Statistical issues with microarrays: processing and analysis

Robert Nadon and Jennifer Shoemaker

The study of gene expression with printed arrays and prefabricated chips is evolving from a qualitative to a quantitative science. Statistical procedures for determining quality control, differential expression, and reproducibility of findings are a natural consequence of this evolution. However, problems inherent to the technologies have raised important issues of how to apply adequate statistical tests. As a consequence, statistical approaches to microarray research are not yet as routine as they are in other sciences. Statistical methods, tailored to microarrays, continue to be adapted and developed. We present an overview of these methods and of outstanding issues in their use and validation.

The current interest in gene expression data has increasingly focused on image quantification, data processing, and data analysis. Data processing includes background subtraction, normalization, and detection of outliers. Data analysis can be viewed in two broad categories: (1) pattern recognition, which can be unsupervised (cluster analysis, class discovery) or supervised (discriminant analysis, class prediction); or (2) detection of differential expression on a probe-by-probe basis. This article discusses conceptual statistical issues concerning data processing and differential expression.

The current bottleneck in the processing of microarray data occurs after the data are generated; the magnitude of the problem is proving to be on a par with developing the technology itself. The difficulties stem primarily from the myriad potential sources of random and systematic measurement error in the microarray process [1,2] and from the small number of samples (e.g. cell lines, patients) relative to the large number of variables (probes). This raises questions about the validity of many of the microarray findings reported to date [3]. Not enough information is available to allow an adequate estimation of all parameters of interest within the limits of standard analysis. Confronting these issues is crucial to providing useful microarray data and drawing informative conclusions.

This article focuses on the elucidation of issues rather than an exhaustive review of the burgeoning technical literature. Although differing in their particulars, issues and solutions are common to all array technologies (cDNA versus oligonucleotide-spotted arrays; radioisotopic versus fluorescent labeling; nylon membrane versus glass slide versus biochip substrates). Other sources contain introductions to microarrays [4–6], data acquisition and analysis [7], data mining [1,8,9], quantitative biology [10], and statistical concepts [11,12].

Robert Nadon*
Imaging Research Inc., Brock University, 500 Glenridge Ave, St Catharines, Ontario, Canada L2S 3A1.
*e-mail: Robert.Nadon@imagingresearch.com

Jennifer Shoemaker
Duke University Medical Center, 208 Hanes House, Box 3958, Durham, NC 27710, USA.

Statistical and scientific inference

Consider the problem of detecting a difference in expression between two groups. Contributing to this difference is a biological (treatment) component, which is negligible in the case of no difference, and an error component, which is divided into systematic and random components. The purpose of the statistical analysis in this context is to detect whether there is a reliable, biologically relevant difference in expression level. Difficulties in detecting a difference in expression stem from problems such as CONFOUNDING (see Glossary) of the systematic error and the treatment components (which might result in a statistically significant result that has no biological underpinning) and small sample sizes (which engender large random error and potential violation of PARAMETRIC assumptions, leading to difficulties in interpreting results).

True expression values are unknown in real-world applications and must be inferred from measured data. Observed differences in expression that exceed a threshold defined jointly by random error and by the probability of a false positive are considered 'statistically significant', a minimum requirement for biological significance. If chance factors provide a reasonable explanation for a putative effect, then biological significance is moot. In both instances, statistical results must be interpreted within the context of the experimental design and the purpose of the study:

- Statistically significant effects can reflect the biasing effects of EXTRANEOUS FACTORS rather than biology (see the section on systematic error, below).
- Lack of statistical significance can reflect low experimental SENSITIVITY, rather than absence of a biological effect. Low sensitivity can be caused by an inadequate number of REPLICATES, failure to control extraneous factors that contribute to random error, or both (see the section on errors of inference, below).

Gene expression data can be analyzed in several ways; here we discuss univariate procedures, which are critical for planning experiments, determining the function of specific genes, and producing high quality data for subsequent analyses. Figure 1 shows how univariate procedures can be incorporated profitably into the early stages of microarray analysis.

Errors of measurement

There are two types of measurement error: random and systematic. Random error is minimized by controlling extraneous factors and by obtaining more repeated measurements (replicates). Systematic errors (bias) are

 0168-9525/02/$ – see front matter PII: S0168-9525(02)02665-3

Glossary

Coefficient of variation: Standard deviation divided by the mean.
Confounded: Experimental factors are confounded when their potential effects cannot be distinguished. For example, if control and experimental measurements are obtained on different days, treatment and day effects are confounded.
Efficiency: A statistical test or measure is efficient when its sampling distribution has smaller variance (i.e. is more reliable) than other similar tests or measures. For example, in estimating the mean of a normal population, the sample mean is more efficient than the sample median. There is often a tradeoff, however, between efficiency and robustness. The sample mean is less robust than the sample median.
Extraneous factor: A characteristic of an experiment that is not of interest but that might influence the outcome (also called a 'nuisance parameter').
***P*-value:** The probability of observing a statistical result at least as discrepant as the one actually observed, assuming that the null hypothesis is true. The smaller the *P*-value, the less probable the statistical result is due to 'chance'.
Parametric versus nonparametric statistical tests: Parameters are quantities that are constant for particular distributions but that take on different values for different members of families of the same kind of distribution. For example, the population mean and variance are parameters of the normal distribution. One or more population parameters are estimated when using a parametric test. Nonparametric tests (sometimes called 'distribution-free' tests) have weaker distributional assumptions and do not involve parameter estimation. When parametric assumptions are reasonable (or when the parametric statistic is robust with respect to their violation), parametric tests are generally statistically more powerful than their nonparametric counterparts.
Population: The (hypothetical) universe of numbers of interest. For example, the expression intensities of replicate measurements of cancerous tissue for a particular gene come from the same statistical population as intensities from non-cancerous tissue if the two expression intensity distributions are identical, despite the fact that measurements come from different physical populations.
Resampling procedures: Computer-intensive procedures that sample repeatedly from observed data to generate empirical estimates of results that would be expected by 'chance'. Bootstrapping and permutation tests are examples.
Replicate: A repeated measurement of an attribute (or process) of interest. Ideally, replicates should be obtained in a manner that provides broad inference, which is determined by experimental design and statistical sampling factors. Consider an experiment with spotted arrays of cDNA. To examine whether gene expression is related to a certain tumor type, tumor and normal tissue replicates might be obtained from: (1) different patients; (2) different RNA samples from one patient; or (3) different aliquots of the same RNA sample from one patient. In the case of (3), a gene's differential expression might be due to specifics of the patient, to the way the RNA was extracted, to the tumor, or to some combination of these factors. Interpretation in (2) is less ambiguous because specifics of RNA preparation are less likely to underlie the differential expression, although patient or patient–tumor combination effects might still be the underlying causes. Interpretation is clearest and results generalize most broadly in (1) because the replicates are sampled from a more representative set of observations. Additional issues arise when multiple oligonucleotides are used to characterize specific genes. Affymetrix chips, for example, provide varying numbers of oligonucleotide probes (e.g. 16, 20) for each gene. Expression values associated with these probes from the same chip can be considered replicates in that they are intended to represent the expression values of the same gene. In practice, however, oligonucleotide-specific expression effects reveal that the multiple oligonucleotides are being sampled from different statistical populations, preventing the expression values from being used as replicates. Fortunately, treatment:control ratios (or log ratios) of the various oligonucleotides for the same gene can be considered replicates because treatment effects are generally not oligonucleotide-specific. In this sense, the current Affymetrix chips provide a large number of replicated treatment:control ratios. Because these ratios are obtained from the same RNA sample, however, the generalizability of single chip studies is limited.
Robustness: A statistical test or measure is robust either when violation of assumptions has little effect on the distribution of the statistic (and consequently on the false-positive and false-negative rates) or when it is based on weaker assumptions. For example, the sample median is more robust than the sample mean. There is often a tradeoff, however, between robustness and efficiency. The sample median is less efficient than the sample mean.
Sensitivity: An inferential procedure that produces few false negatives has high sensitivity.
Specificity: An inferential procedure that produces few false positives has high specificity.
Stochastic model: A description of an event or process within a probabilistic framework that includes 'chance' events in the form of random measurement error. It is contrasted by a deterministic model in which random error is inconsequential or nonexistent.

controlled experimentally as far as possible, although additional statistical correction is invariably necessary with current microarray technology.

A partial list of extraneous factors that contribute to random error and/or to bias of microarray expression values includes: the time of day that arrays are processed [13]; target accessibility, which is affected by variations in the absorbency of discrete nylon membranes [14]; target fixation to glass slides [2]; and variations in washing procedures [15].

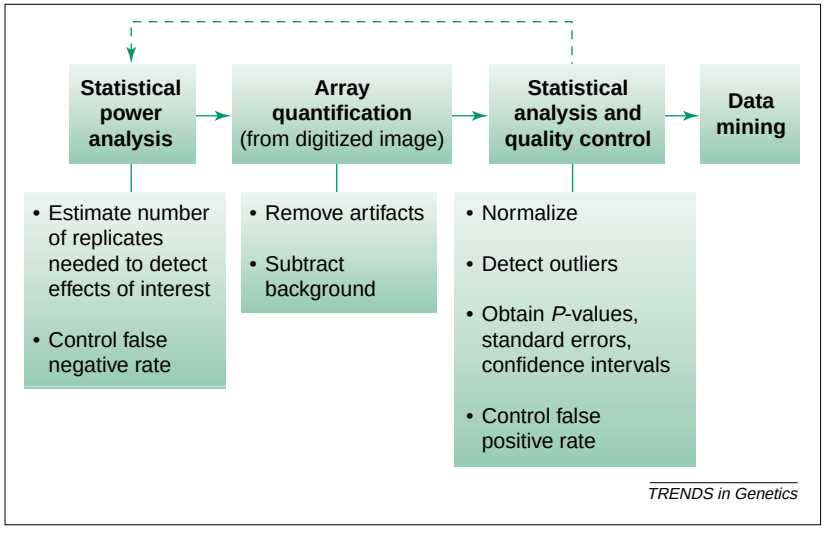

Fig. 1. Data analysis workflow.

Random error

Random error is a measure of uncertainty in the measurement and is therefore central to statistical inference. Random errors are not 'mistakes' in the colloquial sense. Rather, they reflect inevitable uncertainties in all scientific measurements, making statistical procedures necessary. Random error cannot be eliminated, but instead is estimated from observed data.

For example, consider the case of a probe that is not differentially expressed. Because of random measurement error, its measured differential expression ratio will deviate from its true value of 1:1. Deviations from this 1:1 ratio (or, more typically, from the log ratio) are due to 'chance', reflecting a STOCHASTIC model (Boxes 1,3).

Depending on the purpose of the experiment, replicates can be obtained from the same or different cell lines, patients, and so on. Random error across replicates obtained from aliquots of the same RNA sample restricts sources of error to technical aspects of the process. Obtaining replicates from different biological samples (e.g. patients) increases random error [16] but produces results that have better external validity and broader applicability (wider inference). A minimum of three or four replicates per group or experimental condition has been recommended to account for random variation and to provide good sensitivity (R. Nadon, unpublished; Refs [17,18]).

Box 1. Ratios and log transformation

Ratios of raw expression values outputted by image quantification software are usually not appropriate for statistical analysis. As described below, log-transformed data are generally preferred to ratios of raw signal values. Note, however, that an alternative transformation has recently been proposed, which takes into account different error characteristics of low and high signals to produce a common error variance across the entire data range [a].

Why not use ratios of raw expression values?

Random error of replicate raw expression values is approximately proportional to signal intensity; hence, equivalent-fold changes are not equally reproducible. For example, a twofold change derived from large expression values is less reproducible than a twofold change derived from average expression values, producing lower confidence in the former. Moreover, most parametric statistical tests assume an additive rather than a proportional error model.

Distributions of replicated raw expression values (and consequently of differential expression ratios) tend to be asymmetric (skewed). This violates the normality assumption of many statistical tests. The central limit theorem affords little protection for most microarray studies because of the typically small sample sizes, which produce incorrect *P*-values associated with parametric tests like the *t*-test and ANOVA.

Summary statistics of replicated ratios yield different quantities, depending on the numerator/denominator assignment. Consider a single treatment:control ratio of 2:1. Because its inverse is 1:2, the same quantitative conclusion is reached whether we divide treatment by control or vice versa. This is not the case when using multiple ratios. Now, consider the following treatment:control ratios obtained from three replicates: 2:1.1, 5:1.4, and 15:5 (mean = 2.80, standard deviation = 0.89, COEFFICIENT OF VARIATION = 0.32). Now consider the inverted (control:treatment) ratios: 1.1:2, 1.4:5 and 5:15 (mean = 0.39, standard deviation = 0.14, coefficient of variation = 0.37). These quantities are different and, unlike the single ratio case, the inverse of the averaged treatment:control ratio is not equal to the inverse of the averaged control:treatment ratios ($1 \div 2.80 \neq 0.39$). Also, using the standard deviation as an index of reproducibility, it appears that greater confidence should be placed in the control:treatment ratios (smaller standard deviation); using the coefficient of variation, the reverse is true. Because the numerator/denominator assignment is arbitrary, these quantification differences are scientifically unacceptable.

Why use log transformed expression values?

Transforming expression data to a log scale (any base) removes much of the proportional relationship between random error and signal intensity. (As discussed in the main body of the text, low signals are often an exception. Random error of log-transformed data is often inversely proportional to signal in the low signal range because of the proportionally nontrivial error associated with background correction.) Moreover, log transformation has the advantage of transforming the error model from a proportional to an additive one because log(a/b) = log(a) – log(b), although non-logged (raw) difference scores (a–b) have also been recommended for low signals [b].

Distributions of replicated logged expression values (and consequently of log ratios) tend to be normal.

Summary statistics of log ratios yield the same quantities, regardless of the numerator/denominator assignment. Consider the following treatment:control $\log_{10}$ ratios obtained from the same three replicates above: log (2:1.1), log (5:1.4), and log (15:5) (mean = 0.43, standard deviation = 0.15). Now consider the inverted (control:treatment) ratios: log (1.1:2), log (1.4:5) and log (5:15) (mean = –0.43, standard deviation = 0.15). The difference in sign for the means reflects whether on average the numerator is larger (+) or smaller (–) than the denominator. Also, taking the antilog of the average log ratios returns the data to a fold-metric. For example, $10^{0.43}$ yields the geometric mean of the raw values (2.69), which can be interpreted as an average-fold change. Note that the geometric mean is always less than the arithmetic mean, except when all numbers are equal, when identity holds.

References

a Durbin, B. *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data (www.cipic.ucdavis.edu/~dmrocke/preprints.html)

b Rocke, D.M. and Durbin, B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.* 8, 557–569

Random variation across replicates can be estimated on a probe-by-probe basis, assuming that the true random error associated with each replicate is the same. The problem with this approach is the small number of replicates per gene that is typical of array studies. Random error estimates with small sample sizes have large random error. This results in poor outlier detection (which in turn produces biased expression estimates) and low sensitivity for statistical tests. More accurate random error estimates can be obtained by pooling (averaging) error variances across all probes, thereby circumventing problems inherent to small sample sizes [19–21]. However, this latter approach requires the assumption that one random error is common to all probes, which is not always reasonable, especially at low intensity levels. This is because random error associated with background correction can be large relative to low expression values (see the section on systematic error, below). Methods for addressing this issue include using ROBUST methods for pooling error estimates locally according to expression intensities (T.B. Kepler, unpublished [see Box 2]; Refs [7,20,22]) and modeling the random error associated with background correction directly [23,24]. However obtained, estimates of random error can be used to conduct statistical tests of differential expression for individual probes (see the section on statistical tests of differential expression, below) and to assess the reliability of data mining results [24–26].

Systematic error

Systematic errors are biases; they result in a constant tendency to over- or underestimate true values, thereby decreasing accuracy. Biasing factors come in many forms and are partially dependent on spotting, scanning and labeling technologies. Bias can affect all expression values on an array equally or depend on other factors (e.g. spatial location, spotting pins, signal intensity) [2,27]. The ubiquitous potential for inaccurate expression values is a major impediment to creating public expression databases derived from different laboratories [28]. Indeed, without stringent controls, the accuracy of intra-laboratory comparisons is often questionable.

Sources of bias are in theory identifiable by quality control studies. However, biasing effects from various sources can be nonorthogonal and are often nonlinear. This, and the typically few replicates available for estimation, complicate quantifying the specific sources of bias. None the less, experimental design and robust statistical approaches have contributed extensively to correcting biases in array data (T.B. Kepler, unpublished; M. Sapir and G.A. Churchill, unpublished [see Box 2]; Refs [2,19,27]).

Background on the substrate presents a special case of bias. On the assumption of additive error, an estimate of background is usually subtracted from the measured expression value before log transformation and before correcting other systematic errors. Proportional error models have also been proposed [29]. Estimates of background intensity can be obtained from low pixel intensities within spots [30], from areas outside the spots [7], or from negative controls that contain either no DNA or nonspecific DNA [29].

Other sources of systematic error are considered proportional to signal intensity and are often

Box 2. Data from unpublished work

Some of the unpublished data are from conference and technical papers that are available on the authors' websites.

- Dudoit, S. *et al.* (2000) Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. (stat-www.berkeley.edu/users/terry/zarray/html/papersindex.html)
- Kepler, T.B. *et al.* (2000) Normalization and analysis of DNA microarray data by self-consistency and local regression. (www.santafe.edu/sfi/publications/00wplist.html)
- Sapir, M. and Churchill, G.A. (2000) Estimating the posterior probability of differential gene expression from microarray data. (www.jax.org/research/churchill/pubs/index.html)
- Storey, J.D. (2001) A New Approach to False Discovery Rates and Multiple Hypothesis Testing. Technical Report No. 18. (www-stat.stanford.edu/research/)

corrected by dividing the raw expression value by an estimate of the systematic error, although a preferred method is to log-transform both the error estimate and the expression value and to correct the values by subtraction. Either way, corrected expression values are said to be 'normalized'.

Various normalization methods have been proposed. Global methods divide expression values by an estimate of systematic error (mean, quantile) for each array, controlling for proportional differences across slides (membranes or chips). To be effective, most probes must not be affected by the treatment [31], or the normalization procedure confounds error and treatment effect, potentially masking the effect of differentially expressed probes or creating differential effects where none exist.

Reference standards that are unaffected by treatment are needed to circumvent this problem. Housekeeping genes, which in theory should show little treatment effect, have largely been unsatisfactory [32,33] and heterologous synthetic DNA has been advanced as a possible alternative [34]. As with the global methods, however, this type of systematic error correction assumes that bias is constant across the entire range of the data.

As a step towards resolving these issues, titration series of spiked heterologous DNA look promising. However, calibrating to spiked standards is difficult, partly because label intensities might not reflect absolute message levels [31] and partly because uncertainties in fluid-handling and RNA preparations could bias the standards [35]. As a result, even calibrated expression arrays provide only relative measurements (e.g. this spot is 50% brighter than that one, instead of specifying 150 labeled molecules versus 100).

This problem is most marked when two fluors are used to construct ratios. The magnitude of the ratio will be biased by what is usually a fluor-specific relation between intensity and hybridization, although this might be partially gene specific [16,36]. Compounding the difficulties in two-color arrays, fluorescent dyes differ in mean brightness and background noise [37]. A standard design to control for these effects is to use the same reference sample for various experimental conditions. For example, three different tissues could be compared with the same reference sample. The same goal can sometimes be achieved more efficiently with known experimental designs [38,39].

Two-color arrays present an additional normalization problem. The extent of systematic error caused by differences in fluorescent dyes typically depends on expression level. Nonlinear statistical regression procedures have been developed to address this problem (T.M. Houts, unpublished; Ref. [27]), although replicate experiments that alternate the dyes across the treatment and the reference samples can sometimes make nonlinear normalization unnecessary [40].

Outliers

Outliers are extreme values in a distribution of replicates. Poor reproducibility can be caused by uncorrected image artifacts (e.g. dust on fluorescent arrays or 'blooming' of adjoining spots on radioisotopic arrays). They can also be caused by factors that are undetectable by image analysis, such as cross-hybridization or failure of one probe to hybridize adequately. Outliers can number as high as 15% in typical microarray studies [17] and are revealed only by the extreme deviance of their expression values relative to other replicates. Undetected outliers bias the estimation of both the expression value and its associated random error, reducing both SPECIFICITY and sensitivity. They thus compromise individual tests of differential expression and data mining classification (R. Nadon, unpublished).

Although numerous methods are available for statistical outlier detection [41], they are generally inadequate given the small number of replicates typical of microarray studies. Because they estimate random error on a probe-by-probe basis, these methods often falsely identify replicates as outliers and fail to detect true outliers. For example, given the large number of probes in an array, it is not unusual for the expression level for two of three replicates for a particular probe to be very close together because of random sampling error. In this circumstance, the third seemingly extreme value is incorrectly identified as an outlier. By contrast, because random error estimation based on small sample sizes is imprecise, many true outliers go undetected. Larger sample sizes are needed to detect outliers more accurately and precisely. One method is to pool standardized residuals for all probes [20]. Alternatively, standard errors can be estimated from a training set of numerous arrays and applied to the data at hand [42].

Errors of inference

All statistical inferences have a probability of being incorrect. False positives (Type I errors) are incorrect inferences of differential expression; false negatives (Type II errors) are failures to detect true differential expressions. In standard practice, the false-positive rate (α) is set in advance. The false-negative rate (β) is a function of various parameters, including α, and can

Box 3. Hypothesis testing

Statistical hypothesis testing provides a benchmark (probability of a false positive) against which to assess observed effects. A research hypothesis is usually stated in positive terms (e.g. a particular gene or group of genes is related to a disease state). By contrast, the statistical null hypothesis is stated in negative terms (e.g. a particular gene or group of genes is *not* related to a disease state).

Under a stochastic model, observed data will deviate somewhat from the null hypothesis simply by chance. Observed values must exceed a predefined chance threshold (α level) for the null hypothesis to be regarded as improbable, in which case the null hypothesis is rejected in favor of the alternative (research) hypothesis. The probability framework of statistical null hypotheses is particularly useful in exploratory research because, without it, chance results of analyses with many variables are readily and erroneously interpreted as meaningful [a].

The null hypothesis for a difference between two conditions and its mutually exclusive alternative hypothesis can be stated as follows:

$H_0: \mu_{DISEASE} - \mu_{CONTROL} = 0$

$H_1: \mu_{DISEASE} - \mu_{CONTROL} \neq 0$

The symbol μ represents the POPULATION expression value for the particular probe of interest. A zero difference is usually hypothesized, although a non-zero value can be used also. The hypotheses can also be directional, in which case the = sign in the null hypothesis is replaced by $\geq$ or by $\leq$ and the $\neq$ sign in the alternative hypothesis is replaced by < or >, respectively.

Reference

a Armstrong, J.S. (1967) Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *Am. Statistician* 21, 17–21

be estimated by statistical power analysis (Box 4). Given the large number of probes in array studies, failure to consider the false-positive rate can lead to hundreds of false leads relative to a small number of true effects (low specificity:sensitivity ratio). Alternatively, imposing conservative requirements for judging a probe to be differentially expressed increases the false-negative rate, typically producing a low sensitivity:specificity ratio.

To set the false-positive rate for one statistical test, an acceptable α must be chosen in advance (e.g. the familiar *P*-VALUE < 0.05 criterion). However, it becomes more complicated when large numbers of tests are conducted, as is the case with microarrays. If none of the probes are differentially expressed, 5% are expected to reach 'statistical significance' with the 0.05 criterion. In instances where few probes are expected to show differential effects, the number of false positives overwhelms the correct differential expression inferences. Discussions of false-positive control in microarray studies are available (S. Dudoit *et al.*, unpublished [see Box 2]; Refs [20,43–45]).

The single-step Bonferroni correction is the best-known procedure for controlling the false-positive rate when multiple tests are conducted [46]. The nominal false-positive rate is divided by the number of tests to yield the effective rate. For example, in the case of 10 000 probes, a nominal rate of 0.05 is divided by 10 000 to yield a new false-positive threshold of 0.000005 to be used with each statistical test. This procedure ensures that the probability of making at least one false-positive error among the entire set of statistical tests is no more than 0.05. This is a stringent control that drastically increases the false-negative rate. Other procedures involve multiple steps and offer improvements over the single-step procedure [46,47].

Methods that address some of the problems with the Bonferroni procedures include the false discovery rate (FDR) [48] and positive false discovery rate (pFDR) (J.D. Storey, unpublished [see Box 2]), which have more statistical power than the Bonferroni procedures and so provide a better sensitivity:specificity ratio. Still other methods are based on RESAMPLING PROCEDURES (S. Dudoit *et al.*, unpublished [see Box 2]; Refs [44,49]).

All procedures that correct for multiple tests lower sensitivity to improve specificity. The key is achieving the right balance. In early screening phases it might be more important to cast the net widely and minimize false negatives at the expense of a large number of false positives. In later phases, when more intensive follow-ups of individual genes are contemplated, minimizing false positives could become more important. Either way, a little forethought can provide substantial gains in sensitivity. For example, it is not necessary to control the false-positive rate in the same way for all probes in a study. There could be a subset of probes that is thought (before examining the data) to be especially important. The subset can be tested first, reducing the number of tests and lessening the stringency of the correction among the more promising probes [50].

Box 4. Power analysis

Statistical power analysis provides a benchmark (probability of a false negative) against which to assess the *lack* of observed effects. High power is desirable and it implies a low probability of a false negative.

Power analysis can be used to estimate how many replicates are needed to have a specified probability of finding a minimum effect size. A researcher can estimate, for example, how many replicates are needed to detect geometric means of twofold or greater with a 0.80 probability. By the same token, improvements in sensitivity gained by adding replicates can be estimated. A law of diminishing returns usually applies. In microarrays, substantial gains in sensitivity are almost always achieved by adding one replicate when sample sizes are small (e.g. increasing sample size from two to three). Adding one replicate tends to provide lower gains in sensitivity when sample sizes are larger (e.g. increasing sample size from five to six).

Statistical power is a function of four parameters: α level, number of replicates, random measurement error, and population effect size (e.g. $\mu_{DISEASE} - \mu_{CONTROL}$). All things being equal, power increases as α, replicates, and effect size increase and as random error decreases.

Statistical tests of differential expression

The first formal statistical model for assessing significance of differential expression ratios did not use replicates [51]. The central idea is that most probes in a study will not express differentially and that the average treatment:reference ratio of all the probes will be approximately 1; probes with differential expression that deviates substantially from this average are considered significantly differentially expressed. Alternatively, a reference subset of probes that, by definition, do not show differential expression (e.g. housekeeping genes) can be used to generate the distribution of ratios under the null hypothesis and the probes of interest are compared with this.

Disadvantages of studies without replicates include: (1) the inability to distinguish between large ratios

caused by real effects and large ratios caused by outliers; (2) reliance on the assumption of no differential expression among most probes; (3) lower sensitivity; and (4) the exclusion of biological variability across different samples. These disadvantages have led to calls for the routine use of replicates (J. Woodgett, unpublished, Refs [17,18,20,52]).

Specific choices depend on assumptions about the data. If a sufficient number of replicates is available, and if expression values are assumed to be normally distributed with few outliers, then the well-known *t*-tests can readily be calculated with spreadsheet software. Nonparametric alternatives (e.g. the Mann–Whitney U test, Wilcoxon's matched pairs signed rank test) are available but are generally too insensitive to detect moderate-to-small differential effects. Other issues, such as false-positive control and modified random error estimation, require rudimentary programming or specialized software (S. Dudoit *et al*., unpublished [see Box 2]; Refs [24,44]). The underpinning of these approaches, and that of the analysis-of-variance model proposed by Wolfinger *et al*. [39] is that true random error across replicates is probe-specific and consequently should be estimated separately for each probe.

A more general approach assumes that the same true random error applies to all probes within a specific study, or to probes of similar intensities. This approach pools error estimates across probes and permits the use of the *z*-test, a statistical test similar to the *t*-test but that requires fewer replicates to achieve the same sensitivity by virtue of more precise error estimation (T.B. Kepler, unpublished [see Box 2]; Ref. [20]). In a variation of this approach, Kerr *et al*. have used a bootstrapping procedure that assumes a common random error across all probes but that does not assume the normal distribution [19]. In all cases, if the underlying assumptions are warranted, the probability of detecting small expression changes can be an order of magnitude (or more) larger than the *t*-test, especially with small numbers of replicates [20].

A conceptual combination of the two approaches assumes that random error is probe-specific but the same across experiments. One approach obtains random-error estimates based on a modification of Chen *et al*. [51] and/or from replicates [23]. Another approach, designed for Affymetrix data, provides statistical tests based on weighted averages of oligonucleotide-specific random-error estimates [26]. Both approaches require relatively large training sets to generate the error estimates, which can then be applied to new, smaller, sample data sets.

Variations of the above methods are possible. For example, the *t*-test can be used with pooled error estimates. The pooling can vary (e.g. across probes or treatments) and permutations within probes can be used (S. Dudoit *et al*., unpublished [see Box 2]; Ref. [18]). Pooled error estimates can also be generated with training sets and applied to new datasets. Gene-specific error estimates can be generated for unreliable low-expression probes, whereas error for other probes is based on pooled or training estimates.

Throughout, this review has focused on classical (frequentist) statistical methods; however, Bayesian methods, or methods that include a Bayesian component, are also being developed [22,53–56]. Bayesian approaches have been used to study many problems in genetics and molecular biology [57] and are well suited to the field of gene expression. Specifically with respect to testing for differential expression, a Bayesian model of the distribution of differences uses a combination of a prior distribution and the data. Because a fully Bayesian method is computationally intensive, shortcuts can be taken. Thus, both measurement error and error due to gene spotting can be described using an empirical Bayes model [54]. Using simulations, a Bayesian analog to the *t*-test worked better than the *t*-test when the number of replicates was low; when it increased, the *t*-test and the Bayesian analog worked equally well [55]. Bayesian methods are a standard way of analyzing differential expression using serial analysis of gene expression (SAGE) data [53,58,59].

Validation

Most microarray studies are more or less exploratory (hypothesis-generating). Exploration is an important part of the scientific process because it forms the basis for new directions and future experiments. Well-defined questions are important (because they minimize statistical uncertainties caused by unconstrained exploration), as is validation (because it places conclusions on firmer ground). Molecular 'gold-standard' techniques such as RT–PCR or northern blots are often used to validate the results in cases where the primary interest is in which genes are differentially expressed [60–62]. This approach is useful for determining the specificity of microarray analysis but provides no information about the equally critical issue of sensitivity; this is especially important because differential effects are often larger with RT–PCR and northern blots [36,63,64]. Ideally, some nondifferentially expressed probes should also be validated.

Findings can also be validated using various statistical approaches. A subset of probes in the original experiment can be tested in another sample. Alternatively, the study can be repeated to determine whether differential effects are reproducible. One way to do this is to determine whether the same probes are found to be statistically significant in each study. For a two-condition comparison, a better way would be to conduct a two-way (condition × study) factorial analysis of variance; failure to cross-validate the initial results would be evidenced by a statistical interaction between the factors.

Putting it all together

Although microarray technology itself is now well established, statistical analysis of microarrays is still in its infancy, and methods are not yet in place to allow

automated high-throughput analysis without human intervention. Data processing and analysis remain the bottlenecks. Although these problems will be eased by new software tools that implement some of the methods discussed here, scientists and statisticians will need to be familiar with each other's areas of expertise at all stages of microarray analysis (experimental design, analysis, and interpretation) for optimal collaboration.

References

1 Bassett, D.E. *et al.* (1999) Gene expression informatics: It's all in your mine. *Nat. Genet.* 21, 51–55
2 Schuchardt, J. *et al.* (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28, E47
3 Miller, R.A. *et al.* (2001) Interpretation, design, and analysis of gene array expression experiments. *J. Gerontol. Series A – Biol. Sci. Med. Sci.* 56, B52–B57
4 Bowtell, D.D.L. (1999) Options available – from start to finish – for obtaining expression data by microarray. *Nat. Genet.* 21, 25–32
5 Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays – a technology review. *Nat. Cell Biol.* 3, E190–E195
6 Van Hal, N.L.W. *et al.* (2000) The application of DNA microarrays in gene expression analysis. *J. Biotechnol.* 78, 271–280
7 Hess, K.R. *et al.* (2001) Microarrays: Handling the deluge of data and extracting reliable information. *Trends Biotechnol.* 19, 463–468
8 Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427
9 Sherlock, G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* 12, 201–205
10 Burton, R.F. (1998) *Biology by Numbers: An Encouragement to Quantitative Thinking*, Cambridge University Press
11 Holland, B.K. (1998) *Probability Without Equations: Concepts for Clinicians*, Johns Hopkins University Press
12 Motulsky, H. (1995) *Intuitive Biostatistics*, Oxford University Press
13 Lander, E.S. (1999) Array of hope. *Nat. Genet.* 21, 3–4
14 Perret, E. *et al.* (1998) Improved differential screening approach to analyse transcriptional variations in organized cDNA libraries. *Gene* 208, 103–115
15 Shalon, D. *et al.* (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645
16 Bartosiewicz, M. *et al.* (2000) Development of a toxicological gene array and quantitative assessment of this technology. *Arch. Biochem. Biophys.* 376, 66–73
17 Lee, M.L.T. *et al.* (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9834–9839
18 Li, Y. *et al.* (2001) Evaluation of current methods of testing differential gene expressions and beyond. *Proceedings of the CAMDA 2001 conference* (Johnson, K.F. and Lin, S.M., eds), Kluwer
19 Kerr, M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819–837
20 Nadon, R. *et al.* (2001) Statistical inference methods for gene expression arrays. In *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE* (Vol. 4266) (Bittner, M.L. *et al.*, eds), pp. 46–55, SPIE
21 Nadon, R. *et al.* Statistical inference in array genomics. In *Microarrays for the Neurosciences: An Essential Guide* (Geschwind, D. and Gregg, J., eds), pp. 109–140, MIT Press
22 Long, A.D. *et al.* (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework – analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.* 276, 19937–19944
23 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
24 Rocke, D.M. and Durbin, B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.* 8, 557–569
25 Kerr, M.K. and Churchill, G.A. (2001) Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. U. S. A.* 98, 8961–8965
26 Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol.* 2, research0032.1–0032.11
27 Yang, Y.H. *et al.* (2001) Normalization for cDNA microarray data. In *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE* (Vol. 4266), (Bittner, M.L. *et al.*, eds) pp. 141–152, SPIE
28 Aach, J. *et al.* (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.* 10, 431–445
29 Wu, W. *et al.* (2001) Chemometric strategies for normalisation of gene expression data obtained from cDNA microarrays. *Anal. Chim. Acta* 446, 451–466
30 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
31 Duggan, D.J. *et al.* (1999) Expression profiling using cDNA microarrays. *Nat. Genet.* 21, 10–14
32 Ermolaeva, O. *et al.* (1998) Data management and analysis for gene expression arrays. *Nat. Genet.* 20, 19–23
33 Thellin, O. *et al.* (1999) Housekeeping genes as internal standards: Use and limits. *J. Biotechnol.* 75, 291–295
34 Eickhoff, B. *et al.* (1999) Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res. Methods* 27, 22–24
35 Hill, A.A. *et al.* (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* 2, 0055.1–0055.13
36 Taniguchi, M. *et al.* (2001) Quantitative assessment of DNA microarrays – comparison with Northern blot analyses. *Genomics* 71, 34–39
37 Geiss, G.K. *et al.* (2000) Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology* 266, 8–16
38 Kerr, M.K. and Churchill, G.A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.* 77, 123–128
39 Wolfinger, R.D. *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8, 625–637
40 Tseng, G.C. *et al.* (2001) Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* 29, 2549–2557
41 Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, John Wiley
42 Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* 98, 31–36
43 Mills, J.C. and Gordon, J.I. (2001) A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res.* 29, U5–U17
44 Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121
45 Sabatti, C. *et al.* Thresholding rules for recovering a sparse signal from microarray experiments. *Math. Biosci.* (in press)
46 Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 800–803
47 Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statistics* 6, 65–70
48 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statistical Soc. Ser. B - Methodological* 57, 289–300
49 Westfall, P.H. and Young, S.S. (1993) On adjusting p-values for multiplicity. *Biometrics* 49, 941–944
50 Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates
51 Chen, Y. *et al.* (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* 2, 364–374
52 Dougherty, E.R. *et al.* (2001) Time series inference from clustering. In *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE* (Vol. 4266), (Bittner, M.L. *et al.*, eds), pp. 222–227, SPIE
53 Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Res.* 7, 986–995
54 Newton, M.A. *et al.* (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 8, 37–52
55 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
56 Theilhaber, J. *et al.* (2001) Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *J. Comput. Biol.* 8, 585–614
57 Shoemaker, J.S. *et al.* (1999) Bayesian statistics in genetics – a guide for the uninitiated. *Trends Genet.* 15, 354–358
58 Chen, H.X. *et al.* (1998) Characterization of gene expression in resting and activated mast cells. *J. Exp. Med.* 188, 1657–1668
59 Lal, A. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res.* 59, 5403–5407
60 Bustin, S.A. (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* 25, 169–193
61 Chaib, H. *et al.* (2001) Profiling and verification of gene expression patterns in normal and malignant human prostate tissues by cDNA microarray analysis. *Neoplasia* 3, 43–52
62 Piétu, G. *et al.* (1999) The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Res.* 9, 195–209
63 Rajeevan, M.S. *et al.* (2001) Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR. *J. Mol. Diagnostics* 3, 26–31
64 Wurmbach, E. *et al.* (2001) Gonadotropin-releasing hormone R-coupled gene network organization. *J. Biol. Chem.* 276, 47195–47201