DNA AND CELL BIOLOGY
Volume 22, Number 6, 2003

Pp. 357–394

The Design and Analysis of Microarray Experiments: Applications in Parasitology

DAVID A. MORRISON[1] and JOHN T. ELLIS[2]

ABSTRACT

Microarray experiments can generate enormous amounts of data, but large datasets are usually inherently complex, and the relevant information they contain can be difficult to extract. For the practicing biologist, we provide an overview of what we believe to be the most important issues that need to be addressed when dealing with microarray data. In a microarray experiment we are simply trying to identify which genes are the most "interesting" in terms of our experimental question, and these will usually be those that are either overexpressed or underexpressed (upregulated or downregulated) under the experimental conditions. Analysis of the data to find these genes involves first preprocessing of the raw data for quality control, including filtering of the data (e.g., detection of outlying values) followed by standardization of the data (i.e., making the data uniformly comparable throughout the dataset). This is followed by the formal quantitative analysis of the data, which will involve either statistical hypothesis testing or multivariate pattern recognition. Statistical hypothesis testing is the usual approach to "class comparison," where several experimental groups are being directly compared. The best approach to this problem is to use analysis of variance, although issues related to multiple hypothesis testing and probability estimation still need to be evaluated. Pattern recognition can involve "class prediction," for which a range of supervised multivariate techniques are available, or "class discovery," for which an even broader range of unsupervised multivariate techniques have been developed. Each technique has its own limitations, which need to be kept in mind when making a choice from among them. To put these ideas in context, we provide a detailed examination of two specific examples of the analysis of microarray data, both from parasitology, covering many of the most important points raised.

INTRODUCTION

THE RECENT PUBLICATION of the malarial genome sequence (Gardner *et al.*, 2002), along with the current efforts to sequence the genomes of many other major protozoan pathogens, represents an important historical landmark (reviewed by Ellis *et al.*, 2003). One of the most important outcomes of this era is the opportunity to define and understand the entire genetic organization of these taxa and to analyze gene expression during their complex life cycles. The enormous amount of sequence data being put into the public domain is daunting; nevertheless, it requires, if not demands, attention be given to methods of data analysis.

Microarray analysis is one technology available that allows the potential to profile gene expression on a genome-wide scale (Schena *et al.*, 1995), and only recently has it been applied to the analysis of gene expression in parasitic organisms (Hayward *et al.*, 2000; Blader *et al.*, 2001; Mamoun *et al.*, 2001; Matrajt *et al.*, 2002; Rathod *et al.*, 2002; Singh *et al.*, 2002). Microarray technology is attractive for application to parasitology for several reasons. In the first instance, the increasing amount of sequence data available facilitates the production of microarrays. Second, the ability to analyze changes in gene expression during transitions from different life-cycle stages of parasites presents the opportunity to view globally the nature of the changes occurring, and so facilitates the raising of hypotheses to explain the differentiation pathways. Third, the potential to analyze both parasite and host responses under the same experimental conditions represents an unsurpassed method to investigate parasite–host interactions.

Microarray experiments can thus generate enormous amounts of data, which is a relatively recent phenomenon in biology but has been traditionally more common in fields such as physics and chemistry. Unfortunately, this potential bounty

[1]Department of Parasitology (SWEPAR), National Veterinary Institute and Swedish University of Agricultural Sciences, Uppsala, Sweden.
[2]Institute for the Biotechnology of Infectious Diseases, University of Technology, Sydney, Gore Hill NSW, Australia.

comes at a cost. Large datasets are usually inherently complex, and the relevant information they contain can therefore be difficult to extract. More to the point, biologists are often not experts in mathematics (in many cases that is admittedly why they became biologists in the first place), and they are therefore prone to either misuse or even abuse mathematical analyses. Analysis of microarray data is thus a potential minefield, with the possibility of the data analyses being too complex to either perform properly or to understand. There may be as much thinking and time required for data analysis as there is for all other parts of the experiment put together.

Many biologists seem to want data analysis to be like a laboratory protocol—a series of steps that, if followed faithfully, guarantee to produce the correct answer to their experimental question. Unfortunately, that is rarely possible. Data analysis involves detecting and displaying whatever patterns are present in the data, and you don't necessarily know exactly what these patterns are beforehand.

What is needed is the most suitable method for highlighting the patterns in each particular situation. So, data analysis can often be a form of trial and error, where several potentially appropriate analyses are tried and their results are evaluated. It can be a very big mistake habitually to follow a series of predefined data analysis steps, as this will only be an effective strategy if they just happen to reveal the true answer.

However, it is not all gloom and doom. In a microarray experiment we are simply trying to identify which genes are the most "interesting" in terms of our experimental question, and these will usually be those that are either overexpressed or underexpressed (upregulated or downregulated) under the experimental conditions being studied. Conceptually, therefore, there is nothing new in the analysis of microarray data. It is always worthwhile for biologists to learn as much as possible about their data analyses (Leung, 2002), and so here we provide an overview of what we believe to be the most important issues for dealing with microarray data.

Our aim is to provide an introduction to each of a broad range of topics that have arisen in the data analysis of microarray studies, hoping to put them into context for you. Therefore, we do not provide a lot of details about the various methods that we cover, referring instead to the pertinent primary literature. A more detailed discussion of many of the issues can be found in the introductory books by Knudsen (2002) and Baldi and Hatfield (2002), with a briefer overview provided by Smyth *et al.* (2002). Several important reviews about specific aspects of microarray data analysis are mentioned at the appropriate places below. Two specific examples of the analysis of microarray data, covering many of the most important points raised in the following sections, are provided in the final section. Most of what we have to say applies to both complementary DNA arrays and oligonucleotide arrays (see Tefferi *et al.*, 2002), but we concentrate on the former technology.

The analysis of gene expression data collected using microarrays involves two distinct issues: (1) preprocessing of the raw data—this answers the question "How do I ensure that I have high-quality data?"; and (2) quantitative analysis of the data—this answers the question "Having got high-quality data, how do I now answer my experimental question?"

Preprocessing involves filtering of the data (e.g., detection of outlying values) followed by standardization of the data (i.e., making the data uniformly comparable throughout the dataset). Quantitative analysis involves the assessment of mathematical patterns in the data, which we subsequently interpret as having biological meaning. This analysis will involve one or both of: (a) statistical hypothesis testing; and (b) multivariate pattern recognition.

There are fundamental differences between these two types of data analysis, mostly related to the distinction between univariate and multivariate data. For univariate data analysis, we examine one mathematical pattern at a time (e.g., we might assess the pattern shown by each gene separately), while for multivariate data we are examining common patterns (e.g., we assess whether there is a single common pattern shown by a large group of genes). You might like to think of univariate analysis as a reductionist approach to studying the patterns and multivariate analysis as a holistic approach to analysis. The mathematical techniques developed for the analysis of univariate and multivariate data are quite different, and the purposes to which they are put can be quite distinct. These issues are thus treated separately in the following sections.

In this review, the value of a single piece of data collected on a single entity is called an "observation." A collection of observations in an experiment is called a "sample," while a "population" is all of the observations that *could* have been included in the experiment. The type of characteristic that is being measured is called a "variable," and the entity on which the characteristic is measured is called a "sampling unit." In microarray studies (Fig. 1) the observations will be the measured expression levels. There will be many variables, which are usually genes or expressed sequence tags (ESTs). There will be one or more sample units, which will be arrays subject to extractions from a single or several experimental conditions (e.g., different tissues, a time sequence of development, different experimental manipulations). This imbalance between the many variables and the relatively few sampling units can create problems for data analysis; indeed, this is likely to be the biggest challenge for the analysis of microarray data in the foreseeable future. There may also be additional information about each variable or sample unit, which are called "covariables." The covariables for the sample units will consist of information about extra variables (e.g., other known characteristics related to the tissues, such as disease pathology), while the covariables for the variables will consist of information about extra sample units (e.g., unknown tissues about which predictions are to be made), as shown in Figure 1. Also, for ease we use the term "gene" to refer to the spotted cDNA sequences irrespective of whether they are actually genes, ESTs, or other DNA sequences.

DATA ANALYSIS

Preprocessing

Preprocessing involves filtering of the data, followed by standardization. It is based on the assumption that you have already collected appropriate data. This is itself not necessarily straightforward, because we are not measuring gene expression directly but instead are measuring a dye intensity (Wu, 2001). Retrieving the data from microarray images has its own data analysis problems, which we will not be covering here. Useful reviews

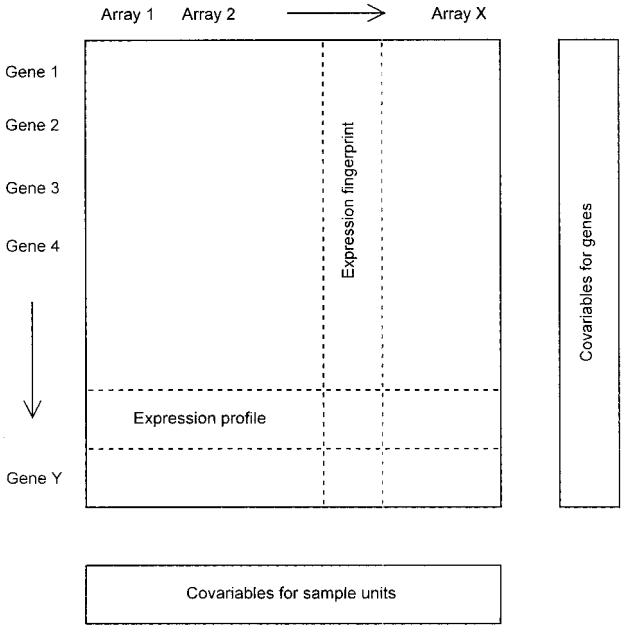


FIG. 1. Schematic representation of expression data from a microarray experiment. The observations are arranged in a data matrix of rows and columns. Here, the columns represent the sampling units (e.g., arrays, tissues, times, experimental conditions) and the rows represent the variables (e.g., genes, ESTs). There will be one observation for each variable for each sample unit. The collection of observations for any one variable is called an "expression profile," while the collection of observations for any one sample unit is called an "expression fingerprint" or "expression signature." There may also be additional information about each variable or sample unit, which are called covariables.

of this topic are provided by Y.H. Yang *et al.* (2002a) and Glasbey and Ghazal (2003), with spot quality being discussed by Y. Chen *et al.* (2002).

Filtering of the microarray data is necessary because the laboratory techniques have not yet proved capable of producing reliable data consistently (Bakay *et al.*, 2002; Kothapalli *et al.*, 2002; Kuo *et al.*, 2002; Li *et al.*, 2002; Novak *et al.*, 2002; Wierling *et al.*, 2002). This is particularly so given the very large nature of the datasets generated. Consequently, it would be naive to analyze data without first assessing its quality, particularly with respect to what statisticians call "outlying values," which are those observations that appear to be out of line with the other observations in the data. These can be irreproducible duplicate spots/probes on the same array or duplicate spots/probes on replicate arrays (Quackenbush, 2002).

Outliers can occur for a number of possible reasons: (1) the observation is a mistake or error during data acquisition; (2) the observation is real but is not representative of the population from which the sample came; and (3) the observation is representative of a very variable population. Clearly, mistakes should be corrected, which is the role of the specialist imaging software used (see Hess *et al.*, 2001), or else the observation must be deleted from the sample if the correction can no longer be made. Representative observations should be left alone, and the consequences for the data analysis accepted. Nonrepresentative observations can be trimmed from the dataset, provided some explicit and repeatable method is used. There are many methods available for this (Barnett and Lewis, 1978), and Hess *et al.* (2001), Tseng *et al.* (2001) and Nadon and Shoemaker (2002) make some suggestions specifically for microarray data. The consequences of leaving outliers in microarray datasets (e.g., discrepant results from the data analysis) are discussed by Chu *et al.* (2002).

An important distinction can be made between observations that appear to be below some detection limit of the experimental method and observations of replicated samples that appear to deviate strongly from each other. In the former case (e.g., when the foreground pixels have smaller values than the background) the observations are often replaced in the analysis by some prespecified minimum value (or by optimizing some mathematical criterion; Wernisch *et al.*, 2003), while in the latter case one of the discrepant values is omitted from the analysis (I.V. Yang *et al.*, 2002). You must be vary careful when dealing with these situations, especially if the number of affected observations is large (e.g., $>5\%$), as they can materially alter the conclusions of the subsequent analyses (Chu *et al.*, 2002; Grant *et al.*, 2002). It is probably better to mathematically model the background and foreground signals to reduce the extent of these problems (Kooperberg *et al.*, 2002a). There may also be truncation at the upper end of the data, due to saturation (Ramdas *et al.*, 2001; Wu, 2001).

Standardization can consist of either or both of two processes: (a) transformation; and (b) normalization. This standardization step is necessary to adjust the contribution of each data variable and/or sampling unit to the data summary, if this is required. It is possible that some of the data variables or some of the sampling units will contribute more to the form of the final pattern detected than will other variables and/or units, and this may or may not be desirable. Note that it is the procedures used for standardization that make the biggest difference between the analysis of complementary DNA arrays and oligonucleotide arrays—most poststandardization procedures apply equally well to both technologies.

Transformation changes the scale on which the data are measured. For example, it is usual to measure H^+ concentration on a logarithmic scale, which we then call pH ($-\log_{10}H^+$). This is because experience has taught us that the biological characteristics that are related to H^+ concentration are best studied and analyzed using this particular scale. A similar argument applies to gene expression data. Such data usually have a lognormal (rather than normal) frequency distribution and their behavior usually shows multiplicative (rather than additive) effects (Kerr *et al.*, 2000; M.L.-T. Lee *et al.*, 2000; Wolfinger *et al.*, 2001; Hoyle *et al.*, 2002; Speed and Yang, 2002; Tsodikov *et al.*, 2002). Both of these features mean that any microarray dataset should be analyzed using a logarithmic scale, so that the data are comparable in a biologically meaningful way. It is thus standard to transform microarray observations to $\log_2$ as part of the preprocessing (Dudoit *et al.*, 2002c; Nadon and Shoemaker, 2002), so that linear changes in the observations will represent fold-changes in expression level (i.e., 0 represents normal expression, $+1$ represents twofold everexpression, -1 represents twofold underexpression, etc.; Dopazo *et al.*, 2001).

Normalization is used to make sure that all of the variables

being compared are being measured on roughly the same scale of variation. The data are adjusted so that experimental variability has been accounted for, thus eliminating bias—we don't want variation caused by the technology to interfere with our study of biological variation. The sort of variability we are talking about here is that between spots or probes on an array and between arrays, including: background noise, production and detection efficiency (e.g., printing, labeling and scanning differences), probe/dye-specific effects, and differences in RNA quantity and quality (Schuchhardt *et al.*, 2000; Wu, 2001). We want the quantitative data analysis to be about biological variability (i.e., differential expression) rather than about this experimental variability (i.e., experimental artefacts), and normalization tries to achieve this. This means that we can focus on the "interesting variation" rather than the "obscuring variation" (Bolstad *et al.*, 2003). This is not a trivial issue, because the results of the data analysis can be influenced more by different methods of normalization than by different methods of statistical analysis (Hoffmann *et al.*, 2002), at least for oligonucleotide arrays. However, oligonucleotide arrays do seem to require much stronger normalization than do cDNA arrays (Workman *et al.*, 2002).

A number of general normalization techniques have been suggested, and several of these have been summarized by Schuchhardt *et al.* (2000), Hess *et al.* (2001), Kroll and Wölfl (2002), Nadon and Shoemaker (2002), Quackenbush (2002), and Tsodikov *et al.* (2002). All of the techniques have their own strengths and weaknesses. Perhaps the most important distinction is between the "global" methods, based on the average value for an array (e.g., the median spot intensity for the array), and the use of reference standards (e.g., a reference treatment that is repeated several times on each array) or gene subsets (e.g., housekeeping genes). The former are strongly influenced by the particular experimental conditions being used (e.g., there may be a nonlinear relationship between the red and green fluorescence values, and this cannot be corrected by using only a single calibration value). If global methods are used, then it is probably best to subtract the median and divide by the interquartile range (Nadon and Shoemaker, 2002; Pan, 2002), or use a trimmed mean (Kroll and Wölfl, 2002), as these will be robust to variability in the type of data. The use of gene subsets also does not deal with intensity-dependent dye biases, nor does it deal with sample-specific biases.

However, dye or probe biases can depend strongly on signal intensity and spatial location within the array, particularly for cDNA arrays (Tseng *et al.*, 2001; Workman *et al.*, 2002; Wernisch *et al.*, 2003), which can be caused by the printing device, conditions during printing, or the scanning device. Therefore, instead of global methods it is probably better to use methods based on MA plots (Dudoit *et al.*, 2002c; Workman *et al.*, 2002; Bolstad *et al.*, 2003). These scatterplots show the ratio of two dye/probe intensities (vertically) and their mean intensity (horizontally), which is a better graphical method than plotting one intensity against the other because it attributes uncertainty to both intensities. If there are no intensity-dependent effects in the array then the points on the MA graph (sometimes also called an R-I graph) would form a horizontal straight line. A nonlinear regression, for example, locally weighted smoothing (LOWESS) or cubic quantile splines (qsplines), can be used to normalize the data so that they do form such a line. Spatial variation is dealt with by normalizing the means of print-tip groups (Dudoit *et al.*, 2002c; Workman *et al.*, 2002), or by fitting a two-dimensional trend surface (Workman *et al.*, 2002; Wernisch *et al.*, 2003).

There is also variation between arrays, which can be substantial, due to different manufacturing characteristics or different experimental setups. This variation can be dealt with by scaling the various arrays in an experiment to some common standard. This could be done by global methods such as the same median or (better) the same variance (Quackenbush, 2002) or median average deviation (Dudoit *et al.*, 2002c; Y.H. Yang *et al.*, 2002b). Alternatively, it can be done by using a nonlinear regression to linearize the relationships, either using all genes (Bolstad *et al.*, 2003) or a subset of genes that is rank invariant across the arrays (Schadt *et al.*, 2001; Tseng *et al.*, 2001; Kepler *et al.*, 2002).

All of these normalization methods have the potential problem of "over fitting" the data (Kerr *et al.*, 2002). That is, because the parameters of the adjustment are estimated from the dataset itself, there can be unnecessary adjustments of the data for problems that don't actually exist, or even introduction of biases greater than the ones removed. However, studies to date indicate that this does not happen often in practice (Y.H. Yang *et al.*, 2002b). If this is a concern, then a simple shift-log transformation may be more appropriate (Kerr *et al.*, 2002), although this also requires the parameter to be estimated from the dataset itself.

Unfortunately, variation in normalization requirements among experiments will probably mean that there will be no universally applied standardization method for microarrays. In each experiment we will have to investigate all three known sources of unwanted variation: global, signal-dependent and spatially dependent variation. Each source of variation can then be dealt with on its own merits, as some will have additive effects (which can be corrected by the global methods), some will have multiplicative effects (dealt with by the log transformation) and some will have nonlinear effects (adjusted by the nonlinear regression and trend surfaces).

Hypothesis testing

The detection of increased or decreased expression in genes is often demonstrated by using some arbitrarily chosen level of expression (e.g., twofold change from some nominated reference sample). This is a rational and logical thing to do because it is explicit and repeatable, but it is nevertheless probably unacceptable as a general principle in science because it is completely arbitrary—that is, there is no theoretical basis for choosing any particular fold change (despite several attempts to provide this; Y. Chen *et al.*, 2002; Quackenbush, 2002; I.V. Yang *et al.*, 2002), and the reliability of fold changes may depend on spot intensity. Scientists have long preferred more objective methods of data analysis, and in particular, they have used statistical methods to help provide the evidence that they seek in their experiments. The quantitative analysis of microarray data should be no different, and there is no real reason why standard techniques cannot be applied.

Univariate data are thus best analyzed using statistical hypothesis testing. The experimental question is turned into a sta-

tistical question, and the answer to the statistical question is used to guide our answer to the original experimental question. In statistics this is referred to as "confirmatory data analysis," because we wish to confirm (or deny) our prespecified mathematical pattern, or "inferential data analysis," because we wish to use our samples to infer a pattern about the population from which they came. The basic intention is to assess whether an observed change in expression is likely to represent a real biological pattern rather than a random accident.

In the usual hypothesis-testing framework an explicit pattern is predicted to occur in the dataset (e.g., that the averages of the observations in two experimental groups will be different from each other) and a single, repeatable mathematical test is used to evaluate whether this pattern exists or not (e.g., at some specified probability level). The procedure for statistical hypothesis testing is exactly the same no matter what the type of analysis is. The experimental null hypothesis is treated as a statistical null hypothesis for the purpose of the analysis—this latter hypothesis says that the observed pattern is entirely due to random chance. The likelihood of the statistical null hypothesis being false is assessed by applying a statistical test to the experimental data. This test produces a test statistic, which is then compared to the critical region of the frequency distribution for that statistic. If the value of the test statistic falls within this critical region then the statistical null hypothesis is rejected; otherwise it is accepted.

The logic of statistical hypothesis testing is thus an inductive argument (from consequences back to an hypothesis) rather than a deductive argument—that is, we are going from a specific instance (our sample) to the general (our population). Inductive arguments do not constitute proof in a formal philosophical sense, but they *can* provide very convincing arguments. The basic problem with induction is that no matter how much evidence we gather in support of a particular hypothesis, we can never be certain that this same evidence would not equally support any number of other unknown hypotheses. It seems strange to many people that, recognizing the virtues of deductive logic in experimentation, we often have to resort to inductive logic to analyse the data. That is, we are expecting a yes/no answer to our experimental question, and we have to get this by assessing a probability. This apparent paradox is simply a product of using a sample in our experiment rather than the population. If there was no variability among the sampling units then this would not be a problem; but whilever (1) there is biological variability and (2) we have only a sample, then the use of induction will be a necessary consequence.

Statistical hypothesis testing is best developed for univariate data. That is, we specify a statistical pattern for one variable at a time and test it separately, rather than specifying a common pattern for a series of variables. For example, if we are comparing a series of genes across two experimental groups, then we would test the pattern for each gene separately, rather than testing some common pattern across all genes. This makes each analysis conceptually quite simple.

If there are two experimental groups being compared (e.g., an experimental manipulation and a control, or two developmental stages of some organism) then the most common statistical test is a *t*-test; an analysis of variance is needed if there are more than two groups. This can be thought of as *class comparison*, because we are comparing two classes of objects. Alternatively, if the experimental objective is the comparison of gene expression levels to other biological characteristics then the most common tests are correlation or regression. These tests are described by all introductory statistical books. Unfortunately, there is as yet no single test that has achieved widespread acceptance or use for microassays. This is mainly because all such tests require replication of the experimental treatments (see below), which has in the past been problematic for microarray studies. We expect this situation to change.

A specific problem arises with the use of these univariate tests for microarray data because a large series of statistical hypotheses are being tested (i.e., one for each of the hundreds or thousands of genes). This is the multiple-testing problem (Shaffer, 1995), and it has long been recognized as a serious bugbear for microarray data analysis (e.g., Claverie, 1999; Dudoit *et al.*, 2002b). If a large number of statistical comparisons are made then some of the decisions to reject the null hypothesis will be due to errors—the more null hypotheses that are tested using the one dataset, then the more likely it becomes that at least one of these hypotheses will be rejected by random chance. For example, if we test each of 100 genes at $P = 0.05$, then we would expect that (on average) we will falsely reject the null hypothesis five times (this is conceptually what $P = 0.05$ means). These errors are called "Type I errors" by statisticians, or "false positives" or "false discoveries" by many biologists. It is thus necessary to distinguish between the probability of a Type I error for each individual hypothesis test (the individual error rate) and the probability of a Type I error for the entire collection of hypothesis tests (the setwise, familywise, or experimentwise error rate). For example, if the individual error rate is $P = 0.05$, then for 100 independent comparisons the setwise error rate will be $P = 0.994$, which is clearly unacceptable. This obviously has the potential to lead the experimenter to an unjustified conclusion (see Morrison, 2002a, for a specific example), and the problem needs to be dealt with.

There seem to be two basic approaches to solving this problem. The first is to carry out the analysis by specifying the setwise (or familywise) error rate and then adjusting the individual error rate in an appropriate manner (Dudoit *et al.*, 2002b). The most widely used adjustment is the Bonferroni correction, in which the probability for each separate hypothesis test is calculated as the setwise error rate divided by the number of tests (Morrison, 2002a). This approach and several of its variants, such as the sequential methods of Holm and Hochberg, are discussed in more detail by Nadon and Shoemaker (2002) and Dudoit *et al.* (2002c). There will, in practice, be very little difference in the outcome of these variants, especially if very few of the null hypotheses are rejected.

Furthermore, this is rather a conservative approach, especially when applied to microarray data with thousands of genes (Xiao *et al.*, 2002; Ge *et al.*, 2003), and so alternatives have been proposed (Dudoit *et al.*, 2002b). These include permutation (see below) versions of these procedures, as well as novel approaches such as those of Westfall and Young, all of which increase their statistical power by taking into account the non-independence resulting from the covariance structure of gene expression patterns (Slonim, 2002; Ge *et al.*, 2003). That is, they are expected to be better tests because they explicitly take into account the inevitable correlations that exist among expression levels of different genes (e.g., due to participation in

related biochemical pathways). The underlying premise of all of these methods is to rank the probabilities and to apply a different correction for each probability depending on its rank. The method that is currently most popular involves controlling the false discovery rate rather than the setwise error rate (Tusher *et al.*, 2001; Efron and Tibshirani, 2002; Ge *et al.*, 2003), which is the probability that a rejected null hypothesis is false. However, it should be noted that the most commonly used "implementation" of this procedures, in significance analysis of microarrays (SAM) (Tusher *et al.*, 2001), actually estimates the individual error rate rather than explicitly controlling the false discovery rate.

The second approach to dealing with the multiple-testing problem is to incorporate the multiple comparisons into the analysis. For example, instead of doing a simple *t*-test or analysis of variance (ANOVA), a multifactorial ANOVA can be used, in which the genes form the levels of a second factor. The calculation of the degrees of freedom in the ANOVA will effectively deal with the multiple comparison tests. This idea is explored in more detail in the context of microarray data by Kerr *et al.* (2000), Kerr and Churchill (2001b), and Dobbin and Simon (2002). Incorporating several factors into the data analysis is an important issue, and so it is worth considering further here.

An analysis of variance is simply a statistical test of the differences between two or more groups of replicate observations defined by a particular experimental "treatment"—a grouping variable is called a "factor" and the groups are called "levels." The calculations of the analysis produce an *F*-ratio (the test statistic) for each factor, which is used to decide whether there is a statistically significant difference between the means of two or more of the groups. The *F*-ratio is calculated using the mean-squares, which measure the amount of variability both within and between the groups. When an ANOVA involves more than one factor, there is a separate *F*-ratio calculated for each factor and also potentially also for all of the interactions between the factors. Multifactorial ANOVA is thus a general technique, in that it subsumes all possible simpler analyses. This means that the *t*-test, paired *t*-test, and *1*-factor ANOVA are all simply special cases of the more general multifactorial analysis (e.g., the results of a *t*-test and an ANOVA of the same data are directly related by their test statistics: $F = t^2$). There is a long tradition of the use of ANOVA models in multifactorial experiments, and their application to microarray studies has not yet really begun to make use of the potential that is clearly there. Chu *et al.* (2002) provide a good overview of the issues involved in the use ANOVA in microarray studies, albeit focussed on the application to oligonucleotide arrays.

The point of wanting to have several factors in an ANOVA is clearly that several statistical hypotheses are tested simultaneously in a valid manner. Thus, we can analyze experiments involving multiple experimental conditions, which takes us beyond the constraints of the simple treatment/control type of experiment. The degrees of freedom associated with the analysis are used to make sure that the various comparisons among groups are carried out appropriately, in the sense that each comparison takes into account the result of every other comparison that has been included in the analysis. So, if an experiment is designed such that there are multiple influences being studied, then each influence should be incorporated into the analysis as a separate factor—that is, the analysis should match the experimental design. This is the principal property of analysis of variance, that it allows us to focus on the features of the experimental design to carry out an appropriate analysis of our experimental hypotheses, rather than trying to subdivide the analysis into a series of simpler analyses that are often independent of our original experimental question. (As an aside, please do not confuse multifactorial analyses with multivariate analyses. The latter refers to the number of variables measured on each experimental unit, while the former refers to the number of factors incorporated into the analysis. It is possible for an analysis to be either multifactorial or multivariate independently of the other.)

The practical issues associated with including multiple factors in an ANOVA are covered in books such as those of Zar (1999) and Glantz and Slinker (2001). Perhaps the most important of these are the distinctions between nested and orthogonal factors and between fixed and random factors (Morrison, 2002b). Some of the issues specifically related to their application to microarrays are discussed by Wolfinger *et al.* (2001), Kerr *et al.* (2002), Y.H. Yang and Speed (2002) and Wernisch *et al.* (2003). Also, it is important to assess the mathematical assumptions on which the analysis is based (see Black and Doerge, 2002, and Chu *et al.*, 2002, for some examples), particularly normality (examined using a normal probability, or QQ, plot) and homogeneity of the variances (examined using a plot of residuals). There have, unfortunately, been few applications of multifactorial ANOVA to microarray data to date (e.g., Jin *et al.*, 2001; Boldrick *et al.*, 2002; Kerr *et al.*, 2002), but the ideas are standard in the rest of biology and so will presumably become more common here as well. Perhaps one limitation is that, due to the enormous number of genes involved, it is usually impossible to carry out the actual calculations using common statistical computer programs (i.e., the number of levels for the gene factor is too large for them to handle). Therefore, specialist computer programs have been developed specifically for microarray data (e.g., Didier *et al.*, 2002).

There are three further potential advantages to the use of multifactorial ANOVA in statistical hypothesis testing of microarray data. First, global normalization of the data can be incorporated directly into the analysis rather than requiring a separate preprocessing step. That is, for cDNA arrays the red and green spot intensities are kept separate in the data analysis, rather than being combined into a ratio, and a factor with two levels is used to include them both in the calculations. This idea is explored in more detail by Kerr *et al.* (2000), Wolfinger *et al.* (2001), and Dobbin and Simon (2002), and has been applied by Jin *et al.* (2001). However, it may actually be preferable to use more flexible normalization procedures, leaving the ANOVA to deal with the remaining sources of experimental variation (Dudoit *et al.*, 2002c).

Second, the variances that are used for detecting over- or underexpressed genes in different experimental groups are estimated from the pooled collection of genes across all groups rather than from each group individually. Estimating a variance separately for each gene for each experimental group is potentially risky because of the small sample size that is usually involved (i.e., the few replicate arrays). For example, estimating a variance from only two observations is not a very accurate procedure, because even a small change in the observation for

either replicate can greatly affect the calculated variance. This effect obviously decreases as the sample size increases. It is therefore preferable to use the pooled estimate of the within-group variance. This, however, does place more reliance on the data meeting the assumptions of the ANOVA procedure.

Third, ANOVA allows the estimation of variance components in addition to the hypothesis tests (Gibson, 2002; Wernisch *et al.*, 2003). The variance components estimate how much of the total experimental variation is accounted for by each of the factors and interactions included in the analysis. This information can be a useful adjunct to the simple consideration of statistical probability. For example, those factors contributing most of the variation may be of biological significance even if they are not statistically significant. Assessment of variance components also helps in making decisions about what aspects of the experiment need replication (e.g., should it be the samples, the arrays, or the spots?) (Churchill, 2002). This idea has been applied to microarray data by, for example, Jin *et al.* (2001) and Wernisch *et al.* (2003).

One potential problem with the use of ANOVA for the analysis of microarray data is the calculation of the probability that is associated with each hypothesis test. The usual parametric calculation associated with ANOVA (i.e., based on the assumption that the data have a normal frequency distribution with equal variances) is likely to be invalid for microarray data (i.e., the data will violate the assumptions of the analysis), and so these possibly should not be used (Thomas *et al.*, 2001; Troyanskaya *et al.*, 2002; Xiao *et al.*, 2002). Alternative parametric strategies have therefore been used to deal with estimating the correct probabilities for microarray data, all based on trying to somehow model the actual underlying error distribution (e.g., using "penalized" versions of standard tests such as the *t*-test). These strategies include regression modeling (Thomas *et al.*, 2001; Spang *et al.*, 2002), mixture modeling (Pan, 2002), associative analysis (Dozmorov and Centola, 2003), Bayesian analysis (Baldi and Long, 2001; Long *et al.*, 2001; Tseng *et al.*, 2001; Townsend and Hartl, 2002), empirical Bayesian methods (Efron *et al.*, 2001; Efron and Tibshirani, 2002; Lönnstedt and Speed, 2002), permutation-validated principal components analysis (Landgrebe *et al.*, 2002), limit fold change (Mutch *et al.*, 2002) and significance analysis of microarrays (SAM) (Tusher *et al.*, 2001). It is not yet clear how successful these strategies are. Many of these methods are currently restricted to comparing two experimental groups only, but many of them can probably be used in a hierarchical factorial context similar to analysis of variance (e.g., Long *et al.*, 2001; Lönnstedt *et al.*, 2001; Tusher *et al.*, 2001; Townsend and Hartl, 2002). Unfortunately, this has not been done in most of the computer programs released to date. When this is done, there will need to be some quantitative method to compare the results of the different models for analysing the data (Kooperberg *et al.*, 2002b).

Alternatively, many authors replace the parametric techniques with nonparametric alternatives (Troyanskaya *et al.*, 2002). For example, a *t*-test can be replaced with a Mann-Whitney *U*-test or a one-factor analysis of variance with a Kruskal-Wallis test. These have been shown to be viable alternatives for gene expression data (P.J. Park *et al.*, 2001; Liu *et al.*, 2002; Troyanskaya *et al.*, 2002; Tsodikov *et al.*, 2002), although they are conservative tests (i.e., they have a tendency to miss some of the "real" patterns, resulting in what are called "false negatives" by many biologists). This is because these tests are based on ranking the observations, which necessarily loses information, especially for small sample sizes (Thomas *et al.*, 2001). The expression level of a gene may change in different experimental treatments without changing its ranking, and the rank may change without a change in expression level, because the rank depends entirely on the behaviour of the other genes while the absolute expression level does not necessarily do so (other than the biological dependence resulting from the biochemical pathways).

Therefore, it is might be better to use nonparametric randomization procedures (Troyanskaya *et al.*, 2002), as described in more detail for multivariate data in the next section. The usual approach to assessing the statistical significance of the factors is the use of permutation testing (Good, 1999; Lunneborg, 1999), as used by Tusher *et al.* (2001), Dudoit *et al.* (2002c), Grant *et al.* (2002), and Kooperberg *et al.* (2002b) for microarray data. However, this does not necessarily tell us which particular genes are showing over- or underexpression in the ANOVA, and for this purpose Kerr *et al.* (2000, 2002) have suggested the use of bootstrapping to create confidence intervals that can then be used for hypothesis testing (a similar procedure is used by Wernisch *et al.*, 2003), while P.J. Park *et al.* (2001) have used permutation of nonparametric scores to create probabilities, and Landgrebe *et al.* (2002) have used permutation for variance testing. There are currently some limitations on the use of permutation testing, including the potential need for equal sample sizes in the groups being compared, especially for multifactorial ANOVAs (which is then called a balanced design). It is for this reason that missing data need to be dealt with, as discussed in the next section (e.g., the SAM data analysis imputes missing data in order to balance the sample sizes; although Bayesian analysis can potentially sidestep the potential problems; Townsend and Hartl, 2002). Permutation procedures can also be used to calculate probabilities that have been adjusted for the multiple hypothesis testing problem (Dudoit *et al.*, 2002b; Ge *et al.*, 2003).

It is also worth pointing out that statistical probabilities should not be overinterpreted. Biological scientists have long been accused by statisticians of naive statistical hypothesis testing, with overzealous attention to *statistical* hypothesis testing often taking attention away from more important aspects of the *experimental* hypothesis test. The statistical null hypothesis is not the experimental null hypothesis, and statistical significance is not biological significance. Furthermore, no calculation of probabilities can ever be exact, and so microarray probabilities should never be treated as though they are particularly precise. More to the point, given that the observations have been filtered, transformed and normalized, microarray data have been artificially homogenized, thus tending to decrease the resulting probabilities (Wu, 2001; Xiao *et al.*, 2002). This can only increase the number of Type I errors. It is therefore always a good idea to plot what are known as volcano plots (Lönnstedt *et al.*, 2001; Chu *et al.*, 2002; Gibson, 2002; Townsend and Hartl, 2002). These show the strength of the statistical pattern on the vertical axis and the strength of the biological pattern on the horizontal axis (the reason for the name will be obvious when you see the example below)—this makes clear the sometimes unequal relationship between the two concepts, and highlights observations that show a large pattern for only one of the two characteristics.

Another form of hypothesis testing involves examining the relationship between the microarray dataset and any continuous covariables for which we might also have data. We have already discussed the analysis of grouping covariables above, when we discussed the role of multiple factors in an ANOVA. However, different analysis techniques are needed if we have covariables that are measured on a continuous scale.

Correlation analysis is the most common technique for looking at the relationship between continuous covariables and either expression profiles or expression fingerprints (Wu, 2001). This is usually done in a pairwise fashion, testing each fingerprint against each covariable. However, care must be taken, because we immediately come to the same multiple-testing problem highlighted above. Furthermore, you must remember that the normal correlation coefficient only tests for linear relationships, and will therefore not detect any nonlinear ones.

Similar comments apply to the use of regression analysis for the same purpose, including logistic regression (Dopazo *et al.*, 2001; Shannon *et al.*, 2002). The important difference between regression and correlation analyses is that correlation is concerned with assessing the closeness of the relationship between two data variables, while regression is concerned with estimating the equation of the line that best describes the relationship between two data variables. Correlation thus seeks relationships between two variables, while regression quantifies suspected relationships between them. That is, regression is about the *form* of the relationship, while correlation is about the *strength* of the relationship. The two analyses are actually based on mutually exclusive sets of assumptions about the data, and you should not treat them as interchangeable. Tibshirani and Efron (2002) use a variant of jackknifing for assessing the validity of regression equations as predictors.

Finally, the analysis of time sequences involving microarray data has been poorly addressed to date (Filkov *et al.*, 2002). Time-series analysis is a well-recognized technique within statistics, and its power has not yet been fully applied to finding and testing patterns in microarray data (Slonim, 2002). Recent suggestions for different approaches to the problems include: using a multifactorial ANOVA with time as one of the factors, with permutation testing and adjustment for multiple hypothesis testing (T. Park *et al.*, 2003); information-theory analysis (Kasturi *et al.*, 2003); piecewise nonparametric regression (de Hoon *et al.*, 2002); clustering based on time-series analysis (Ramoni *et al.*, 2002) or nonlinear curve fitting (Luan and Li, 2003); and partial least-squares regression (Johansson *et al.*, 2003). The relative efficacy of these alternatives remains to be assessed.

Pattern recognition and prediction

Multivariate data analysis does not easily fit into the usual framework of statistical hypothesis testing. In the usual hypothesis-testing framework an explicit pattern is predicted to occur in the dataset and a single mathematical test is used to evaluate whether this pattern exists or not. However, because multivariate data are inherently complex, we frequently do not know what pattern to predict in the dataset, and so we cannot be explicit beforehand about what pattern or patterns we would interpret as biologically meaningful. Instead, we wish to *search* for any patterns that might exist in the dataset, which we might then interpret *post hoc* as being potentially meaningful. So, the answer to our experimental question arises from the result of our search for patterns in the dataset rather than from the answer to an explicit statistical question.

Multivariate data analysis is thus frequently in the nature of a fishing exercise, or what is more technically referred to as "data mining" or "pattern analysis." In statistics this is referred to as "exploratory data analysis," because we are exploring the dataset for mathematical patterns, or "descriptive data analysis," because we wish to summarize those patterns rather than make explicit inferences about them. The important thing to recognize is that the dataset consists of measurements for a series of variables for each sample unit, and it is the whole series of measurements that we wish to summarize and explore. For microarray data this means having multiple measurements of expression for each gene, either under different experimental conditions, from different tissues or as part of a time series. These multiple measurements are sometimes referred to as an "expression profile" for each gene.

The most depressing thing about multivariate pattern analysis for most biologists is that there is no single mathematical technique that can be universally recommended. This situation arises from the fact that there is no single pattern that can be expected in the data—if there are many possible patterns in the data then there must be many possible mathematical techniques for finding those patterns, as each technique looks for a different type of pattern. Consequently, there is a myriad of possible analysis techniques available, and many of them have been applied to microarray data at one time or another.

More to the point, there is no way of knowing a priori which technique(s) will find those pattern(s) that happen to be in your data. Choosing the technique that is suggested to be best under the widest range of possible circumstances sounds like a reasonable criterion of choice, but this is not necessarily a good idea. For example, if your data always have certain characteristics (e.g., because of the experimental conditions and the way the data are quantified) then what you actually want is the technique that performs best under those precise circumstances, irrespective of how it performs under other circumstances. So, what we need is critical assessments of the techniques specifically for microarray-type data.

It is thus currently not easy to present multivariate pattern analysis as part of a rigid protocol, similar to a laboratory protocol—one fishes in the data and sees what fish one catches. This may seem rather nonscientific, but it is the end result of the experimenter not having a fixed idea about exactly what pattern will exist in the data. There is no point in blaming statisticians for this situation—if we could tell them exactly what we want then they could do something about providing it.

Since there are a lots of techniques, it is necessary to summarize them and their characteristics in some way. The traditional way is to recognize two main types of multivariate analysis: (1) *supervised* (or machine learning) techniques, in which the analyst "supervises" the search for patterns by making suggestions about what patterns to look for; and (2) *unsupervised* techniques, in which it is solely the mathematical algorithm that determines the search for pattern.

Both techniques have a multitude of alternative possibilities. Sadly, the proponents of the many techniques have rarely made any attempt to systematically demonstrate that their proposed

method is superior—they merely demonstrate that their method will work (or, in technical terms "provides competitive performance"). So, a detailed comparison and critique of the techniques as applied to microarray data is currently lacking, other than the empirical comparison of several supervised techniques by Dudoit *et al.* (2002a) and several unsupervised techniques by Yeung *et al.* (2001), G. Chen *et al.* (2002), Dougherty *et al.* (2002), and Datta and Datta (2003) (although several unpublished comparisons are also extant, see Slonim, 2002). One major problem with performing a worthwhile comparison of the techniques is providing large enough training and testing datasets. The normal method when confronted with small datasets is to use crossvalidation, but to date this has usually been implemented in a manner (i.e., leave one out) that results in overly optimistic estimates of prediction accuracy (Ambroise and McLachlan, 2002; Bø and Jonassen, 2002; Kim *et al.*, 2002).

The supervised techniques are designed to act as predictive classification tools (i.e., *class prediction*). That is, they allow us to predict the characteristics of unknown specimens based on the characteristics of known specimens—we could then, for example, assign an unknown DNA sequence to one of a set of known gene classes. Among the techniques, we have traditional techniques such as: discriminant function (or canonical variates) analysis; logistic regression; nearest-neighbor classifiers, or weighted voting; classification, regression or decision trees (recursive partitioning analysis); and support vector machines or artificial neural networks. Some of these techniques are briefly described in the context of microarray data analysis by Quackenbush (2001), Raychaudhuri *et al.* (2001), and Slonim (2002), and details can be found in the book by Legendre and Legendre (1998). There are also more recent developments such as: between-group analysis (Culhane *et al.*, 2002); compound covariate prediction (Radmacher *et al.*, 2002); pairwise feature subset selection (Bø and Jonassen, 2002); sufficient dimension reduction (Chiaromonte and Martinelli, 2002); nearest shrunken centroids (Tibshirani *et al.*, 2002a); Bayesian variable selection (K.E. Lee *et al.*, 2003); independently consistent expression discriminators (Bijlani *et al.*, 2003); strong-feature classifiers (Kim *et al.*, 2002); disjoint principal components analysis (Bicciato *et al.*, 2003); and various genetic algorithms (Deutsch, 2003; Ooi and Tan, 2003).

Such supervised techniques essentially use training groups to quantify the expected patterns in the dataset, and then use this "learned" information to analyze the unknown patterns. The training groups are based on existing biological information, such as knowledge of gene function, tissue origin, cell type, or experimental treatment. The learned information is encapsulated in a set of mathematical equations or rules, designed to reliably assign new sampling units to the groups. The techniques differ in the type of discrimination rule that they use. So, the usual use of such methods is for rapidly identifying diagnostic genes, or for selecting a minimally predictive set of genes. Having too few predictive genes is as bad as having too many genes, the former because the discrimination will be inadequate and the latter because some of the genes will be irrelevant to the prediction. Most of the techniques seem to provide rather similar results, at least based on the limited number of datasets for which they have been compared.

Dudoit *et al.* (2002a) concluded that some of the simpler methods actually work better than most of the more complex ones. Much of the reason for this appears to be to do with the number of genes compared to the number of biological samples. Selection of classifiers tends to become unreliable when the number of potential classifiers (genes) is large relative to the number of objects (samples) (Dougherty, 2001), and in microarray studies to date the number of genes exceeds the number of samples by several orders of magnitude. Class prediction is thus very much a search for a subset of genes that will act as good predictors of the classes of biological samples. This is no trivial task, because it essentially involves either trying every possible combination of genes, which is impossible in practice for large numbers of genes (mathematically, it is called NP-hard), or finding some heuristic search strategy, which will depend for its success on the appropriateness of the heuristics. There are two basic requirements if the genes are to act as good predictors (Dougherty, 2001; Ooi and Tan, 2003): the genes need to be highly correlated with the distinction between the classes (to be good predictors), and the genes should *not* be correlated with each other (because they will then provide redundant information). Finding subsets of uncorrelated genes seems to become more problematic the more genes that are considered (Ooi and Tan, 2003). Furthermore, because the number of genes is much greater than the number of arrays, perfect discrimination of the groups in the training set is always possible. This is the same overfitting problem as referred to above (Kim *et al.*, 2002; Radmacher *et al.*, 2002; Szabo *et al.*, 2002), and it means that the prediction may be perfect for the training dataset but useless for any other set of data. So, only simple classifiers can be used, which may entail a compromise between simplicity and classification accuracy (Dougherty, 2001). This simplicity means that there may be very many gene subsets that are equally optimal with respect to prediction (Kim *et al.*, 2002).

The supervised analyses will only be as good as the information used in the training sets, and it is necessary for the analyst to make a series of decisions that can have quite a profound influence on the outcome of the analysis. There can also be quite severe mathematical restrictions assumed by the analyses. For example, linear discriminant functions assume linear relationships among the groups and strict multivariate normality. These constraints can become more serious the more classes there are for consideration. Recently, therefore, techniques have been suggested based on constrained or canonical ordination analysis of binary variables (Culhane *et al.*, 2002), such as redundancy analysis (RDA, using principal components ordination) and canonical correspondence analysis (CCA, using correspondence ordination), which have been successful in other fields of biology that face similar problems (e.g., Jongman *et al.*, 1995).

In contrast, the unsupervised multivariate techniques are designed merely to act as summaries of the data structure (i.e., *class discovery*). That is, they try to find and highlight whatever general mathematical patterns there are, which we might then interpret as having some biological meaning—we could, for example, distinguish between different functional classes of genes based on their expression patterns. There is a veritable arsenal of such techniques, but they can be broken into two main groups: ordination; and clustering. Such techniques are basically a two-step process: (a) summarize the patterns in the dataset by calculating some measure of the relationship between

all possible pairs of the sampling units, or between all possible pairs of the data variables; and (b) display the data summary. There are several possibilities for each of these two steps (e.g., correlation coefficient and Euclidean distance for step a, and ordination and clustering for step b). The differences between these various possibilities mean that each variant is best at finding a particular type of pattern. Some of these techniques are briefly described in the context of microarray data analysis by Sherlock (2000), Dopazo *et al.* (2001), Hess *et al.* (2001), Quackenbush (2001), and Slonim (2002), and details can be found in books such as those of Podani (1994), Jongman *et al.* (1995), Legendre and Legendre (1998), and Theodoridis and Koutroumbas (1999).

More recently, unsupervised methods tailored specifically to microarray data have been proposed (e.g., Lazzeroni and Owen, 2002; Mavroudi *et al.*, 2002; Medvedovic and Sivaganesan, 2002). A particular trend has been to propose methods that use some of these unsupervised techniques as the first step of a two-step procedure, in which the second step is a supervised pattern-prediction technique or an hypothesis-testing technique. Thus, the unsupervised technique is used as a data summary or dimension reduction procedure, and it is the summarized (or reduced) data that is analyzed for class prediction or class comparison. Combinations include: ordination and discriminant analysis (Chiaromonte and Martinelli, 2002; Méndez *et al.*, 2002); partial least-squares and discriminant analysis (Nguyen and Rocke, 2002a, 2002b) or nonlinear regression (Ghosh, 2002); self-organizing maps and cluster analysis (Wang *et al.*, 2002a); correlation/clustering and support vector machines (Jaeger *et al.*, 2003); ordination and regression (West *et al.*, 2001; Spang *et al.*, 2002); ordination and ANOVA (Landgrebe *et al.*, 2002); and adaptive dimension reduction and regression (Antoniadis *et al.*, 2003). These combinations form new variants of techniques that have been used in other fields of biology, all of which use step (a) above and replace step (b) with a further data analysis (e.g., Legendre and Anderson, 1999). However, this second step may be unnecessary, because the results of step (a) can actually be examined directly (Anderson, 2001; McArdle and Anderson, 2001), an approach that does not yet seem to have been examined for microarray data.

It may be worth briefly emphasizing the differences between ordination and clustering analyses, since they are far and away the most commonly used techniques for class discovery in microarray data. Both techniques display the data summary as a graph, with ordination using a scatterplot and clustering a line graph (or tree). Ordination methods seek to arrange (or order) the sampling units so as to place similar ones close together on the graph, while grouping techniques try to sort the sampling units into meaningful groups, with similar sampling units being placed in the same group. Ordination is therefore more appropriate when the sampling units are thought of as being related along a gradient, and clustering is more appropriate when the sampling units are thought of as comprising a number of partly dissociated subpopulations. Which of these two possible approaches should be used can thus only be decided by reference to the original experimental question. In fact, it may be a good idea to apply both techniques and to compare the results, as the comparison may provide more insights into the data than either analysis does alone (Quackenbush, 2001)—for example, ordinations can be used to assess the distinctness of clusters, or they may reveal outliers (outliers may be mistakes or they may indicate interesting biological phenomena).

As far as visualizing the patterns of the complex data is concerned, several useful strategies have been applied to microarray data. For example, Eisen *et al.* (1998) have championed the use of ordered two-way diagrams for displaying the results of clustering analyses. This involves clustering both the genes and the expression profiles, and showing the two results simultaneously. There are objective criteria for sorting such two-way diagrams so that the order maximizes the similarity of adjacent elements (e.g., Hill, 1979), but in general, this has been done on a more *ad hoc* basis for microarray data (see some suggested criteria in Eisen *et al.*, 1998). The results of these techniques can be visualized using coloring to represent different levels of expression (sometimes called a heat map). Sadly, red and green have been the chosen colors for over- and underexpressed genes (Eisen *et al.*, 1998), which makes it difficult for those people who are red–green colour blind (i.e., 5% of the male population)—blue and yellow would have been much better choices.

It is important to recognize the serious limitations that can apply to multivariate pattern analysis. They are attempts to summarize complex data into a simpler form, and therefore (by definition), they lose information by making assumptions about what aspects of the data can be usefully lost—their very strength (the ability to reduce complexity) is also their greatest weakness (the loss of potentially valuable information). The most important of these limitations is therefore the existence of sometimes quite critical mathematical constraints within particular analysis techniques, to which the analyses are not robust. Each analysis technique is based on a set of strict assumptions, and if these assumptions are violated then the output of the analysis is unreliable. At worst, the patterns revealed by the analysis may be nothing more than mathematical artefacts rather than representations of real biological phenomena. Three examples are worth considering here, because they mean that suboptimal methods are often being used as a result of the sheer size of the datasets that are generated by microarray experiments.

First, some ordination techniques, particularly eigenanalysis (i.e., singular value decomposition) techniques such as principal components analysis (PCA; Hilsenbeck *et al.*, 1999; Alter *et al.*, 2000; Raychaudhuri *et al.*, 2000) and correspondence analysis (CA; Fellenberg *et al.*, 2001), are based on strict ideas about the relationships among the data variables (e.g., that they are linear or unimodal). None of the proponents of these techniques for the analysis of microarray data seem to have seriously addressed this issue. Nevertheless, these are quite serious assumptions that are likely to violated by many biological datasets (e.g., the relationships are often curvilinear), and they have long been known to potentially introduce distortions into the ordination diagrams (e.g., Gauch, 1982; Ludwig and Reynolds, 1988). An example of the distortion is shown in Figure 2. It is for this reason that multidimensional scaling (MDS) is usually recommended as an ordination technique in biology (e.g., Morrison *et al.*, 1994), as it does not require the same underlying assumptions. However, PCA is far more frequently used than MDS in microarray studies, because the mathematical calculations are easier to perform on large datasets. In fact, MDS may be infeasible for many microarray datasets, although it has been successfully used (e.g., Bittner *et al.*, 2000; Khan *et al.*, 2001).

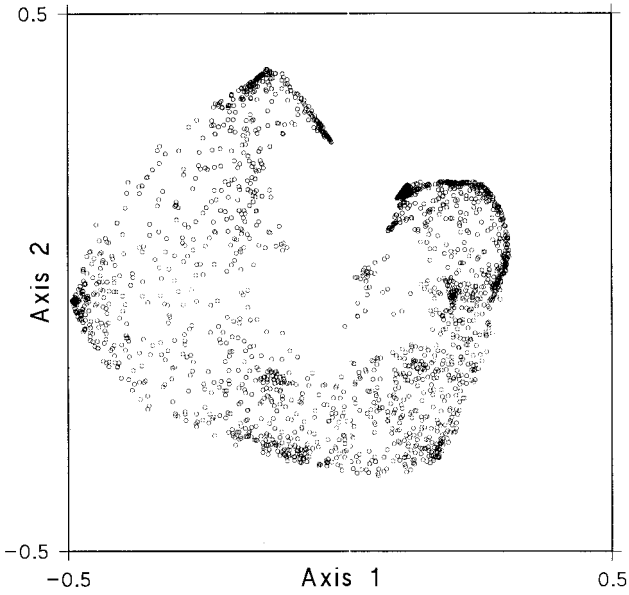


FIG. 2. Eigenanalysis ordination of the 2066 ESTs from the Matrajt *et al.* (2002) data set. Each symbol in the scatterplot represents one of the ESTs, and their spatial proximity indicates how similar they are in terms of their expression profiles. The relationship between all possible pairs of the ESTs was estimated via a Bayesian infinite mixture model, calculated using the Infinite-Gaussian program of Medvedovic and Sivaganesan (2002), which specifically takes into account the replication of the arrays in the experiment. The data summary is displayed as a principal coordinates analysis, calculated using the PATN (v. 3.5+) program of Belbin (1995). Only the first two dimensions of the ordination analysis are shown (out of the four available). Note that there are two distinct distortions of the relationships among the ESTs in this diagram. First, there is an unwanted curvilinear relationship between the two axes—the second axis is merely a quadratic function of the first axis (usually called a "horseshoe shape"). The second axis thus provides no more summary of the data than does the first axis, but merely distorts the same pattern. Second, there is misrepresentation of the positions of the points along the ordination axes—there is contraction at the ends of the horseshoe so that many hundreds of points are clustered together, while the points near the middle of the horseshoe are reasonably well separated. When looking at such an arch and contraction, we cannot tell whether it is a real pattern or an artefact. It is safer to assume the latter, because these distortions are well-known potential by-products of eigenanalysis ordinations. Furthermore, an agglomerative clustering analysis of the same data indicates that three of the four main clusters are within the right-hand contraction, so that this analysis is certainly finding something that the ordination is not. Note that the distortion is known to be an effect of the eigenanalysis itself (i.e., the decomposition of the association matrix into eigenvectors, often called generalized singular value decomposition) rather than of the similarity measure or standardization.

Second, agglomerative clustering is far more frequently used than divisive clustering (in fact, it is the single most commonly used multivariate technique). Divisive techniques start with the whole set of sampling units in one group, and divide this set into two or more subgroups; each of these subgroups is then analyzed and divided in its turn. Agglomerative techniques, on the other hand, start with each sampling unit in a separate group, and then start grouping them together, thus building up larger and larger groups. In principle, divisive techniques should perform better at the important higher levels of the grouping hierarchy, because the divisive techniques are most reliable at the level of few groups whereas agglomerative techniques are least reliable at this level. However, divisive clustering is probably infeasible for most microarray datasets, although it has been used (e.g., Alon *et al.*, 1999). To circumvent this problem, nonhierarchical techniques such as k-means clustering, quality clustering, and self-organizing maps have been used (see Quackenbush, 2001), but it is not clear yet how appropriate they might be because these have their own limitations. Furthermore, there is no consensus about how to make the decision concerning how many clusters should be recognized in any particular microarray dataset, irrespective of how those clusters were derived, mainly because there is actually no simple definition of the word "cluster" by which we could recognize one if we saw one (e.g., see the competing suggestions by Yeung *et al.*, 2001; Tibshirani *et al.*, 2001a, 2001b; Méndez *et al.*, 2002).

Third, relatively little attention seems to have been paid to step (a) of the unsupervised techniques, which involves calculating some measure of the relationship between all possible pairs of sampling units or variables. The choice of metric to be used to estimate the size of the relationship can have serious effects on the analysis results, as can any prior standardization that is applied to the metric (e.g., Quackenbush, 2001). One intuitively appealing idea, which has been commonly used in microarray studies, is to use the usual parametric (Pearson) correlation coefficient (Dopazo *et al.*, 2001; Szabo *et al.*, 2002). The use of the correlation coefficient as a measure of similarity is based on the idea that concordant (i.e., interdependent) changes have occurred in the objects being compared—that is, any genes that are different between the two arrays have changed simultaneously. On the other hand, the alternative Manhattan coefficient (Kasturi *et al.*, 2003) assumes that all changes have occurred independently of each other. Unfortunately, biological reality is likely to be somewhere between these two extremes for any one dataset. Nevertheless, which of these two metrics we might choose should be based on what we are prepared to assume about our data.

For those of you who are familiar with distance and similarity measures, the correlation coefficient is simply the Euclidean distance that has been standardized, while the Gower coefficient is simply the Manhattan distance that has been standardized. There are thus close connections among the various metrics. Other proposed metrics, such as the rank correlation, the angle between two vectors and mutual information, have not received much attention (Kim *et al.*, 2002). Interestingly, the Gower measure appears not to have been used for microarray data, although it has been a highly recommended metric in detailed critiques of other fields of biology (e.g., Faith *et al.*, 1987).

It is important to recognize that each pattern analysis technique is designed to examine a particular type of pattern in the dataset even if that particular type of pattern is absent from the data (McShane *et al.*, 2002). The most egregious example of this is occurs when using clustering techniques, all of which will put the sampling units into clusters irrespective of whether there are distinct clusters in the dataset or not (except for the

plaid technique of Lazzeroni and Owen, 2002, which allows objects to be in more than one group or not in a group at all). More to the point, both agglomerative and divisive clustering are based on the assumption that the relationships among the objects are hierarchical, and there is actually no evidence for the existence of such relationships in biological functions of different genes (Szabo *et al.*, 2002). Consequently, it is inadvisable to interpret the details of multivariate pattern analyses too closely—pattern analyses are a preliminary (exploratory) step rather than an end in themselves. In particular, confirmatory evidence from some other experimental procedure (e.g., Northern blots, real-time PCR, RNase protection assays, *in situ* hybridization; Chuaqui *et al.*, 2002) needs to be acquired for all microarray experiments, as the technology itself is still not very robust (Bakay *et al.*, 2002; Kothapalli *et al.*, 2002; Kuo *et al.*, 2002; Li et al, 2002; Novak et al, 2002; Wierling *et al.*, 2002). If nothing else, technical issues such as crosshybridization and nonspecific binding mean that all microarray data should be treated with some caution (Chuaqui *et al.*, 2002).

Recently, therefore, statisticians have turned their attention to the possible assessment of statistical significance for the patterns found by multivariate analyses, irrespective of whether they are supervised or unsupervised. These analyses are not amenable to the traditional parametric approaches, because we cannot specify the necessary information required for calculating the probabilities (e.g., the underlying frequency distributions). So, the approaches are necessarily nonparametric. The approach that has shown most promise is the use of a *randomization test.* The name derives from the fact that we are generating the statistical null hypothesis (i.e., a random pattern) by randomizing the dataset itself.

A randomization test generates the statistical null hypothesis by calculating all of the possible permutations of the dataset, which is all of the possible arrangements of the observations into patterns that could arise from the original experimental design. These permutations are what would be expected if the null hypothesis was true. So, the mathematical calculations are performed on the original dataset; then the set of permutations is produced, and the mathematical calculations are performed on each of these permutations. The pattern produced by these permutations is treated as being the relevant one for the null hypothesis (i.e., that expected by random chance); and so the calculations from the original dataset are compared to those from the permutations. If the original data produce a pattern that is more extreme (e.g., stronger) than those produced by the permutations, then it can be considered to be statistically significant (e.g., if the original data has a pattern that is stronger than those produced by $>95\%$ of the permutations then the pattern is statistically significant at $P < 0.05$).

The basic practical problem with randomization tests is that the number of possible permutations of the dataset increases exponentially with increases in the number of observations. It is thus a time-consuming exercise to generate all possible permutations, and for large datasets it is completely impractical. The usual solution to this problem is to generate only a sample of the possible permutations, using Monte Carlo simulations. The randomization tests are then called *resampling tests.* A sufficiently large number of samples needs to be taken in order to generate a reasonable approximation to the set of all permutations, usually in the order of 2–5000 (Manly, 1997). Clearly, randomization tests were not possible in practice before the advent of computers.

Three distinct but related approaches to resampling have been developed (see Manly, 1997, for the details):

1. *permutations*—this generates samples that are exactly the same size as the original dataset but which have relationships among the data variables randomly jumbled. So, to generate each resampled dataset the observations for each data variable are randomly rearranged among the sampling units (i.e., sampling without replacement);
2. *bootstrapping*—this also generates samples that are exactly the same size as the original dataset but which have each of the data variables sampled with replacement. So, to generate each resampled dataset the observations for each data variable are randomly sampled from among the sampling units in such a manner that each observation could appear more than once in each resampled dataset (while others might not appear at all); and
3. *jackknifing*—this generates samples that are smaller than the original dataset and which are sampled without replacement. So, the proportional size of the resampled dataset needs to be specified. The most common versions are delete-one sampling, in which each resampled dataset has one less sampling unit than in the original set, and delete-half sampling, in which each resampled dataset is half the size of the original set.

In general, statisticians use permutations when performing hypothesis tests (such as *t*-tests, analysis of variance, or correlation), bootstrapping for creating summaries of variability or accuracy (such as standard deviations, standard errors, or confidence intervals), and jackknifing for crossvalidation assessments.

All three techniques have been applied to microarray data at one time or another. For example, Heyer *et al.* (1999) use jackknifing to deal with outlying values in multivariate data analysis, Kerr and Churchill (2001c) use bootstrapping to assess the reliability of clusters derived from clustering analyses (technically, they bootstrap the residuals from an analysis of variance rather than bootstrapping the variables) and Yeung *et al.* (2001) use jackknifing for the same purpose, Tibshirani *et al.* (2001b) use permutations to assess the "optimal" number of clusters in a dataset, while Tibshirani *et al.* (2001a) use jackknifing for the same purpose, Radmacher *et al.* (2002) employ permutation to quantify the significance of prediction results, and McShane *et al.* (2002) use permutation to assess the degree of clustering present in the dataset (technically, they permute nearest-neighbor distances in a PCA ordination rather than permuting the variables).

An important point to remember is that it is not straightforward to apply statistical analyses to unsupervised clustering or ordination, because statistical hypothesis testing requires independence. For example, in the hypothesis-testing scenario described above the pattern being tested is specified before the data are collected, and so the data can be used validly to test the pattern (because the pattern and the data are independent of each other). However, in unsupervised techniques the pattern is derived from the data and therefore the same data cannot be used to test that pattern. Hence, neither the bootstrapping method

of Kerr and Churchill (2001c) nor the permutation method of McShane *et al.* (2002) can be used to estimate the probability that two clusters are "significantly different" from each other, for example—the methods are designed to assess the stability of the clusters instead.

It is, however, possible to statistically compare two datasets to assess whether they contain the same pattern (e.g., gene expression data in one dataset and some other biological information in the other dataset), provided the two datasets are independent of each other (Jongman *et al.*, 1995). So, if we wish to examine the relationship between the microarray dataset and any covariables for which we might also have data, then this can be done in a multivariate fashion. That is, we can treat the covariables as a second dataset, and we can look for general patterns that the two datasets have in common. For microarray data, this can be done most effectively using the Mantel test (Shannon *et al.*, 2002). This is a generalization of correlation (and regression) analysis, where a single correlation value is used to describe the overall relationships between the two datasets rather than between each pair of variables. The correlation value can be statistically tested using permutation testing. Alternatively, Chapman *et al.* (2001) and Landgrebe *et al.* (2002) have shown the advantages of using biplots for displaying the results of both unconstrained (e.g., PCA, CA) and constrained (e.g., RDA or CCA using continuous variables) ordination information, which is an alternative approach to the same problem, and which can also be tested using permutations. Similarly, Tibshirani *et al.* (2002b) using clustering in combination with correlation analysis, followed by permutation testing. Alternatively, LeBlanc *et al.* (2003) take a more direct approach by using a reference gene, which is highly related to the external biological information, to rank the other genes.

As a final point, it is worth noting that most of the multivariate techniques cannot deal with "missing" data, unlike the hypothesis-testing techniques discussed in the previous section. Data are missing if an observation is not collected for every variable for every sample unit, and this is a common occurrence in microarray studies because there can be numerous technical reasons why some genes are not detectably expressed on some arrays (Troyanskaya *et al.*, 2001). Unfortunately, the mathematical calculations cannot proceed under these circumstances, and so this situation must be explicitly dealt with. It is possible to drop either the offending genes or arrays from the analysis, and then to proceed. However, this is rather extreme if it is the array that is to be dropped (why drop all of the data from an array just because a few genes were not expressed?), and so this is not recommended. Dropping those genes with missing data is more feasible, but this will mean that potentially important genes may be missed in the analysis. Therefore, many imputation methods have been developed that will supply estimates for the values of any missing observations (Schafer, 1997; Allison, 2001), so that the data analysis can then proceed normally. Several of these methods have been evaluated for microarray data (Troyanskaya *et al.*, 2001), with another proposal from Wernisch *et al.* (2003).

However, it is important to recognize that all imputation methods assume that the data are missing at random from the dataset, and this may not be true for microarray data. For example, lowly expressed genes may have many missing data because the expression levels are hard to detect, or one of the arrays might be faulty. Data that are below the detection limit may not be missing at random (and most evaluations of imputation methods have been based on data that are missing completely at random). The best strategy may therefore be some combination of gene deletion and imputation, deleting those genes with more than one to two missing observations and then imputing any remaining missing data. A rough guideline may be that 0–5% missing data are unlikely to cause analysis problems even if the data are not missing at random, while 5–10% missing data can be imputed safely using a reputable technique, and >10% missing data will require specialist techniques for analysis. Faulty arrays should always be deleted from the data analysis, of course.

EXPERIMENTAL DESIGN

The sheer size of datasets that can be collected in microarray experiments has apparently tempted some people to assume that the data can be collected in an *ad hoc* manner, based on the naive assumption that the "truth" will be revealed by the sheer weight of numbers (Simon *et al.*, 2002). This is unlikely to be a realistic expectation, and it will always be best for experiments to be carefully planned in the light of an explicit hypothesis. If the data are to form useful scientific evidence then they must be collected in a "designed" manner, and experimental design is as important for microarray experiments as it is for any other type of experiment (Churchill, 2002).

Clearly, the analysis of data is intimately related to the design of the experiment from which the data came—for the analysis to be successful its characteristics and assumptions must match those of the experimental data. Thus, statisticians have traditionally seen experimental design as a subbranch of statistics (dating from the seminal works of Fisher, 1935, and Yates, 1937). While biologists are likely to think that this attitude is a bit extreme (e.g., most biologists can design successful experiments with only the remotest knowledge of statistics), it is undeniable that close attention during the design phase to the form of the intended data analysis will greatly improve the chances that the experiment will successfully answer the question being examined. Sadly, this simple concept has not yet become standard practice in microarray experiments (Y.H. Yang and Speed, 2002), and so a brief discussion of the important issues is worthwhile here.

It is worth pointing out before we start that there are two distinct types of scientific experiments: manipulative (or intervention) experiments and descriptive (or observational) experiments. Descriptive experiments are those where the scientist does not try to modify the world in the process of the experiment (i.e., by manipulating the experimental conditions), whereas in manipulative experiments the scientist does modify the experimental conditions. Thus, manipulative experiments test hypotheses about processes while descriptive experiments test hypotheses about patterns. Most microarray experiments to date have been descriptive, but this does not mean that they must be so. Indeed, the two experiments for which we provide an example data analysis in the final section of this paper are clearly manipulative ones. Even though much is often made in the literature about the differences between these two types of experiments, from the point of view

of successful experimental design there is sometimes very little to choose between them.

Design

Experimental design is all about deciding how to deal with unwanted sources of variation in an experiment (Simon *et al.*, 2002). That is, we have an experimental question that concerns some particular source of biological variation (e.g., the relative effects of some experimental treatments, changing phenotypic expression during development, etc.) and to study this we wish to isolate it from the seemingly infinite other types of variation that occur in the universe—if we can separate out its pattern then we can see it clearly and thereby evaluate its relative importance. The other (unwanted) types of variation can include both biological variation (e.g., between tissues, between individuals, between populations, between species) as well as experimental or technical variation (e.g., between probes, between reactions, between dyes, between extractions). The sources of variation that we want to study are the "interesting variation" while the unwanted sources are the "obscuring variation."

There are three main ways that scientists have developed for dealing with unwanted sources of variation: (1) eliminate them, either by controlling the experimental conditions or by having an explicit control treatment; (2) randomize their effect, by having replicate samples; and (3) incorporating their effect into the data analysis, either as part of the preprocessing or as part of the quantitative analysis.

These strategies are not mutually exclusive, and they all should ideally play a part in most, if not all, biological experiments. As an example of the distinction between them, consider the unwanted variation caused by spatial irregularities within an array, which can affect the measurement of intensity for each gene. To deal with this problem we could try to improve the technology to manufacture near-perfect arrays, thus eliminating the problem—this would be strategy (1). Alternatively, we could also use several arrays for each experimental group, expecting that the spatial variation would "average out" across the replicates when we take their mean—this would be strategy (2). Finally, we could collect the data from a single imperfect array and adjust them mathematically afterwards to correct for the irregularities—this would be strategy (3). Currently, we use the latter approach, which is what preprocessing the data by normalization is all about. This does not, of course, stop the manufacturers from working on option (1), or us from also using option (2).

Laboratory scientists have always favored the first of these three strategies (which is why they devised laboratories in the first place). However, this strategy has the limitation that it is logically invalid to extrapolate the experimental results beyond the confines of the particular laboratory conditions used for the experiment. Many molecular biologists have apparently not worried much about this problem, as they seem to be quite happy to *assume* that their results can be generalized without further quantitative demonstration that this assumption is valid. In contrast, the advantage of the second strategy (replication) is that it explicitly demonstrates the generality of the experimental results, and it can be argued that proper replication is therefore a feature of all worthwhile experiments. This issue will be dealt with below. The third strategy (data processing) has been discussed in some detail in the preceding sections, and only one further comment will be made below.

Replication

Replication is at the heart of all statistical analyses, as it is the key to reliable inference from a sample to a population. It is the repetition of a pattern across a series of replicates that provides the convincing evidence that the pattern is real, because the repetition is unlikely to occur by chance. Most data analyses perform badly with only small amounts of replication, and if microarray experiments are poorly replicated then the tests are best treated as exploratory tools rather than as useful tests of scientific hypotheses (Long *et al.*, 2001). However, even when used as exploratory tools replication of arrays is beneficial (Dougherty *et al.*, 2002; Medvedovic and Sivaganesan, 2002).

Lack of replication has been a recurring comment with respect to microarrays. In fact, most of the hypothesis-testing techniques that have been proposed are designed specifically to try to get patterns out of the datasets in the face of extremely poor replication (or even entire lack of it). This situation seems to have arisen because of the focus of attention on the genomic aspects of the technology—the fascination with the sheer number of genes that can be incorporated into an experiment has drawn attention away from the other aspects of the experimental design. Furthermore, attention has focussed on sampling as many experimental conditions as possible with the limited number of arrays available, which must be at the expense of using those arrays for replicate sampling of the conditions. Presumably things will settle down soon, and replication will be seen as just as fundamental for array experiments as it is in all other areas of biology.

If replication is to be used then it is necessary to understand the essential features of what a replicate must be. The most important feature for true replication is that the replicated objects must be independent of each other—that is, a replicate is the smallest unit to which an independent application of the experimental procedures is applied. In microarray experiments it is frequently assumed that the individual spots on the slide or chip are the fundamental units that need to be replicated, because each spot is a separate application of a particular nucleotide sequence. However, repeated placement of the same sequence on a single slide/chip is not true independent replication, because the entire chip/slide is treated as a single experimental unit at some subsequent stages of the laboratory protocol (e.g., application of the test sample, and the ensuing hybridization, washing, and scanning). Consequently, it is the arrays that also need to be replicated, not just the spots/probes, because it is the arrays that are the "independent experiments" that can be used as "true replicates" in the data analyses.

The spots and probes can also be replicated if desired, but these are not independent replicates. This is often referred to as pseudoreplication (Hurlbert, 1984), to distinguish it from proper independent replication such as is assumed by the data analysis techniques discussed in the preceding sections. The most common form of experimental design for microarrays usually has no true replication (Y.H. Yang and Speed, 2002), and so the importance of this design feature is apparently either not understood by many biologists or is undervalued. One appar-

ent disincentive to proper replication is the financial cost of replicating arrays, but this cannot be allowed to outweigh scientific rigour (Firestein and Pisetsky, 2002).

Another way of thinking about replication (i.e., other than independence) is that it is supposed to be dealing with sources of variation. Microarray experiments have several distinct sources of variation, including variation among spots within an array (which is measurement error) and variation among arrays, both of which are technical errors, and variation among samples, which is biological variation (Churchill, 2002). Replication can be used to dispense with all three sources of variation, but different things need to be replicated in each case.

In a microarray experiment there are several sources of technical variation to be dealt with (summarized by Li *et al.*, 2002), and replicating spots on a single array will only deal with some of them while replicating arrays will deal with others. Since it is recognized that arrays are inherently variable (Bakay *et al.*, 2002; Li *et al.*, 2002; Nadon and Shoemaker, 2002; Novak *et al.*, 2002; Simon *et al.*, 2002), the arrays must be replicated to deal with this source of variation, because many of the sources of technical variation are beyond the experimenter's control. This is particularly true for genes expressed at low levels, which are almost always observed to be poorly reproducible in replicate arrays.

This raises the question of how many replicates should be used in a microarray experiment to deal with technical variation. There is no simple answer to this question, as the answer will vary from experiment to experiment because the relative expression levels vary from gene to gene and sample to sample, and there are actually several competing criteria to be considered. However, the consensus seems to be that at least three arrays is probably a good compromise between the various competing criteria (M.L.-T. Lee *et al.*, 2000; Black and Doerge, 2002; Grant *et al.*, 2002; Novak *et al.*, 2002), and five arrays might be sufficient to detect a 1.5-fold change in expression level (Gibson, 2002; Wierling *et al.*, 2002). It is unlikely that a twofold change could ever be reliably detected with a single array, given the amount of technical variation involved (Li *et al.*, 2002; Wierling *et al.*, 2002).

Perhaps a more important issue is the biological variation that may need to be dealt with by replication. If we wish to generalize to a large biological population (such as a tissue or a cell line) then taking a single extraction is clearly inadequate, as it will not deal with physiological variation. Furthermore, if we wish to generalize to an even larger biological population (such as a species) then taking a sample from a single individual is clearly inadequate—we will have eliminated interindividual variation (i.e., phenotypic and genotypic variation) from the experiment and thus cannot validly generalize beyond that single individual. Replicating biological samples, as well as experimental samples, is thus an issue that deserves closer attention, because it is likely to be the biggest source of variation in microarray studies (Bakay *et al.*, 2002; Novak *et al.*, 2002). Perhaps the most important limitation here is that biological variation can result in heterogeneous samples (e.g., mixed tissues from a single individual). Current data analysis techniques are ill-equipped to deal with heterogeneous samples, and the interpretation of their results can therefore be problematic (Grant *et al.*, 2002). It is thus sometimes recommended that only homogeneous cell populations should be used (Firestein and Pisetsky, 2002).

The question of how many replicates should be used to deal with biological variation is even more problematic to answer than is the question of technical variation. The standard statistical technique is to use what is called "power analysis," and this is discussed in more detail below.

A related question is how best to evaluate the repeatability of experimental replicates, either replicate spots on a single array or replicate arrays. One common suggestion is to plot one replicate against the other and then to measure the statistical correlation between them. However, this has long been recognized to be a poor method (Bland and Altman 1986), despite its very common use in microarray studies. It is likely to be an invalid use of correlation analysis because it will overestimate the reproducibility of the two replicates. A far better alternative (Hess *et al.*, 2001; Wang *et al.*, 2002b) is to produce a mean-difference plot for each pair of replicates, which is a scatterplot showing the difference between the paired observations (vertically) and their mean (horizontally). This is conceptually the same as the MA plot described above. It involves a 45° rotation of the scatterplot and a rescaling, compared to the standard plot. The rotation makes the plot more easily interpreted, because the differences between the replicates are confined to the vertical dimension (rather than being confounded across both dimensions as they are on the standard plot), and the rescaling helps emphasize differences at the equally important low expression levels (which in the standard plot get swamped by the higher expression levels).

Treatments

A different but related design issue is that if only one experimental treatment is applied per array then the array clearly needs to be replicated. That is, the experimental treatment needs to be applied separately to several arrays. If this is not done then it will be impossible logically to distinguish between the effects of the experimental treatments and the effects of differences between individual arrays. That is, if the expression of a gene is observed to be different between two arrays each containing a separate treatment, then is this difference caused by the inherent variability between arrays or by the influence of the experimental treatments?—we can't know. When it is impossible to distinguish between two effects in an experiment this is referred to as a *confounded* experimental design, and unwanted confounding is the cardinal sin that can be committed by an experimenter. A confounded design cannot be used to answer the experimental question convincingly, so it is of little value to a scientist.

In this regard, it is important to note that the experimental unit in a microarray experiment is an array, which may also be the unit used to deal with biological variation (since each array has a single biological sample applied to it). This can cause confusion about replication, because replicating arrays does not necessarily mean that there is replication of the experimental treatments. Replication of the treatments involves independent application of the experimental manipulation, as explained above, and replication of biological samples does not necessarily achieve this. For example, if several biological samples are subjected to a single experimental manipulation simultaneously, and each sample is then applied to a separate array, there will be biological replication but no replication of the experimental treatment.

If two experimental treatments can be applied per array (as they must be when using both red and green dyes in a cDNA array, for example), then clearly the comparison of those two experimental treatments can be made without the confounding influence of variability between arrays (i.e., the unwanted effect of interarray variation has been eliminated from the comparison). However, this does not relieve the experimenter of the necessity for replicating arrays. If arrays are inherently variable, then how do we know that the experimental result is not unique to this one particular slide? If we wish to generalize beyond the confines of that one array, then replicate arrays are necessary to demonstrate the generality. In particular, dye-swap (or reverse labelled) pairs are the best design for two experimental treatments (i.e., pairs of arrays in which the dye assignments are oppositein each member).

For cDNA arrays, an equally pressing point if two experimental treatments are applied per array is deciding which two treatments should be hybridized together on the same slide when there are more than two treatments being included in the experiment. This is important because there is more experimental variation between arrays than there is within arrays. This general issue has been addressed in detail by statisticians, and the application of the relevant points has been made to microarray experiments (e.g., Kerr and Churchill, 2001a; Churchill, 2002; Dobbin and Simon, 2002; Glonek and Solomon, 2002; Speed and Yang, 2002).

The important design issues are the "power" and "statistical efficiency" of the subsequent data analysis. The first criterion refers to the ability of the analysis to detect a pattern in the data when it is really there, and the second refers to minimizing the amount of resources used for achieving that power. These are clearly important criteria for experimenters, as failing to find a real pattern would be disastrous, and we all want to maximize the use of our limited resources. The most common form of design for microarray experiments is still the *reference design*, in which the second experimental treatment on each slide is merely the reference DNA sample that is being used for data standardization (see Fig. 3a). However, this is not necessarily the most efficient way to organize things, as it will usually produce a larger variance than will a *direct design* (see Fig. 3b), in which the comparisons of direct importance for the experimental hypothesis are made on each array (Glonek and Solomon, 2002; Speed and Yang, 2002). Therefore, more useful information will be obtained from the same number of arrays for a direct design compared to a reference design, because the technical variation is dealt with more efficiently by the direct design.

To a biologist, however, there are a number of advantages to using a reference design as opposed to a direct design. For example, the data derived from a reference design have intuitive and obvious biological interpretation because the experimental effects are all independently measured relative to the same standard. That is, if we are comparing two experimental treatments, A and B, using a reference design, then in the experiment we measure A and B independently and we subsequently work out their relative effects by calculating their difference (B − A). Using a direct design, however, we work out B − A directly, without ever estimating either B or A on their own. This means that we know their relative effects, which is the point of interest in the experiment, and we will have more efficient estimates of the appropriate numbers. However, biologists are also used to knowing both A and B, as these may have additional biological interest, and they seem to feel more comfortable with this approach, even if it is statistically less efficient. In addition, data analysis of a reference design is usually straightforward, being in principle no different to what biologists are used to. A direct design, on the other hand, requires modeling of the effects sizes (i.e., B − A), which is not necessarily straightforward for microarray studies and has not yet been resolved in any really satisfactory manner (Lönnstedt *et al.*, 2001). Furthermore, using reference samples allows the experiment to be conducted over an extended period of time, because all of the experimental samples do not have to be ready at the same time (which they obviously do if direct comparisons are involved).

When using a reference, it is important to note that unless a dye-swap design is used then the treatment effects and the dye effects will be confounded (Kerr and Churchill, 2001a; Churchill, 2002). This involves using two replicate arrays for each comparison, one using the red dye for the experimental sample and the green dye for the reference sample, and the other using the red dye for the reference sample and the green dye for the experimental sample.

Churchill (2002), Dobbin and Simon (2002), Simon *et al.* (2002), and Yang and Speed (2002) provide recent summaries and critiques of possible experimental designs that achieve maximum statistical efficiency, including extended direct designs such as loop designs. In this regard, if is worth noting that the reference design makes all comparisons between experimental

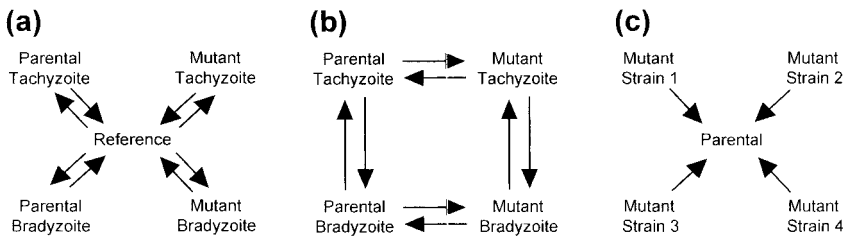


FIG. 3. Experimental designs for the hybridizations of the samples onto the cDNA microarrays in (**a**) the experiment of Matrajt *et al.* (2002), (**b**) a statistically more efficient design for the experiment of Matrajt *et al.* (2002), and (**c**) the experiment of Singh *et al.* (2002). The labels indicate the source of the cDNA samples (e.g., "Parental Tachyzoite" indicates the parental strain grown under tachyzoite conditions). Each arrow represents a single microarray, with the head of the arrow indicating the sample that was labeled with the green dye and the tail indicating the sample that was labeled with the red dye.

samples with the same statistical efficiency, even if that efficiency is lower than for some other designs that might be applied in particular situations. The efficiency of loop designs will vary from one situation to another.

Power analysis

The question of how many replicates should be used to deal with biological variation is an important one to answer. In statistics, this is dealt with via power analysis (Nadon and Shoemaker, 2002).

Power analysis is the investigation of the relationships between five aspects of statistical hypothesis tests: the probability of falsely rejecting a true null hypothesis (i.e., Type I error mentioned above); the probability of correctly accepting a true null hypothesis (called power); the sample size (i.e., the number of replicates); the variability of the data (i.e., the variation among sampling units); and the size of the pattern to be detected (i.e., the biological effect size).

So, the power of a specific statistical test can be estimated—given specified values for the size of the biological effect (e.g., the means), the variation among the sampling units (e.g., the standard deviations), the number of sampling units, and the probability of a Type I error, then the power can be calculated. In fact, given values for any four of these five characteristics then the fifth one can be calculated. This allows us to estimate, for example, the sample size necessary to detect a particular biological pattern, or the minimum biological effect size that can be detected with a given sample size. It is the estimation of sample size that is of most relevance here.

Power analysis has two main uses. First, it can be used to aid in the design and planning of the experiment (i.e., a prospective analysis). If there is prior knowledge about the likely size of the pattern and the likely variability of the data (e.g., from previous studies), then the power analysis will indicate the sample size needed to detect the pattern. Alternatively, given the same information and a specified sample size, the power of the analysis can be evaluated; or given the power and the sample size, the size of the pattern capable of being detected can be calculated. The experimenter can then evaluate the potential usefulness of conducting the proposed experiment.

Second, power analysis can be used to evaluate any nonsignificant results from the statistical analyses in an experiment (i.e., a retrospective analysis). The actual power inherent in the analysis can be calculated, and consequently, the ability of the experiment to detect the hypothesized pattern can be evaluated. Note that the "hypothesized pattern" should be one that is thought to be biologically important, rather than merely the one observed in the experiment. The experimenter can then objectively assess the success of the experiment—was the failure to detect a pattern because there really was no pattern or was it because the experimental design made it unlikely that the real pattern would be detected?

Power analysis can be difficult to apply to microarray experiments because not all of the necessary information will necessarily be available (e.g., we would need to know the variability of the expected gene expression levels). Black and Doerge (2002), Yang and Speed (2002), and Simon *et al.* (2002) discuss some possible strategies. Furthermore, if nonparametric procedures are to be used for the data analysis then nonparametric power analysis needs to be used, and this is more complicated to carry out than parametric power analysis (requiring permutations or bootstraps; see Black and Doerge, 2002). More general books about power analysis and sample-size estimation include those of Kraemer and Thiemann (1987), Cohen (1988), Desu and Raghavarao (1990), and Lipsey (1990).

Methods of power analysis exist for most statistical hypothesis-testing procedures. However, a particular problem is the application of power analysis to multivariate datasets, as there is no intuitively obvious way to do it. Measuring the size of the biological effect is not straightforward for multivariate data, and there are several aspects of the data analysis that are affected by the variation among the sampling units. Therefore, sample size affects the precision of the analysis (i.e., how representative the sample is of the true population) and the stability of the results (i.e., variability across repeated sampling) in complex ways (MacCallum *et al.*, 1999), and it is rare that general suggestions can be made about sample size and power for multivariate datasets. Nevertheless, Hwang *et al.* (2002) have made one specific suggestion for microarray data.

EXAMPLE ANALYSES OF MICROARRAY DATA

Many of the points made above might be clearer if we look at a specific example or two. Here we will particularly consider: unsupervised multivariate analysis, multifactorial analysis of variance for complex experimental designs, the use of permutations in relation to violation of statistical assumptions, multiple hypothesis testing, and power analysis with reference to replication.

We will consider the analysis of two published datasets. These data can be analysed in various ways to illustrate some of the theoretical points made in previous sections. The two sets of data are based on quite similar experiments, but the problems encountered in the data analyses are quite different, and they will be used to illustrate different topics. This will highlight the need for careful thought about how to proceed with the analysis of microarray data.

We have restricted ourselves almost entirely to standard statistical techniques, as no really good reason has yet been given for deviating from them for the analysis of microarray experiments. In particular, we have used the standard framework of analysis of variance, which is specifically designed for replicated multifactorial experimental designs. Almost all of the "problems" identified in the literature with respect to the analysis of microarray data stem from a lack of replication in the experiments, not from any inherent characteristics of microarrays themselves. We are therefore looking to the future rather than the past, on the assumption that replication will shortly increase to an acceptable level.

Note that what we are doing here is attempting to show some of the possibilities for microarray data analysis, and to make some important points about the design of array experiments. We are not attempting a reanalysis of the data from these experiments, nor are we trying to provide a definitive analysis of these data. There is no necessary reason why our analyses should produce the same conclusions as those reached by the original authors. In particular, we have restricted ourselves

solely to those subsets of the data that appeared to us to be "complete" for the analyses that we wished to do. Our emphasis is very much on evaluating the quality of the data used to draw scientific inferences. Furthermore, we cannot give a detailed biological interpretation of the results for either experiment, because the information accompanying the datasets is incomplete. This means that we may also have misinterpreted some aspects of the data and how it was collected. Standards have recently been proposed for the information content of published microarray datasets that are to be archived (e.g., Brazma *et al.*, 2001; Spellman *et al.*, 2002), and we recommend that everyone support these attempts at appropriate standardization.

Finally, the data analyses that we discuss here can be carried out via a wide variety of computer programs. We have used an arbitrary selection here, and our use of these particular programs should not be seen as a recommendation. Far more convenient forms of microarray data analysis are likely to be in the pipeline. However, in science, an explicit statement should always be made about which programs have been used, and which versions of these programs, just as the details of equipment and chemicals are provided for laboratory protocols.

First dataset

Matrajt *et al.* (2002) describe a study conducted with *Toxoplasma gondii* (Apicomplexa), in which the biology of conversion between the tachyzoite and bradyzoite life cycle stages *in vitro* were studied by gene expression profiling. This study included mutant strains which lacked the ability to form bradyzoites compared to a parental line. Thus, the main purpose of this study was to investigate the changes in gene expression during stage conversion, thereby providing an insight into the mechanisms of this process. As part of their experimental plan they performed a microarray experiment to examine stage-specific expression on a large number of genes. They used a reference design, in which each of the experimental samples was compared to a reference sample on each array slide. The processed data from this experiment are available at www.blackwell-science.com/products/suppmat/mole/mole2904/mmi2904sm.htm, with the original array data in the Stanford Microarray Database (http://genome5-www.stanford.edu).

We can start the discussion by thinking about the experimental design itself. Part of this design process involves consideration of which cDNA samples should be hybridized together on each array. This particular experiment involved a dye-swap reference design, as shown in Figure 3a. That is, each array involved a comparison between a cDNA sample from one of the experimental conditions and a reference sample. Furthermore, each of these comparisons involved two arrays, one with the reference sample labeled with the green dye and one with it labeled using the red dye. So, there was a total of eight arrays used for the experiment (i.e., the eight arrows shown in the figure). Statistical considerations suggest that there is actually a more efficient way to design this sort of experiment using the same number of arrays, as shown in Figure 3b. This design does away with the reference sample, and each array is then used for a direct comparison between two of the experimental cDNA samples. This design is then a dye-swap direct design. It can actually be performed with only six arrays, if necessary, as explained by Gronek and Solomon (2002).

Let's start the data analysis by taking a brief look at the necessity for preprocessing the output from the scanning of the arrays themselves. Preprocessing involves filtering of the data followed by standardization. Standardization itself can consist of transformation and normalization of the intensity values.

As far as filtering is concerned, we are told for this experiment that "no value was entered for poor-quality spots (based on measurements of pixel distribution and background fluorescence)." The unwanted effect of background fluorescence is usually dealt with by subtracting the median background fluorescence intensity from each of the spot intensities for both of the dyes. At this stage some of the spots will be considered unsuable because the spot intentsity is less than the median background intensity. As an example, array number 17081 (which involved a comparison of cDNA from the parental strain grown under bradyzoite conditions and the reference cDNA) has 5376 spots, 206 of which have intensities less than the median background intensity for one of the two dyes. This leaves 5170 potentially usable spots.

Transformation of the intensities usually involves taking the $\log_2$ value of the ratio of the background-corrected intensity of the sample dye to that of the reference dye. This corrects for the lognormal frequency distribution of the intensities, which results from the multiplicative effects of changes in expression among the ESTs. Normalization then involves correcting for unwanted spatial variation in intensity across each microarray. This can be done most effectively by assessing an MA plot, as shown in Figure 4 for array number 17081. If there is no unwanted spatial variation in intensities across the array then the ESTs would be randomly scattered around a horizontal line on such a plot. This is clearly not the case for this array. At low average intensities the red dye is stronger than the green dye, as it also is at high intensities, while the reverse is true at medium intensities. The normalization procedure involves "straightening out" the curvilinear line shown on the graph, which is done by subtracting from each spot intensity ratio the value predicted by the LOESS line (i.e., the normalized y-value equals the observed y-value minus the y-value predicted by the line). Note that the spots that are a long way from the LOESS line are the ones that represent ESTs that are either overexpressed (above the line) or underexpressed (below the line).

Having dealt with the normalization of each array, we can now proceed to consider the biological variability between the two independent experimental inductions (i.e., the replicate arrays for each experimental condition). Currently, the best way of comparing the results of replicate microarray experiments is via a mean-difference plot, as shown in Figure 5 for the *T. gondii* data. If the two experiments produced identical results then the points on the graph would form a horizontal straight line (the dotted line at $y = 0$ on the graph), and vertical deviations from this line represent the between-experiments variability. In this example, there are two points to note. First, there is quite a degree of scatter in the points around the line, indicating that there is considerable interexperiment variability. Therefore, in this system it seems essential to have replicate experiments, as the results of a single experiment might not be highly repeatable. It is for this reason that we decided to restrict all of our subsequent analyses to only those ESTs for

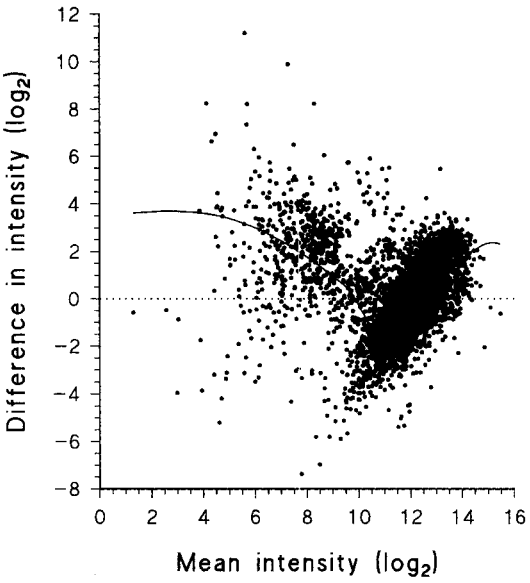


FIG. 4. MA plot of the two dye intensities from one of the arrays (no. 17081, parental strain grown under bradyzoite conditions) from the experiment of Matrajt *et al.* (2002). Each symbol represents one of the 5170 spots, with the horizontal axis being the average of the background-corrected $\log_2$ intensity values for that EST and the vertical axis being their difference (red minus green). The solid line is a LOESS smoother (span = 50%), calculated using the SYSTAT (v. 9.01) program of SPSS (1998), which shows the nonlinear relationship between the two variables caused by spatial variation in intensities. The dotted line is the expected relationship between the two variables if there was no spatial variation of intensity within the microarray.

which there are data for two experiments (i.e., 48% of the original number of ESTs).

Note that we have therefore explicitly dealt with missing data by excluding ESTs with observations lacking for any of the eight experimental samples. This is quite drastic for this dataset because it excludes more than half of the ESTs. As an alternative strategy, we might have included the 908 ESTs for which only one of their eight observations was missing, by imputing the missing value for each of these ESTs. However, this would mean that we were imputing 3.8% of the data, which may be an unduly large amount of "fake" data. Furthermore, the overall data are clearly not missing at random, but are concentrated in two of the experimental replicates, which have 28 and 37% of their data missing (the other replicates have only 1–15% missing data). Also, of the 908 single-missing ESTs 61% have the missing observation in the same replicate, which is clearly not a random arrangement and thus makes us doubtful of the utility of any imputation procedure in this case.

The second point to note about the mean-difference graph is that the points are evenly scattered around the "expected" line except in the bottom left-hand corner. This phenomenon is quite common in microarray experiments, and indicates that small expression levels (to the left on the graph) are likely to be measured in a biased fashion. Therefore, small expression values (underexpressed ESTs) are not to be trusted too much.

An initial experimental question concerning the microarray data concerns a comparison of the expression fingerprints (or expression signatures) of the various experimental conditions (see Fig. 1). That is, we wish to know whether the parental and mutant parasite strains produce the same relative EST expression levels under the bradyzoite and tachyzoite cultivation conditions. For illustrative purposes, this may be done by an exploratory multivariate pattern analysis. To do this, we first calculate some measure of the pairwise similarity among the four possible expression fingerprints, and then we display the resulting patterns in a graph. This is a multivariate data analysis because we are searching for a general (or common) pattern across a large number of variables (the ESTs). That is, we are summarizing the pattern shown on average by all of the ESTs, the pattern being the relationships among the experimental conditions. It is an unsupervised analysis because we are not specifying exactly what pattern to look for.

A commonly used measure of similarity is the correlation coefficient, and some possible calculations for this measure are shown in Table 1 for the *T. gondii* data. Before we proceed to display these data in a graph, we can pause to further consider the effect of the experimental replications on these calculations. In the top half of the table are shown the correlation values based on both experiments, using all of the data (in the upper triangle) and only our preferred subset (in the lower triangle). Note that while these two datasets produce quite similar results they are not identical. More importantly, in the bottom half of the table are shown the correlation values based on each of the experiments individually. These two datasets do not produce very similar results at all, thus reemphasizing our claim that replicate data are needed for the ESTs in this system.

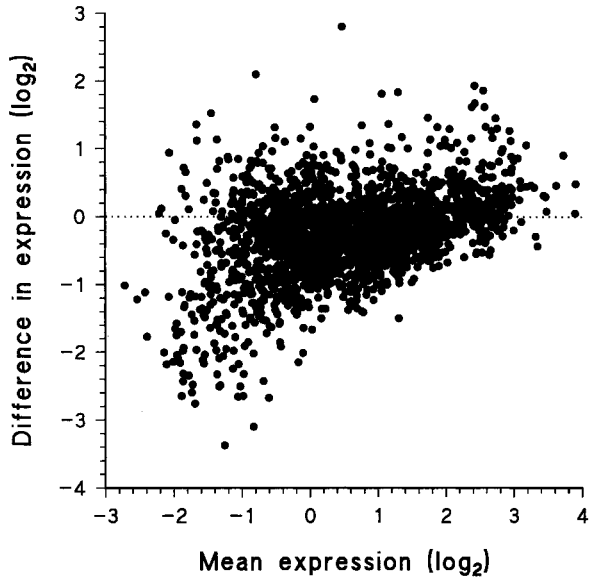


FIG. 5. Mean-difference plot of the ESTs from the parental strain grown under bradyzoite conditions from the experiment of Matrajt *et al.* (2002). Each symbol represents one of the 2066 ESTs, with the horizontal axis being the average of the $\log_2$ expression values for that EST from the two experiments and the vertical axis being their difference.

The patterns are not presented very well visually in such a table, however. A graphical display of the pattern shown by the correlation values based on a multidimensional scaling ordination is therefore shown in Figure 6a. Each point on the graph represents a single experimental condition, and the closeness of the points shows how similar are their expression fingerprints. From this ordination analysis we might conclude that the parasite strains grown under tachyzoite conditions produce very similar expression fingerprints, as expected from the experimental hypothesis being tested, while the parasite strains grown under bradyzoite conditions produce ones that are quite different from each other as well as from the tachyzoite conditions. Note that these conclusions are not easy to reach just from a perusal of the original correlation values in the table.

However, this graph does not really allow us easily to test the original experimental prediction, which was that the mutant strain grown under bradyzoite conditions (the symbol in the top left corner of the graph) will produce an expression fingerprint that is more like that of the strains when grown under tachyzoite conditions (the two symbols at the bottom of the graph) than will the parental strain grown under bradyzoite conditions (the symbol in the top right corner of the graph). So, we performed a minimum spanning tree analysis (a type of clustering analysis) of the correlation values as well, to see if this helps us make this decision, because a combination of ordination and clustering can often reveal patterns that neither analysis can alone. The result of this tree analysis is shown on the graph (Fig. 6a) as the dotted lines. The relative length of the two relevant lines indicates some support for the hypothesis (i.e., the line on the left is slightly shorter than the one on the right), but it is hardly convincing evidence.

More to the point, we need to evaluate the robustness of our conclusions from this analysis. This is an important consideration for all exploratory data analyses, when there is no formal statistical test of the validity of the interpretation. There are two possible components to the evaluation here. First, we have data from two experimental replicates, and it would thus be better to analyze these separately to see if they produce the same pattern, rather than averaging them as was done above. Second, we need to know if our conclusions depend on the choice of similarity measure used (i.e., the metric). We might do this by also analyzing the data using the Gower coefficient (this is the standardized version of the Manhattan distance just as the correlation coefficient is the standardized version of the Euclidean distance).

The results of this robustness evaluation are shown in Figure 6b. The symbols represent the same things as before, but this time there is a separate symbol for each of the two experimental replicates. The pairs of similar symbols are clearly near each other, but they are not in exactly the same spots and this indicates that there is considerable experimental variation contributing to any conclusions that we might derive from the multivariate analysis. More importantly, the use of the different similarity measure has drastically altered the apparent pattern in the data. The pattern is still not strong, but both the ordination and the tree analyses indicate that we should reject our hypothesis: the mutant strain grown under bradyzoite conditions has an expression fingerprint that is *less* like that of the strains when grown under tachyzoite conditions than does the parental strain grown under bradyzoite conditions. Once again, there is certainly *some* evidence in the data but it is equivocal, because our result appears to be sensitive to the choice of similarity measure.

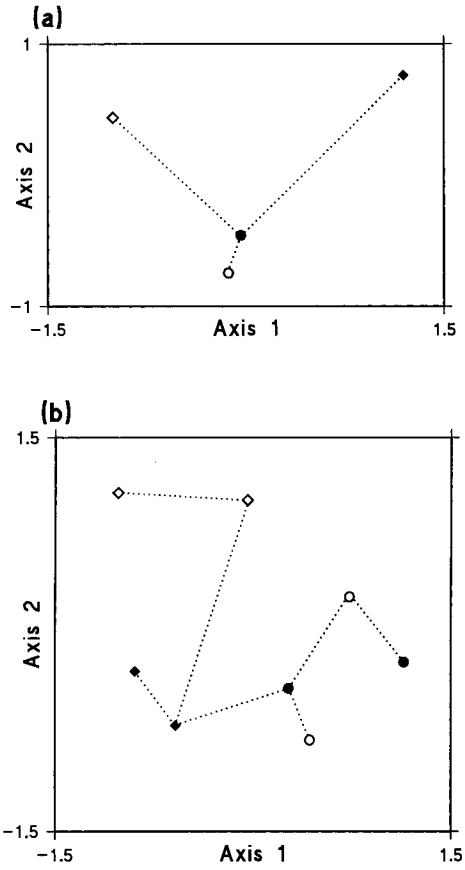


FIG. 6. Results of the two-dimensional metric multidimensional scaling (MDS) ordination and minimum spanning tree (MST) analyses, based on the 2066 ESTs from the Matrajt *et al.* (2002) data set, of (**a**) the Pearson correlation coefficients of the four combinations of the experimental conditions, and (**b**) the Gower coefficients of the eight experimental units. For the ordination analyses, each symbol represents one of the objects analyzed, and their spatial proximity indicates how similar they are in terms of their EST expression fingerprints. For the tree analyses the dotted lines join the most similar objects based on their EST expression fingerprints. The solid symbols represent the parental strain and the closed symbols the mutant strain; the ○ symbols represent the tachyzoite culture conditions and the ◇ symbols the bradyzoite culture conditions. Calculated using the PATN (v. 3.5+) program of Belbin (1995).

This brings us to the important issue of power analysis. Part of the problem encountered in discerning any clear pattern in these data might be due to the relatively large amount of variability between the replicate arrays. We might therefore ask ourselves how many replicate experiments we would need to make the pattern clearer. This is no easy matter to address for multivariate pattern analyses, but using the method suggested by Hwang *et al.* (2002), we can make a stab at it.

The results of this power analysis are shown in Figure 7. This graph shows the general form of this type of power anal-

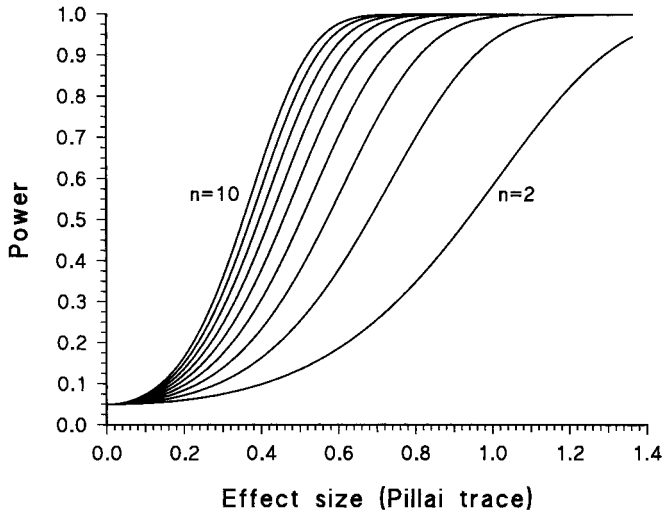


FIG. 7. Power analysis plots of the multivariate data analysis for all sample sizes from 2–10 replicates. The effect size is the magnitude of the pattern in the dataset, measured as the Pillai trace from a discriminant function analysis of the groups representing the four combinations of experimental conditions. The power is the probability of correctly accepting a true null hypothesis. Alpha is the probability of falsely rejecting a true null hypothesis, and was set at $P = 0.05$. Calculated using the GPower (v. 2.0f) program of Erdfelder *et al.* (1996).

ysis. Each of the lines represents a different number of replicates (varying from 2–10 in our example), and shows how the power (on the vertical axis) changes with variation in the size of the "biological pattern" to be detected in the data (on the horizontal axis). Naturally, the power increases as the magnitude of the pattern increases, but what we are particularly interested in is how this relationship is affected by the number of replicates that we use. The graph shows that there are major improvements in power if we increase the number of replicates from 2 to 3 or 4, but there are only relatively minor improvements after that. So, we would receive considerable benefits from even one more replicate (although this is a 50% increase in the size of the whole experiment!).

Note that the assessment of replication in power analyses applies only to the sources of variation taken into account by the replicates actually used—if other sources of variation need to be dealt with then this may require more replicates. In this example, the replicates are apparently replicate experiments performed on the same culture from a single parasite strain. If there is variability in response to the growing conditions, either between cultures or between strains, then the experiment cannot reveal this. The replicates are thus in some ways closer to technical replicates than to biological replicates.

This brings us to an alternative, and much more appropriate, set of analyses of the same data. This is actually a "designed" experiment, in the sense that the experimental manipulations were carried out by the experimenters. The pattern analysis that we performed above ignored this fact, merely summarizing the data and hoping that the data summary was relevant to our experimental question. We could pursue this form of analysis, for example, by using one of the multivariate class-prediction analyses (i.e., a supervised technique) to assess which ESTs are the ones that show the biggest difference in expression between the culture conditions and parasite strains. However, such an approach is not likely to be productive because there are only two experimental replicates for each of our groups.

Furthermore, while such an approach can be effective for the exploration of a complex biological situation, it cannot be a substitute for the rigorous testing of experimental hypotheses in a designed experiment. It would be better to explicitly take the experimental design into account in our analysis, in which case the experiment should properly use hypothesis-testing analyses rather than exploratory ones. In particular, we chose to use this example because the experimental design is a bit more sophisticated than a simple two-sample treatment/control comparison, such as is most common in biology. It therefore has interesting potential statistical data analyses, including multifactorial analysis of variance (ANOVA).

The important point about the experimental design is that it must match the data analysis that is to be used. That is, the statistical question being tested should match the experimental question. In this example, the form of the analysis of variance should match the experimental question being asked. One of the most common problems with the use of ANOVA in the biological literature is using a much simpler design for the analysis of variance than that specified by the experimental design.

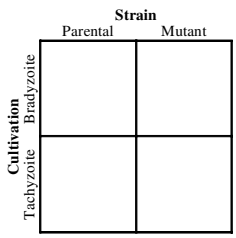
To look at possible hypothesis-testing analyses, we can consider a comparison of the expression profiles of the various ESTs (see Fig. 1). Each of these profiles is made up of eight observations, one for each of the replicates of each of the experimental conditions. Furthermore, these eight observations can be grouped together in various ways. For example, some of the observations naturally group together because they come from the same cultivation condition and others naturally group together because they come from the same parasite strain. In a statistical analysis, these various ways of grouping the observations are called "factors." Thus, factors are covariables that form groups of genes or arrays.

The *T. gondii* experiment was designed so that there are three factors:

1. Strain—this factor forms two groups, one for data that came from the parental strain and one for data from the mutant strain;
2. Cultivation—this factor forms two groups, one for data that came from the tachyzoite cultivation conditions and one for data from the bradyzoite cultivation conditions;
3. EST—this factor forms 2066 groups, one for each of the ESTs.

It is therefore possible to use a statistical test to assess whether the average expression levels are different between these various possible groups, which constitute the statistical hypotheses being tested. The appropriate statistical test is a three-factor orthogonal analysis of variance, which is a univariate statistical analysis that will simultaneously test several hypotheses about differences between groups.

In addition, it is possible for each of these three factors to form interaction groups, and for these interactions also to be tested in the same analysis. For example, the interaction between the first two factors forms four groups, one for each of the four experimental conditions, and it is possible to separately test whether these four groups differ:



There are three possible two-factor interactions that can be tested by the ANOVA, as well as one three-factor interaction. This means that a total of seven statistical hypothesis tests are carried out as part of the one analysis (three single factors, three two-factor interactions, and one three-factor interaction).

As discussed in a previous section, when there are multiple factors in an ANOVA then it is necessary to consider whether each of the factors is random or fixed. We think that the EST factor should be treated as random in a screening study such as this, since the ESTs chosen are an arbitrary selection of all of the available possibilities (the factor might be fixed for an array with a small number of genes or a focus on a specific metabolic pathway). The other two factors in the experiment are fixed, since in each case they represent the only two possible groupings that are of interest in this particular experiment.

Before proceeding to the results it is important to emphasize two points. First, we are doing the analysis of variance as a screening procedure, to control for the problems of performing multiple hypothesis tests on the 2066 ESTs. We will only proceed to test the ESTs individually if the ANOVA indicates that it is worthwhile to do so. This helps provide protection against unwarranted conclusions. Second, the most popular forms of statistical analysis for microarray data, such as significance analysis of microarrays (SAM), are not currently designed to handle experimental designs with multiple factors. Consequently, they are inappropriate for analyzing the data here, because we need to match the statistical analysis to the actual experimental design.

As an aside, it may be worth noting that, properly speaking, a factor dealing with variation between the arrays should be included in this ANOVA. This is because the ESTs for each array are not independent of each other, and so the arrays are actually not independent replicates with respect to the EST factor. This array factor would be a random factor nested within the Strain*Cultivation interaction, and would thus be orthogonal to the EST factor. We have not included it here because it complicates the analysis unnecessarily, and its inclusion does not change any of our substantive conclusions.

TABLE 1. PEARSON CORRELATION COEFFICIENTS AMONG THE ESTS FROM THE MATRAJT *ET AL.* (2002) DATA SET, BASED ON FOUR SUBSETS OF THE ENTIRE MICROARRAY DATA

Strain	*Cultivation conditions*	*Parental*		*Mutant*	
		Tachyzoite	*Bradyzoite*	*Tachyzoite*	*Bradyzoite*
Based on the average of both experiments[a]					
Parental	Tachyzoite	—	0.754	0.942	0.818
	Bradyzoite	0.781	—	0.746	0.706
Mutant	Tachyzoite	0.916	0.739	—	0.810
	Bradyzoite	0.834	0.727	0.825	—
Based on individual experiments[b]					
Parental	Tachyzoite	—	0.761	0.859	0.727
	Bradyzoite	0.680	—	0.712	0.740
Mutant	Tachyzoite	0.832	0.669	—	0.670
	Bradyzoite	0.807	0.626	0.753	—

For each of the subsets, the coefficients are shown for all pairwise combinations of the experimental conditions of parasite strain (parental versus mutant) and cultivation conditions (tachyzoite versus bradyzoite).

[a]Above the diagonal are shown the correlation coefficients for all 4307 ESTs, averaged across both experiments; below the diagonal are shown the correlation coefficients for only those 2066 ESTs that have data for both experiments, averaged across both experiments.

[b]Below the diagonal are shown the correlation coefficients for the subset of 2066 ESTs based on experiment 1 only, with those for experiment 2 only above the diagonal.

The results of the analysis of variance are shown in Table 2a, in standard format. This analysis indicates that we should reject the null hypothesis (i.e., $P < 0.05$) for each of the seven hypothesis tests. We are actually most interested in the Strain*Cultivation interaction, because this is the comparison of the four groups formed by the experimental conditions. A summary of the actual expression data analysed for this interaction is shown in Figure 8. The ANOVA indicates that the differences among these four groups are statistically significant (as can also be seen from the 95% confidence intervals on the graph, which do not overlap). The pattern shown here confirms what we saw from the multivariate pattern analysis: the mutant strain grown under bradyzoite conditions has an average expression profile that is *less* like that of the strains when grown under tachyzoite conditions than does the parental strain grown under bradyzoite conditions.

However, the ANOVA also indicates that we must place a caveat on this conclusion. This is because the three-factor interaction is significant. What this means is that only *some* of the ESTs show this pattern (i.e., the one seen in Fig. 8), while others do not. These other ESTs may show the opposite pattern or they may show no pattern of differences among the experimental conditions at all. In other words, the results of the experiment are not quite as simple as we have suggested.

To investigate this pattern in more detail we have a number of options open to us. The basic problem is that we now need to examine each EST individually, and this brings into play the multiple-testing problem referred to in a previous section. It also raises the issue of exactly what testing procedure we might use to examine each EST. For illustrative purposes, we have chosen simply to analyse each EST with a two-factor orthogonal analysis of variance, with Strain and Cultivation as the two factors. This is a somewhat rough-and-ready way of doing it, but few better procedures have yet been proposed (and it is conceptually quite close to the recommended approach of Dudoit *et al.*, 2002c). Furthermore, for illustrative purposes we have performed the probability calculations using the standard (normal-theory) procedure and also using permutation testing.

A summary of the results of these ANOVAs is shown in Table 3. Only 48 of the ESTs are shown in the table, leaving out those for which we have concluded there is not likely to be a significant pattern. As before, we are only interested in the two-factor interaction, and thus the table shows those ESTs that possibly show a significant difference between the four groups formed by the experimental conditions.

The first thing to note from the table is that the normal-theory probabilities are usually much smaller than are the permutation probabilities. We believe that the permutation probabilities are likely to be more realistic. However, there are in fact only 105 possible permutations from which to calculate these probabilities, while it is usually recommended that 5000 permutations be used, and so there may be some limitation here. Furthermore, we couldn't use permutations to estimate the probabilities for the three-factor ANOVA above because there are practical problems, which are discussed further below.

Second, if we are to take the multiple-testing problem seriously, then we would probably conclude that none of these ESTs individually show any significant pattern at all. This is because the smallest probability is only $p = 0.00007$, which is not very small given the large number of hypothesis tests that were performed. This issue of multiple hypothesis testing is also discussed further below. Finally, if we do accept that some of these ESTs show a significant pattern, then we can see that the majority of them do indeed fit the general pattern, although as

TABLE 2. RESULTS OF THE ANALYSES OF VARIANCE OF THE TWO EST DATA SETS FROM THE MATRAJT *ET AL.* (2002) DATA SET

Source of variation	*Degrees of freedom*	*Mean-square*	F-*value*	P
(a) Analysis of the expression of the 2066 ESTs[a]				
Strain	1	29.8571	47.129	<0.00001
Cultivation	1	1094.1318	1642.391	<0.00001
EST	2065	9.8685	28.804	<0.00001
Strain*Cultivation	1	121.8546	231.274	<0.00001
Strain*EST	2065	0.6323	1.846	<0.00001
Cultivation*EST	2065	0.6662	1.944	<0.00001
Strain*Cultivation*EST	2065	0.5269	1.538	<0.00001
Residual	8264	0.3426		
(b) Analysis of the induction of the 459 ESTs[b]				
Strain	1	3.1088	4.513	0.03665
Cluster	82	1.5539	3.927	<0.00001
Strain*Cluster	82	0.6889	1.581	0.00241
EST(Cluster)	376	0.3957	0.908	0.82512
Residual	376	0.4358		

[a]Three-factor orthogonal analysis of the expression data, with Strain and Cultivation as fixed factors and EST as a random factor. Calculated using the GeneANOVA program of Didier *et al.* (2002), based on the expected mean-squares derived by the DESIGN (v. 3.0) program of Dallal (1988).

[b]Three-factor mixed analysis of the induction data, with Strain as a fixed factor, Cluster and EST as random factors, and EST nested within Cluster. Calculated using the SYSTAT (v. 9.01) program of SPSS (1998), based on the expected mean-squares derived by the DESIGN (v. 3.0) program of Dallal (1988).

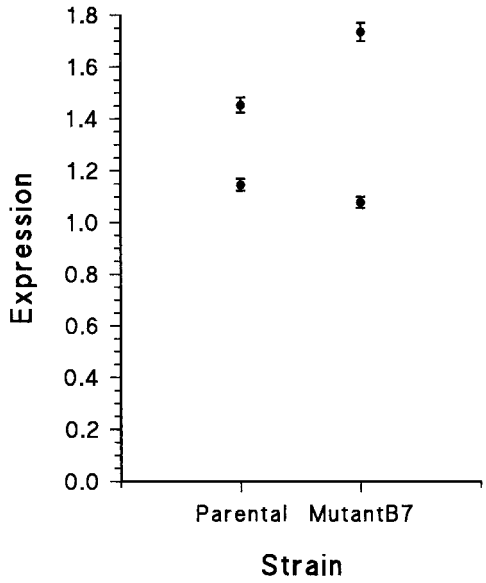


FIG. 8. Expression levels for each of the four combinations of experimental conditions averaged across all 2066 ESTs from the Matrajt *et al.* (2002) data set. The lower pair of symbols represent the tachyzoite culture conditions and the upper pair the bradyzoite culture conditions. The error bars are the pooled 95% confidence limits as derived from the orthogonal analysis of variance (Table 2a).

expected there are also some that show the opposite pattern (i.e., the pattern that we originally hypothesized).

It is, of course, good to also have a picture of the full dataset, with which to compare these "significant" results. The best picture for this situation is known as a volcano plot, as shown in Figure 9. ESTs that fit the originally hypothesized interaction pattern are shown on the left of this graph, while those that do not fit are to the right, and those showing no difference between the parental and mutant strains are in the middle. The statistically significant ESTs are at the top of the graph and the nonsignificant ones are at the bottom. Note, first, that there are relatively few ESTs fitting the hypothesized pattern compared to those that contradict it, just as we would expect from the subset of statistically significant results shown in Table 3. Second, note that the results form a continuum, so that the boundary of the "statistically significant" region is completely arbitrary. This is a reminder that probabilities should not be overinterpreted. Third, note that there are very many genes with "strong" patterns (i.e., outside the dotted lines) that are not statistically significant (i.e., below the dashed line). Indeed, the strongest patterns in both directions are actually not statistically significant at all. This is because of the variability among the replicate arrays for these genes; and therefore, a larger number of replicate arrays would probably result in many more statistically significant patterns. Similarly, there are several genes with relatively "weak" patterns (i.e., inside the dotted lines) that are statistically significant (i.e., above the dashed line). This is because of the lack of variability among the replicate arrays for these genes, and therefore, we should treat these particular results with some caution because of the small number of replicates (e.g., if their weak patterns are confirmed by increased replication then they may not be of much biological interest).

This brings us to a consideration of power analysis for this type of statistical hypothesis testing. Power analysis is much better developed for statistical tests than for multivariate analysis, and so the analysis is relatively straightforward. Let us consider the two-factor ANOVA that we have just been looking at, where we consider that the tests might not be very powerful (i.e., we are doubtful that many of the tests should be considered to be significant). Here, we have a single specified biological pattern that we are testing (i.e., we have a prediction about what the pattern will be), which in this case is that the relative microarray expression levels will be 1 (parental parasite strain under the tachyzoite cultivation conditions): 1 (mutant strain under tachyzoite conditions): 1.5 (mutant strain under bradyzoite conditions): 2 (parental strain under bradyzoite conditions).

The results of this power analysis are shown in Figure 10, as the solid line. This graph shows how the power (on the vertical axis) changes with variation in the number of replicates (on the horizontal axis). Clearly, as expected, our two replicates are completely inadequate for this analysis, as their power is very small. To get a decent power size, say 0.8, would require six replicate experiments. This may sound outrageous, but these are the facts of the case—if you want to do serious amounts of testing of individual ESTs, based on relatively small differences in expression level, then you have to have a serious number of replicates.

Note that, unlike the first power graph, the results of this power analysis (and the next ones) apply solely to the experiment conducted here—the estimated number of replicates required cannot necessarily be used as a guideline for any other experiments. Furthermore, we have used parametric statistical tests for the data, and so we have used parametric power analyses. It is often possible to perform the power analyses using permutations, as well, which would be more appropriate if permutation tests are being used for the formal analyses.

As an aside, before we proceed to the next substantive analysis, it might be worth pointing out that it is also possible to statistically test the multivariate pattern shown in Figure 6b. This pattern was based on estimating the pairwise similarity among the expression fingerprints, with the pattern displayed showing the relative similarity among the eight experimental treatments. If we take the complement of the similarity measures then they are distances (equivalent to the visual distances between the points shown on the graph), and these distances can actually be analyzed using analysis of variance, as discussed by Anderson (2001) and McArdle and Anderson (2001). This would be a two-factor orthogonal analysis of variance, with Strain and Cultivation as the two factors, as we have just discussed. The only difference is that the F-value produced from a distance-based analysis is not a true F-value (i.e., its frequency distribution will not follow the usual F-distribution), and so the probability must be calculated using permutations.

The results of this distance-based analysis using both the Gower coefficient and the correlation coefficient are shown in Table 4, in standard format. Note that both analyses agree that there is little evidence to reject the null hypothesis for the interaction term (i.e., $P > 0.05$), in contrast to the results of the three-factor analysis shown in Table 2a. This is because of the

TABLE 3. RESULTS OF THE TESTING THE EXPRESSION OF EACH OF THE 2066 ESTS FROM THE MATRAJT *ET AL.* (2002) DATA SET INDIVIDUALLY USING A TWO-FACTOR ORTHOGONAL ANALYSIS OF VARIANCE

EST clone	*Interaction* F	*Normal* P^a	*Permutation* P^b	*Pattern*[c]	*EST clone*	*Interaction* F	*Normal* P^a	*Permutation* P^b	*Pattern*[c]
tgzz67h10.r1	285.768	0.00007	0.0108	Alt	tgzz53g01.r1	32.969	0.00456	0.0060	Alt
tgzz58g04.r1	102.250	0.00054	0.0132	Alt	tgzz63h04.r1	31.453	0.00497	0.1032	Alt
tgzz68g05.r1	83.839	0.00079	0.0372	Alt	tgzz64g08.r1	30.764	0.00517	0.0036	Alt
tgzz64h06.r1	71.863	0.00106	0.0040	Alt	tgzz59c05.r1	30.207	0.00534	0.0010	Alt
tgzz64e07.r1	68.173	0.00117	0.0128	Alt	tgzz38h01.r1	29.768	0.00549	0.0074	Exp
tgzz60c05.r1	67.656	0.00119	0.0072	Alt	tgzy48d04.r1	28.211	0.00604	0.0266	Alt
tgzz68h02.r1	54.972	0.00177	0.0126	Alt	tgzz59d01.r1	27.496	0.00632	0.0238	Alt
tgzz67h02.r1	53.040	0.00189	0.0350	Alt	tgzy37b05.r1	27.268	0.00642	0.0148	Alt
tgzy71c07.r1	50.742	0.00205	0.0578	Alt	tgzz59g07.r1	27.113	0.00649	0.0216	Alt
tgzz63h07.r1	49.784	0.00213	0.0360	Alt	tgzz29a10.r1	26.976	0.00654	0.0610	Alt
tgzy50d07.r1	47.253	0.00235	0.0126	Alt	tgzy52d07.r1	26.974	0.00655	0.0128	Exp
tgzz49g05.r1	47.029	0.00237	0.0210	Alt	tgzz61g09.r1	26.471	0.00677	0.0720	Alt
tgzy67c07.r1	46.497	0.00242	0.0164	Alt	tgzz69e04.r1	26.430	0.00679	0.0590	Exp
tgzy46e04.r1	41.387	0.00300	0.0046	Alt	tgzz60h10.r1	26.363	0.00682	0.0148	Alt
tgzz43g03.s1	37.413	0.00362	0.0046	Alt	tgzy68b04.r1	25.953	0.00701	0.0174	Alt
tgzz64h08.r1	35.718	0.00394	0.0764	Alt	tgzy59c02.r1	25.782	0.00709	0.0892	Alt
tgzz50g01.r1	35.661	0.00395	0.0076	Alt	tgzz67g12.r1	25.502	0.00723	0.0410	Alt
tgzz60h04.r1	35.659	0.00395	0.0318	Alt	tgzz46g05.r1	25.377	0.00729	0.1118	Alt
tgzz60d11.r1	35.569	0.00397	0.0030	Alt	tgzz67h12.r1	24.390	0.00782	0.0226	Alt
tgzz30a09.r1	34.926	0.00410	0.0072	Alt	tgzz63g06.r1	23.330	0.00846	0.0144	Exp
tgzz27b02.r1	34.705	0.00415	0.0050	Alt	tgzz42h07.s1	22.636	0.00892	0.0050	Exp
tgzz53g02.r1	34.604	0.00417	0.0518	Alt	tgzz61e07.r1	22.062	0.00933	0.0184	Alt
tgzz37c10.r1	33.446	0.00444	0.0200	Exp	tgzy04d12.r1	21.768	0.00955	0.0010	Alt
tgzz64d12.r1	33.002	0.00455	0.0132	Alt	tgzz56h07.r1	21.465	0.00979	0.0148	Alt

Only the results of testing the interaction between parasite strain (parental versus mutant) and cultivation conditions (tachyzoite versus bradyzoite) are shown, and only those results for which the normal-theory is $P < 0.01$.

[a]This is the Type I error probability assuming that the data are normally distributed, calculated using the SYSTAT (v. 9.01) program of SPSS (1998).

[b]This is the Type I error probability based on permutation testing, calculated using the NPMANOVA program of Anderson (2001).

[c]This represents which of the two alternative patterns the data support if we reject the null hypothesis of no pattern. Exp = the parental tachyzoite and bradyzoite conditions have different expression levels but the mutant ones do not; Alt = the mutant tachyzoite and bradyzoite conditions have different expression levels but the parental ones do not.

low power of the two-factor analysis compared to the three-factor analysis—the information provided by each gene individually is ignored in the distance-based analysis, which uses only a summary of the gene information. This point is obvious just from comparing the degrees of freedom of the analyses shown in Table 2a and Table 4. The three-factor analysis is thus to be preferred in this case.

As a final way of analysing the data, Matrajt *et al.* (2002) converted their EST expression data into induction/repression data by calculating the bradyzoite-to-tachyzoite (B/T) ratio for each EST, thus highlighting constitutively expressed genes. Furthermore, they formed gene clusters of ESTs, by grouping together those ESTs known to be associated with a particular gene. This subset of the data involves only 459 of the ESTs.

This new approach is an important suggestion from the point of view of experimental design. The inferences drawn about the gene clusters will be based on the data from replicate ESTs, which are, in turn, based on data from replicate arrays. This should produce better-quality data and therefore more reliable conclusions, even though it involves using less of the data collected. Even better would have been to have replicate mutant and parental strains—we would then have as good an experiment as could be reasonably expected in this situation.

Under these revised circumstances, the experimental design has three factors:

1. Strain—this factor forms two groups, one for data that came from the parental strain and one for data from the mutant strain;
2. Cluster—this factor forms 83 groups, one for data that came from each of the genes;
3. EST—this factor forms 459 groups, one for each of the ESTs.

However, the relationship between these factors is no longer completely orthogonal. In particular, the EST factor is nested within the Cluster factor, because each EST must come from only one gene cluster, and thus cannot interact with it. The appropriate statistical test is still a three-factor analysis of variance, but it is now a mixed analysis, and there is only one interaction that can be tested (the two-factor one between Strain and Cluster). The EST factor is still random and the Strain fac-

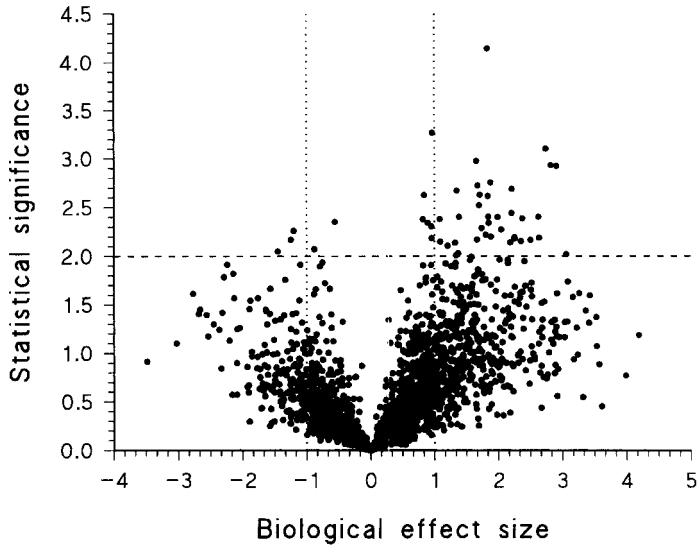


FIG. 9. Volcano plot of the results from the testing of the expression of each of the 2066 ESTs from the Matrajt *et al.* (2002) data set individually using a two-factor orthogonal analysis of variance. Only the results of testing the interaction between parasite strain (parental versus mutant) and cultivation conditions (tachyzoite versus bradyzoite) are shown. Each symbol in the scatterplot represents one of the ESTs, with the horizontal axis representing the magnitude of the biological pattern and the vertical axis representing the magnitude of the statistical pattern. The statistical pattern is measured as the $-\log_{10}$ probability (based on the standard normal-theory assumption), so that increasing values indicate increasing statistical significance; values above the dashed line therefore have $P < 0.01$. The biological pattern is measured as the relative difference in response to the cultivation conditions between the parental and mutant strains, so that negative values indicate that the parental tachyzoite and bradyzoite conditions have different expression levels but the mutant ones do not (the predicted pattern), while positive values indicate that the mutant tachyzoite and bradyzoite conditions have different expression levels but the parental ones do not. Values outside the dotted lines have a twofold difference in behavior between the parental and mutant strains.

tor is fixed, while the Cluster factor is best treated as random (since the genes chosen are an arbitrary selection of all of the available possibilities). For the analyses, the ratio data need to be $\log_2$ transformed, for the same reasons as for the expression ratio data.

Furthermore, we need to assess whether the assumptions of the proposed analysis are met by the data at hand (we should have done so for the previous ANOVA as well, of course, but we decided to defer discussion of the topic until this analysis). The two most important assumptions for analysis of variance are normality and equal variances. Both of these assumptions can be tested using the results of the ANOVA itself, in particular, the residuals from that analysis, which are just the difference between the observed value of each observation and the value predicted by the analysis. Normality is assessed using a normal probability plot, which is a scatterplot showing the relationship between the residuals (horizontally) and what would be expected if they did come from a normal distribution (vertically). The points on this plot should form a straight line. They do for our analysis, and so we may assume that our data are normally distributed. Variance homogeneity can be assessed using a residual plot, which is a scatterplot showing the relationship between the predicted observations (horizontally) and the residuals (vertically). The points on this plot should not show any regular patterns, including curvilinear trends, systematic trends of vertical spread (e.g., increasing residual values with increasing predicted values), or nonsymmetric vertical distribution. For our analysis the points are randomly scattered, and so we can assume that the data have homogeneous variances.

The results of the appropriate analysis of variance are shown in Table 2b, in standard format. This analysis indicates that we should accept the null hypothesis (i.e., $P > 0.05$) for the EST factor (i.e., there is no significant difference between the induction levels of the ESTs within each gene cluster), but that we should reject each of the other three hypothesis tests (i.e., $P < 0.05$). The result for the EST factor is to be expected if the ESTs have been correctly assigned to their genes, since all of the ESTs in any particular gene cluster should either be expressed together or not expressed at all—the large probability indicates that this is so. However, we are actually most interested in the Strain*Cluster interaction, because this tells us that some genes are significantly induced or repressed in either the parental or mutant strain (but not in the other).

Once again, to investigate this pattern in more detail we have a number of options open to us, with the same potential problems as before. For illustrative purposes, we performed two analyses. First, we simply calculated the 95% confidence intervals for each gene for each parasite strain. If these intervals do not overlap between the two strains for a particular gene, then we can declare the induction/repression levels for the two strains to be significantly different from each other. Furthermore, if the 95% confidence interval for a particular gene does not overlap zero (on our $\log_2$-transformed scale), then we can declare that it is significantly induced or repressed for that parasite strain. This provides an explicit statistical criterion for induction, as opposed to merely choosing an arbitrary level such as two-fold. Note that we can base the calculation of the confidence intervals on the results from the ANOVA because the

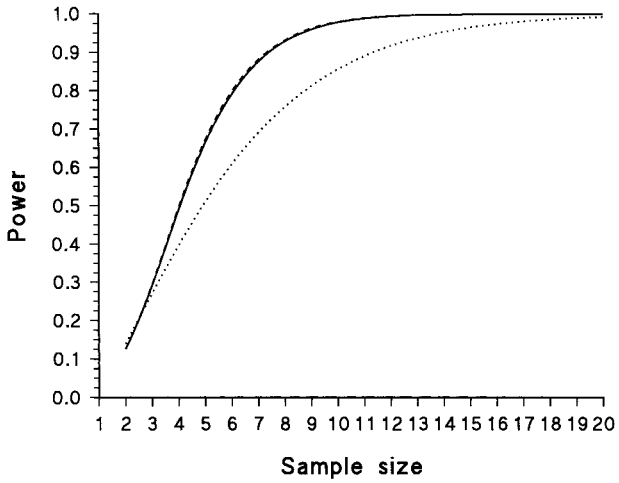


FIG. 10. Power analysis plots of the multifactorial analyses of variance for all sample sizes from 2–20 replicates from the Matrajt *et al.* (2002) data set. The power is the probability of correctly accepting a true null hypothesis. Alpha is the probability of falsely rejecting a true null hypothesis, and was set at $P = 0.05$. The solid line is from the two-factor analysis of expression for each EST, with the effect size for the magnitude of the pattern in the dataset specified as group mean expression levels of 1, 1, 1.5, and 2 for the four combinations of experimental conditions, along with the within-group variance calculated from the three-factor orthogonal analysis of variance (Table 2a). This estimates the number of experimental replicated required per EST. The dotted line is from the one-factor analysis of induction for each EST cluster, detecting a difference between the parental and mutant strains, based on the within-group variance calculated from the three-factor mixed analysis of variance (Table 2b). The dashed line is from the same analysis, detecting a significant induction or repression level (i.e., B/T $\neq$ 1). Calculated using the DESIGN (v. 3.0) program of Dallal (1988). These latter two lines estimate the number of replicate ESTs required per gene cluster.

assumptions of the analysis have been met—consequently, we can create pooled confidence intervals rather than creating separate confidence intervals for each gene, and this should provide more statistical power.

Second, we performed a one-factor ANOVA for each gene cluster, comparing the two parasite strains (this is the analogous procedure to what we did above), and we then used the false discovery rate to correct the probabilities to deal with the multiple-testing problem (setting the rate to 5%). Note that we do a one-factor ANOVA here rather than a two-factor nested ANOVA because the EST factor was not statistically significant in the main analysis, so that we can pool the replicate experiments and the "replicate" ESTs for each gene cluster, giving a more powerful statistical test.

A summary of the results of these analyses is shown in Table 5. Only 20 of the gene clusters are shown, leaving out those for which we have concluded there is no significant induction or repression. First, the boldface values >1 indicate those gene clusters with significant induction in that parasite strain, while boldface values <1 indicate those gene clusters with significant repression in that parasite strain. Note that most of the detected patterns refer to repression rather than to induction. Apparently, induction is much harder to detect in this dataset. Second, the corrected probabilities from the ANOVA detected only one significant difference between the parasite strains, while the 95% confidence intervals detected two differences. As is becoming a familiar theme, there seems to be very little power in these statistical tests.

The results of the simple power analyses for these two analyses are shown in Figure 10. However, in this case we are not examining the number of experimental replicates needed, but are instead looking at how many "replicate" ESTs are needed for each gene cluster. The power line for detecting significant induction/repression (the dashed line) is fortuitously nearly the same as that for the previous analysis, indicating that we really should have six replicate ESTs to have a decent chance to detect induced genes. The power line for detecting a significant difference between the two parasite strains (the dotted line) is much lower, as would be expected from the results of the sta-

TABLE 4. RESULTS OF THE DISTANCE-BASED ANALYSES OF VARIANCE OF THE EXPRESSION DATA FROM THE 2066 ESTS FROM THE MATRAJT *ET AL.* (2002) DATA SET

		Analysis of variance[a]		
Source of variation	*Degrees of freedom*	*Mean-square*	*pseudo-F-value*	P
(a) Gower coefficient				
Strain	1	0.0908	1.861	0.1563
Cultivation	1	0.2029	4.157	0.0162
Strain*Cultivation	1	0.0797	1.633	0.2010
Residual	4	0.0488		
(b) Correlation coefficient				
Strain	1	0.0160	3.846	0.0363
Cultivation	1	0.0174	4.193	0.0280
Strain*Cultivation	1	0.0106	2.558	0.0995
Residual	4	0.6549		

[a]Two-factor orthogonal analysis, with Strain and Cultivation as fixed factors. Calculated using the DISTLM (v. 2) program of McArdle and Anderson (2001), based on 50,000 permutations and the Monte Carlo *P*-value.

TABLE 5. TRANSCRIPTS THAT WERE SIGNIFICANTLY INDUCED OR REPRESSED IN EITHER THE PARENTAL OR MUTANT STRAINS, BASED ON THEIR BRADYZOITE/TACHYZOITE (B/T) RATIO, AS DETERMINED BY A THREE-FACTOR MIXED-MODEL ANALYSIS OF VARIANCE OF THE 459 ESTS FROM THE MATRAJT *ET AL.* (2002) DATA SET

Cluster ID[a]	*Parental B/T*[b]	*Mutant B/T*[b]
Induced in parent		
4054	**1.70**	0.89
4184	**1.72**	1.09
4303	**2.26**	1.64
Induced in mutant		
4276	1.53	**2.03**
Repressed in parent		
502	**0.60**	0.89
1016	**0.58**	0.88
4059	**0.55**	1.16
2900	**0.54**	0.68
3613	**0.50**	0.71
4152	**0.41**	0.80
4307	**0.39**	0.84
Repressed in mutant		
4287	0.98	**0.56**
1776	0.88	**0.54**
481	1.01	**0.54**
1101	0.80	**0.50**
3626	1.24	**0.47**[c]
4093	1.16	**0.46**[d]
1577	0.80	**0.46**
Repressed in parent and mutant		
2496	**0.47**	**0.40**
4396	**0.57**	**0.59**

[a]Refers to the "Toxoqual3" clusters describing overlapping EST clones in the Parasite Databases of Clustered Sequences.

[b]Bold values indicate that the induction or repression is statistically significant (i.e., B/T $\neq$ 1) based on the 95% confidence intervals as derived from the mixed-model analysis of variance of all 83 clusters.

[c]The parental and mutant levels of induction are statistically different based on their 95% confidence intervals as derived from the mixed-model analysis of variance of all 83 clusters.

[d]The parental and mutant levels of induction are statistically different based on their 95% confidence intervals as derived from the mixed-model analysis of variance of all 83 clusters, as well as based on a one-factor analysis of variance with the probability corrected for the false discovery rate of all 83 clusters.

tistical tests (i.e., we found 20 gene clusters with significant induction/repression but only one to two gene clusters with a significant difference between the strains). In this case, we would need nine replicate ESTs to have a respectable chance of detecting real differences between the two parasite strains. The average number of ESTs per gene in the experiment is 5.5, indicating why most of the gene cluster tests have relatively low power.

Second dataset

Singh *et al.* (2002) also describe a study involving the analysis of changes in gene expression during stage conversion of *T. gondii* tachyzoites into bradyzoites *in vitro*. As part of their experimental plan they performed a microarray experiment to examine expression on approximately 4000 ESTs. They used a direct design, in which each of the experimental samples was compared directly to a sample from the wild-type strain on each array slide. The processed data from this experiment are available at www.stanford.edu/~blader/Usinghetalwebfig1.xls, with the original array data in the Stanford Microarray Database (http://genome5-www.stanford.edu).

We can start the discussion here by thinking about the experimental design itself, as before. In contrast to the previous experiment, this experiment involved an unreplicated direct design, as shown in Figure 3c. In this case, each array involved a comparison between a cDNA sample from one of the mutants and from the parental strain, which is the comparison of direct interest in the experiment. Note, however, that using this design we do not end up with any data concerning expression levels in the parental strain, but only comparisons of the mutants to the parents.

As far as data analysis is concerned, the dataset here is conceptually the same as that used for the final analysis discussed in the previous section (i.e., the analysis of induction). That is, we have ratio data for a collection of ESTs within a set of gene clusters. Last time, the observations represented the bradyzoite-to-tachyzoite ratio for both a mutant and a wild-type strain; this time the observations represent the mutant-to-wild-type ratio for bradyzoites of four different mutant strains. However, this is a difference in the *biological interpretation* of what the numbers mean, it does not represent any mathematical difference in how the data should be *treated* mathematically. So, in fact, we are faced with a rather similar data analysis situation here. Nevertheless, it will turn out that the analyses need to be quite different, due to the consequences of some significant differences in the experimental design.

Before we begin, there is one general point that can be made. In the discussion of the previous experiment it was pointed out that a better experimental design would be to have replicate mutant and parental strains, as well as replicate experiments. So, it is worthwhile to note that in this experiment we have replicate mutant strains but not replicate parental strains or replicate arrays.

We begin the analysis by proceeding in a similar fashion to what we did before. First, we selected only that subset of the data for which there were two replicate ESTs from the same gene cluster for each of the four mutant parasite strains. This turns out to be an important consideration here, because the arrays used for mutants TBD-1 and TBD-2 seem to be quite different from those used for mutants TBD-3 and TBD-4. In particular, the ESTs recorded for each of these two groups of mutants are rarely the same in the dataset. Consequently, the data can be expected to have quite different properties from the previous dataset, where the same ESTs were recorded for both strains.

This point has several consequences for the data analysis. First, we could not impute missing values in this dataset even if we wanted to, because the values are very much missing not at random. Which ESTs are missing is determined mostly by which array is being studied. Second, we must therefore treat each EST as a replicate estimate of induction for each gene cluster, independently of each other EST. The ESTs are thus

best treated as a random sample of the estimated induction level for each gene cluster for each mutant strain. This is not an ideal situation, because all of the "replicates" come from a single array, but this seems to be the only acceptable approach to the data analysis. For the analysis, we therefore have observations for 202 gene clusters for each of the four mutant strains, spread across 4200 EST measurements (i.e., an average of ~5 replicate ESTs per gene cluster for each mutant).

Under these circumstances, the experimental design has two factors:

1. Strain—this factor forms four groups, one for data that came from each of the mutant strains;
2. Cluster—this factor forms 202 groups, one for data that came from each of the genes.

Note that we do not have replicate experiments in this case, and so we do not have any replication of induction estimates within ESTs. Consequently, "EST" does not appear as a separate factor in the analysis, as it did last time (see Table 2b). The appropriate statistical analysis here is a two-factor orthogonal analysis of variance, with one interaction that can be tested (the two-factor one between Strain and Cluster). As before, the Strain factor is fixed, while the Cluster factor is best treated as random (since the genes chosen are an arbitrary selection of all of the available possibilities); and the ratio data need to be $\log_2$ transformed, so that a transformed ratio of zero means no induction or repression.

The results of the appropriate analysis of variance are shown in Table 6, in standard format. This analysis indicates that we should reject the null hypothesis only for the Cluster factor (i.e., $P < 0.05$). We are actually most interested in the Strain*Cluster interaction. In this case, the analysis tells us that there is no evidence of differences in induction or repression among the mutant strains for different genes—that is, each gene behaves the same across all four of the mutants.

However, as before, we need to assess whether the assumptions of this analysis are met by the data. This time, the points of the normal probability plot do not form a straight line; instead, they form a definite sigmoid shape. We must thus conclude that our data are *not* normally distributed. In the residual plot the points are randomly scattered, and so we can assume that the data have homogeneous variances. Nevertheless, the results of the ANOVA cannot be trusted, due to the non-normality.

We should therefore try a nonparametric analysis instead, to evaluate whether the ANOVA is leading us to a wrong conclusion. We could, for example, use permutation testing to produce the probabilities. This was not necessary for the analysis shown in Table 2b, since the ANOVA assumptions were met. However, just for the exercise we actually did perform that analysis using permutation tests, and in that case we got probabilities that were the same as those shown in Table 2b to three decimal places (using only 2000 permutations). This is what would be expected when the assumptions are met. However, this is *not* what we would expect for the analysis shown in Table 6.

Unfortunately, permutation testing cannot be done for this dataset, just as it couldn't for the analysis in Table 2a. There are some practical restrictions on the use of permutation tests for multifactorial analysis of variance, compared to the usual parametric method. For example, most of the computer programs available assume that: (1) there are only two factors, and/or (2) there are relatively few levels per factor, and/or (3) it is a balanced experimental design (i.e., there are equal numbers of replicates in all of the levels), and/or (4) there are relatively few total observations. Microarray data do not meet these constraints when there are thousands of genes (and it was difficult enough to do the permutation testing of Table 2b, which is a much smaller dataset).

So, instead we can try a Kruskal-Wallis test, which analyzes the ranks of the observations as opposed to their original values. This is thus a nonparametric test rather than a parametric one. Parametric tests are generally to be preferred because they will be more powerful, as they use the full information in the data rather than merely the rank-order. However, they are only powerful if the data meet their assumptions, and they can be quite sensitive to departures from those assumptions. Since the assumptions are doubtfully met in this case, we need to assess whether the departure from the assumptions is important here.

The results of the appropriate Kruskal-Wallis analysis are also shown in Table 6. The probabilities from this analysis are quite different to those from the ANOVA, but they still lead us to the same conclusions for each of the three hypothesis tests. This is thus confirmation of the ANOVA results, and we should perhaps trust them after all. We can thus safely conclude that there is no evidence that any of the genes behaves differently among the four mutants. The genes do, however, behave quite differently from each other. Therefore, if a gene is induced then it is induced in all four mutant strains, and if it is repressed then it is repressed in all four mutants.

TABLE 6. RESULTS OF THE ANALYSES OF VARIANCE AND KRUSKAL-WALLIS TEST OF THE INDUCTION/REPRESSION DATA FROM THE 4200 ESTS IN 202 GENE CLUSTERS FROM THE SINGH *ET AL.* (2002) DATA SET

Source of variation	*Degrees of freedom*	*Analysis of variance*[a]			*Kruskal-Wallis test*[b]	
		Mean-square	F-*value*	P	H-*value*	P
Strain	3	0.7173	1.205	0.3498	2.855	0.5855
Cluster	201	14.8447	22.668	<0.0001	1931.746	<0.0001
Strain*Cluster	603	0.5953	0.909	0.9325	337.999	>0.9999
Residual	3392	0.6549				

[a]Two-factor orthogonal analysis, with Strain as a fixed factor and Cluster as a random factor. Calculated using the SYSTAT (v. 9.01) program of SPSS (1998), based on the expected mean-squares suggested by Zar (1999).

[b]Two-factor orthogonal analysis. Calculated using the SYSTAT (v. 9.01) program of SPSS (1998), based on the calculations indicated by Zar (1999).

As another aside, we could consider the possibility of dealing with the unlikeliness of the ANOVA assumptions by using alternative estimation procedures for this factorial experimental design, such as weighted least-squares (e.g., Kerr *et al.*, 2002), maximum likelihood (Chu *et al.*, 2002) or even empirical Bayesian analysis (Lönnstedt *et al.*, 2001). However, we decided to keep things simple and compare the ordinary least-squares method to a rank-order equivalent.

These ANOVA results are interesting, but they do not tell us everything that we want to know. The ANOVA compares the average induction/repression levels between groups of genes and groups of mutants, but it does not tell us which genes are induced or repressed. To find this out, we can proceed as we did last time, by calculating the confidence intervals for each gene for each parasite strain. If this interval does not overlap zero for a particular gene then we can declare that it is significantly induced or repressed in that parasite strain. Note that we cannot base the calculation of the confidence intervals on the results from the ANOVA (as we did before) because the assumptions of the ANOVA analysis are not met. Consequently, we cannot create pooled confidence intervals but must instead create separate confidence intervals for each gene. This means that the statistical power may vary dramatically between the various confidence intervals, whereas last time it was a constant powerful test. In particular, any ESTs that have small confidence intervals by chance (e.g., as a result of having only two replicates) will appear to be spuriously "significant." This is a common problem in microarray studies, which is why the pooled estimates from the ANOVA are to be preferred.

Alternatively, assessing whether a confidence interval overlaps zero or not is conceptually the same as performing a one-sample *t*-test to test the null hypothesis that the mean induction is zero, and this would actually be the standard statistical approach to this problem. If we use this approach instead, then it allows us to try to deal explicitly with the multiple-testing problem, since we are calculating actual probabilities. This is an important point here, because we are doing 202 genes $\times$ 4 strains = 808 tests, rather than the 166 tests that we did before. So, for the analysis we used the false discovery rate to correct the probabilities (again setting the rate to 5%). To make this testing procedure a bit more comparable to the approach using confidence intervals, we calculated 99% confidence intervals rather than 95% ones.

However, we still have the problem of demonstrated non-normality in the dataset, and both of these procedures are also based on the assumption of normality. It will thus be best to have some confirmatory evidence from other analyses; and for comparison, we should therefore also perform these two procedures using nonparametric methods. The permutation equivalent of the one-sample *t*-test is the Fisher randomization test, and the bootstrap-*t* procedure can be used to create 99% confidence intervals. The limitation of both of these procedures is that they are only effective for samples with at least seven observations. This means that they could only be performed for 167 of the 808 samples. We must then conclude that no statistical evidence can be provided for the other samples. Furthermore, the normal confidence intervals use fixed levels for detecting both induction and repression (i.e., the intervals are symmetrical about the mean, so that the same cutoff is used for both induction and repression), whereas the bootstrap-*t* intervals are not necessarily symmetrical. Whether this difference is a good or bad thing is a moot point.

Before proceeding to the comparison of the results of these four analyses, it may be interesting to show what happens when we adjust probabilities to deal with the multiple-testing problem, which is one of the biggest problems in the analysis of microarray data. The results of the 808 one-sample *t*-tests are shown in Figure 11, where they are compared to four different criteria for assessing statistical significance.

The dotted line represents $P = 0.05$, which is the conventional statistical criterion unadjusted for multiple hypothesis tests. Using this criterion, we would reject the null hypothesis (that the mean induction is zero) for 182 of the 808 tests, which is clearly rather liberal. We would be expecting $808 \times 0.05 = 40.4$ Type I errors (false discoveries) in the complete set of tests. The two solid lines represent corrections for multiple hypotheses so that the familywise error rate for the collection of 808 hypothesis tests is maintained at $P = 0.05$. The lower line is the Bonferroni correction, which is the most conservative possibility, while the upper line is the sequential Bonferroni (i.e., Holm/Hochberg) correction, which is more powerful. In this case, both procedures produce the same result, with 27 of the null hypotheses being rejected. Under these circumstances we would be 95% confident that none of these rejections are Type I errors. Finally, the dashed line represents correction for multiple hypotheses based on a false discovery rate of $P = 0.05$, thus allowing 5% of the hypothesis rejections to be Type I errors (false discoveries). Using this procedure, we would reject 86 of the 808 null hypotheses, with the expectation that $86 \times 0.05 = 4.3$ of these rejections are Type I errors. This is thus a more powerful criterion than the use of the Bonferroni probabilities, because we have rejected 59 more hypotheses in exchange for only ~5 mistakes. This is quite a good swap, in this case, and it is thus probably the preferred solution.

There will usually be little to choose between the sequential Bonferroni and the false discovery procedures for small numbers of hypotheses, but in situations such as the one discussed here the difference can be quite remarkable. Note, also, that our use of a 99% confidence interval produces results that are quite similar to the use of the false discovery rate for this example, as it indicates that 107 of the samples have induction/repression levels that differ from zero (on the $\log_2$-transformed scale).

A summary of the results of our four analyses of the induction/repression data are shown in Table 7. Only 52 of the gene clusters are shown, leaving out those for which we have concluded that there is no significant induction or repression based on any of the four tests. The different letters in the table represent those genes and strains for which each of the four tests detected significant induction/repression. Several points can be noted. First, the "bcde" results are those that are confirmed by all four tests, and are therefore clearly the most reliable results. This comprises 47% of the tests shown, as well as all of the tests that have not been shown, suggesting that this has been quite a successful experiment. Second, the most liberal of the tests is the normal 99% confidence interval, as almost all of the other results are a subset of these results. Third, the "bc" results are those detected by the parametric tests but not by the non-parametric ones. This is the second-largest group of results, because of those tests where the sample size was insufficient to perform the nonparametric tests. These test results are not

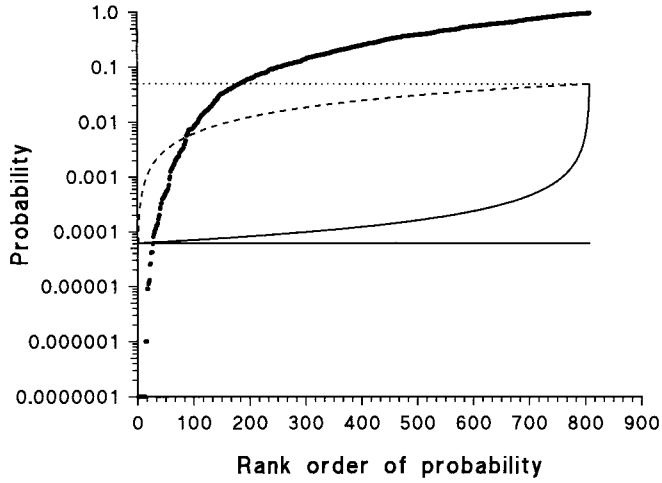


FIG. 11. Rank-order probability plot of the results of the one-sample *t*-tests from the Singh *et al.* (2002) data set. The symbols represent the 808 probabilities in increasing rank order, calculated using the SYSTAT (v. 9.01) program of SPSS (1998). The lines show the boundary of the critical region based on four different criteria for statistical significance, so that the null hypothesis of no induction/repression would be rejected for any of the symbols lying below the respective line (and accepted for those lying above the line). The dotted line represents the 95% confidence interval (i.e., $P = 0.05$), which is the conventional statistical criterion unadjusted for multiple hypothesis tests. The two solid lines represent corrections for multiple hypotheses so that the setwise (or familywise) error rate for the collection of 808 hypothesis tests is maintained at $P = 0.05$; the lower line is the Bonferroni correction, which is the most conservative possibility, while the upper line is the Holm/Hochberg correction, which is more powerful. The dashed line represents correction for multiple hypotheses based on a false discovery rate of $P = 0.05$, thus allowing 5% of the hypothesis rejections to be Type I errors (or false discoveries).

reliable, because of the small sample size and the known non-normality of the data, and they need confirmation before they can be accepted. Fourth, there are two "de" results, which are those detected by the nonparametric tests but not by the parametric ones. These two samples presumably violate the parametric assumptions so much that the parametric tests fail. This emphasizes the value of nonparametric tests in the analysis of microarray data.

Finally, note that the results here contradict those of the main two-factor ANOVA, because the genes are clearly shown as having different behaviour among the mutant strains (i.e., few of the genes show significant induction or repression across all four mutants). This outcome is mostly because of the unequal sample sizes among the groups. Many of the strains have small numbers of ESTs recorded for a particular gene while the other strains have larger numbers of ESTs recorded for the same gene, and this affects the power of the statistical tests. This situation is having a strong effect on the individual analyses but does not affect the ANOVA as strongly (i.e., the two-factor ANOVA is more robust than are the individual tests). This is why the main ANOVA should be used prior to performing the individual tests, as it will help avoid spurious conclusions that can arise from looking at the details before looking at the overall picture.

Performing a power analysis in this situation is a bit problematical, but we can have a go at it. The problem is that power analysis is only a well-developed discipline where the assumption of normality holds, which is not the case here. However, given the congruence of the ANOVA and Kruskal-Wallis tests, along with the congruence of most of the individual parametric and nonparametric tests, a parametric power analysis will probably not be misleading, although it is likely to overestimate the power (see Black and Doerge, 2002). We can base such an analysis on the averaged results of the individual tests, and look at how many "replicate" ESTs are needed for each gene cluster to detect significant induction/repression.

The results of this power analysis are shown in Figure 12, based on several different scenarios. The three solid lines represent the situation for detecting a two-fold induction or repression, while the three dashed lines represent the situation for detecting a 1.5-fold induction or repression. As expected, we need more replicate ESTs for detecting the smaller pattern. Within each of these two line types, the three lines represent different significance levels, with increasing statistical stringency from left to right. The increased stringency helps us to understand what happens when dealing with the multiple-testing problem—we need to set a more stringent criterion when we have lots of tests to perform. For example, our use of the 99% confidence intervals was too liberal, and we were thus dealing with a situation somewhere between the $P = 0.01$ and $P = 0.001$ lines on the graph. In this case, we would need about eight replicate ESTs to have a respectable chance of detecting a real two-fold induction or repression, and about 15 replicate ESTs for detecting a real 1.5-fold induction or repression. Only 16% of the samples meet the former criterion and only 6% meet the latter, indicating that most of the gene cluster tests probably have relatively low power in this experiment.

Alternatively, we can perform a nonparametric power analysis by choosing some of the samples as examples, and we can then compare these to a parametric power analysis of those same samples. This will help us to understand the circumstances under which the main parametric power analysis might be misleading. We have chosen three samples each with 10 replicate

TABLE 7. TRANSCRIPTS THAT WERE SIGNIFICANTLY INDUCED OR REPRESSED IN ANY OF THE FOUR MUTANT STRAINS, BASED ON THEIR BRADYZOITE RATIO TO THE WILD TYPE, AS DETERMINED BY VARIOUS TESTS OF THE 202 GENE CLUSTERS FROM THE SINGH *ET AL.* (2002) DATA SET

Cluster ID[a]	*TBD-1*	*TBD-2*	*TBD-3*	*TBD-4*
Induced				
20	bc	bc		
449	b de	bcde		
546	bc			
604	bcde	bcde	bc e	e
619	bcde	bcde	bcde	bcde
622		bc		
830	bcde	bcde		
854	b			
1101	bcde	de		
1406	bcde	bcde	b de	bcde
1589	bc	bc		
1776	bc			
1854				bcde
1886		b		
2199	bc	bc	bc	b
2303				bc
2376	bcde	bc		
2486	bcde	bcde		
3429		bc		
3622		bcde		b de
3735		bc		
3949	b			
4091	bc			
4240	bcde	bcde	bc e	bc e
4253	bcde	bcde		
4294		bc		
4365				bc
4387	b	b		
4395				bc
Repressed				
101	bcde	bcde		
750	bc e	e		
1577		b		
2161	bc	bc		b
2329			bc	
3906	bcde	bcde	bcde	bcde
3919	bcde	de	bcde	bcde
3985			bc	bc
3993			bc	
4018		bc		
4034	b de			
4054	bcde	bcde	bcde	bcde
4130	b de	b d	bcde	bcde
4131	bcde	bcde	bcde	bcde
4135		bc	bcde	bc
4144		bc		
4192		b	bc	b
4196	bcde		bcde	bcde
4219			b e	
4243	b	b	bcde	bcde
4303	bcde		b e	
4432	bcde	bc e	bcde	bcde
4436	b de	bcde		

[a]Refers to the "Toxoqual3" clusters describing overlapping EST clones in the Parasite Databases of Clustered Sequences.

[b]Induction or repression is statistically significant based on the individual 99% confidence interval. Calculated using the SYSTAT (v. 9.01) program of SPSS (1998).

[c]Induction or repression is statistically significant based on the individual one-sample *t*-test, adjusted for a false discovery rate of 5% among the 808 tests. Calculated using the SYSTAT (v. 9.01) program of SPSS (1998).

[d]Induction or repression is statistically significant based on the individual bootstrapped-t 99% confidence interval using 10,000 resamples. Calculated using the Resampling Procedures (v. 1.3) program of Howell (2002).

[e]Induction or repression is statistically significant based on the individual Fisher randomization test using complete enumeration, adjusted for a false discovery rate of 5% among the 167 tests. Calculated using the RT (v. 2.1) program of Manly (1997).

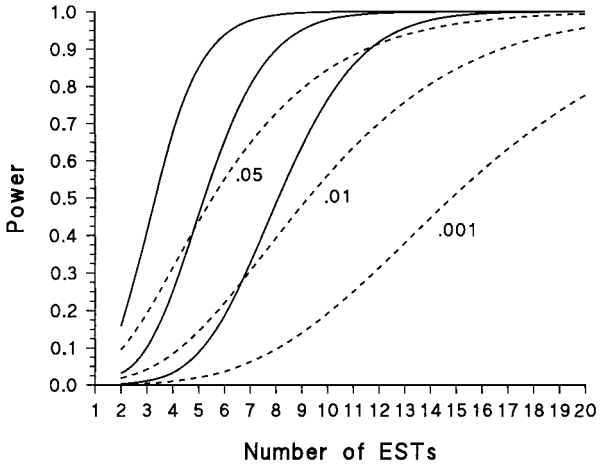


FIG. 12. Power analysis plots of the one-sample tests for all sample sizes from 2–20 ESTs within each cluster from the Singh *et al.* (2002) data set. The power is the probability of correctly accepting a true null hypothesis. Alpha is the probability of falsely rejecting a true null hypothesis, and three different probabilities are shown for each of the two scenarios. The solid lines are for the effect size (the magnitude of the pattern in the dataset) specified as a twofold induction or repression level, while the dashed lines are for a 1.5-fold induction/repression. Calculated using the DSTPLAN (v. 4.2) program of Brown *et al.* (2000).

observations, which seem to represent the extremes of the samples in the experiment. For each sample we performed a parametric power analysis, using the assumption of a normal frequency distribution for the samples (i.e., an exact method based on the one-sample *t*-test), as well as a nonparametric power analysis using all possible permutations (i.e., an exact method based on the Fisher randomization test). The results of the six power analyses are shown in Figure 13, with the parametric analyses as solid lines and the nonparametric analyses as dashed lines.

One of the samples (shown to the far left on the graph) has a normal distribution, and the parametric and nonparametric analyses are in close agreement about the power of the data to detect induction. This sample is an "average" (i.e., typical) one for the dataset, with an average standard deviation (the measure of variability in the power analysis), and so it represents the estimated overall parametric power analysis for the dataset.

The other two samples are both non-normal, as determined by a Kolmogorov-Smirnov test. The non-normal sample on the left of the graph has a single outlying observation, which creates a slightly larger standard deviation, and hence, lower power compared to the normally distributed sample. The non-normal sample on the right of the graph has several divergent observations (creating a platykurtic distribution), which results in a very large standard deviation, and therefore, much lower power compared to the other two samples. Furthermore, in both of these cases the parametric and nonparametric analyses disagree about the power, due to the fact that the assumptions of the parametric power analysis are being violated. The nonparametric analysis indicates lower power for both of these samples compared to the parametric analysis, and this is likely to be a more realistic assessment of the true situation.

Clearly, the parametric power analysis might be quite misleading for some samples in this type of experiment. If 10 replicates are used, then any samples like the first one will have quite high power, and any samples like the second one will be not too bad either. However, samples like the third one do exist (since there was at least one in this dataset), although they are rare, and any induction shown by this type of gene will probably not be detected in an experiment with only 10 replicates.

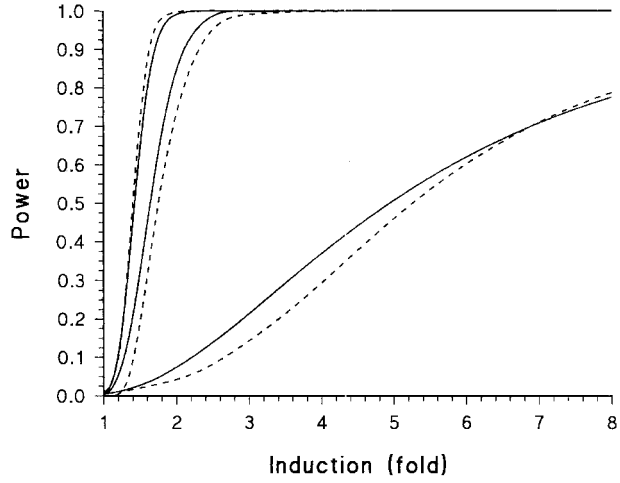


FIG. 13. Power analysis plots of the one-sample tests for three of the EST samples from the Singh *et al.* (2002) data set. The power is the probability of correctly accepting a true null hypothesis. Alpha is the probability of falsely rejecting a true null hypothesis, and was set at $P = 0.01$. The three samples chosen for the analysis fix the estimates of variability and sample size (chosen as $n = 10$), and the biological effect size was varied for each analysis. The solid lines are from the parametric analysis, calculated using the DSTPLAN (v. 4.2) program of Brown *et al.* (2000). The dashed lines are from the nonparametric analysis, calculated using the RT (v. 2.1) program of Manly (1997). The sample shown to the far left has a normal distribution, while the other two samples are non-normal.

We will finish on this note about having a decent experimental sample size. Where there is replication of the experimental units and replication of the ESTs then we have *high-quality* data. This is probably one of the biggest current problems with most microarray experiments, that they have insufficient statistical power (due to small sample sizes) to detect the biological patterns that the experimenters are looking for. This issue needs to be seriously addressed.

REFERENCES

ALLISON, P.D. (2001). *Missing Data*. (Sage Publications, London).

ALON, U., BARKAI, N., NOTTERMAN, D.A., GISH, K., YBARRA, S., MACK, D., and LEVINE, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA **96,** 6745–6750.

ALTER, O., BROWN, P.O., and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA **97,** 10101–10106.

AMBROISE, C., and MCLACHLAN, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA **99,** 6562–6566.

ANDERSON, M.J. (2001). A new method for non-parametric analysis of variance. Aust. Ecol. **26,** 32–46.

ANTONIADIS, A., LAMBERT-LACROIX, S., and LEBLANC, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics **19,** 563–570.

BAKAY, M., CHEN, Y.-W., BORUP, R., ZHAO, P., NAGARAJU, K., and HOFFMAN, E.P. (2002). Sources of variability and effect of experimental approach on expression profiling data interpretation. BMC Bioinformat. **3,** 4.

BALDI, P., and HATFIELD, G.W. (2002). *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. (Cambridge University Press, Cambridge).

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.

BARNETT, V., and LEWIS, T. (1978). *Outliers in Statistical Data*. (John Wiley & Sons, New York).

BELBIN, L. (1995). *PATN Pattern Analysis Package. Technical Reference.* Technical Report (CSIRO, Canberra, Australia).

BICCIATO, S., LUCHINI, A., and DI BELLO, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinformatics **19,** 571–578.

BIJLANI, R., CHENG, Y., PEARCE, D.A., BROOKS, A.I., and OGIHARA, M. (2003). Prediction of biologically significant components from microarray data: Independently consistent expression discriminator (ICED). Bioinformatics **19,** 62–70.

BITTNER, M., MELTZER, P., CHEN, Y., JIANG, Y., SEFTOR, E., HENDRIX, M., RADMACHER, M., SIMON, R., YAKHINI, Z., BENDOR, A., *et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature **406,** 536–540.

BLACK, M.A., and DOERGE, R.W. (2002). Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics **18,** 1609–1616.

BLADER, I.J., MANGER, I.D., and BOOTHROYD, J.C. (2001). Microarray analysis reveals previously unknown changes in *Toxoplasma gondii*-infected human cells. J. Biol. Chem. **276,** 24223–24231.

BLAND, J.M., and ALTMAN, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet **1,** 307–310.

BØ, T.H., and JONASSEN, I. (2002). New feature subset selection procedures for classification of expression profiles. Genome Biol. **3,** research0017.

BOLDRICK, J.C., ALIZADEH, A.A., DIEHN, M., DUDOIT, S., LIU, C.L., BELCHER, C.E., BOTSTEIN, D., STAUDT, L.M., BROWN, P.O., and RELMAN, D.A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. Proc. Natl. Acad. Sci. USA **99,** 972–977.

BOLSTAD, B.M., IRIZARRY, R.A., ÅSTRAND, M., and SPEED, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19,** 185–193.

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., SHERLOCK, G., SPELLMAN, P., STOECKERT, C., AACH, J., ANSORGE, W., BALL, C.A., CAUSTON, H.C., *et al.* (2001). Minimum information about a microarray experiment (MIAME)—Towards standards for microarray data. Nat. Genet. **29,** 365–371.

BROWN, B.W., BRAUNER, C., CHAN, A., GUTIERREZ, D., HERSON, J., LOVATO, J., POLSLEY, J., RUSSELL, K., and VENIER, J. (2000). *DSTPLAN, Version 4.2. Calculations for Sample Sizes and Related Problems.* Technical Report (M.D. Anderson Cancer Center, Department of Biomathematics, University of Texas, Houston, TX).

CHAPMAN, S., SCHENK, P., KAZAN, K., and MANNERS, J. (2001). Using biplots to interpret gene expression patterns in plants. Bioinformatics **18,** 202–204.

CHEN, G., JARADAT, S.A., BANERJEE, N., TANAKA, T.S., KO, M.S.H., and ZHANG, M.Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. Stat. Sinica **12,** 241–262.

CHEN, Y., KAMAT, V., DOUGHERTY, E.R., BITTNER, M.L., MELTZER, P.S., and TRENT, J.M. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis. Bioinformatics **18,** 1207–1215.

CHIAROMONTE, F., and MARTINELLI, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. Math. Biosci. **176,** 123–144.

CHU, T.-M., WEIR, B., and WOLFINGER, R. (2002). A systematic statistical linear modelling approach to oligonucleotide array experiments. Math. Biosci. **176,** 35–51.

CHUAQUI, R.F., BONNER, R.F., BEST, C.J.M., GILLESPIE, J.W., FLAIG, M.J., HEWITT, S.M., PHILLIPS, J.L., KRIZMAN, D.B., TANGREA, M.A., AHRAM, M., *et al.* (2002). Post-analysis follow-up and validation of microarray experiments. Nat. Genet. Suppl. **32,** 509–514.

CHURCHILL, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. Nat. Genet. Suppl. **32,** 490–495.

CLAVERIE, J.-M. (1999). Computational methods for the identification of differential and coordinated gene expression. Hum. Mol. Genet. **8,** 1821–1832.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ).

CULHANE, A.C., PERRIÈRE, G., CONSIDINE, E.C., COTTER, T.G., and HIGGINS, D.G. (2002). Between-group analysis of microarray data. Bioinformatics **18,** 1600–1608.

DALLAL, G.E. (1988). *DESIGN: A Supplementary Module for SYSTAT and SYGRAPH* (SYSTAT Inc., Evanston, IL).

DATTA, S., and DATTA, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics **19,** 459–466.

DE HOON, M.J.L., IMOTO, S., and MIYANO, S. (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics **18,** 1477–1485.

DESU, M.M., and RAGHAVARAO, D. (1990). *Sample Size Methodology* (Academic Press, San Diego, CA).

DEUTSCH, J.M. (2003). Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics **19,** 45–52.

DIDIER, G., BRÉZELLEC, P., REMY, E., and HÉNAUT, A. (2002). GeneANOVA—Gene expression analysis of variance. Bioinformatics **18,** 490–491.

DOBBIN, K., and SIMON, R. (2002). Comparison of microarray designs for class comparison and class discovery. Bioinformatics **18,** 1438–1445.

DOPAZO, J., ZANDERS, E., DRAGONI, I., AMPHLETT, G., and FALCANI, F. (2001). Methods and approaches in the analysis of gene expression data. J. Immunol. Methods **250,** 93–112.

DOUGHERTY, E.R. (2001). Small sample issues for microarray-based classification. Comp. Funct. Genom. **2,** 28–34.

DOUGHERTY, E.R., BARRERA, J., BRUN, M., KIM, S., CESAR, R.M., CHEN, Y., BITTNER, M., and TRENT, J.M. (2002). Inference from clustering with application to gene-expression microarrays. J. Comput. Biol. **9,** 105–126.

DOZMOROV, I., and CENTOLA, M. (2003). An associative analysis of gene expression array data. Bioinformatics **19,** 204–211.

DUDOIT, S., FRIDLYAND, J., and SPEED, T.P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. **97,** 77–87.

DUDOIT, S., SHAFFER, J.P., and BOLDRICK, J.C. (2002b) *Multiple Hypothesis Testing in Microarray Experiments*. Technical Report (Division of Biostatistics, University of California, Berkeley, CA).

DUDOIT, S. YANG, Y.H., CALLOW, M.J., and SPEED, T.P. (2002c). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sinica **12,** 111–139.

EFRON, B., and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. Genet. Epidemiol. **23,** 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J.D., and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. **96,** 1151–1160.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95,** 14863–14868.

ELLIS, J.T., MORRISON, D.A., and REICHEL, M.P. (2003). Genomics and its impact on parasitology and the potential for development of new parasite control methods. DNA Cell Biol. **22,** 395–403.

ERDFELDER, E., FAUL, F., and BUCHNER, A. (1996). GPower: A general power analysis program. Behav. Res. Methods, Instrum. Comput. **28,** 1–11.

FAITH, D.P., MINCHIN, P.R., and BELBIN, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. Vegetation **69,** 57–68.

FELLENBERG, K., HAUSER, N.C., BRORS, B., NEUTZNER, A., HOHEISEL, J.D., and VINGRON, M. (2001). Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. USA **98,** 10781–10786.

FILKOV, V., SKIENA, S., and ZHI, J. (2002). Analysis techniques for microarray time-series data. J. Comput. Biol. **9,** 317–330.

FIRESTEIN, G.S., and PISETSKY, D.S. (2002). DNA microarrays: Boundless technology or bound by technology? Guidelines for studies using microarray technology. Arthrit. Rheum. **46,** 859–861.

FISHER, R.A. (1935). *The Design of Experiments* (Oliver & Boyd, Edinburgh).

GARDNER, M.J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R.W., CARLTON, J.M., PAIN, A., NELSON, K.E., BOWMAN, S., *et al*. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature **419,** 498–511.

GAUCH, H.G. (1982). *Multivariate Analysis in Community Ecology* (Cambridge University Press, Cambridge).

GE, Y., DUDOIT, S., and SPEED, T.P. (2003). *Resampling-Based Multiple Testing for Microarray Data Analysis*. Technical Report (Department of Statistics, Stanford University, Stanford, CA).

GHOSH, D. (2002). Singular value decomposition regression models for classification of tumors from microarray experiments. Pacific Symp. Biocomput. **7,** 18–29.

GIBSON, G. (2002). Microarrays in ecology and evolution: A preview. Mol. Ecol. **11,** 17–24.

GLANTZ, S.A., and SLINKER, B.K. (2001). *Primer of Applied Regression and Analysis of Variance*, 2nd ed. (McGraw-Hill, New York).

GLASBEY, C.A., and GHAZAL, P. (2003). Combinatorial image analysis of DNA microarray features. Bioinformatics **19,** 194–203.

GLONEK, G.F.V., and SOLOMON, P.J. (2002). *Factorial and Time Course Designs for cDNA Microarray Experiments*. Technical Report (Department of Mathematics, University of Adelaide, Adelaide, Australia).

GOOD, P.I. (1999). *Resampling Methods. A Practical Guide to Data Analysis* (Springer Verlag, Berlin).

GRANT, G.R., MANDUCHI, E., and STOECKERT, C.J. (2002). Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. In *Methods in Microarray Analysis*. S.M. Lin and K.F. Johnson, eds. (Kluwer Academic, Dordrecht) pp. 37–56.

HAYWARD, R.E., DERISI, J.L., ALFADHLI, S., KASLOW, D.C., BROWN, P.O., and RATHOD, P.K. (2000). Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. Mol. Microbiol. **35,** 6–14.

HESS, K.R., ZHANG, W., BAGGERLY, K.A., STIVERS, D.N., and COOMBES, K.R. (2001). Microarrays: handling the deluge of data and extracting reliable information. Trends Biotechnol. **19,** 463–468.

HEYER, L.J., KRUGLYAK, S., and YOOSEPH, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. Genome Res. **9,** 1106–1115.

HILL, M.O. (1979). TWINSPAN—A *Fortran Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes*. Technical Report (Section of Ecology & Systematics, Cornell University, Ithaca, NY).

HILSENBECK, S.G., FRIEDRICHS, W.E., SCHIFF, R., O'CONNELL, P., HANSEN, R.K., OSBORNE, C.K., and FUQUA, S.A.W. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J. Natl. Cancer Inst. **91,** 453–459.

HOFFMANN, R., SEIDL, T., and DUGAS, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol. **3,** research0033.

HOWELL, D.C. (2002). *Statistical Methods for Psychology*, 5th ed. (Duxbury Press, Belmont, CA).

HOYLE, D.C., RATTRAY, M., JUPP, R., and BRASS, A. (2002). Making sense of microarray data distributions. Bioinformatics **18,** 576–584.

HURLBERT, S.H. (1984). Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. **54,** 187–211.

HWANG, D., SCHMITT, W.A., STEPHANOPOULOS, G., and STEPHANOPOULOS, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics **18,** 1184–1193.

JAEGER, J., SENGUPTA, R., and RUZZO, W.L. (2003). Improved gene selection for classification of microarrays. Pacific Symp. Biocomput. **8,** 53–64.

JIN, W., RILEY, R., WOLFINGER, R., WHITE, K.P., PASSADOR-GURGEL, G., and GIBSON, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29,** 389–395.

JOHANSSON, D., LINDGREN, P., and BERGLUND, A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics **19,** 467–473.

JONGMAN, R.H.G., TER BRAAK, C.J.F., and VAN TONGEREN,

O.F.R. (eds.) (1995). *Data Analysis in Community and Landscape Ecology*, 2nd ed. (Cambridge University Press, Cambridge).

KASTURI, J., ACHARYA, R., and RAMANATHAN, M. (2003). An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics **19,** 449–458.

KEPLER, T.B., CROSBY, L., and MORGAN, K.T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biol. **3,** research0037.

KERR, M.K., AFSHARI, C.A., BENNETT, L., BUSHEL, P., MARTINEZ, J., WALKER, N.J., and CHURCHILL, G.A. (2002). Statistical analysis of a gene expression microarray experiment. Stat. Sinica **12,** 203–218.

KERR, M.K., and CHURCHILL, G.A. (2001a). Experimental design for gene expression microarrays. Biostatistics **2,** 183–201.

KERR, M.K., and CHURCHILL, G.A. (2001b). Statistical design and the analysis of gene expression microarray data. Genet. Res. **77,** 123–128.

KERR, M.K., and CHURCHILL, G.A. (2001c). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proc. Natl. Acad. Sci. USA **98,** 8961–8965.

KERR, M.K., MARTIN, M., and CHURCHILL, G.A. (2000). Analysis of variance for gene expression microarray data. J. Comput. Biol. **7,** 819–837.

KHAN, J., WEI, J.S., RINGNÉR, M., SAAL, L.H., LADANYI, M., WESTERMANN, F., BERTHOLD, F., SCHWAB, M., ANTONESCU, C.R., PETERSON, C., *et al.* (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7,** 673–679.

KIM, S., DOUGHERTY, E.R., BARRERA, J., CHEN, Y., BITTNER, M.L., and TRENT, J.M. (2002). Strong feature sets from small samples. J. Comput. Biol. **9,** 127–146.

KNUDSEN, S. (2002). *A Biologist's Guide to Analysis of DNA Microarray Data* (John Wiley & Sons, New York).

KOOPERBERG, C., FAZZIO, T.G., DELROW, J.J., and TSUKIYAMA, T. (2002a). Improved background correction for spotted DNA microarrays. J. Comput. Biol. **9,** 55–66.

KOOPERBERG, C., SIPIONE, S., LEBLANC, M., STRAND A.D., CATTANEO, E., and OLSON, J.M. (2002b). Evaluating test statistics to select interesting genes in microarray experiments. Hum. Mol. Genet. **11,** 2223–2232.

KOTHAPALLI, R., YODER, S.J., MANE, S., and LOUGHRAN, T.P. (2002). Microarray results: How accurate are they? BMC Bioinformat. **3,** 22.

KRAEMER, H.C., and THIEMANN, S. (1987). *How Many Subjects? Statistical Power Analysis in Research* (Sage Publications, London).

KROLL, T.C., and WÖLFL, S. (2002). Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. Nucleic Acids Res. **30,** e50.

KUO, W.P., JENSSEN, T., BUTTE, A.J., OHNO-MACHADO, L., and KOHANE, I.S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics **18,** 405–412.

LANDGREBE, J., WURST, W., and WELZL, G. (2002). Permutation-validated principal components analysis of microarray data. Genome Biol. **3,** research0019.

LAZZERONI, L., and OWEN, A. (2002). Plaid models for gene expression data. Stat. Sinica **12,** 61–86.

LEBLANC, M., KOOPERBERG, C., GROGAN, T.M., and MILLER, T.P. (2003). Directed indices for exploring gene expression data. Bioinformatics **19,** 686–693.

LEE, K.E., SHA, N., DOUGHERTY, E.R., VANNUCCI, M., and MALLICK, B.K. (2003). Gene selection: A Bayesian variable selection approach. Bioinformatics **19,** 90–97.

LEE, M.-L.T., KUO, F.C., WHITMORE, G.A., and SKLAR, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA **97,** 9834–9839.

LEGENDRE, P., and ANDERSON, M.J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecol. Monogr. **69,** 1–24.

LEGENDRE, P., and LEGENDRE, L. (1998). *Numerical Ecology*, 2nd ed. (Elsevier, Amsterdam).

LEUNG, Y.F. (2002). Unravelling the mysteries of microarray data analysis. Trends Biotechnol. **20,** 366–368.

LI, X., GU, W., MOHAN, S., and BAYLINK, D.J. (2002). DNA microarrays: Their use and misuse. Microcirculation **9,** 13–22.

LIPSEY, M.W. (1990). *Design Sensitivity. Statistical Power for Experimental Research* (Sage Publications, London).

LIU, W.-M., MEI, R., DI, X., RYDER, T.B., HUBBELL, E., DEE, S., WEBSTER, T.A., HARRINGTON, C.A., HO, M.-H., BAID, J., *et al.* (2002). Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics **18,** 1593–1599.

LONG, A.D., MANGALAM, H.J., CHAN, B.Y.P., TOLLERI, L., HATFIELD, G.W., and BALDI, P. (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.

LÖNNSTEDT, I., GRANT, S., BEGLEY, G., and SPEED, T. (2001). *Microarray Analysis of Two Interacting Treatments: A Linear Model and Trends in Expression Over Time*. Technical Report (Department of Mathematics, Uppsala University, Uppsala, Sweden).

LÖNNSTEDT, I., and SPEED, T.P. (2002). Replicated microarray data. Stat. Sinica **12,** 31–46.

LUAN, Y., and LI, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics **19,** 474–482.

LUDWIG, J.A., and REYNOLDS, J.F. (1988). *Statistical Ecology. A Primer on Methods and Computing* (John Wiley & Sons, New York).

LUNNEBORG, C.E. (1999). *Data Analysis by Resampling. Concepts and Applications* (Duxbury Press, Boston, MA).

MACCALLUM, R.C., WIDAMAN, K.F., ZHANG, S., and HONG, S. (1999). Sample size in factor analysis. Psychological Methods **4,** 84–99.

MAMOUN, C.B., GLUZMAN, I.Y., HOTT, C., MACMILLAN, S.K., AMARAKONE, A.S., ANDERSON, D.L., CARLTON, J.M.R., DAME, J.B., CHAKRABARTI, D., MARTIN, R.K., *et al.* (2001). Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. Mol. Microbiol. **39,** 26–36.

MANLY, B.J.F. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd ed. (Chapman & Hall, London).

MATRAJT, M., DONALD, R.G.K., SINGH, U., and ROOS, D.S. (2002). Identification and characterization of differentiation mutants in the protozoan parasite *Toxoplasma gondii*. Mol. Microbiol. **44,** 735–747.

MAVROUDI, S., PAPADIMITRIOU, S., and BEZERIANOS, A. (2002). Gene expression data analysis with a dynamically extended self-organized map that exploits class information. Bioinformatics **18,** 1446–1453.

MCARDLE, B.H., and ANDERSON, M.J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. Ecology **82,** 290–297.

MCSHANE, L.M., RADMACHER, M.D., FREIDLIN, B., YU, R., LI, M.C., and SIMON, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics **18,** 1462–1469.

MEDVEDOVIC, M., and SIVAGANESAN, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics **18,** 1194–1206.

MÉNDEZ, M.A., HÖDAR, C., VULPE, C., GONZÁLEZ, M., and CAMBIAZO, V. (2002). Discriminant analysis to evaluate clustering of gene expression data. FEBS Lett. **522,** 24–28.

MORRISON, D.A. (2002a). How to improve statistical analysis in parasitology research publications. Int. J. Parasitol. **32,** 1065–1070.

MORRISON, D.A. (2002b). Further difficulties with multifactorial analysis of variance: random and nested factors and independence of data. Infect. Genet. Evol. **2,** 149–152.

MORRISON, D.A., ELLIS, J., and JOHNSON, A.M. (1994). An empirical comparison of distance matrix techniques for estimating codon usage divergence. J. Mol. Evol. **39,** 533–536.

MUTCH, D.M., BERGER, A., MANSOURIAN, R., RYTZ, A., and ROBERTS, M.-A. (2002). The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. BMC Bioinformat. **3,** 17.

NADON, R., and SHOEMAKER, J. (2002). Statistical issues with microarrays: processing and analysis. Trends Genet. **18,** 265–271.

NGUYEN, D.V., and ROCKE, D.M. (2002a). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics **18,** 39–50.

NGUYEN, D.V., and ROCKE, D.M. (2002b). Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics **18,** 1216–1226.

NOVAK, J.P., SLADEK, R., and HUDSON, T.J. (2002). Characterization of variability in large-scale gene expression data: Implications for study design. Genomics **79,** 104–113.

OOI, C.H., and TAN, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics **19,** 37–44.

PAN, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics **18,** 546–554.

PARK, P.J., PAGANO, M., and BONETTI, M. (2001). A nonparametric scoring algorithm for identifying informative genes from microarray data. Pacific Symp. Biocomput. **6,** 52–63.

PARK, T., YI, S.-G., LEE, S., LEE, S.Y., YOO, D.-H., AHN, J.-I., and LEE, Y.-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics **19,** 694–703.

PODANI, J. (1994). *Multivariate Analysis in Ecology and Systematics* (SPB Academic Publishing, The Hague).

QUACKENBUSH, J. (2001). Computational analysis of microarray data. Nat. Rev. Genet. **2,** 418–427.

QUACKENBUSH, J. (2002). Microarray data normalization and transformation. Nature Genet. Suppl. **32,** 496–501.

RADMACHER, M.D., MCSHANE, L.M., and SIMON, R. (2002). A paradigm for class prediction using gene expression profiles. J. Comput. Biol. **9,** 505–511.

RAMDAS, L., COOMBES, K.R., BAGGERLY, K., ABRUZZO, L., HIGHSMITH, W.E., KROGMANN, T., HAMILTON, S.R., and ZHANG, W. (2001). Sources of nonlinearity in cDNA microarray expression measurements. Genome Biol. **2,** research0047.

RAMONI, M., SEBASTIANI, P., and KOHANE, I. (2002). Cluster analysis of gene expression dynamics. Proc. Natl. Acad. Sci. USA **99,** 9121–9126.

RATHOD, P.K., GANESAN, K., HAYWARD, R.E., BOZDECH, Z., and DERISI, J.L. (2002). DNA microarrays for malaria. Trends Parasitol. **18,** 39–45.

RAYCHAUDHURI, S., STUART, J., and ALTMAN, R.B. (2000). Principal components analysis to summarize microarray experiments: Application to sporulation time series. Pacific Symp. Biocomput. **5,** 455–466.

RAYCHAUDHURI, S., SUTPHIN, P.D., CHANG, J.T., and ALTMAN, R.B. (2001). Basic microarray analysis: Grouping and feature reduction. Trends Biotechnol. **19,** 189–193.

SCHADT, E.E., LI, C., ELLIS, B., and WONG, W.H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J. Cell. Biochem. Suppl. **37,** 120–125.

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data* (Chapman & Hall, London).

SCHENA, M., SHALON, D., DAVIS, R.W., and BROWN, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270,** 467–470.

SCHUCHHARDT, J., BEULE, D., MALIK, A., WOLSKI, E., EICKHOFF, H., LEHRACH, H., and HERZEL, H. (2000). Normalization strategies for cDNA microarrays. Nucleic. Acids Res. **28,** e47.

SHAFFER, J.P. (1995). Multiple hypothesis testing. Annu. Rev. Psychol. **46,** 561–576.

SHANNON, W.D., WATSON, M.A., PERRY, A., and RICH, K. (2002). Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. Genet. Epidemiol. **23,** 87–96.

SHERLOCK, G. (2000). Analysis of large-scale gene expression data. Curr. Opin. Immunol. **12,** 201–205.

SIMON, R., RADMACHER, M.D., and DOBBIN, K. (2002). Design of studies using DNA microarrays. Genet. Epidemiol. **23,** 21–36.

SINGH, U., BREWER, J.L., and BOOTHROYD, J.C. (2002). Genetic analysis of tachyzoite to bradyzoite differentiation mutants in *Toxoplasma gondii* reveals a hierarchy of gene induction. Mol. Microbiol. **44,** 721–733.

SLONIM, D.K. (2002). From patterns to pathways: Gene expression data analysis comes of age. Nat. Genet. Suppl. **32,** 502–508.

SMYTH, G.K., YANG, Y.H., and SPEED, T. (2002). Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols.* M.J. Brownstein and A.B. Khodursky, eds. (Humana Press, Totowa, NJ) pp. 111–136.

SPANG, R., ZUZAN, H., WEST, M., NEVINS, J., BLANCHETTE, C., and MARKS, J.R. (2002). Prediction and uncertainty in the analysis of gene expression profiles. In Silico Biol. **2,** 0033.

SPEED, T.P., and YANG, Y.H. (2002). Direct versus indirect designs for cDNA microarray experiments. Sankhyā Ser. A **64,** 707–721.

SPELLMAN, P.T., MILLER, M., STEWART, J., TROUP, C., SARKANS, U., CHERVITZ, S., BERNHARDT, D., SHERLOCK, G., BALL, C., LEPAGE, M., *et al.* (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. **3,** research0046.

SPSS INC. (1998). *SYSTAT 9 for Windows* (SPSS Inc., Chicago, IL).

SZABO, A., BOUCHER, K., CARROLL, W.L., KLEBANOV, L.B., TSODIKOV, A.D., and YAKOVLEV, A.Y. (2002). Variable selection and pattern recognition with gene expression data generated by microarray technology. Math. Biosci. **176,** 71–98.

TEFFERI, A., BOLANDER, M.E., ANSELL, S.M., WIEBEN, E.D., and SPELSBERG, T.C. (2002). Primer on medical genomics, part III: Microarray experiments and data analysis. Mayo Clinic Proc. **77,** 927–940.

THEODORIDIS, S., and KOUTROUMBAS, K. (1999). *Pattern Recognition* (Academic Press, New York).

THOMAS, J.G., OLSON, J.M., TAPSCOTT, S.J., and ZHAO, L.P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. **11,** 1227–1236.

TIBSHIRANI, R.J., and EFRON, B. (2002). Pre-validation and inference in microarrays. Stat. Applic. Genet. Mol. Biol. **1,** 1.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., and CHU, G. (2002a). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA **99,** 6567–6572.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., EISEN, M., SHERLOCK, G., BROWN, P., and BOTSTEIN, D. (2002b). Exploratory screening of genes and clusters from microarray experiments. Statist. Sinica **12,** 47–59.

TIBSHIRANI, R., WALTHER, G. BOTSTEIN, D., and BROWN, P. (2001a). *Cluster Validation by Prediction Strength.* Technical Report (Department of Statistics, Stanford University, Stanford, CA).

TIBSHIRANI, R., WALTHER, G., and HASTIE, T. (2001b). Esti-

mating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B **63,** 411–423.

TOWNSEND, J.P., and HARTL, D. (2002). Bayesian analysis of gene expression levels: Statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biol. **3,** research0071.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., and ALTMAN, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics **17,** 520–525.

TROYANSKAYA, O.G., GARBER, M.E., BROWN, P.O., BOTSTEIN, D., and ALTMAN, R.B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics **18,** 1454–1461.

TSENG, G.C., OH, M.-K., ROHLIN, L., LIAO, J.C., and WONG, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. **29,** 2549–2557.

TSODIKOV, A., SZABO, A., and JONES, D. (2002). Adjustments and measures of differential expression for microarray data. Bioinformatics **18,** 251–260.

TUSHER, V.G., TIBSHIRANI, R., and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98,** 5116–5121.

WANG, J., DELABIE, J., AASHEIM, H.C., SMELAND, E., and MYKLEBOST, O. (2002a). Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformat. **3,** 36.

WANG, J., NYGAARD, V., SMITH-SØRENSEN, B., HOVIG, E., and MYKLEBOST, O. (2002b). MArray: Analysing single, replicated or reversed microarray experiments. Bioinformatics **18,** 1139–1140.

WERNISCH, L., KENDALL, S.L., SONEJI, S., WIETZORREK, A., PARISH, T., HINDS, J., BUTCHER, P.D., and STOKER, N.G. (2003). Analysis of whole-genome microarray replicates using mixed models. Bioinformatics **19,** 53–61.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, J.A., MARKS, J.R., and NEVINS, J.R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA **98,** 11462–11467.

WIERLING, C.K., STEINFATH, M., ELGE, T., SCHULZE-KREMER, S., AANSTAD, P., CLARK, M., LEHRACH, H., and HERWIG, R. (2002). Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. BMC Bioinformat. **3,** 29.

WOLFINGER, R.D., GIBSON, G., WOLFINGER, E.D., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C., and PAULES, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8,** 625–637.

WORKMAN, C., JENSEN, L.J., JARMER, H., BERKA, R., GAUTIER, L., NIELSEN, H.B., SAXILD, H.-H., NIELSEN, C., BRUNAK, S., and KNUDSEN, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. **3,** research0048.

WU, T.D. (2001). Analysing gene expression data from DNA microarrays to identify candidate genes. J. Pathol. **195,** 53–65.

XIAO, Y., SEGAL, M.R., RABERT, D., AHN, A.H., ANAND, P., SANGAMESWARAN, L., HU, D., and HUNT, C.A. (2002). Assessment of differential gene expression in human peripheral nerve injury. BMC Genom. **3,** 28.

YANG, I.V., CHEN, E., HASSEMAN, J.P., LIANG, W., FRANK, B.C., WANG, S., SHAROV, V., SAEED, A.I., WHITE, J., LI, J., *et al.* (2002). Within the fold: Assessing differential expression measures and reproducibility in microarray assays. Genome Biol. **3,** research0062.

YANG, Y.H., BUCKLEY, M.J., DUDOIT, S., and SPEED, T.P. (2002a). Comparison of methods for image analysis on cDNA microarray data. J. Comput. Graph. Stat. **11,** 108–136.

YANG, Y.H, DUDOIT, S., LUU, P., LIN, D.M., PENG, V., NGAI, J., and SPEED, T.P. (2002b). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. **30,** e15.

YANG, Y.H., and SPEED, T. (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. **3,** 579–588.

YATES, F. (1937). *The Design and Analysis of Factorial Experiments* (Imperial Bureau of Soil Science, Harpenden, UK).

YEUNG, K.Y., HAYNOR, D.R., and RUZZO, W.L. (2001). Validating clustering for gene expression data. Bioinformatics **17,** 309–318.

ZAR, J.H. (1999). *Biostatistical Analysis*, 4th ed. (Prentice-Hall, Upper Saddle River, NJ).

Address reprint requests to:
John T. Ellis, Ph.D.
Department of Cell & Molecular Biology
University of Technology, Sydney
Westbourne St.
Gore Hill, NSW 2065, Australia

E-mail: john.ellis@uts.edu.au