles is the same. This is also true for gene *D* and *E*. As a result, the correlation coefficients for both *A* and *C* and gene *D* and *E* are 1. This result suggests gene *A* and *C*, gene *D* and *E* are likely to be co-regulated, and analyzing their promoters can sometimes identify common regulatory elements. Further insight is provided using absolute correlation coefficient with centering as a similarity measure (Figure 4d). This time *A*, C, *D* and *E* are clustered perfectly together, leaving *B* separate. It is because the shape of the expression profiles of *A* and *C* are a mirror image of *D* and *E*. Although their correlation coefficient is -1, which will place them in two separate clusters as shown in Figure 4b, the absolute value of their correlation coefficient is the same and will place them in the same cluster. Therefore it is very likely that *A*, *C*, *D*, *E* are regulated by a same factor or mechanism, which represses the expression *A* and *C* while enhancing the expression of *D* and *E*, and vice versa. The same principle also applies to the choice of clustering algorithms [31].

Hence, it is always advisable to apply several unsupervised analyses and different parameters to explore the data. Nonetheless, there must be a balance between the time spent on data analysis and the time spent on subsequent experimental confirmation. Unsupervised analysis is a useful method for generating new hypotheses. The validity of the result has to be built upon both statistical significance and biological knowledge.

Class prediction – using gene expression profiles as a means to classify samples

Another intriguing type of data analysis is to train a classifier algorithm using the expression profiles of predefined sample groups, so that the classifier can best assign any new sample to the respective group. This type of analysis is also known as supervised data analysis, which has great promise in clinical diagnostics [31] and has been used successfully in several recent studies [34–36]. Examples of such analysis include support vector machines [37], artificial neural networks [38], k-nearest neighbor [39] and various discrimination methods (http://stat-www.berkeley.edu/users/terry/zarray/Html/discr.html). The ultimate goal is to generalize the trained classifier as a routine diagnostic tool for differentiating between the samples that are difficult or even impossible to classify by available methods.

The challenge for supervised data analysis is to generalize the classifier for all situations. The training samples are often limited in number that might not be sufficiently representative for their classes in general. Over-training on the same dataset would result in a situation called 'over-fitting', in which the classifier is very effective in classifying the training samples but not accurate enough for new samples. A balance between accuracy and generalizability has to be established by validation of the trained classifier. Several approaches are available for this purpose. For example, the training samples are divided into two individual sets, one for training and one for validation. The training of the classifier will be stopped when the prediction error on the validation set reaches a minimum. More sophisticated cross-validation methods divide the training dataset into several subsets. Each subset will be the validation set in turn. The overall accuracy therefore is the average accuracy across all validation trials. An extreme case of cross-validation is called leave-one-out cross-validation, in which one sample is taken away from the training set to be a validation sample each time. An investigation of several supervised analyses, their performance, and cross-validation was detailed previously [40].

An emerging approach – pathway analysis

Genes never act alone in a biological system – they are working in a cascade of networks. As a result, analyzing the microarray data in a pathway perspective could lead to a higher level of understanding of the system. There are at least three interesting approaches in this area. The first is a natural extension of the exploratory cluster analysis described above. If several genes are assigned to the same group by cluster analysis, as discussed above, they might be co-regulated or involved in the same signaling pathway. Analyzing the promoters of this group of genes can often reveal common regulatory motifs and unveil a higher level of network organization in the biological system [41]. The second is to reverse-engineer the global genetic pathways, the identification of the global regulatory network architecture from microarray data. It can be done by a systematic targeted perturbation like mutation or chemical treatment [42], and time series experiments [43]. The assumption here is that the perturbation will cause a change in expression of other proteins in the network. This

Figure 4. Different distance measures provide different views of the data. Line graphs of a hypothetical time series experiment with five genes and seven time points (upper panel). Hierarchical clustering of the data using six common distance or similarity measures (lower panel): **(a)** correlation coefficient without centering, **(b)** correlation coefficient with centering, **(c)** absolute correlation coefficient without centering, **(d)** absolute correlation coefficient with centering, **(e)** Euclidean distance, **(f)** Manhattan distance. Clustering was performed using unweighted pair group method with arithmetic mean algorithm (UPGMA).

change in the expression profiles should be able to capture the underlying architecture of the network. Various methods have been proposed for constructing a network from this kind of microarray data, such as a Boolean network that simplifies gene expression as a binary logical value to infer the induction of a gene as a deterministic function of the state of a group of other genes [44–46] and a Bayesian network that models interactions among genes, evaluates different models and assigns them probability scores [47,48] (readers are referred to two excellent reviews on these and other methods for reverse engineering of networks [49,50]). The final approach is to study the expression data on a pathway perspective. Our group has recently developed a method called Pathway Processor (http://cgr.harvard.edu/cavalieri/pp.html) that can map expression data onto metabolic pathways and evaluate which metabolic pathways are most affected by transcriptional changes in whole-genome expression experiments [51]. We used the Fisher Exact Test to score biochemical pathways according to the probability that as many or more genes in a pathway would be significantly altered in a given experiment than by chance alone. Results from multiple experiments can be compared, reducing the analysis from the full set of individual genes to a limited number of pathways of interest.

Conclusion

Microarray analysis is evolving rapidly. New and more complex analyses appear everyday, making it easy for the researcher to get lost in endless new methods and software. Collaborating with statisticians and mathematicians is often advisable for performing a proper microarray analysis. Nonetheless, this will not replace biological expertise, a good foundation for statistical methods and meticulousness in conducting experiments.

Acknowledgements

YFL is supported by a Croucher Foundation Postdoctoral Fellowship. We thank Alice Yu Ming Lee and Abel Chiu Shun Chun for their critical comments on this manuscript.

References

1 Leung, Y.F. *et al.* (2002) Microarray software review. In *A practical approach to microarray data analysis* (Berrar, D.P. *et al.*, eds), Kluwer academic
2 Box, G.E.P. *et al.* (1978) *Statistics for experimenters – an introduction to design, data analysis, and model building*, John Wiley & Sons
3 Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32 (Suppl. 2), 490–495
4 Yang, Y.H. and Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3, 579–588
5 Simon, R.M. and Dobbin, K. (2003) Experimental design of DNA microarray experiments. *Biotechniques*, S16–S21
6 Perou, C.M. (2001) Show me the data!. *Nat. Genet.* 29, 373
7 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371
8 Anonymous, (2002) Microarray standards at last. *Nature* 419, 323
9 Yang, Y.H. *et al.* (2001) Analysis of cDNA microarray images. *Brief. Bioinform.* 2, 341–349
10 Jain, A.N. *et al.* (2002) Fully automatic quantification of microarray image data. *Genome Res.* 12, 325–332
11 Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* 32 (Suppl.), 496–501
12 Lee, P.D. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12, 292–297
13 Novak, J.P. *et al.* (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics* 79, 104–113
14 Pritchard, C.C. *et al.* (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13266–13271
15 Nadon, R. and Shoemaker, J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.* 18, 265–271
16 Lee, M.L. *et al.* (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9834–9839
17 Lönnstedt, I. and Speed, T.P. (2002) Replicated Microarray Data. *Stat. Sinica* 12, 31–46
18 Storey, J.D. and Tibshirani, R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (Parmigiani, G. *et al.*, eds), Springer
19 Kerr, M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819–837
20 Long, A.D. *et al.* (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.* 276, 19937–19944
21 Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t* test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519
22 Wu, T.D. (2001) Analysing gene expression data from DNA microarrays to identify candidate genes. *J. Pathol.* 195, 53–65
23 Dudoit, S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica* 12, 111–139
24 Chuaqui, R.F. *et al.* (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.* 32 (Suppl. 2), 509–514
25 Reiner, A. *et al.* (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 368–375
26 Raychaudhuri, S. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466
27 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106
28 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
29 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285
30 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912
31 Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427
32 Sherlock, G. (2001) Analysis of large-scale gene expression data. *Brief. Bioinform.* 2, 350–362
33 Valafar, F. (2002) Pattern recognition techniques in microarray data analysis: a survey. *Ann. N. Y. Acad. Sci.* 980, 41–64
34 Pomeroy, S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436–442
35 Shipp, M.A. *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8, 68–74
36 Khan, J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673–679
37 Brown, M.P. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
38 Vohradsky, J. (2001) Neural network model of gene expression. *FASEB J.* 15, 846–854
39 Theilhaber, J. *et al.* (2002) Finding genes in the C2C12 osteogenic

pathway by k-nearest-neighbor classification of expression data. *Genome Res.* 12, 165–176
40 Ben-Dor, A. *et al.* (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.* 7, 559–583
41 Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159
42 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
43 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285
44 Liang, S. *et al.* (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 18–29
45 Akutsu, T. *et al.* (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.* 7, 331–343
46 Maki, Y. *et al.* (2001) Development of a system for the inference of large scale genetic networks. *Pac. Symp. Biocomput.*, 446–458
47 Friedman, N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620
48 Hartemink, A.J. *et al.* (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, 422–433
49 de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103
50 D'haeseleer, P. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726
51 Grosu, P. *et al.* (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.* 12, 1121–1126
52 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
53 Lipshutz, R.J. *et al.* (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* 21 (Suppl. 1), 20–24
54 Zhou, Y. and Abagyan, R. (2003) Algorithms for high-density oligonucleotide array. *Curr. Opin. Drug Discov. Devel.* 6, 339–345
55 Schadt, E.E. *et al.* (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 37 (Suppl.), 120–125
56 Schadt, E.E. *et al.* (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 80, 192–202
57 Sasik, R. *et al.* (2002) Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics* 18, 1633–1640
58 Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* 98, 31–36