Microarray data analysis: from disarray to consolidation and consensus

David B. Allison[*‡§], *Xiangqin Cui*[*§], *Grier P. Page*[*] and *Mahyar Sabripour*[*]

Abstract | In just a few years, microarrays have gone from obscurity to being almost ubiquitous in biological research. At the same time, the statistical methodology for microarray analysis has progressed from simple visual assessments of results to a weekly deluge of papers that describe purportedly novel algorithms for analysing changes in gene expression. Although the many procedures that are available might be bewildering to biologists who wish to apply them, statistical geneticists are recognizing commonalities among the different methods. Many are special cases of more general models, and points of consensus are emerging about the general approaches that warrant use and elaboration.

Fold change
A metric for comparing a gene's mRNA-expression level between two distinct experimental conditions. Its arithmetic definition differs between investigators.

Case
In a microarray experiment, a case is the biological unit under study; for example, one soybean, one mouse or one human.

Gene-expression microarrays have become almost as widely used as measurement tools in biological research as western blots (BOX 1). A wide range of methods for microarray data analysis have evolved, ranging from simple fold-change (FC) approaches to testing for differential expression, to many complex and computationally demanding techniques[1]. The result might seem like a statistical tower of Babel, but many methods are in fact special cases of general approaches. Recognizing this allows investigators to choose procedures more judiciously and methodologists to direct their efforts more efficiently.

Here we examine five key components of microarray analysis — design, preprocessing, inference, classification and validation (BOX 1) — and address important areas where consensus has emerged or seems imminent, and key areas where questions remain. The methods we discuss are often supplemented with graphical representations, which serve as important interpretive aids (FIG. 1). We focus on aspects that are relevant to the widest range of microarray users, in typical small or moderately sized, single-laboratory studies. We also note that other issues might apply to larger, multi-site studies. Additionally, some currently rapidly expanding areas, such as graphical modelling for gene networks and pathways, are not discussed here.

Design

Experimental design affects the efficiency and internal validity of experiments[1–7]. Here we discuss points that are relevant to optimizing microarray experiments for most design strategies. The relative merits of specific designs are discussed elsewhere[1,4–7].

Consensus point 1: Biological replication is essential. In microarray analysis, two types of replication can be carried out: technical replication, when mRNA from a single biological case is used on multiple microarrays, and biological replication, when measurements are taken from multiple cases. Although early microarray experiments used few or no biological replicates, their necessity is now undisputed[1,6,7]. Technical replicates allow only the effects of measurement variability to be estimated and reduced, whereas biological replicates allow this to be done for both measurement variability and biological differences between cases. Consequently, although almost all experiments that use statistical inference (BOX 1) require biological replication, technical replicates are almost never required when the aim is to make inferences about populations that are based on sample data, as is the case in most microarray studies.

However, there are some situations where technical replication is needed, such as quality-control studies. Additionally, if the number of cases available is finite or small, or if the cost of obtaining another case exceeds the cost of an array, then technical replication might be useful in addition to biological replicates[8].

Consensus point 2: There is strength in numbers — power and sample size. How many biological replicates are needed? Traditional approaches to analysing statistical power are ill-suited to microarray studies, which

*Section on Statistical Genetics, Department of Biostatistics, Ryals Public Health Building, 1665 University Avenue, University of Alabama at Birmingham, Alabama 35294-0022, USA.
‡Clinical Nutrition Research Center, University of Alabama at Birmingham.
§Department of Medicine, University of Alabama at Birmingham.
Correspondence to D.B.A.
e-mail: Dallison@uab.edu
doi:10.1038/nrg1749

© 2005 Nature Publishing Group

Box 1 | Principles of microarray experiments and analysis

In microarrays, thousands of probes are fixed to a surface, and RNA samples (the targets) are labelled with fluorescent dyes for hybridization. After hybridization, laser light is used to excite the fluorescent dye; the hybridization intensity is represented by the amount of fluorescent emission, which gives an estimate of the relative amounts of the different transcripts that are represented. There are many microarray platforms that are different in array fabrication and dye selection.

In cDNA microarrays, both the probes and the targets are cDNAs. mRNA from biological samples is reverse transcribed and simultaneously labelled with Cy3 and Cy5. After hybridization, Cy3 and Cy5 fluorescence is measured separately, and captured in two images. These are merged to produce a composite image, which goes though preprocessing (see below) before expression values are analysed. Long-oligonucleotide microarrays are similar to cDNA microarrays, but the probes are derived from genomic or EST sequences. High-density oligonucleotide microarrays involve probe pairs that each consist of 25-nt oligonucleotides. Each probe pair has a perfect-match (PM) probe and a mismatch (MM) probe. The MM probe has identical sequence to the PM probe, except at the central base and functions as an internal control. Unlike cDNA microarrays, the mRNA sample is converted to biotinylated cRNA and only one target is hybridized to each array — therefore, only a single colour of fluorescence is used.

Statistical components

The statistical components of a microarray experiment involve the following steps:

- *Design*. The development of an experimental plan to maximize the quality and quantity of information obtained.
- *Preprocessing*. Processing of the microarray image and normalization of the data to remove systematic variation. Other potential preprocessing steps include transformation of data[94], data filtering[95] and, in the case of two-colour arrays, background subtraction (although there is some emerging consensus that background subtraction is not helpful[29]).
- *Inference and/or classification*. Inference entails testing statistical hypotheses (these are usually about which genes are differentially expressed). Classification refers to analytical approaches that attempt to divide data into classes with no prior information (unsupervised classification) or into predefined classes (supervised classification).
- *Validation of findings*. The process of confirming the veracity of the inferences and conclusions drawn in the study.

test many hypotheses, use false-discovery-rate (FDR) estimates for inference, and often use classification techniques that have thousands of transcripts (BOX 1; and see later). Statisticians have therefore begun to provide tailor-made solutions to calculate power and sample-size requirements for microarrays.

For common designs, which we refer to as designs in which two groups of cases are evaluated for differential expression, evidence indicates that a minimum of 5 biological cases per group should be analysed[9–11]. We emphasize that this applies only when differential expression testing — not classification — is the primary goal, and that this is a minimum, not an optimum. Power analyses often indicate that larger sample sizes are warranted. Several methods have been put forward recently to address the optimal number of replicates[12–16]. Methods have also been developed for estimating sample sizes for classification studies[17,18]. Moreover, tools have been developed that allow investigators to estimate optimal sample sizes on the basis of public data sets (for example, the PowerAtlas software). Although there is no consensus about which sample-size determination procedures are best, there is consensus that power analyses should be done, that newer methods specifically for microarray research should be used, and that more replicates generally provide greater power.

Consensus point 3: Pooling biological samples can be useful. Variability among arrays can be reduced by pooling mRNA from biological replicates. Many investigators favour this strategy because sample size can be increased without purchasing more microarrays. For example, 15 cases divided into 5 pools of 3, with each pool run on a separate array, should have more power than 5 cases run on different arrays, although the power will be less than when 15 cases are run separately.

However, there are caveats that apply to mRNA pooling [1,6,19,20]. First, pooling is not always beneficial — for example, in the context of classification, pooling interferes with the ability to accurately assess inter-individual variation and covariation. Second, one cannot simply analyse one pool per group — analysing multiple pools is required to estimate variance for inference testing. A corollary to this observation corrects a common misapprehension that pooling RNA from *n* cases and creating *n* technical replicates is a better strategy than hybridizing *n* arrays to the *n* individual RNA samples. Third, we note the potential problem of the 'poisoned pool' — that is, one outlier can yield misleading results. Finally, measurements from pools do not necessarily correspond to mathematical averages of measurements from individuals comprising the pool[20]. Nevertheless, pooling can be beneficial when identifying differential expression is the sole goal, when biological variability is high relative to measurement error, and when biological samples are inexpensive relative to array cost[20].

Consensus point 4: Avoiding confounding by extraneous factors is crucial. Microarray measurements can be greatly influenced by extraneous factors. If such factors covary with the independent variable — for example, with different treatments that are applied to two sets of samples — this might confound the study and yield erroneous conclusions. Therefore it is crucial that such factors are minimized or, ideally, eliminated. For example, arrays should be used from a single batch and processed by one technician on the same day. However, this is difficult with large experiments and it is therefore important to orthogonalize extraneous factors (for example, by analysing equal numbers of samples from two groups under assessment on each day of analysis), or to randomize cases to levels of these factors[1].

Preprocessing

Data preprocessing, including image analysis, normalization and data transformation, remains an active research area (BOX 1). In terms of image analysis, how to appropriately quantify spots on microarrays is a topic of vigorous inquiry. Many image-processing approaches have been developed[21–25], among which the main differences relate to how spot segmentation — distinguishing foreground from background intensities — is carried

Power
This is classically defined as the probability of rejecting a null hypothesis that is false. However, power has been defined in several ways for microarray studies.

False-discovery rate
(FDR). The expected proportion of rejected null hypotheses that are false positives. When no null hypotheses are rejected, FDR is taken to be zero.

Normalization
The process by which microarray spot intensities are adjusted to take into account the variability across different experiments and platforms.

© 2005 Nature Publishing Group

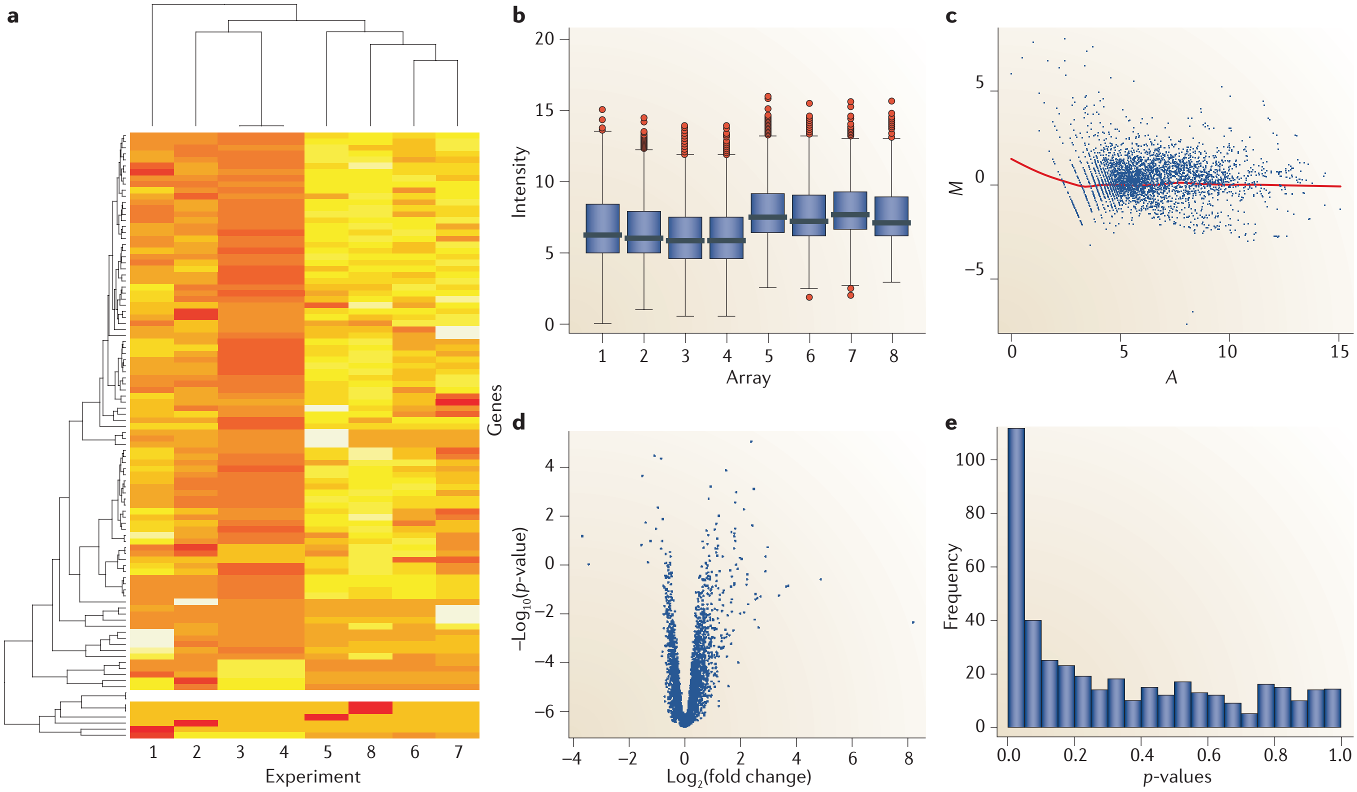


Figure 1 | **Visualization tools for microarray analysis.** Many visualization tools are available that are of great assistance in interpreting the results of microarray experiments. The most commonly used of these are illustrated. **a** | Heatmaps consist of small cells, each consisting of a colour, which represent relative expression values. Heatmaps are often generated from hierarchical cluster analyses of both samples and genes. Often the rows represent genes of similar expression values, whereas the columns indicate different biological samples. Heatmaps offer a quick overview of clusters of genes that show similar expression values. **b** | Box plots present various statistics for a given data set. The plots consist of boxes with a central line and two tails. The central line represents the median of the data, whereas the tails represent the upper (seventy-fifth percentile) and lower quartile (twenty-fifth percentile). Such plots are often used in describing the range of log ratios that is associated with replicate spots. **c** | *MA* plots are used to detect artefacts in the array that are intensity dependent. They are often used as an aid when normalizing two-colour cDNA microarrays. The data consist of intensity measurements that correspond to both red (R) and green (G) dyes. A Cartesian plot is constructed with *M* on the ordinate and *A* on the abscissa, where $M = \log_2(R/G)$ and $A = \log_2(\sqrt{(R \times G)})$. The data are often fitted with a lowess curve, which is used to normalize the gene-expression measurements. **d** | Volcano plots are used to look at fold change and statistical significance simultaneously. Cartesian plots typically show $-\log_{10}$(*p*-values) or log odds on the ordinate and fold-change values on the abscissa for all genes in a data set. The name stems from the volcano shape of the plots. The upper corners of the plot represent genes that show both statistical significance and large fold changes. **e** | *p*-value histograms have abscissae that range from 0 to 1 and contain the *p*-values for a test of differential expression for each gene. They are common supplements to the formal mixture models that enable the popular calculation of false-discovery rates.

Transformation
The application of a specific mathematical function so that data are changed into a different form. Often, the new form of the data satisfies assumptions of statistical tests. The most common transformation in microarray studies is $\log_2$.

Plasmode
A real (not computer simulated) data set for which the true structure is known and is used as a way of testing a proposed analytical method.

out. Another important preprocessing step is normalization, which allows comparisons between microarray experiments and the control of extraneous variation among experiments. It also generally makes data more consistent with the assumptions that underlie many inferential procedures. Several normalization approaches have been introduced, and are discussed elsewhere[26–29].

Remaining question 1: What is the best image-processing algorithm? Several image-processing methods have been developed for Affymetrix arrays[30], which are the most commonly used oligonucleotide microarrays. These methods estimate the amount of RNA from fluorescent array images, while trying to minimize the extraneous variation that occurs owing to technical artefacts[31,32]. Plasmode data sets[33] have been used to evaluate different image-processing normalization methods. One method, robust multi-array average (RMA), corrects arrays for background using a transformation, normalizes them using a formula that is based on a normal distribution, and uses a linear model to estimate expression values on a log scale. RMA and a modification of this method, GCRMA, often perform as well or better than competitors, although there is some controversy about which method is best[32,34,35]. It is also unclear whether there is an ideal way of defining which method produces the best results.

© 2005 Nature Publishing Group

Box 2 | **Shrinkage**

To 'borrow information' across genes, many differential expression tests manipulate the denominator of a test statistic that, practically speaking, is a *t*-test (for example, the 'SAM' procedure[4] and the regularized *t*-test[5]). Most if not all of these *t*-like statistics (including Student's *t*-statistic itself) can be seen as special cases of a statistic that combines the gene-specific variance estimate and a predicted variance (often the microarray-wide average) in different ways. They can be roughly generalized in the form (equation 1):

$$\frac{\hat{\delta}}{\sqrt{B\bar{\sigma}^2 + (1-B)\,\hat{\sigma}^2}} \qquad (1)$$

where $\hat{\delta}$ is some sample statistic that quantifies group differences in gene expression, $\bar{\sigma}^2$ is a predicted variance of $\hat{\delta}$ that is based on all genes on the array, $\hat{\sigma}^2$ is an estimated variance of $\hat{\delta}$ that is based only on the data from the gene being tested, and B is a 'shrinkage factor' that ranges between 1 and 0. The main differences between the many procedures that are offered are the value to which B is set and whether $\hat{\sigma}^2$ is transformed before shrinkage. When B is 1, we are essentially putting a constant in the denominator and letting all inference information come from $\hat{\delta}$ — that is, basically fold change. When B is 0, we have a standard *t*-test without any shrinkage.

To allow developers to submit preprocessing methods for evaluation in a similar fashion, the web-based framework AffyComp has been established[36], although the plasmodes that are available are limited. Most algorithms have been developed and evaluated using one or two small data sets, with a single array type in a single species (human). There is concern that the methods are optimized for these data sets, but might not perform in the same way for others. Further publicly available plasmode data sets are needed to overcome this problem. Finally, for non-Affymetrix platforms, image-processing and normalization algorithms abound and vary substantially in their approaches[21–24,28]; a clear 'winner' has not yet emerged.

Remaining question 2: How should data quality be evaluated? Many researchers recognize the need for microarray quality-control (QC) measures that quantify the measurement quality for any particular array, and several have been proposed (for example, REF. 37). However, the usefulness of most QC measures is unsubstantiated and no specific QC method has been embraced by the community.

Inference

Inference involves making conclusions about the truth of hypotheses that involve unobserved parameters about whole populations, which are based on statistics obtained from samples. An example hypothesis is: 'there is a difference in gene expression between mice that are exposed to conditions A and B in the theoretical population of all mice that could have been exposed to conditions A and B.' Importantly, there is a clear distinction between inference and the simple ranking of findings for follow-up.

Methods are needed to minimize inferential errors — that is, type 1 error (false-positive error) and type 2 error (false-negative error) – and that estimate the long-range error rate. Here we do not discuss specific tests in detail, but focus on points that are applicable to most inferential analyses, including commonly used methods such as *t*-tests, ANOVA, logistic regression and survival analysis.

Consensus point 1: Using fold change alone as a differential expression test is not valid. FC was the first method used to evaluate whether genes are differentially expressed, and is a reasonable measure of effect size. However, it is widely considered to be an inadequate test statistic[38,39] because it does not incorporate variance and offers no associated level of 'confidence'[38,40]. Using FC alone with a fixed cut off, regardless of sample size or variance, results in type 1 error rates that are either unknown or depend on sample size, and even tests for which power can decrease with increasing sample size.

The popularity of FC stems primarily from its simplicity. However, many researchers use it because it often seems to work reasonably well for ranking results (but not for true inference). This is presumably because all transcripts go through the same processing together, and therefore have similar variances. Thinking about FC is conceptually useful because FC essentially assumes a constant variance across transcripts and, therefore, occupies one pole of a continuum of variance shrinkage (an important concept that is discussed below). Nevertheless, FC alone is not valid as an inferential statistic because it does not produce known and controllable long-range error rates, which are essential for inference.

Consensus point 2: 'Shrinkage' is a good thing. Considering each gene separately when conducting statistical tests is inefficient. That is, it does not use all the information that is available to increase the power of tests. Usually, few data points are available for each gene and, therefore, gene-specific variance estimates are imprecise. By using all the data simultaneously, better estimates of variance can be obtained, resulting in more powerful testing. Capitalizing on the parallel nature of microarrays, information can be 'borrowed' across genes to improve variance estimates and thereby increase statistical power. A weighted combination of data from the specific gene and data from all genes can be used in a procedure called variance shrinkage (BOX 2). This has the greatest benefit when sample sizes are small, decreasing as the sample size increases. Different procedures weight the gene-specific and combined elements differently[41–43], but all seem to work reasonably well[41]; future research should aim to find the optimal weighting. Some researchers also shrink the differential expression estimate itself[44], but this approach is less popular.

Consensus point 3: False-discovery rate is a good alternative to conventional multiple-testing approaches. Microarrays involve multiple testing — the testing of many hypotheses within a single study — which presents important challenges. Testing tens of thousands of transcripts is likely to produce hundreds of false positives if α-values that are commonly applied in other types of statistical analysis are

Parameter
A quantity (for example, mean) that characterizes some aspect of a (usually theoretically infinite) population.

Type 1 error
A false positive, or the rejection of a true null hypothesis; for example, declaring a gene to be differentially expressed when it is not.

Type 2 error
A false negative, or failing to reject a false null hypothesis; for example, not declaring a gene to be differentially expressed when it is.

Long-range error rate
The expected error rate if experiments and analyses of the type under consideration were repeated an infinite number of times.

t-tests
Statistical tests that are used to determine a statistically significant difference between two groups by looking at differences between two independent means.

ANOVA
Analysis of variance. A statistical test for determining differences in mean values between two or more groups.

Logistic regression
A regression technique that is used in cases where the outcome variable is binary (dichotomous).

© 2005 Nature Publishing Group

Survival analysis
A statistical methodology for analysing time-to-event data.

α-value
The nominal probability (set by the investigator) of making a type 1 error.

Bonferroni correction
A family-wise error rate (FWER) control procedure that sets the α-value level for each test and strongly controls the FWER for any dependency structure among the tests.

Bayesian probability
The probability of a proposition being true, which is conditional on the observed data.

Gene Ontology
A way of describing gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Null hypothesis
The hypothesis that is being tested in a statistical test. Typically in a microarray setting it is the hypothesis that states: there is no difference between gene-expression levels across groups or conditions.

p-value
The probability, were the null hypothesis true, of obtaining results that are as discrepant or more discrepant from those expected under the null hypothesis than those actually obtained.

Permutation test
A statistical hypothesis test in which some elements of the data are permuted (shuffled) to create multiple new pseudo-data sets. One then evaluates whether a statistic quantifying departure from the null hypothesis is greater in the observed data than a large proportion of the corresponding statistics calculated on the multiple pseudo-data sets.

Intersection-union tests
Multicomponent tests in which the compound null hypothesis consists of the union of two or more component null hypotheses.

used (for example, 0.05) (REF. 45). Methodologists initially reacted to this problem in a way that was reflexive, draconian and largely unresponsive to the goals of biologists by providing family-wise error rate (FWER) control methods such as the Bonferroni correction. These methods limit the probability of making one or more type 1 errors to less than the α-value across the entire experiment; however, most biologists seem willing to accept that some errors will occur, as long as this allows findings to be made. For example, an investigator might specify that it is acceptable for a small proportion of findings (for example, a maximum of 10%) to be wrong. Such an investigator is expressing interest in FDR control or estimation, not FWER.

This difference between the needs of biologists and the tools that are provided by methodologists ushered in a whole new approach to inference. Benjamini and Hochberg[46] first coined the term FDR, and provided a procedure for its control. Subsequently, much new methodology and accompanying jargon emerged[11,47,48] (for clarification of some of these terms see REF. 49). We introduced mixture-models, which treat genes as being composed of two or more populations — one represents those genes that are differentially expressed, and the other(s) those genes that are not differentially expressed[9]. Many related mixture-model methods (MMMs) were subsequently devised[50–52]. MMMs estimate FDRs for genes that are declared differentially expressed, whereas the original Benjamini and Hochberg approach controls the FDR at or below a certain level. Consequently, MMMs tend to be much more powerful. Although there are subtle distinctions between different MMM approaches[49], they all estimate a 'gene-specific' FDR that is interpreted as the Bayesian probability that a gene that is declared to be differentially expressed is a false positive.

FDR is equivalent or related to several other metrics that quantify the confidence we can have that any particular gene is differentially expressed[49,53–55]. The methods for estimating these quantities all seem to perform reasonably well under some circumstances. Therefore, there is a consensus that FDR-estimation procedures are preferred to both FDR- and FWER-controlling procedures, although there is no consensus as to which FDR-estimation procedure is best. However, limited evidence indicates that differences in performance might not be profound[52].

Finally, questions remain about accommodating dependence among genes in FDR estimation. This relates to the possibility that the amount of one transcript in a biological specimen might be related to the amount of other transcripts, and whether and how this should be tackled is unresolved.

Consensus point 4: Gene-class testing is desirable. Small sample sizes, coupled with efforts to maintain low FDRs, often result in low power to detect differential expression. Therefore, obtaining a long list of genes that can be confidently declared as differentially expressed is, initially, a triumph. However, this often subsequently leaves investigators bewildered by a myriad of unorganized findings. In response to the dual need to increase power to detect differential expression and to reduce the interpretive challenge that is posed by a long list of differentially expressed genes[56,57], gene-class testing (GCT) has become a popular and widely accepted analytical tool. Gene classes are usually based on Gene Ontology (GO) categories (for example, genes that are involved in organ growth, or genes that are involved in feeding behaviour), but alternatives exist[57]. Several GCT methods and software packages are available[58–61], most of which use statistical tests to compare the number of genes in a class that are 'significant' with the number that are expected under a particular null hypothesis.

One problem with GCT is that the null hypotheses that are tested are often poorly defined (if defined at all). There are also other unresolved issues. First, the typical reliance on a list of significant genes ignores the continuity of available evidence. Rather than using the continuous distribution of *p*-values, which quantifies the strength of evidence, GCT arbitrarily dichotomizes them at some threshold and loses information. To our knowledge, only two GCT methods (erminej[60,61] and GoDist[62]) use continuous evidence. Second, most methods treat the gene, rather than the case, as the unit of analysis, and this is also the case when such methods are used in permutation testing (see later). This is inappropriate for several reasons, including the fact that it assumes that transcripts are expressed independently[63]. We know of only one method — gene-set-enrichment analysis (GSEA)[56] — that correctly permutes across cases.

Other unresolved issues include how to handle multiple testing[58] and the curious fact that many approaches (for example, GSEA[56]) 'penalize' some gene classes when other gene classes are highly differentially expressed in the same data set[64]. This occurs in a 'zero-sum-game' manner — that is, gene classes are pitted against one another such that the stronger the evidence in support of differential expression is for one class, the weaker the evidence for differential expression is judged to be for a second class. This occurs even though the data for the second class have not changed.

In summary, we believe GCT is valuable, although all current approaches suffer from at least one major flaw. GSEA[56] and erminej[60,61] suffer the least from these problems and merit use, and improved methods are likely to become available soon.

Remaining question 1: How should intersections between sets of findings be assessed? Some of the issues surrounding the testing of multiple hypotheses in microarray analysis have begun to be addressed. Intersection-union testing (IUT) is useful when asking 'and' or 'all' questions, such as which genes are differentially expressed or correlated with each other in all tissues analysed (or all species analysed, or all conditions analysed). For example, Persson *et al.*[65] sought to identify *Arabidopsis thaliana* genes for which expression correlated with all genes that are known to be involved in cell-wall formation.

A typical IUT approach was used in a study by Kyng *et al.*[66] The authors evaluated differential expression

© 2005 Nature Publishing Group

between cell lines from normal young and old people, and also between those from normal young people and young people with Werner syndrome (WRN, also known as WS, which is an accelerated ageing disease). They reported that "alterations in WS were strikingly similar to those in normal aging: 91% of annotated genes displayed similar expression changes in WS and in normal aging, 3% were unique to WS, and 6% were unique to normal aging." To determine whether the degree of overlap in genes that was found to be differentially expressed in the WRN versus young comparison and the old versus young comparison was significantly greater than expected by chance, it would be tempting to treat the gene as the unit of analysis. A chi-square test of independence of two dichotomous variables could then be calculated, which would reveal whether a gene was differentially expressed in old versus young, and whether it was differentially expressed in the normal young versus the WRN young group. This would probably produce a small p-value.

However, there are at least three problems with this approach. First, it is valid only if all transcript levels are independent, apart from the effects that are putatively induced by independent variables (in this case, age and WRN). However, this is unlikely to be true. Second, this approach ignores the fact that both old and WRN cell lines are compared with the same young cell lines. Therefore, even if there were no true population-level correlation in the degree of differential expression across the two comparisons, a sample correlation would still exist given the common comparitor[67]. This will make it seem that there is a significantly greater consistency of effects across the two comparisons than is expected by chance because the expectation will be erroneous.

Finally, by imposing cut offs for declaring differential expression, the approach described above ignores the continuity of the available evidence. If the classic IUT min-test[68] was used, a gene with p-values of 0.009 for both the old versus young comparison and the WRN versus young comparison would be declared differentially expressed at the 0.01 level. The same would apply to a second gene with p-values of 0.009 for the old versus young comparison and 10^{-20} for the WRN versus young comparison. The min-test offers no ability to distinguish between these different degrees of evidence and therefore does not yield a distribution of p-values that will be suitable for FDR-estimation procedures. These limitations have led us to consider Bayesian approaches in which posterior probabilities rather than p-values are used to quantify evidence against the union of several null hypotheses (K. Kim *et al.*, personal communication).

Remaining question 2: How should computationally intensive resampling-based inference be used? Many inference methods that are used for microarray analysis are parametric tests, which rely on specific assumptions about the distribution of the variables studied, and derive properties of theoretical distributions to make inferences. By contrast, instead of using parametric tests, resampling-based inference (RBI) methods rely on resampling data. Compared with parametric testing, RBI has the advantage of being robust and flexible enough to accommodate almost any new statistic (for example, the statistic that is obtained after shrinkage of variance), without the need for methodologists to mathematically derive a statistic's distribution. RBI has the disadvantage of being computationally intensive, but with modern computational tools it is now feasible in most cases and is widely used[69–73]. However, there are marked differences in how such approaches are implemented, and some confusion and uncertainty remains. Microarray investigators who use RBI rarely discuss these issues or state why one RBI approach (for example, bootstrap analysis or permutation testing) is chosen over another. Different RBI procedures can yield markedly different results in two-group microarray studies[74], so choice of procedure is important.

Problems, which are often unrecognized, can arise when using RBI for complex experiments. For example, consider the use of permutation testing in an experiment to test the difference between two groups (for example, old and young mice) after controlling for some other factor (for example, body fat). There are several ways to permute the data in such circumstances, and only some will produce valid inferences[75]. Another issue is the sampling unit; a common error is to treat the gene rather than the case as the unit of analysis (for example, in GCT). This type of resampling effectively ignores both sample size and non-independence across genes, and can result in completely nonsensical results (for example, tests in which power does not increase with sample size).

Another issue that faces RBI is that, because of the small samples that are typically used in microarray experiments, RBI p-value distributions can be coarse or 'granular', and it will often be algebraically impossible to obtain p-values that are below some specified level[76]. This greatly reduces the ability to follow RBI with the popular FDR procedures that are described above. To overcome this problem some software (for example, the SAM algorithm[42]) combines all resampled test statistics across all genes to obtain very small p-values. This is based on two assumptions: that the null distribution of the test statistic is the same for all transcripts; and that all transcripts are independent. Unfortunately, neither assumption is necessarily correct. Therefore, some software offers a choice of whether the resampled test statistics are combined across genes[41]. Consequently, how to obtain the benefit of combining RBI statistics across transcripts without requiring the two assumptions is an important question meriting research.

Classification

The process of classification entails either placing objects (for example, genes) into pre-existing categories (supervised classification), or developing a set of categories into which objects can subsequently be placed (unsupervised classification). Many classification algorithms are extensively used in microarray research (BOXES 1,3).

Chi-square test of independence
A test of the independence of two categorical variables that is based on the chi-square distribution. The test is valid only under the assumption that all cases are independent.

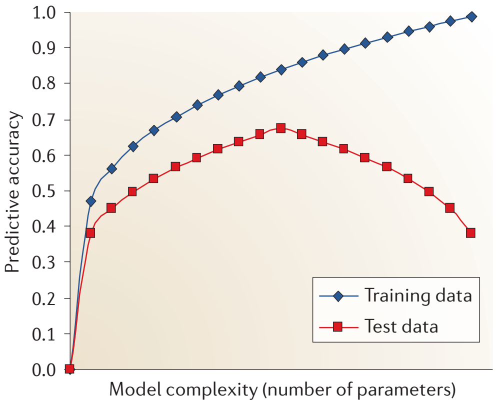
Min-test
A statistical IUT test in which the union of a null hypotheses is rejected if, and only if, for each component null hypothesis the p-value $<\alpha$.

Posterior probability
The Bayesian probability that a hypothesis is correct, which is conditional on the observed data.

Bootstrap analysis
A form of computer-intensive resampling-based inference. Pseudo-data sets are created by sampling from the observed data with replacement (that is, after a case is resampled, it is returned to the original data and can, potentially, be drawn again).

© 2005 Nature Publishing Group

Box 3 | Classification



Classification algorithms are used either to discover new categories within a data set (class discovery; unsupervised classification) or assign cases to a given category (class prediction; supervised classification).

Supervised classification

Supervised classification (often called 'class assignment', 'prediction' or 'discrimination') entails developing algorithms to assign objects to *a priori*-defined categories. Algorithms are typically developed and evaluated on a 'training' data set and an independent 'test' data set, respectively, in which the categories to which objects belong are known before they are used in practical applications. Many supervised classification algorithms are available, but all are susceptible to overfitting to some degree. The phenomenon of overfitting is shown in the figure, which shows the effect of the complexity of the model used on its predictive accuracy. The smaller the sample and the larger the number of transcripts modelled, the more algorithms will capitalize on chance sample patterns and obtain predictive functions that perform well with training data but poorly with new data. The great challenge is to determine the optimal degree of model complexity that a given data set can support. A common misconception is that the set of the most differentially expressed genes will necessarily give the best predictive accuracy. The gene list that is obtained from hypothesis testing does not necessarily give the best prediction. No one method for constructing prediction algorithms is widely accepted as superior or optimal. However, experience and intuition suggest that with the sample sizes that are typically available in microarray studies, simpler methods might out-perform more complex approaches.

Unsupervised classification

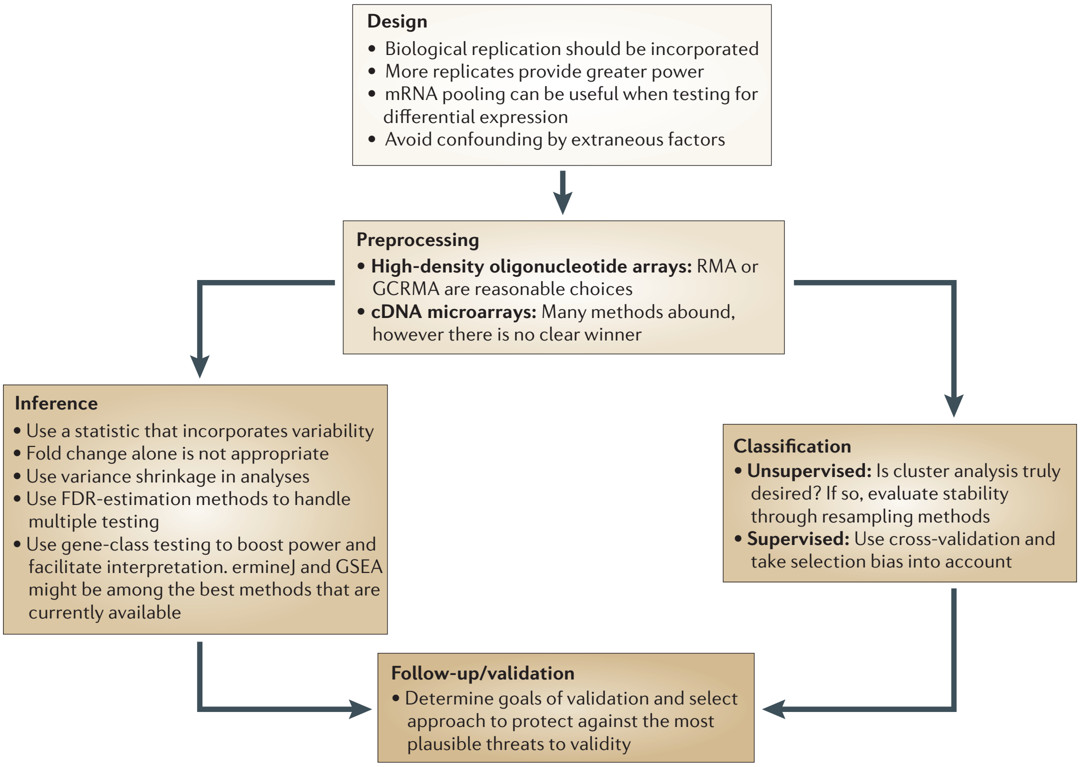
Algorithms for unsupervised classification or cluster analysis abound. Unfortunately however, algorithm development seems to be a preferred activity to algorithm evaluation among methodologists. Cluster-analysis algorithms group objects on the basis of some sort of similarity metric that is computed for one or more 'features' or variables. For example, genes (biological objects) can be grouped into classes on the basis of the similarity in their expression profiles across tissues, cases or conditions. Hierarchical cluster analysis graphically presents results in a tree diagram (dendrogram), and is probably the most common unsupervised classification algorithm in microarray analysis. Non-hierarchical clustering methods divide the cases (samples or genes) into a predetermined number of groups in a manner that maximizes a specific function (for example, the ratio of variability between and within clusters). Cluster-analysis approaches entail making several choices, such as which metric to use to quantify the distance or similarity among pairs of objects, what criteria to optimize in determining the cluster solution, and how many clusters to include in the solution. No consensus or clear guidelines exist to guide these decisions. Cluster analysis always produces clustering, but whether a pattern observed in the sample data characterizes a pattern present in the population remains an open question. Resampling-based methods can address this last point, but results indicate that most clusterings in microarray data sets are unlikely to reflect reproducible patterns or patterns in the overall population[18].

Consensus point 1: Unsupervised classification is overused. Unsupervised classification was one of the first statistical techniques to be applied to microarray analysis, and is one of the most popular. Its popularity is understandable: no hypotheses and almost no data assumptions are required, but the researcher is guaranteed to obtain a clustering of genes, irrespective of sample size, data quality or experimental design — or indeed any biological validity that is associated with the clustering.

We believe that unsupervised classification is overused; first, little information is available about the absolute validity or relative merits of clustering procedures[77,78]; second, the evidence indicates that the clusterings that are produced with typical sample sizes (<50) are generally not reproducible[18,79,80]; third, and most importantly, unsupervised classification rarely seems to address the questions that are asked by biologists, who are usually interested in identifying differential expression. However, it is important to note that there might be cases where clustering is warranted — for example, if the goal is to simply obtain a general description of how genes covary with respect to their gene-expression levels within a population.

Consensus point 2: Unsupervised classification should be validated using resampling-based procedures. In situations where unsupervised classification is warranted, some reproducibility measure should be provided. Standard unsupervised-classification procedures provide no information about the extent to which the results reflect a pattern that exists in the population rather than random sampling variation[81]. A consensus has emerged that resampling techniques can assess the reproducibility of unsupervised classification and should be used[18,78,81–84]. In these procedures, subsets are resampled from the original sample, unsupervised classification is applied, and the consistency of results

© 2005 Nature Publishing Group



Design

- Biological replication should be incorporated
- More replicates provide greater power
- mRNA pooling can be useful when testing for differential expression
- Avoid confounding by extraneous factors

Preprocessing

- **High-density oligonucleotide arrays:** RMA or GCRMA are reasonable choices
- **cDNA microarrays:** Many methods abound, however there is no clear winner

Inference

- Use a statistic that incorporates variability
- Fold change alone is not appropriate
- Use variance shrinkage in analyses
- Use FDR-estimation methods to handle multiple testing
- Use gene-class testing to boost power and facilitate interpretation. ermineJ and GSEA might be among the best methods that are currently available

Classification

- **Unsupervised:** Is cluster analysis truly desired? If so, evaluate stability through resampling methods
- **Supervised:** Use cross-validation and take selection bias into account

Follow-up/validation

- Determine goals of validation and select approach to protect against the most plausible threats to validity

Figure 2 | **Guidelines for the statistical analysis of microarray experiments.** The flow chart indicates the guidelines for each relevant stage of a microarray study. FDR, false-discovery rate; RMA, robust multi-array average (includes a modification, CGRMA).

across the resamples is quantified. Those procedures that resample at the level of the case — rather than the gene — all perform reasonably well, and none is widely recognized as the best.

Consensus point 3: Supervised-classification procedures require independent cross-validation. In supervised classification, the aim is to obtain a function or rule that uses expression data to predict whether a case is of one type or another (for example, drought-resistant versus non-drought-resistant). A computer algorithm finds the rule that best classifies a set of available cases for which the correct type is known. Because of this attempted optimization, overfitting is a concern (BOX 3). To estimate how well the rule will perform on fresh data, one must cross-validate it on test data that are completely independent from the data from which the classification rule was derived, and there are many approaches to this[85].

One key point is the need to avoid selection bias. This requires cross-validation procedures that separate the validation data from all aspects of the rule-derivation process, including the selection of initial transcripts to include in the model[86–88]. Early microarray papers failed to account for selection bias and thereby radically overestimated prediction accuracy[86]. Of course, effective cross-validation requires an adequate sample size, and methods for estimating sample sizes for supervised classification studies have been developed[89–91].

Validation

Many researchers have called for the 'validation' of microarray findings (for example, see REF. 92). But exactly when and how should validation be carried out[93], which error types should be protected against, and what are the criteria by which one can say that a finding has been validated? Here we consider only the operational validation of efforts to detect differentially expressed genes, not constructive validation. Operational validation can arise from at least two potential sources of error: measurement error and sampling error.

Remaining question 1: Is it necessary to validate against measurement error? For genes that are not declared differentially expressed, it is possible that random measurement error has reduced the ability to detect true differences and has produced an erroneous inference. So, one could argue that the genes that should be 'validated' are those for which the test statistic was almost significant. This could be done by using a more precise gene-expression measure and/or a larger sample size. It makes sense to do this because random measurement error does not bias results away from, but only towards, the null hypothesis.

However, this strategy is not generally used. Investigators seem more concerned about false positives, and might therefore seek to 'validate' significant results by taking fresh aliquots from the same specimens using a different mRNA-measurement procedure (for example,

Sampling variation
The variability in statistics that occurs among random samples from the same population and is due solely to the process of random sampling.

Overfitting
This occurs when an excessively complex model with too many parameters is developed from a small sample of 'training' data. The model fits those data well, but does so by capitalizing on chance variations and, therefore, will fit a fresh set 'test' data poorly.

Selection bias
This occurs when the prediction accuracy of a rule is estimated using cases that had some role in the derivation of the rule. It is an upward bias — that is, one that overestimates the predictive accuracy.

Operational validation
Re-testing a hypothesis using the original methodology (also referred to as operational replication).

© 2005 Nature Publishing Group

Box 4 | **Recommendations for future microarray analysis methods**

- Empirical evidence about the performance of pooling is based on a small number of experiments in a few model organisms. Extensions to other situations and species are recommended.
- More research is needed on how to best examine intersections between sets of findings and evaluate complex multi-component hypotheses. Bayesian approaches might be especially advantageous here.
- More careful consideration of how to best use resampling-based inference is warranted, as are guidelines for its use for the field. If and how the vast number of genes assayed in microarray experiments can be used to partially compensate for small sample sizes when using resampling-based inference requires further study.
- Methods for microarray quality-control assessment are needed, as are approaches to the validation of such methods.
- For all statistical procedures, the fact that transcripts are not necessarily independent should be considered. The potential impact of this on the performance of procedures should be assessed, and ways to accommodate this are needed.
- Whether microarrays require any validation guidelines that are fundamentally different from other types of study deserves questioning. If this is found to be the case, the exact goals of such validation should be defined, and an assessment of which procedures will meet those goals will be needed.
- Development of standardized testing platforms, similar to AffyComp, for platforms other than Affymetrix would be useful.
- In many areas (for example, cluster-analysis algorithms, normalization algorithms and false-discovery-rate estimation procedures), the need for thoroughly evaluating existing techniques currently seems to outweigh the need to develop new techniques.
- Well-curated publicly available archives of plasmode data sets are needed to test the validity of various methods.

reverse-transcriptase PCR). Assuming that this is more accurate than the initial measurement procedure (which is debatable), this might protect against erroneous inferences that are due to poor measurement quality. However, there is no reason to suspect that measurement errors cause false positives unless the measurement error in microarrays is not random but systematic. Moreover, it would have to be systematic in a manner such that, under the null hypothesis, the measurement errors are correlated with the study's independent variable (for example, different treatment of samples). Therefore, for genes that are declared to be differentially expressed, we believe repeating measurements on aliquots from the same biological specimen with a new measurement technique is a highly questionable practice that stems more from tradition than careful thought.

Remaining question 2: How should validation against sampling variability be conducted? Sometimes, by chance alone, a sample from one population will have a substantially different mean level of expression for a particular transcript than a sample from another population, even though the two populations do not differ in terms of the mean expression of that transcript. This is more likely to occur when many transcripts are tested. Such an event would be considered a type 1 error that is due to random sampling variability. The evaluation of whether specific significant results obtained in microarray studies are actually false positives can be carried out by using new cases to test for differential expression using the same experimental model that was used in the initial experiment. Repeating the analyses on the same biological specimens with a new measurement procedure would not help. We believe that this distinction is not generally made in the literature.

Constructive validation
Testing a hypothesis through a different methodology (also referred to as constructive replication).

Remaining question 3: What are the criteria under which a finding can be said to be validated? What result must be obtained in a second set of observations to state that one has validated a result obtained in a first set of observations? Must a particular p-value be achieved and, if so, which value? Is it sufficient that the effect-size obtained in the second set is not significantly different from that obtained in the first? If so, must a certain level of power be obtained before the conclusion that there is no difference between results is accepted? Investigators are only just beginning to address these important questions about validation, which methodologists and applied researchers will need to address in the future.

Implications and future directions

We have attempted to distil the statistical literature to provide a guide to choosing methods at each stage of microarray analysis (FIG. 2) and identifying unresolved issues that merit further research (BOX 4). The many current methods can be reduced to a modest number of categories, and there is often little evidence to support one method within a category over others. In such situations, although investigators should be encouraged to pick a method from within an important category, there should be no dogma about which methods are chosen until further data emerge. Methodologists should now focus as much on assessing the comparative merits of various procedures within classes as on developing new

© 2005 Nature Publishing Group

procedures. Moreover, the need for new procedures seems highly questionable in those categories in which there are already several options, whereas it is acute in other categories for which few valid alternatives are available. We hope that our discussion serves to highlight those areas in which further research is needed, as well as those in which consensus has been reached.

1. Kerr, M. K. Design considerations for efficient and effective microarray studies. *Biometrics* **59**, 822–828 (2003).
2. Page, G. P., Edwards, J. W., Barnes, S., Weindruch, R. & Allison, D. B. A design and statistical perspective on microarray gene expression studies in nutrition: the need for playful creativity and scientific hard-mindedness. *Nutrition* **19**, 997–1000 (2003).
3. Yang, M. C., Yang, J. J., McIndoe, R. A. & She, J. X. Microarray experimental design: power and sample size considerations. *Physiol. Genomics* **16**, 24–28 (2003).
4. Kerr, M. K. & Churchill, G. A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201 (2001).
5. Dobbin, K., Shih, J. H. & Simon, R. Statistical design of reverse dye microarrays. *Bioinformatics* **19**, 803–810 (2003).
6. Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genet.* **32**, S490–S495 (2002).
7. Yang, Y. H. & Speed, T. Design issues for cDNA microarray experiments. *Nature Rev. Genet.* **3**, 579–588 (2002).
8. Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F. & Pi-Sunyer, F. X. Power and money: designing statistically powerful studies while minimizing financial costs. *Psychol. Methods* **2**, 20–33 (1997).
9. Allison, D. B. *et al.* A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Analysis* **39**, 1–20 (2002).
This was the first paper in the field of microarray research to introduce mixture modelling.
10. Pavlidis, P., Li, Q. & Noble, W. S. The effect of replication on gene expression microarray experiments. *Bioinformatics* **19**, 1620–1627 (2003).
11. Tsai, C. A., Hsueh, H. M. & Chen, J. J. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* **59**, 1071–1081 (2003).
12. Pan, W., Lin, J. & Le, C. T. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.* **3**, research0022 (2002).
13. Zien, A., Fluck, J., Zimmer, R. & Lengauer, T. Microarrays: how many do you need? *J. Comput. Biol.* **10**, 653–667 (2003).
14. Gadbury, G. L. *et al.* Power analysis and sample size estimation in the age of high dimensional biology: a parametric bootstrap approach and examples from microarray research. *Stat. Methods Med. Res.* **13**, 325–338 (2004).
This paper offers convenient FDR-based methods for power analysis and sample-size estimation in microarray and other high-dimensional testing situations.
15. Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. & Ploner, A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21**, 3017–3024 (2005).
16. Muller, P., Parmigiani, G., Robert, C. & Rousseau, J. Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Am. Stat. Assoc.* **99**, 990–1001 (2004).
17. Dobbin, K. & Simon, R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics.* **6**, 27–38 (2005).
18. Garge, N., Page, G. P., Spargue, A. P., Gorman, B. S. & Allison, D. B. Reproducible clusters from microarray research: whither? *BMC Bioinformatics* **6** (Suppl. 2), S10 (2005).
The authors evaluate clustering techniques using real data, and find that with sample sizes of less than 50, the reproducibility of results is poor.
19. Kendziorski, C. M., Zhang, Y., Lan, H. & Attie, A. D. The efficiency of pooling mRNA in microarray experiments. *Biostatistics* **4**, 465–477 (2003).
This paper clarifies concepts and statistical design issues that are involved with mRNA pooling in microarray experiments.
20. Kendziorski, C., Irizarry, R. A., Chen, K. S., Haag, J. D. & Gould, M. N. On the utility of pooling biological samples in microarray experiments. *Proc. Natl Acad. Sci. USA* **102**, 4252–4257 (2005).
21. Chen, Y., Dougherty, E. R. & Bittner, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364–374 (1997).
22. Schadt, E. E., Li, C., Ellis, B. & Wong, W. H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem.* **Suppl. 37**, 120–125 (2001).
23. Ekstrom, C. T., Bak, S., Kristensen, C. & Rudemo, M. Spot shape modelling and data transformations for microarrays. *Bioinformatics* **20**, 2270–2278 (2004).
24. Steinfath, M. *et al.* Automated image analysis for array hybridization experiments. *Bioinformatics* **17**, 634–641 (2001).
25. Yang, Y. H., Buckley, M. J. & Speed, T. P. Analysis of cDNA microarray images. *Brief Bioinform.* **2**, 341–349 (2001).
26. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32**, 496–501 (2002).
27. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
28. Smyth, G. K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265–273 (2003).
29. Qin, L. X. & Kerr, K. F. Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.* **32**, 5471–5479 (2004).
This article presents the effect of different image-processing and normalization techniques on microarray analysis conclusions.
30. Affymetrix. *Affymetrix Expression Analysis Technical Manual* (Affymetrix, Santa Clara, California, 2004).
31. Nielsen, H. B., Gautier, L. & Knudsen, S. Implementation of a gene expression index calculation method based on the PDNN model. *Bioinformatics* **21**, 687–688 (2005).
32. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
33. Mehta, T., Tanik, M. & Allison, D. B. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genet.* **36**, 943–947 (2004).
This paper clarifies the importance of methods for evaluating the validity of proposed statistical methodologies in high-dimensional biology, with an emphasis on microarray research.
34. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
35. Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* **6**, R16 (2005).
36. Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. & Speed, T. P. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**, 323–331 (2004).
37. Chen, D. T. A graphical approach for quality control of oligonucleotide array data. *J. Biopharm. Stat.* **14**, 591–606 (2004).
38. Hsiao, A., Worrall, D. S., Olefsky, J. M. & Subramaniam, S. Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. *Bioinformatics* **20**, 3108–3127 (2004).
39. Miller, R. A., Galecki, A. & Shmookler-Reis, R. J. Interpretation, design, and analysis of gene array expression experiments. *J. Gerontol. A* **56**, B52–B57 (2001).
40. Budhraja, V., Spitznagel, E., Schaiff, W. T. & Sadovsky, Y. Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays. *BMC Biol.* **1**, 1 (2003).
41. Cui, X., Hwang, J. T., Qiu, J., Blades, N. J. & Churchill, G. A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75 (2005).
This article provides one method of shrinkage and compares its performance with other variance shrinkage methods.
42. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci USA* **98**, 5116–5121 (2001).
43. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
44. Edwards, J. W. *et al.* Empirical Bayes estimation of gene-specific effects in micro-array research. *Funct. Integr. Genomics* **5**, 32–39 (2005).
45. Ge, Y. C., Dudoit, S. & Speed, T. P. Resampling-based multiple testing for microarray data analysis. *Test* **12**, 1–77 (2003).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate — a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
47. Hsueh, H. M., Chen, J. J. & Kodell, R. L. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Stat.* **13**, 675–689 (2003).
48. van der Lann, M. J., Dudoit, S. & Pollard, K. S. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* **3**, A15 (2004).
49. Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the *q*-value. *Ann. Stat.* **31**, 2013–2035 (2003).
This paper clarifies the key terminology and concepts used in FDR-related methods.
50. Do, K. A., Mueller, P. & Tang, F. A nonparametric Bayesian mixture model for gene expression. *J. R. Stat. Soc. Ser. C* **54**, 1–18 (2005).
51. Pounds, S. & Morris, S. W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics* **19**, 1236–1242 (2003).
52. Datta, S. & Datta, S. Empirical Bayes screening of many *p*-values with applications to microarray studies. *Bioinformatics* **21**, 1987–1994 (2005).
53. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. G. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
54. Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176 (2004).
55. Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K. W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8**, 37–52 (2001).
56. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
57. Osier, M. V. in *DNA Microarrays and Statistical Genomic Techniques: Design, Analysis, and Interpretation of Experiments* (Marcel Dekker, New York, 2005).
58. Osier, M. V., Zhao, H. & Cheung, K. H. Handling multiple testing while interpreting microarrays with the Gene Ontology Database. *BMC Bioinformatics* **5**, 124 (2004).

© 2005 Nature Publishing Group

59. Khatri, P., Draghici, S., Ostermeier, G. C. & Krawetz, S. A. Profiling gene expression using onto-express. *Genomics* **79**, 266–270 (2002).
60. Pavlidis, P., Weston, J., Cai, J. & Noble, W. S. Learning gene functional classifications from multiple data types. *J. Comput. Biol.* **9**, 401–411 (2002).
61. Pavlidis, P., Qin, J., Arango, V., Mann, J. J. & Sibille, E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.* **29**, 1213–1222 (2004). **This study introduces a gene-class testing method that uses the full continuous evidence that is available within *p*-values.**
62. Ben Shaul, Y., Bergman, H. & Soreq, H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics* **21**, 1129–1137 (2005).
63. Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
64. Damian, D. & Gorfine, M. Statistical concerns about the GSEA procedure. *Nature Genet.* **36**, 663 (2004).
65. Persson, S., Wei, H., Milne, J., Page, G. P. & Somerville, C. R. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl Acad. Sci. USA* **102**, 8633–8638 (2005).
66. Kyng, K. J., May, A., Kolvraa, S. & Bohr, V. A. Gene expression profiling in Werner syndrome closely resembles that of normal aging. *Proc. Natl Acad. Sci. USA* **100**, 12259–12264 (2003).
67. Schmid, C. H., Lau, J., McIntosh, M. W. & Cappelleri, J. C. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat. Med.* **17**, 1923–1942 (1998).
68. Berger, R. L. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295–300 (1982).
69. Neuhauser, M., Boes, T. & Jockel, K. H. Two-part permutation tests for DNA methylation and microarray data. *BMC Bioinformatics* **6**, 35 (2005).
70. Barry, W. T., Nobel, A. B. & Wright, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943–1949 (2005).
71. Pan, W. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* **19**, 1333–1340 (2003).
72. Xu, R. H. & Li, X. C. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* **19**, 1284–1289 (2003).
73. Landgrebe, J., Wurst, W. & Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol.* **3**, RESEARCH0019 (2002).
74. Troendle, J. F., Korn, E. L. & McShame, L. M. An example of slow convergence of the bootstrap in high dimensions. *Am. Stat.* **58**, 25–29 (2004). **This presents an excellent overview of the nuances of resampling methodology that is used in microarray research, and discusses the fact that such methods are not assumption-free panaceas that are valid under all circumstances.**
75. Kennedy, P. E. & Cade, B. S. Randomization tests for multiple regression. *Commun. Stat.* **25**, 923–936 (1996).
76. Gadbury, G. L., Page, G. P., Heo, M., Mountz, J. D. & Allison, D. B. Randomization tests for small samples: an application for genetic expression data. *J. R. Stat. Soc. Ser. C* **52**, 365–376 (2003).
77. Yeung, K. Y., Haynor, D. R. & Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics* **17**, 309–318 (2001).
78. Datta, S. & Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**, 459–466 (2003).
79. Shih, J. H. *et al.* Effects of pooling mRNA in microarray class comparisons. *Bioinformatics* **20**, 3318–3325 (2004).
80. Yeung, K. Y., Medvedovic, M. & Bumgarner, R. E. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol.* **5**, R48 (2004).
81. Bryan, J. Problems in gene clustering based on gene expression data. *J. Multivariate Analysis* **90**, 44–66 (2004). **This is an excellent overview of the methodological and conceptual challenges in the use of cluster analysis in gene-expression studies.**
82. Kerr, M. K. & Churchill, G. A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA* **98**, 8961–8965 (2001).
83. Zhang, K. & Zhao, H. Assessing reliability of gene clusters from gene expression data. *Funct. Integr. Genomics* **1**, 156–173 (2000).
84. Tseng, G. C. & Wong, W. H. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**, 10–16 (2005).
85. Horth, J. *Computer Intensive Statistical Methods Validation, Model Selection and Boostrap* (Chapman and Hall, London, 1994).
86. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002). **This article addresses selection bias in the context of predictive error-estimation and cross-validation for microarray studies.**
87. Furlanello, C., Serafini, M., Merler, S. & Jurman, G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics* **4**, 54 (2003).
88. Fu, W. J., Carroll, R. J. & Wang, S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21**, 1979–1986 (2005).
89. Dobbin, K. & Simon, R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* **6**, 27–38 (2005).
90. Hwang, D., Schmitt, W. A., Stephanopoulos, G. & Stephanopoulos, G. Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics* **18**, 1184–1193 (2002).
91. Mukherjee, S. *et al.* Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.* **10**, 119–142 (2003).
92. Rajeevan, M. S., Ranamukhaarachchi, D. G., Vernon, S. D. & Unger, E. R. Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. *Methods* **25**, 443–451 (2001).
93. Rockett, J. C. & Hellmann, G. M. Confirming microarray data — is it really necessary? *Genomics* **83**, 541–549 (2004).
94. Rocke, D. M. & Durbin, B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**, 966–972 (2003).
95. Pounds, S. & Cheng, C. Statistical development and evaluation of microarray gene expression data filters. *J. Comput. Biol.* **12**, 482–495 (2005).

Acknowledgements

The authors are supported in part by grants from the US National Institutes of Health, National Science Foundation and Department of Defense. We are grateful to C. Kendziorski for her helpful discussion on this document.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Werner syndrome

FURTHER INFORMATION

AffyComp II software: http://affycomp.biostat.jhsph.edu
A free online microarray analysis course from the University of Alabama at Birmingham: http://www.soph.uab.edu/ssg_content.asp?id=1410
ArrayExpress microarray data repository: http://www.ebi.ac.uk/arrayexpress
BioConductor open source software for bioinformatics: http://www.bioconductor.org
Cyber-T statistics program: http://visitor.ics.uci.edu/genex/cybert/index.shtml
ermineJ — Gene Ontology analysis for microarry data: http://microarray.genomecenter.columbia.edu/ermineJ
Gene Expression Omnibus data repository: www.ncbi.nlm.nih.gov/geo
Gene Ontology Database: www.geneontology.org
HDBStat*!* High Dimension Biology Statistical analysis software: http://www.soph.uab.edu/ssg_content.asp?id=1164
MAANOVA 2.0 software: http://www.jax.org/staff/churchill/labsite/software/anova
PowerAtlas software: www.poweratlas.org
Stanford MicroArray Database: http://genome-www5.stanford.edu
Access to this interactive links box is free online.

© 2005 Nature Publishing Group