Review

Statistical tests for differential expression in cDNA microarray experiments

Xiangqin Cui and Gary A Churchill

Address: The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA.

Correspondence: Gary A Churchill. E-mail: garyc@jax.org

Published: 17 March 2003

Genome ***Biology*** 2003, **4**:210

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/4/210

Abstract

Extracting biological information from microarray data requires appropriate statistical methods. The simplest statistical method for detecting differential expression is the *t* test, which can be used to compare two conditions when there is replication of samples. With more than two conditions, analysis of variance (ANOVA) can be used, and the mixed ANOVA model is a general and powerful approach for microarray experiments with multiple factors and/or several sources of variation.

Gene-expression microarrays hold tremendous promise for revealing the patterns of coordinately regulated genes. Because of the large volume and intrinsic variation of the data obtained in each microarray experiment, statistical methods have been used as a way to systematically extract biological information and to assess the associated uncertainty. Here, we review some widely used methods for testing differential expression among conditions. For these purposes, we assume that the data to be used are of good quality and have been appropriately transformed (normalized) to ensure that experimentally introduced biases have been removed [1,2]. See Box 1 for a glossary of terms. For other aspects of microarray data analysis, please refer to recent reviews on experimental design [3,4] and cluster analysis [5].

Comparing two conditions

A simple microarray experiment may be carried out to detect the differences in expression between two conditions. Each condition may be represented by one or more RNA samples. Using two-color cDNA microarrays, samples can be compared directly on the same microarray or indirectly by hybridizing each sample with a common reference sample [4,6]. The null hypothesis being tested is that there is no difference in expression between the conditions; when conditions are compared directly, this implies that the true ratio between the expression of each gene in the two samples should be one. When samples are compared indirectly, the ratios between the test sample and the reference sample should not differ between the two conditions. It is often more convenient to use logarithms of the expression ratios than the ratios themselves because effects on intensity of microarray signals tend be multiplicative; for example, doubling the amount of RNA should double the signal over a wide range of absolute intensities. The logarithm transformation converts these multiplicative effects (ratios) into additive effects (differences), which are easier to model; the log ratio when there is no difference between conditions should thus be zero. If a single-color expression assay is used - such as the Affymetrix system [7] - we are again considering a null hypothesis of no expression-level difference between the two conditions, and the methods described in this article can also be applied directly to this type of experiment.

A distinction should be made between RNA samples obtained from independent biological sources - biological replicates - and those that represent repeated sampling of the same biological material - technical replicates. Ideally, each condition should be represented by multiple independent biological samples in order to conduct statistical tests.

Box 1

Glossary

Analysis of variance (ANOVA): a procedure for constructing statistical tests by partitioning the total variance into different sources.

Biological replicates: biological samples obtained in replicate from independent sources representing the same condition, such as liver tissue from individual mice of the same sex and strain.

Bonferroni correction: a multiple-testing adjustment in which the nominal significance level is divided by the total number of tests.

Broad-sense inference: an inference that applies to the entire population from which biological samples were obtained.

Decomposition: separation of a complex variance term in an ANOVA model into its components. For example, in an experiment that varies sex and treatment, the total variance in the data can be decomposed into components attributable to sex, treatment, interaction, and error.

Degrees of freedom: the number of levels that can vary freely in a term of an ANOVA model. It is typically one less than the number of levels in the factor. For example, the factor sex has two levels, female (F) and male (M). The effects attributed to these levels are deviations from an overall mean and so are constrained to sum to zero. If the effect of F is +1 the effect of M must be -1, and thus there is only one degree of freedom associated with this two-level factor.

Error variance: the variation associated with an estimated quantity. It is the square of the standard error and is commonly used to assess the accuracy of estimation.

False negative rate: the proportion of type II errors among the null hypotheses that were not rejected in multiple testing.

False positive rate: the proportion of type I errors among the rejected null hypotheses in multiple testing.

Fixed effect: a term in an ANOVA model for which the levels are going to be repeated exactly if the experiment is repeated. For example, the factor sex has two levels (F and M) and the same levels will occur again in a replication of the experiment. We are generally interested in the mean values associated with levels of a fixed effect.

Fixed-effects ANOVA: an ANOVA model in which all terms except the residual term are fixed effects. In a fixed-effects model there is only one source of random variation.

Fold change: the ratio of RNA quantities between two samples in a microarray experiment, often estimated by the ratio of fluorescent signal intensities.

Log ratio: logarithm of the fold change.

Mixed-model ANOVA: an ANOVA model in which some terms are treated as random effects and others as fixed effects. In a mixed model there may be multiple sources of random variation.

Narrow-sense inference: an inference that applies only to the biological samples used in the experiment.

Nominal significance level/*p*-value: a significance level/*p*-value to which no multiple-testing adjustment has been applied.

Normalization: the process of removing certain systematic biases from microarray data.

Null hypothesis: a hypothesis for which the effects of interest are assumed to be absent. Commonly used as a basis for constructing statistical tests.

Permutation analysis: a method of simulating data that satisfy a null hypothesis by shuffling the observed data.

Power: the probability that a real effect can be identified by a statistical test. It is one minus the type II error probability.

***p*-value:** a measure of the evidence against the null hypothesis in a statistical test. It is the probability of the occurrence of a test statistic equal to, or more extreme than, the observed value under the assumption that the null hypothesis is true.

Random effect: a term in ANOVA model for which the levels represent a sample from a population of levels. In a replicated experiment the same values will not repeat. For example, the effect of a spot on a microarray slide will vary in repeated experiments because the exact size of spots varies at random. We are generally interested in the variability associated with a random effect.

Residual: the difference between an observed data value and its expectation as predicted by a model. It is the lowest-level term in an ANOVA model and is often denoted as ϵ_i.

Box 1 *(continued)*

Residual sums of squares: the sum of all the residuals squared. It is a measure of the total discrepancy between a model and the observed data.

Restricted maximum likelihood: a numerical method for estimating variance components in a mixed ANOVA model [40,41].

Significance level: the size of a *p*-value that is regarded as providing sufficient evidence against a null hypothesis. If the *p*-value falls below the significance level, the null hypothesis is rejected.

Technical replicates: multiple RNA samples obtained from the same biological source.

Type I error: the event of rejecting a null hypothesis when it is true.

Type II error: the event of failing to reject a null hypothesis when it is false.

Definitions are from [44-46].

If only technical replicates are available, statistical testing is still possible but the scope of any conclusions drawn may be limited [3]. If both technical and biological replicates are available, for example if the same biological samples are measured twice each using a dye-swap assay, the individual log ratios of the technical replicates can be averaged to yield a single measurement for each biological unit in the experiment. Callow *et al.* [8] describe an example of a biologically replicated two-sample comparison, and our group [9] provide an example with technical replication. More complicated settings that involve multiple layers of replication can be handled using the mixed-model analysis of variance techniques described below.

'Fold' change

The simplest method for identifying differentially expressed genes is to evaluate the log ratio between two conditions (or the average of ratios when there are replicates) and consider all genes that differ by more than an arbitrary cut-off value to be differentially expressed [10-12]. For example, if the cut-off value chosen is a two-fold difference, genes are taken to be differentially expressed if the expression under one condition is over two-fold greater or less than that under the other condition. This test, sometimes called 'fold' change, is not a statistical test, and there is no associated value that can indicate the level of confidence in the designation of genes as differentially expressed or not differentially expressed. The fold-change method is subject to bias if the data have not been properly normalized. For example, an excess of low-intensity genes may be identified as being differentially expressed because their fold-change values have a larger variance than the fold-change values of high-intensity genes [13,14]. Intensity-specific thresholds have been proposed as a remedy for this problem [15].

The *t* test

The *t* test is a simple, statistically based method for detecting differentially expressed genes (see Box 2 for details of how it is calculated). In replicated experiments, the error variance (see Box 1) can be estimated for each gene from the log ratios, and a standard *t* test can be conducted for each gene [8]; the resulting *t* statistic can be used to determine which genes are significantly differentially expressed (see below). This gene-specific *t* test is not affected by heterogeneity in variance across genes because it only uses information from one gene at a time. It may, however, have low power because the sample size - the number of RNA samples measured for each condition - is small. In addition, the variances estimated from each gene are not stable: for example, if the estimated variance for one gene is small, by chance, the *t* value can be large even when the corresponding fold change is small. It is possible to compute a global *t* test, using an estimate of error variance that is pooled across all genes, if it is assumed that the variance is homogeneous between different genes [16,17]. This is effectively a fold-change test because the global *t* test ranks genes in an order that is the same as fold change; that is, it does not adjust for individual gene variability. It may therefore suffer from the same biases as a fold-change test if the error variance is not truly constant for all genes.

Modifications of the *t* test

As noted above, the error variance (the square root of which gives the denominator of the *t* tests) is hard to estimate and subject to erratic fluctuations when sample sizes are small. More stable estimates can be obtained by combining data across all genes, but these are subject to bias when the assumption of homogeneous variance is violated. Modified versions of the *t* test (Box 2) find a middle ground that is both powerful and less subject to bias.

In the 'significance analysis of microarrays' (SAM) version of the *t* test (known as the *S* test) [18], a small positive constant is added to the denominator of the gene-specific *t* test. With this modification, genes with small fold changes will not be selected as significant; this removes the problem of stability mentioned above. The regularized *t* test [19] combines information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator for a gene-specific *t* test. The *B* statistic proposed by Lonnstedt and Speed [20] is a log posterior odds ratio of differential expression versus non-differential expression; it allows for gene-specific variances but it also

combines information across many genes and thus should be more stable than the *t* statistic (see Box 2 for details).

The *t* and *B* tests based on log ratios can be found in the Statistics for Microarray Analysis (SMA) package [21]; the *S* test is available in the SAM software package [22]; and the regularized *t* test is in the Cyber T package [23]. In addition, the Bioconductor [24] has a collection of various analysis tools for microarray experiments. Additional modifications of the *t* test are discussed by Pan [25].

Graphical summaries (the 'volcano plot')

The 'volcano plot' is an effective and easy-to-interpret graph that summarizes both fold-change and *t*-test criteria (see Figure 1). It is a scatter-plot of the negative $\log_{10}$-transformed *p*-values from the gene-specific *t* test (calculated as described in the next section) against the $\log_2$ fold change (Figure 1a). Genes with statistically significant differential expression according to the gene-specific *t* test will lie above a horizontal threshold line. Genes with large fold-change values will lie outside a pair of vertical threshold lines. The significant genes identified by the *S*, *B*, and regularized *t* tests will tend to be located in the upper left or upper right parts of the plot.

Significance and multiple testing

Nominal *p*-values

After a test statistic is computed, it is convenient to convert it to a *p*-value. Genes with *p*-values falling below a prescribed level (the 'nominal level') may be regarded as significant. Reporting *p*-values as a measure of evidence allows some flexibility in the interpretation of a statistical test by providing more information than a simple dichotomy of 'significant' or 'not significant' at a predefined level. Standard methods for computing *p*-values are by reference to a statistical distribution table or by permutation analysis. Tabulated *p*-values can be obtained for standard test statistics (such as the *t* test), but they often rely on the assumption that the errors in the data are normally distributed. Permutation analysis involves shuffling the data and does not require such assumptions. If permutation analysis is to be used, the experiment must be large enough that a sufficient number of distinct shuffles can be obtained. Ideally, the labels that identify which condition is represented by each sample are shuffled to simulate data from the null distribution. A minimum of about six replicates per condition (yielding a total of 924 distinct permutations) is recommended for a two-sample comparison. With multiple conditions, fewer replicates are required. If the experiment is too small, permutation analysis can be conducted by shuffling residual values across genes (see Box 1), under the assumption of homogeneous variance [6,25].

When we conduct a single hypothesis test, we may commit one of two types of errors. A type I or false-positive error occurs when we declare a gene to be differentially expressed when in fact it is not. A type II or false-negative error occurs when we fail to detect a differentially expressed gene. A statistical test is usually constructed to control the type I error probability, and we achieve a certain power (which is equal to one minus the type II error probability) that depends on the study design, sample size, and precision of the measurements. In a microarray experiment, we may conduct thousands of statistical tests, one for each gene, and a substantial number of false positives may accumulate. The following are some of the methods available to address this problem, which is called the problem of multiple testing.

Family-wise error-rate control

One approach to multiple testing is to control the family-wise error rate (FWER), which is the probability of accumulating one or more false-positive errors over a number of statistical tests. This is achieved by increasing the stringency that we apply to each individual test. In a list of differentially expressed genes that satisfy an FWER criterion, we can have high confidence that there will be no errors in the entire list. The simplest FWER procedure is the Bonferroni correction: the nominal significance level is divided by the number of tests. The permutation-based one-step correction [26] and the Westfall and Young step-down adjustment [27] provide FWER control and are generally more powerful but more computationally demanding than the Bonferroni procedure. FWER criteria are very stringent, and they may substantially decrease power when the number of tests is large.

False-discovery-rate control

An alternative approach to multiple testing considers the false-discovery rate (FDR), which is the proportion of false positives among all of the genes initially identified as being differentially expressed - that is, among all the rejected null hypotheses [28,29]. An arguably more appropriate variation, the positive false-discovery rate (pFDR) was proposed by Storey [30]. It multiplies the FDR by a factor of π_0, which is the estimated proportion of non-differentially expressed genes among all genes. Because π_0 is between 0 and 1, the estimated pFDR is smaller than the FDR. The FDR is typically computed [31] after a list of differentially expressed genes has been generated. Software for computing FDR and related quantities can be found at [32,33]. Unlike a significance level, which is determined before looking at the data, FDR is a post-data measure of confidence. It uses information available in the data to estimate the proportion of false positive results that have occurred. In a list of differentially expressed genes that satisfies an FDR criterion, one can expect that a known proportion of these will represent false positive results. FDR criteria allow a higher rate of false positive results and thus can achieve more power than FWER procedures.

More than two conditions

Relative expression values

When there are more than two conditions in an experiment, we cannot simply compute ratios; a more general concept of

Box 2

Tests for comparing two conditions

In this box, we define the *t* test and some of its variations. Let R_g be the mean log ratio of the expression levels of one gene and SE_g be its standard error. Let *SE* be the standard error computed by combining data across all genes.

- The global *t*-test statistic is $t = \frac{R_g}{SE}$ and the gene-specific *t*-test statistic is $t = \frac{R_g}{SE_g}$.
- The SAM ('significance analysis of microarrays') test statistic is $S = \frac{R_g}{c + SE_g}$, where the constant *c* can be taken to be the 90th percentile SE_g value [47].
- The regularized *t*-test statistic is $t = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0 + n - 2}}}$

where v_0 is a tunable parameter that determines the relative contributions of gene-specific and global variances and *n* is the number of replicate measurements for each condition [19].

- The *B* statistic requires a somewhat more detailed description than we are able to provide here, but it is spelled out by Lonnstedt and Speed [20]. Essentially it is the logarithm of a ratio of probabilities. The numerator is the probability that a gene is differentially expressed and the denominator is the probability that the gene is not differentially expressed. Both probabilities are estimated in light of the entire data and are called posterior probabilities; thus, the *B* statistic is a logarithm of the posterior odds of differential expression.

The *t* and *B* tests based on log ratios can be found in the Statistics for Microarray Analysis (SMA) package [21]; the *S* test is available in the SAM software package [22]; and the regularized *t* test is in the Cyber T package [23]. In addition, the Bioconductor [24] has a collection of various analysis tools for microarray experiments.

relative expression is needed. One approach that can be applied to cDNA microarray data from any experimental design is to use an analysis of variance (ANOVA) model (Box 3a) to obtain estimates of the relative expression (*VG*) for each gene in each sample [6,34]. In the microarray ANOVA model, the expression level of a gene in a given sample is computed relative to the weighted average expression of that gene over all samples in the experiment (see Box 3a for statistical details). We note that the microarray ANOVA model is not based on ratios but is applied directly to intensity data; the difference between two relative expression values can be interpreted as the mean log ratio for comparing two samples (as $\log A - \log B = \log(A/B)$, where $\log A$ and $\log B$ are two relative expression values). Alternatively, if each sample is compared with a common reference sample, one can use normalized ratios directly. This is an intuitive but less efficient approach to obtaining relative expression values than using the ANOVA estimates. Direct estimates of relative expression can also be obtained from single-color expression assays [35,36].

The set of estimated relative expression values, one for each gene in each RNA sample, is a derived data set that can be subjected to a second level of analysis. There should be one relative expression value for each gene in each independent sample. The distinction between technical replication and biological replication should be kept in mind when interpreting results from the analysis of a derived data. If inference is being made on the basis of biological replicates and there is also technical replication in the experiment, the technical replicates should be averaged to yield a single value for each independent biological unit. The derived data can be analyzed on a gene-by-gene basis using standard ANOVA methods to test for differences among conditions. For example, our group [37] have used a derived data set to test for expression differences between natural populations of fish.

Three flavors of *F* test

The classical ANOVA *F* test is a generalization of the *t* test that allows for the comparison of more than two samples (Box 3). The *F* test is designed to detect any pattern of differential expression among several conditions by comparing the variation among replicated samples within and between conditions. As with the *t* test, there are several variations on the *F* test (Box 3b). The gene-specific *F* test (*F1*), a generalization of

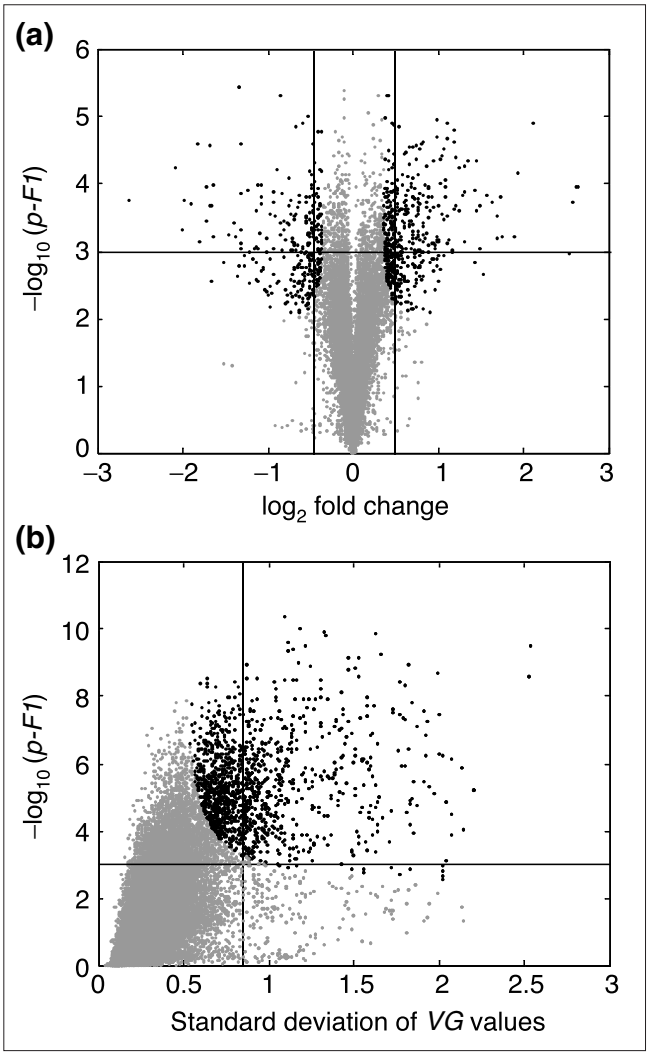

Figure 1
Volcano plots. The negative $\log_{10}$- transformed *p*-values of the *F1* test (see Box 3b) are plotted against **(a)** the log ratios ($\log_2$ fold change) in a two-sample experiment or **(b)** the standard deviations of the variety-by-gene *VG* values (see Box 3a) in a four-sample experiment. The horizontal bars in each plot represent the nominal significant level 0.001 for the *F1* test under the assumption that each gene has a unique variance. The vertical bars represent the one-step family-wise corrected significance level 0.01 for the *F3* test (see Box 3b) under the assumption of constant variance across all genes. Black points represent the significant genes selected by the *F2* test with a compromise of these two variance assumptions.

the gene-specific *t* test, is the usual *F* test and it is computed on a gene-by-gene basis. As with *t* tests, we can also assume a common error variance for all genes and thus arrive at the global variance *F* test (*F3*). A middle ground is achieved by the *F2* test, analogous to the regularized *t* test; this uses a weighted combination of global and gene-specific variance estimates in the denominator. Nominal *p*-values can be obtained for the *F* test, from standard tables, but the *F2* and *F3* statistics do not follow the tabulated *F* distribution and critical values should be established by permutation analysis. Among these tests, the *F3* test is the most powerful, but it is also subject to the same potential biases as the fold-change test. In our experience, *F2* has power comparable to *F3* but it has a lower FDR than either *F1* or *F3*. It is possible to derive a version of the *B* statistic [20] for the case of multiple conditions. This could provide an alternative approach to combine variance estimates across genes in the context of multiple samples. Any of these tests can be applied to a derived data set of relative expression values to make comparisons among two or more conditions.

The results of all three *F* statistics can be summarized simultaneously using a volcano plot, but with a slight twist when there are more than two samples. The standard deviation of the relative expression values is plotted on the *x* axis instead of plotting log fold change; the resulting volcano plot (Figure 1b) is similar to the right-hand half of a standard volcano plot (Figure 1a).

The fixed-effects ANOVA model

The process of creating a derived data set and computing the *F* tests described above can be integrated in one step by applying [20,35] our fixed-effects ANOVA model [9]; further discussion is provided Lee *et al.* [34]. The fixed-effects model assumes independence among all observations and only one source of random variation. Depending on the experimental design, this source of variation could be technical, as in our study [9], or biological if applied to data as was done by Callow *et al.* [8]. Although it is applicable to many microarray experiments, the fixed-effects model does not allow for multiple sources of variation, nor does it account for correlation among the observations that arise as a consequence of different layers of variation. Test statistics from the fixed-effects model are constructed using the lowest level of variation in the experiment: if a design includes both biological and technical replication, tests are based on the technical variance component. If there are replicated spots on the microarrays, the lowest level of variance will be the within-array measurement error. This is rarely appropriate for testing, and the statistical significance of results using within-array error may be artificially inflated. To avoid this problem, replicated spots from the same array can be 'collapsed' by taking the sum or average of their raw intensities. This does not fully utilize the available information, however, and we recommend application of the mixed-effects ANOVA model, described below.

Multiple-factor experiments

In a complex microarray experiment, the set of conditions may have some structure. For example, Jin *et al.* [38] consider eight conditions in a 2 by 2 by 2 factorial design with the factors sex, age, and genotype. There is no biological replication here, but information about biological variance is available because of the factorial design. In other experiments, both biological and technical replicates are included.

Box 3

(a) The microarray analysis of variance model

An analysis of variance (ANOVA) model for microarray data can be specified in two stages. The first stage is the normalization model

$$y_{ijgr} = \mu + A_i + D_j + AD_{ij} + r_{ijgr}.$$

where y_{ijgr} is the logarithm of one signal intensity. The indices track array (*i*), dye (*j*), gene (*g*) and measurement (*r*); μ is the overall mean expression level; *A* is the effect of the array on the measured intensity; *D* is the effect of the dye on the measured intensity; and *AD* is a term accounting for effects of the interaction between the array and the dye. Note that if you have Affymetrix data, the normalization model will be different [36]. The first stage generates the term r_{ijgr} from the measured intensities, and in the second stage, gene-specific effects are modeled in terms of the residuals of the normalization mode. The gene-specific model:

$$r_{ijgr} = G + VG_{ij} + DG_j + AG_i + \epsilon_{ijr}.$$

is applied to the data one gene at a time; the subscript *g* is therefore dropped. In this model *G* is the average intensity associated with a particular gene; AG_i is the effect of the array on that gene DG_j is the effect of the dye on that gene; and ϵ_{ijr} is the residual (see Box 1). The variety-by-gene term *VG* is the term that is of primary interest in our analysis; it captures variations in the expression levels of a gene across samples. We note that VG_{ij} is a 'catch-all' term for the effects associated with the samples. In the simplest case, it is an indicator of the condition represented by the sample on array *i* with dye *j*. In more complex experiments, the design structure at the biological sample level is captured in the *VG* terms. For example, in the Jin *et al.* [38] experiment, where a 2 by 2 by 2 factorial design was used to investigate the effects of age, sex and genotype on RNA expression in *Drosophila*, the *VG* term captures the fixed effects of age, sex and genotype plus a sex by genotype interaction. If there are duplicated spots within an array, additional terms for spot and labeling effects should be included in the model. This two-stage specification of the model was proposed by Wolfinger *et al.* [48] and, when all of the effects are fixed, it is equivalent to our model [49]. The gene-specific model can be modified for one-color data (Affymetrix data) by removing the *DG* and *AG* terms. There is no dye factor and the array effects become part of the residual error term.

(b) Three *F* tests for the fixed-effects ANOVA

Hypothesis testing involves the comparison of two models. In this setting we consider a null model (or null hypothesis) of no differential expression (so that all *VG* values are equal to zero) and an alternative model with differential expression among the conditions (some *VG* values are not equal to zero). *F* statistics are computed on a gene-by-gene basis from the residual sums of squares from fitting each of these models. Thus

$$F1 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1/df_1}$$

$$F2 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\sigma^2_{pool}}$$

$$F1 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{(rss_1/df_1 + \sigma^2_{pool})/2},$$

where rss_0, df_0 and rss_1, df_1, are the residual sums of squares and degrees of freedom for the null and alternative models, respectively (see Box 1). The ratio of *rss1*/*df1* is equal to $2n \cdot SE^2_g$ in Box 2 and σ^2_{pool} is the common error variance pooled across all genes, equal to SE^2 in Box 2 [26].

Box 3 *(continued)*

(c) The mixed ANOVA model

The mixed model has the same structure as the fixed-effects model above; the difference is in the interpretation of terms that are treated as random effects. Typically, the *AG* term will be modeled as a random effect and is assumed to have a normal distribution with a mean of zero ($N(0,\sigma_A^2)$). Additional terms may be required to account for random effects associated with duplicate spotting of clones. In multiple-factor experiments, the *VG* terms may be decomposed into both random and fixed effects. Biological replication should be treated as a random effect. The details will vary according to the particular experiment and it is advisable to work in collaboration with a statistician if there is any doubt about the formulation of an appropriate model. Details for constructing the usual (gene-specific) *F* test can be found in Littell *et al.* [41]. The three variations of the *F tests* can also be computed for mixed-model ANOVA and are implemented in our MAANOVA software package [42].

For example, we [37] considered samples of five fish from each of three populations, and each fish was assayed on two microarrays with duplicated spots. In this study, the conditions of interest are the populations from which the fish were sampled; the fish are biological replicates, and there are two nested levels of technical replication, arrays and spots within arrays. To use fully the information available in experiments with multiple factors and multiple layers of sampling, we require a sophisticated statistical modeling approach.

The mixed-model ANOVA

The mixed model treats some of the factors in an experimental design as random samples from a population. In other words, we assume that if the experiment were to be repeated, the same effects would not be exactly reproduced but that similar effects would be drawn from a hypothetical population of effects. We therefore model these factors as sources of variance.

In a mixed model for two-color microarrays (Box 3c), the gene-specific array effect (*AG* in Box 3a) is treated as a random factor. This captures an important component of technical variation. If the same clone is printed multiple times on each array we should include additional random factors for spot (*S*) and labeling (*L*) effects. Consider an array with duplicate spots of each clone. Four measurements are obtained for each clone, two in the red channel and two in the green channel. Measurements obtained on the same spot (one red and one green) will be correlated because they share common variation in the spot size. Measurement obtained in the same color (both red or both green) will be correlated because they share variation through a common labeling reaction. Failure to account for these correlations can result in underestimation of technical variance and inflated assessments of statistical significance.

In experiments with multiple factors, the *VG* term in the ANOVA model is expanded to have a structure that reflects the experimental design at the level of the biological replicates, that is, independent biological samples obtained from the same conditions such as two mice of the same sex and strain. This may include both fixed and random components. Biological replicates should be treated as a random factor and will be included in the error variance of any tests that make comparisons among conditions. This provides a broad-sense inference (see Box 1) that applies to the biological population from which replicate samples were obtained [3,39].

Constructing tests with the mixed-model ANOVA

The components of variation attributable to each random factor in a mixed model can be estimated by any of several methods [39], of which restricted maximum likelihood (see Box 1) is the most widely used. The presence of random effects in a model can influence the estimation of other effects, including the relative expression values; these will tend to 'shrink' toward zero slightly. This effectively reduces the bias in the extremes of estimated relative expression values.

In the fixed-effects ANOVA model, there is only one variance term and all factors in the model are tested against this variance. In mixed-model ANOVA, there are multiple levels of variance (biological, array, spot, and residual), and the question becomes which level we should use for the testing. The answer depends on what type of inference scope is of interest. If the interest is restricted to the specific materials and procedures used in the experiment, a narrow-sense inference, which applies only to the biological samples used in the experiment, can be made using technical variance. In most instances, however, we will be interested in a broader sense of inference that includes the biological population from which our material was sampled. In this case, all relevant sources of variance should be considered in the test [40]. Constructing an appropriate test statistic using the mixed model can be tricky [41] and falls outside the scope of the present discussion, but software tools are available that can be applied to compute appropriate *F* statistics, such as MAANOVA [42] and SAS [43]. Variations analogous to the *F2* and *F3* statistics are available in the MAANOVA software package [42].

In conclusion, fold change is the simplest method for detecting differential expression, but the arbitrary nature of the cut-off value, the lack of statistical confidence measures, and the potential for biased conclusions all detract from its appeal. The *t* test based on log ratios and variations thereof provide a rigorous statistical framework for comparing two conditions and require replication of samples within each condition. When there are more than two conditions to compare, a more general approach is provided by the application of ANOVA *F* tests. These may be computed from derived sets of estimated relative expression values or directly through the application of a fixed-effects ANOVA model. The mixed ANOVA model provides a general and powerful approach to allow full utilization of the information available in microarray experiments with multiple factors and/or a hierarchy of sources of variation. Modifications of both *t* tests and *F* tests are available to address the problems of gene-to-gene variance heterogeneity and small sample size.

References

1. Cui X, Churchill GA: **Data transformation for cDNA microarray data.** 2002. [http://www.jax.org/staff/churchill/labsite/research/expression/Cui-Transform.pdf]
2. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2:**418-427.
3. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32 Suppl:**490-495.
4. Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3:**579-588.
5. Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brown PO: **Clustering methods for the analysis of DNA microarray data.** *Stanford, Tech report* 1999. [http://www-stat.stanford.edu/~tibs/research.html]
6. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7:**819-837.
7. **Affymetrix** [http://www.affymetrix.com]
8. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10:**2022-2029.
9. Kerr M, Afshari C, Bennett L, Bushel P, Martinez J, Walker N, Churchill G: **Statistical analysis of a gene expression microarray experiment with replication.** *Statistica Sinica* 2000, **12:**203.
10. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93:**10614-10619.
11. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
12. Draghici S: **Statistical intelligence: effective analysis of high-density microarray data.** *Drug Discov Today* 2002, **7:**S55-S63.
13. Rocke DM, Durbin B: **A model for measurement error for gene expression arrays.** *J Comput Biol* 2001, **8:**557-569.
14. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52.
15. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, *et al.*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3:** research0062.1-0062.12.
16. Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, *et al.*: **Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray.** *Proc Natl Acad Sci USA* 2000, **97:**9127-9132.
17. Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW: **Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor.** *J Biol Chem* 2000, **275:**29672-29684.
18. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
19. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17:**509-519.
20. Lonnstedt I, Speed T: **Replicated microarray data.** *Statistica Sinica* 2002, **12:**31.
21. **R package: statistics for microarray analysis** [http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html]
22. **SAM: Significance Analysis of Microarrays** [http://www-stat.stanford.edu/%7Etibs/SAM]
23. **Cyber T** [http://www.igb.uci.edu/servers/cybert/]
24. **Bioconductor** [http://www.bioconductor.org]
25. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics* 2002, **18:**546-554.
26. Wu H, Kerr MK, Cui X, Churchill GA: **MAANOVA: a software package for the analysis of spotted cDNA microarray experiments.** [http://www.jax.org/staff/churchill/labsite/pubs/Wu_maanova.pdf]
27. Dudoit S, Yang Y, Matthew J, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** 2000. [http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html]
28. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc B* 1995, **57:**289-300.
29. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple tesing under dependency.** *Ann Stat* 2001, **29:**1165-1168.
30. Storey J: **A direct approach to false discovery rates.** *J R Statist Soc B* 2002, **64:**479-498.
31. Storey JD, Tibshirani R: **SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays.** 2003. [http://www.stat.berkeley.edu/~storey/papers/storey-springer.pdf]
32. **False Discovery Rate homepage** [http://www.math.tau.ac.il/~roee/index.htm]
33. **q-value** [http://www.stat.berkeley.edu/~storey/qvalue/index.html]
34. Lee ML, Lu W, Whitmore GA, Beier D: **Models for microarray gene expression data.** *J Biopharm Stat* 2002, **12:**1-19.
35. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.** *Genome Biol* 2001, **2:**research0049.1-0049.12.
36. Irizarry RA, Hobbs BG, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed T: **Exploration, normalization and summaries of high density oligonucleotide array probe level data.** 2002. [http://biosun01.biostat.jhsph.edu/~ririzarr/papers/index.html]
37. Oleksiak MF, Churchill GA, Crawford DL: **Variation in gene expression within and among natural populations.** *Nat Genet* 2002, **32:**261-266.
38. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: **The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*.** *Nat Genet* 2001, **29:**389-395.
39. McLean RA, Sanders WL, Stroup WW: **A unified approach to mixed linear models.** *Am Stat* 1991, **45:**54-64.
40. Searle SR, Casella G, McCulloch CE: *Variance Components.* New York, NY: John Wiley and Sons, Inc.; 1992.
41. Littell RC, Milliken GA, Stroup WW, Wolfinger RD: *SAS system for mixed models.* Cary, NC: SAS Institute Inc.; 1996.
42. **R/maanova** [http://www.jax.org/staff/churchill/labsite/software/anova/rmaanova]
43. **SAS microarray solution** [http://www.sas.com/industry/pharma/mas.html]
44. **Statistics glossary** [http://www.statsoftinc.com/textbook/glosfra.html]
45. **Glossary** [http://www.csse.monash.edu.au/~lloyd/tildeMML/Glossary/]
46. **Internet glossary of statistical terms** [http://www.animatedsoftware.com/statglos/statglos.htm]

47. Efron B, Tibshirani R, Goss V, Chu G: **Microarrays and their use in a comparative experiment.** 2000. [http://www-stat.stanford.edu/~tibs/research.html]
48. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models.** *J Comput Biol* 2001, **8:**625-637.
49. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression microarray data.** *Genet Res* 2001, **77:**123-128.