JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 6, 2001
Mary Ann Liebert, Inc.
Pp. 625–637

Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models

RUSSELL D. WOLFINGER,[1] GREG GIBSON,[2] ELIZABETH D. WOLFINGER,[3]
LEE BENNETT,[4] HISHAM HAMADEH,[4] PIERRE BUSHEL,[4]
CYNTHIA AFSHARI,[4] and RICHARD S. PAULES[4]

ABSTRACT

The determination of a list of differentially expressed genes is a basic objective in many cDNA microarray experiments. We present a statistical approach that allows direct control over the percentage of false positives in such a list and, under certain reasonable assumptions, improves on existing methods with respect to the percentage of false negatives. The method accommodates a wide variety of experimental designs and can simultaneously assess significant differences between multiple types of biological samples. Two interconnected mixed linear models are central to the method and provide a flexible means to properly account for variability both across and within genes. The mixed model also provides a convenient framework for evaluating the statistical power of any particular experimental design and thus enables a researcher to a priori select an appropriate number of replicates. We also suggest some basic graphics for visualizing lists of significant genes. Analyses of published experiments studying human cancer and yeast cells illustrate the results.

Key words: ANOVA, cDNA microarray, gene expression, mixed models, statistical significance.

INTRODUCTION

MICROARRAYS ARE BECOMING INCREASINGLY MORE COMMON LABORATORY TOOLS for studying simultaneous changes in expression across a large number of genes. Image data from the arrays lead to gene-specific numerical intensities representing relative expression levels, and these in turn form the input to computational analysis designed to assess significance and relationships across biological samples. This paper presents a general statistical method for analyzing these intensity measurements derived from a potentially large and diverse number of microarray experiments. Our primary goal is to statistically infer significant expression differences in a way that optimally controls both false positives (genes declared to be differentially expressed which in reality are not) and false negatives (genes truly differentially expressed but not declared as such).

[1]Director of Genomics, SAS Institute Inc., Cary, NC 27513.
[2]Department of Genetics, North Carolina State University, Raleigh, NC 27695.
[3]Department of Biology, Meredith College, Raleigh, NC 27607.
[4]National Institute of Environmental Health Sciences, Microarray Center, Research Triangle Park, NC 27709.

The importance of assessing statistical significance cannot be overstressed. Simple rules that eliminate genes with less than two- or three-fold expression changes completely miss very biologically important genes that have a small fold change, but which are highly significant statistically because they can be measured with high precision as a result of replication. Conversely, many genes that have a large fold change in one array may also exhibit high variability across multiple arrays and thus possess little to no statistical significance. Proper determination of significance prevents researchers from "chasing noise" and helps them appropriately distinguish between important biological changes and chance variation (Wittes and Friedman, 1999).

Recent work of Tanaka *et al.* (2000) illustrates the danger of false positives and false negatives when looking strictly at fold change. They investigated placenta and embryo differences in 15K genes measured in triplicate on microarrays producing one observation per spot. Their Fig. 2 depicts how both kinds of errors can be committed when raw fold changes are used as the sole criterion. Unfortunately, the replication and basic t-test they used to assess significance is seldom seen in the cDNA microarray literature. The methods we present provide a direct generalization of simple t-tests to more complex cDNA data sets including those arising from flipped and multiple dye experiments and those designed to detect interactions between treatments.

Our proposed methodology is complementary to various popular clustering methods (refer to Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Claverie, 1999; Hastie *et al.*, 2000; Dudoit *et al.*, 2001a; Lazzeroni and Owen, 2000; Kerr and Churchill, 2001a). Investigators can use it as a precursor to clustering to make sure the inputs are statistically meaningful, or it can be used after clustering to explore and validate implied associations. There is also the interesting possibility of clustering on significance measures such as negative log p-values and on measures of variability instead of on normalized log expression ratios.

Replication of spot measurements either within or between arrays is essential in our approach, as significance levels for each gene are determined on the basis of distinct estimates of intragene variability. This is in contrast to the interesting single-array methods of Chen *et al.* (1997) Newton *et al.* (2000), Sapir and Churchill (2000), Hughes *et al.* (2000), and Rocke and Durbin (2001), who accommodate gene heterogeneity via different global distributional assumptions across all measurements on an array. The importance of replication has been nicely illustrated in recent work by Lee *et al.* (2000), who fit a normal linear mixture model to single-channel data from one slide. They recommend a minimum of three replicates, and although our linear models are of a different form (note "mixture" is different from "mixed"), we concur in this recommendation and describe a method for determining precisely how many replicates are required to achieve desired measures of selectivity and sensitivity.

Our methods are also related to recent ground-breaking work by Kerr *et al.* (2001) and Dudoit *et al.* (2001b). The former propose a general analysis of variance (ANOVA) model for the logs of original fluorescence measurements, never explicitly forming log ratios. The latter recommend a nonlinear smoothing algorithm for the normalization of log ratios and then permutation-based t-statistics for testing the significance of each gene, the p-values for which are suitably adjusted for multiplicity. In addition to its novel features, our approach incorporates and extends some of the best aspects of both of these papers, making a few simplifying assumptions and generalizing other ones. The result is a flexible, unified, and practical approach for assessing gene significance that can be implemented using commercially available software. Our hope is that researchers will adopt this approach as a basic "workhorse" method for the determination of significant genes.

RESULTS

Our approach centers around two interconnected ANOVA models, the "normalization" model and the "gene" model. Both are similar to the overall ANOVA model of Kerr *et al.* (2001) in that they model the logarithms of the original fluorescence measurements, not log ratio values. The normalization model accounts for experiment-wide systematic effects that could bias inferences made on the data from the individual genes. The residuals from this model represent normalized values and are the input data for the gene models. The gene models are fit separately to the normalized data from each gene, allowing inferences to be made using separate estimates of variability. Dudoit *et al.* (2001b) incorporate such heterogeneity in

their t-statistics and recommend this extension to gene-by-gene ANOVA models in their discussion. The two separate models provide a conceptually and computationally efficient means to analyze the data. We now present results of the approach applied to two publicly available data sets.

Yeast data

The *Saccharomyces cerevisiae swi/snf* mutation study of Sudarsanam *et al.* (2000) investigates mutants deleted for a gene encoding one conserved (Snf2) or one unconserved (Swi1) component, each in either rich or minimal media. The four experimental conditions are arrayed in triplicate, and data are available at *http://genome-www.stanford.edu/swisnf* as ScanAlyze files (Eisen *et al.*, 1998). The same wild-type strain is used as a reference sample in all twelve arrays and is labeled with Cy5 in channel 2, while the experimental strains are labeled with Cy3 in channel 1.

Let y_{gij} be the base-2 logarithm of the background-corrected measurement from gene g ($g = 1, \ldots, 6917$), treatment i ($i = 1, \ldots, 5$), and array j ($j = 1, \ldots, 12$). "Treatment" here signifies the type of cDNA sample ($snf2$-rich, $snf2$-mini, $swi1$-rich, $swi1$-mini, wild-type). The 6,917 values for g were determined from unique values of the NAME column in the ScanAlyze files, with blanks replaced by values from the TYPE column. Note we are not forming ratios, but letting the two observations for each gene on each array be indexed by treatment. This scheme assumes no replication of spots within an array, although this could easily be accommodated by adding an additional subscript.

Our normalization model for this example is

$$y_{gij} = \mu + T_i + A_j + (TA)_{ij} + \epsilon_{gij},$$

where μ represents an overall mean value, T is the main effect for treatments, A is the main effect for arrays, TA is the interaction effect of arrays and treatments, and ϵ is stochastic error. This normalization model can be viewed as a modified segment of the global ANOVA model of Kerr *et al.* (2001). Modifications include the use of base-2 logarithms instead of the natural base and the addition of the TA effect to model channels. The latter is usually necessary because of the arbitrary manual intensity scaling done with programs like ScanAlyze (Eisen *et al.*, 1998). Also, we include no main effect for dyes since wild-type was always labeled with Cy5 and therefore the treatment effect T is already accounting for differences between dyes. See the Discussion for disadvantages of this protocol and recommended alternatives.

Let r_{gij} denote the residuals from this model, computed by subtracting the fitted values for the effects from the y_{gij} values. Our gene model is then

$$r_{gij} = G_g + (GT)_{gi} + (GA)_{gj} + \gamma_{gij}.$$

All effects are indexed by g and are assumed to serve similar roles to those from the normalization model, but at the gene level. The GA term models the effects for each spot and is the same as the array by gene interaction effect in Kerr *et al.* (2001). It is crucial to the model, as it serves to account for the insidious spot-to-spot variability inherent in spotted microarray data. The inclusion of this effect allows us to extract appropriate information about the treatment effects and obviates the need to form ratios.

We make standard stochastic assumptions about the preceding linear models. In particular, the effects A_j, $(TA)_{ij}$, ϵ_{gij}, $(GA)_{gj}$, and γ_{gij} are all assumed to be normally distributed random variables with zero means and variance components σ_A^2, σ_{TA}^2, σ_ϵ^2, $\sigma_{GA_g}^2$, $\sigma_{\gamma_g}^2$, respectively. These random effects are assumed to be independent both across their indices and with each other, and note the GA and γ effects have different variances across the gene index g (heterogeneity). The remaining terms in the models are assumed to be fixed effects, and thus both models are mixed models. We exploit these standard mixed-model normality assumptions by using the method of restricted maximum likelihood (REML) to estimate the variance components (refer to Searle *et al.*, 1993, and Littell *et al.*, 1996, for detailed formulas and descriptions). REML also produces estimates of all effects in the model along with appropriate standard errors.

The estimates of primary interest are those of the $(GT)_{gi}$ effects, which measure the treatment effects for each gene. We test for differences between these effects by using mixed-model-based t-tests of all possible pairwise comparisons within a gene. Although more sophisticated methods exist (Littell *et al.*,

1996), for simplicity and ease of interpretation we set the degrees of freedom (DF) for the t-tests equal to the DF for error from the gene model. This is derived from the DF column in a standard ANOVA table:

Source	*DF*
Intercept	1
Spots	11
Treatments	4
Error	8
Total	24

The t-tests are all therefore based on eight degrees of freedom and will likely provide reasonable statistical performance. (Error degrees of freedom of three or less typically result in high statistical uncertainty and few significant differences.) Note that the ANOVA table and error degrees of freedom do not tell the entire story about the performance of a particular design. Equally critical is the replication structure within the design. For example, here the first four treatment values are replicated three times each (corresponding to the three experiments under each condition) and the fifth value (wild-type) is replicated on all twelve arrays. This design produces much more informative inferences about wild-type than the other four conditions simply because it is observed more often.

For each of the 6,917 genes (actually, open reading frames, or ORFs) on the arrays, we construct the ten hypotheses tests corresponding to all possible pairwise differences between the five sample types (*snf*2-rich, *snf*2-mini, *swi*1-rich, *swi*1-mini, wild-type). To adjust for the multiple testing problem (see Methods), we set our p-value cutoff at the Bonferroni value of $0.05/(6917 \times 10) = 1e\text{-}6.14$ to assure an experimentwise false positive rate of 0.05. Figure 1A is a significance plot of the p-values passing the stringent Bonferroni criterion. The x-axis is log2 of estimated fold change and the y-axis negative log10 of the corresponding p-value. The figure includes significant p-values for all ten tests for all genes in order to provide a general impression of the results. Of the 13 genes listed in Tables 1 and 2 of Sudarsanam *et al.* (2000), only three appear in Fig. 1A: PHO5 (not labeled, $-\log10(p)=1.6$), SAG1, and ALPHA1 (same as MATα1).

Figure 1B reveals an interesting discovery we made while analyzing these data. In this plot, the different gene models are determined by sorting the data by their actual gene name, not their ORF name as in Fig. 1A. This produces 7,031 distinct genes instead of 6,917, and results are basically identical except for two genes: ASP3 and ALPHA1. ASP3 has four subunits associated with it in the data set and is therefore

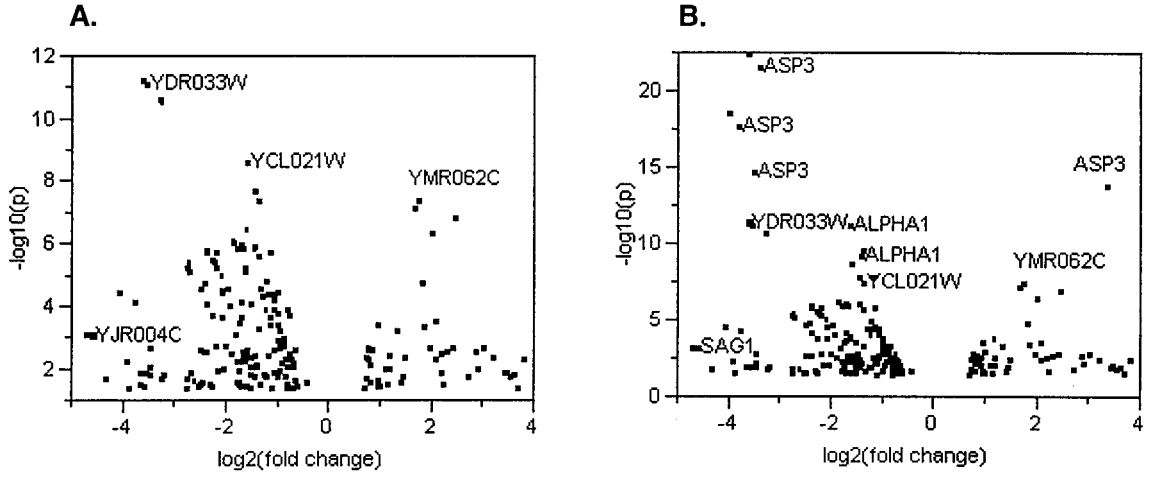

FIG. 1. Gene significance results for yeast data. **A.** Significance plot of all significant pairwise differences from the *snf/swi* yeast mutant study of Sudarsanam *et al.* (2000). A few of the points are labeled by their ORF name. **B.** Same as A, but using gene name as the grouping category instead of ORF name. The significance levels of certain tests from two genes, ASP3 and ALPHA1, increase dramatically because they represent pooled data from four and two ORFs, respectively (note the change in scale of the y-axis).

analyzed in four separate groups for the analysis pictured in Fig. 1A. However, the four groups of data are pooled for the analysis in Fig. 1B, resulting in much higher specificity and significance values for several of the tests associated with ASP3. This can be attributed both to the decrease in estimated standard errors and to the increase in degrees of freedom for the mixed-model t-tests, going from 7 in Fig. 1A to 40 in Fig. 1B. Likewise, ALPHA1 has two subunits, and the degrees of freedom change from 11 in Fig. 1A to 27 in Fig. 1B. In fact, the two subunit analyses did not pass the strict significance criterion used to create Fig. 1A and thus do not appear in the plot.

Lymphoma data

Alizadeh *et al*. (2000) conduct an extensive microarray investigation of distinct types of diffuse large B-cell lymphoma (DLBCL) and present a clustering analysis of 96 normal and malignant lymphocyte samples. ScanAlyze files (Eisen *et al*., 1998) and other supplemental material are available at *http://llmpp.nih.gov/lymphoma*. Part of this work focuses on discovery of two DLBCL subtypes, evidenced by a deep dendrogram split within the 52 separate DLBCL tissue samples studied. The subtypes are also born out by associations with additional survival and clinical risk data. An intriguing feature of the results is that the leaves for germinal center B-cells (GCB) and germinal center centroblasts (GCC) occur directly in the middle of one of the DLBCL branches (refer to their Fig. 3).

We investigated the indicated associations with GCB by testing for statistical significance between it and the first ten DLBCL samples. As controls, we also tested differences of GCB with GCC and with the reference sample prepared from a pool of mRNAs isolated from nine different lymphoma cell lines. The final data set we analyzed consists of base-2 logarithms of background-subtracted fluorescence measurements from 18 arrays, each of which has the reference sample in channel 1 labeled with Cy3 and one of (DLBCL01-DLBCL10,GCB, GCC) in channel 2 labeled with Cy5. Fortunately, two replicated arrays are available for DLBCL02, DLBCL05, DLBCL07, DLBLC09, DLBCL10, and GCB, while the others have only one array of data. After identifying 14,428 distinct clones by match-merging available text files, we fit the same normalization and gene models as used for the yeast data and tested all possible differences with GCB. Had the data been complete, this would have resulted in $14{,}428 \times 12 = 173{,}136$ significance tests, but since roughly 40% of the comparisons are not estimable, the procedure resulted in 98,527 p-values.

Figure 2A plots these p-values and illustrates the substantial difference significance testing can make versus cutoffs made strictly on the basis of fold change. The two vertical reference lines indicate 4-fold cutoffs for either repression or induction and the horizontal line at $p = 10^{-5}$ shows the cutoff for a false positive rate of 1 in 100,000 tests. These reference lines divide the plot into six meaningful sectors. Points in the lower middle sector have low significance and low fold change, and both methods agree that the corresponding changes are not significant. Likewise, points in the upper left and right sectors have high significance and high fold change, and both approaches concur on significant differential expression. The 243 points in the upper middle sector represent potential false negatives when using the simple cutoff method.

Dramatic differences lie in the lower left and right sectors of Fig. 2A. The 4,521 points in these two sectors all represent likely false positives when using a simple 4-fold cutoff rule. A simple response to this situation would be to increase the cutoff, but this still leaves a very large number of potential false positives and greatly increases the number of possible false negatives. Our contention is that the important, significantly expressed genes are those above the horizontal line, and that using this as a selection rule will result in much better sensitivity and selectivity than simple fold-change rules.

Figure 2B shows how significance is related to an estimate of intragene variability. Note how the variability estimates differ by several orders of magnitude and that genes with the smallest amount of variation are not necessarily the most significant. The heterogeneity of the variances explains why a single cutoff value will not work well for all genes; for some it will be too large while simultaneously too small for others. Our approach effectively determines a unique cutoff for each gene based upon the amount of variability it displays.

Figure 2C illustrates that the estimates of log2 fold change from our approach are not exactly the same as those obtained by simple differences of standard log2 ratios. The vertical axis represents differences of log2 ratios of the DLBCL01-DLBCL10, GCC samples and the log2 ratio values for GCB, where ratios are

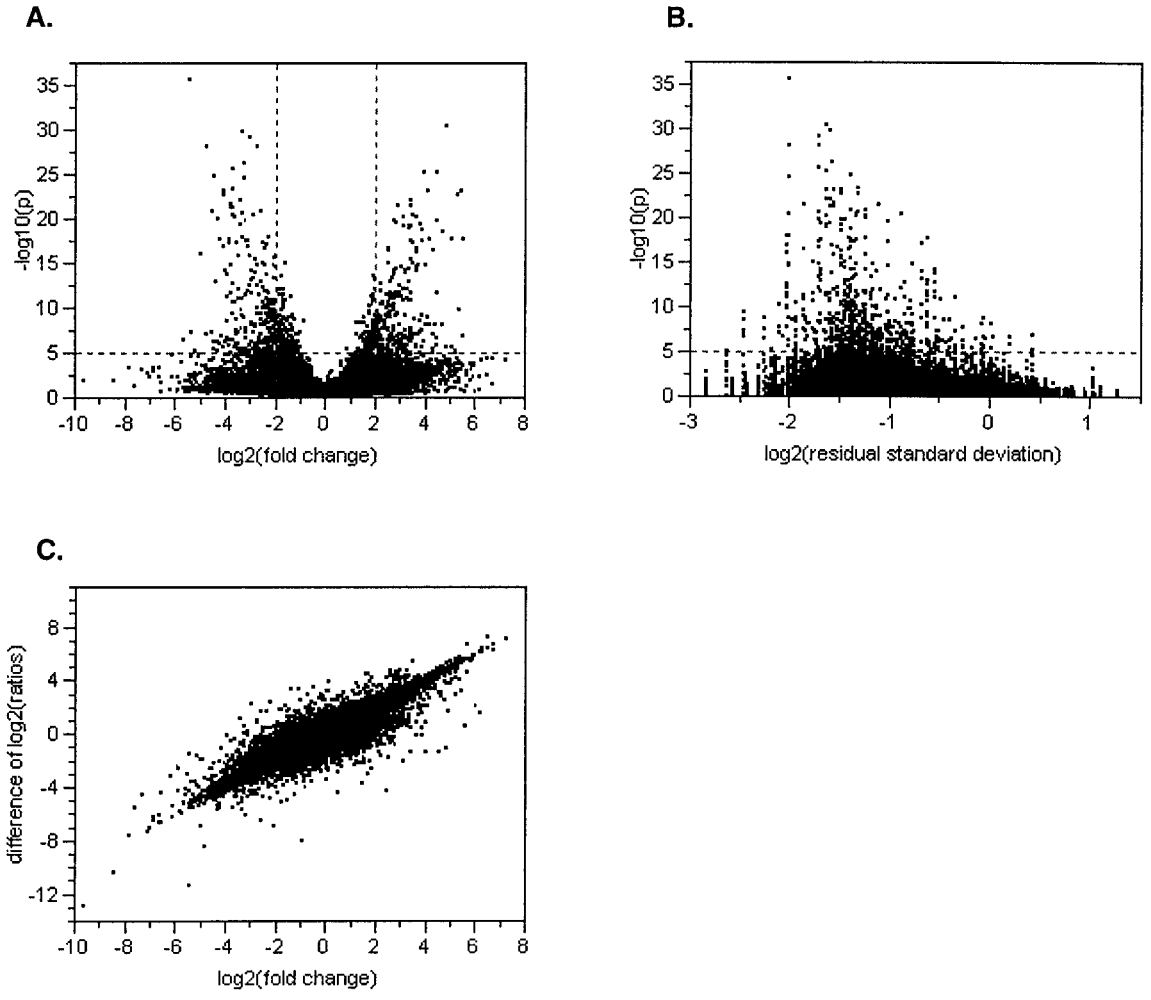

FIG. 2. Gene significance results for lymphoma data. **A.** Significance plot for a subset of the lymphoma data of Alizadeh *et al.* (2000). Plotted on the vertical axis is $-\log_{10}(p\text{-value})$ for contrasts of the germinal center B-cell line with cell lines for ten diffuse large B-cell lymphoma samples, germinal center centroblasts, and the reference sample. Horizontal reference line at 5 represents a false positive rate of 1 in 100,000 tests. The horizontal axis is $\log_2$ of estimated fold change, suitably adjusted for other systematic and random effects in the experiment. Vertical reference lines represent 4-fold changes. **B.** Same as A, but horizontal axis is $\log_2$ of the estimated gene standard deviation. **C.** Horizontal axis is the same as A, and vertical axis is the estimate of fold change obtained by simple differences of $\log_2$ ratios with the reference sample.

taken in the usual fashion with respect to the reference sample on the same spot. Differences of averages are used when there is replication and data from each array are normalized to have mean zero prior to the calculations. The horizontal axis represents the corresponding quantity obtained from the mixed model approach. Although the correlation is strong (0.92), the discrepancies are due primarily to the fact that the mixed model averages across all reference sample values whereas the simple approach uses only those reference sample observations corresponding to the desired difference.

Table 1 explores the relationships of the significance tests with the dendrogram analysis of Alizadeh *et al.* (2000). Shown is a comparison between the dendrogram distance of the samples from GCB (difference in leaf position numbers) and the number of significantly different genes. A very stringent rate of 1 in 10 million is used here to strictly control the likely number of false positives. The table reveals a general agreement between the two procedures, but with some order changes for larger distances. The number and identity of the significantly different genes provides specific quantitative information to accompany the dendrogram. In all, 219 different genes exhibit significant differential expression in one or more cell lines,

Table 1. Dendrogram and Significance Comparison[a]

Sample	*Dendrogram depth*	*No of sig diff genes*
DLBCL03	7	39
DLBCL10	5	59
DLBCL01	2	32
DLBCL08	2	37
GCB	0	0
GCC	1	3
DLBCL04	2	24
DLBCL09	4	47
DLBCL05	9	73
DLBCL07	9	66
DLBCL02	9	103
DLBCL06	9	55
reference	n/a	123

[a]Comparison of dendrogram results from Figure 3 of Alizadeh *et al.* (2000) to those from mixed model significance testing. The second column is the number of splits the sample is away from GCB, and the third column lists the number of significantly different genes. To assure a strict experimentwise false positive rate of 0.05, the Bonferroni's cutoff of 1e-6.3 was used in determining significance.

and this number represents the likely portion of the whole sampled genome involved in class or subtype differences of the kind considered here.

Our analysis allows researchers to determine precisely which genes are most likely driving basic patterns observed in the dendrogram and which ones are "Cy5 herrings." For this subset of data, the most significantly repressed genes are immunoglobulin gamma 3 heavy, kappa light and j chains, early growth response protein 1, and B-lymphocyte CR2-receptor, whereas those genes showing a large but statistically nonsignificant negative log fold change include FUSE binding protein 2, OX-40, acute-phase response factor, and several with unknown functions. On the induction side, the most significant genes from our analysis are cathespin-b, natural killer cell protein 4, src-like adapter protein, and unknown UG Hs.201975 and 161905 whereas potential false positives are ferritin heavy chain, glataminase-2, Exodus-2, and certain osteonectin comparisons.

How many replicates?

The yeast data example illustrates how increasing the number of replications can increase the statistical power of an analysis. To demonstrate how this can be quantified in practice, we perform some prospective calculations based on the lymphoma data design. Recall that in this data set half of the 12 arrays are replicated twice. Using the technique described in Methods, we compute power for mixed-model t-tests conducted from this experimental design, as well as for larger designs that include up to 10 replications on half of the arrays. Table 2 shows the increase in ANOVA degrees of freedom as the number of replications increases.

Figure 3 displays power contour plots for the contrast comparing one of the replicated DLBCL samples with GCB, which is also replicated. Figure 3A assumes a false positive rate of 1 in 200,000, a spot variance of 1, and a residual variance of 0.25. The latter two values are greater than the average value for all of the genes, although the intraspot correlation of 0.8 they produce is the modal value. Under these conditions, even 10 replicates are not enough to detect a 2-fold change with 15% power. Dividing each of the variance components by 4 results in a substantial gain in power, as shown in Fig. 3B. In this case, nine replicates will detect a 2-fold change with 85% power. Figure 3C shows how the plot in 3B changes when the false positive rate is reduced by an order of magnitude. Here only seven replicates are needed to detect a 2-fold change with 85% power under a false positive rate of 1 in 20,000.

Figure 3D displays results from different experimental designs that are based on the same number of microarrays as the previous designs but no longer include a reference sample. In these designs, the

Table 2. Degrees of Freedom for Power Calculations[a]

	Replications				
Source	*2*	*4*	*6*	*8*	*10*
Intercept	1	1	1	1	1
Spots	17	29	41	53	65
Treatments	12	12	12	12	12
Error	6	18	30	42	54
Total	36	60	84	108	132

[a] ANOVA degrees of freedom for designs based on a subset of data from Alizadeh *et al*. (2000). Replications indicates the number of times half of the 12 arrays are repeated experimentally.

replicated samples are paired with each other in a circular fashion as described by Kerr and Churchill (2001b). Figure 3D shows the substantial increase in power that can result by using a more efficient design than the standard reference sample design. Note that even though the error degrees of freedom for these designs is only one greater than the values shown in Table 2, several more replications are included for each point and so the estimated standard errors are roughly 40% smaller. Only 4 replicates are required in this situation to detect a 2-fold change with 85% power assuming a false positive rate of 1 in 20,000.

DISCUSSION

Although most current cDNA experiments waste half of their observations on a reference sample, our linear modeling approach provides a way to properly analyze data from a much wider class of designs which make better use of resources. A nice alternative design is the aforementioned circular one described by Kerr and Churchill (2000b). Regardless of experimental design choice, we recommend that important comparisons be "connected" in the design; that is, there is an unbroken series of other samples between them that have been arrayed together. The reference sample design satisfies this criterion by using the reference sample to connect all other samples, whereas the circular design satisfies it sequentially.

A dye effect is not included in our example models because such an effect is completely confounded with treatment differences unless one employs a "flipped fluor" design. We encourage the use of such designs, as do Kerr and Churchill (2000b). They discuss this effect and mention at least one extreme case in which it was real. (We have also witnessed it in certain arrays processed at the NIEHS Microarray Center.) A significant dye effect in the gene model is disturbing in that it indicates that dyes are behaving differently at the gene level beyond that already accounted at an overall level by the normalization model. Conducting experiments in which the dyes are flipped allows one to check for such an effect and properly adjust for it when it is present.

While the preceding examples were selected to be prototypical, researchers are free to select whatever effects they deem appropriate in constructing both the normalization and gene models. Systematic effects such as pin position and grid location can be included if desired, and some effects may be intentionally omitted to avoid subtracting out true signal (Lazaridis *et al*., 2000). Also, in time-course experiments, one may wish to replace the general treatment effects T with polynomial, trigonometric, or theoretically derived basis functions to more efficiently model the data. Such flexibility places a fair degree of responsibility on the analyst to carefully formulate models appropriate for the cDNA microarray experiment under consideration.

Several reviewers expressed a concern about the statistical assumptions connecting the normalization and gene models. In particular, the residuals from the normalization model are correlated by construction, and yet they are modeled with independent errors in the gene models. While this concern is valid, we argue that it makes little to no difference in practice. Specifically, for experiments with no missing data, the effects fitted by the normalization and gene models are orthogonal, and so results from the two-model approach should be equivalent or at least very similar to those from one large ANOVA model like that

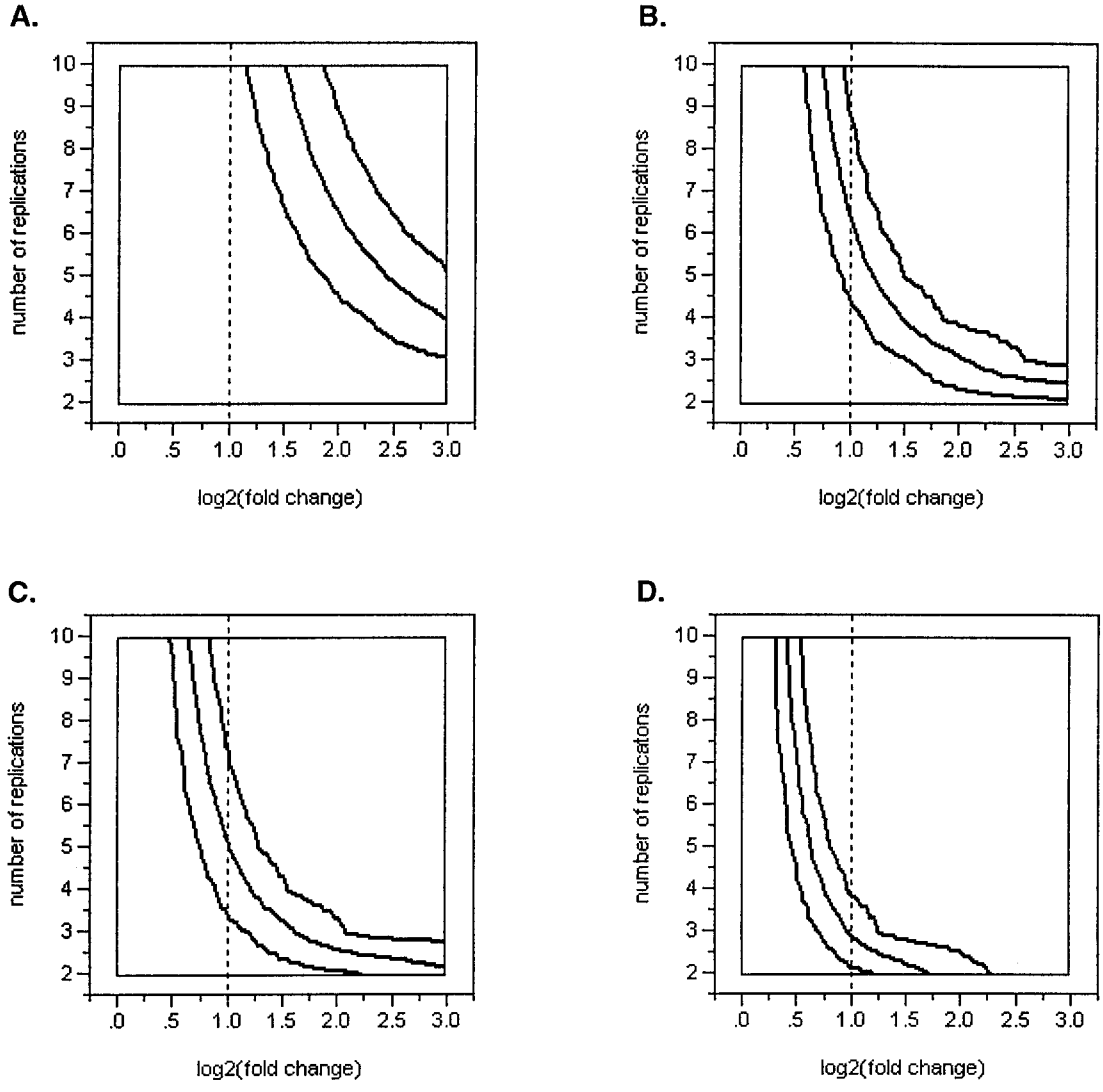

FIG. 3. Statistical power plots. Contour plots of statistical power (1 minus the false negative rate) for experimental designs based on the lymphoma example (see the text for description). The three curves in each plot represent powers of 0.85 (top), 0.50 (middle), and 0.15 (bottom). The horizontal axis is true $\log_2$ fold-change when comparing GCB to a DLCBL sample, both of which are assumed to have been replicated on multiple arrays. The dashed reference line represents a 2-fold change, and the vertical axis is the number of replications. **A.** False-positive rate = 1 in 200,000, spot variance = 1, residual variance = 0.25. **B.** False-positive rate = 1 in 200,000, spot variance = 0.25, residual variance = 0.0625. **C.** False-positive rate = 1 in 20,000, spot variance = 0.25, residual variance = 0.0625. **D.** Same as C., but using a circular design with no reference sample.

proposed by Kerr *et al.* (2001). In addition, effects from the normalization model are estimated with much higher precision than those in the gene model because they are averaged across many more observations, and so the normalization effects can effectively be treated as constants when constructing the data for the gene models.

The heterogeneity in the gene models allows genes to exhibit different degrees of variability, an assumption which is reasonable given the wide variety of known gene functions. Distinct gene-by-gene variance estimates also provide statistically appropriate backdrops by which to assess the significance of corresponding $(GT)_{gi}$ differences and may be of interest in their own right as measures of gene capacity and heritability. Heterogeneous gene variance estimates have been exploited for t-like statistics in Golub *et al.* (1999), Dudoit *et al.* (2001b), and Westfall *et al.* (2001).

While heterogeneity is appealing, the normality assumptions in both the normalization and gene models are subject to criticism. In fact, the aforementioned authors, along with Kerr *et al*. (2001), recommend various resampling-based simulation approaches to circumvent this assumption. In contrast, we suggest that classical statistical procedures based on the assumption of normally distributed errors (e.g., t- and F-tests) will serve microarray analysts well in a large majority of cases. Based upon our experience with a variety of cDNA data sets, assuming normality on the log scale is usually reasonable. However, we do recommend performing standard graphical and statistical checks of this assumption using residuals from the gene models. Normality-based statistical models have been used successfully in nearly all branches of science for decades (Cochran and Cox, 1957; Snedecor and Cochran, 1980; Searle *et al*., 1993; Steel *et al*., 1997; Federer and Wolfinger, 1998).

Linear mixed models can also

- accommodate experiments with missing values for various spots.
- be used with arbitrary methods of background correction.
- handle designs with more than two dyes.
- be applied to data from other kinds of expression data, including those from oligonucleotide chips, SAGE, and EST-counting.
- determine significant genes for class-wide distinctions (e.g., the DLBCL, FL, and CLL classes in Alizadeh *et al*., 2000).
- be used to construct genome-wide quantitative summary measures (Klus *et al*., 2000) and similarity profiles (Hughes *et al*., 2000).

We plan to expound on the details of these points in future publications.

METHODS

Software

We used the Mixed Procedure (SAS Institute Inc., 2000) to perform both the normalization and gene model fits. The gene models are the most computationally intensive portion of the analysis, as a separate REML optimization problem must be solved for each gene. For example, the approximately 7,000 gene models for the yeast data require around four minutes to run on a 400 MHz Pentium II with 128 MB RAM. Although the models are fit sequentially in this instance, the method is highly amenable to parallelization and so is scalable to much larger problems. Example SAS code is available at *http://brooks.statgen.ncsu.edu/ggibson/Pubs.htm*.

The multiple testing problem

Since there are typically thousands of genes in a microarray experiment, the issues of multiple testing and multiple comparisons arise. For example, individual tests carried out at the 5% level will falsely reject a true null hypothesis in 1 out of 20 cases on average, and so the chance of numerous false positives is extremely high when carrying out thousands of tests at this level. Dudoit *et al*. (2001b) provide a nice review of the relevant issues and recommend a resampling-based solution.

As with all traditional statistical hypothesis testing methods, researchers can directly control the false-positive rate α by selecting it a priori and then using it to determine a cutoff for significant p-values. In Fig. 2, $\alpha = 1e-5$, corresponding to a false-positive rate of 1 in 100,000. Alternatively, one can control the false-positive rate over the entire experiment by considering the total number of tests performed. The simplest such adjustment is Bonferonni's method, which sets the cutoff equal to the desired experimentwise false-positive rate divided by the number of tests. For the lymphoma example, a Bonferroni cutoff of 1e-6.3, or about 1 in 2 million, is required to achieve an experimentwise false-positive rate of 0.05. Such a stringent criterion may seem extreme, but results of Westfall *et al*. (2000) suggest that it is not excessively conservative in spite of known dependencies among genes. More complex methods, such as the resampling methods described by Dudoit *et al*. (2001b) and Westfall *et al*. (2000), have been derived only for simple experimental designs.

Statistical power

Specifying an appropriately sized experimental design is critical for assuring that the statistical tests employed in the analysis have power sufficient to detect biologically meaningful differences. "Power" here refers to the probability of declaring statistical significance when a true difference exists, or, equivalently, one minus the probability of a false negative. To determine power, one needs to know the

- experimental design,
- proposed model for analyzing data from the design,
- approximate values for the model parameters,
- hypotheses (contrasts) to be tested,
- desired false positive rate.

The mixed model is well-suited for the determination of statistical power, and we propose the following four-step method that employs analytical formulas based on Muller *et al*. (1992).

1. Specify an exemplary data set corresponding to the proposed experimental design for the gene model. The actual values of the log intensity measurements are unimportant except that they should exhibit enough noise to be able to successfully fit the gene model.
2. Specify the variance components for the proposed gene model and fit the gene model to the exemplary data while holding the variance components at their specified values.
3. From the model output, determine the standard errors of the contrasts of interest. These standard errors are functions of the experimental design and the variance components.
4. Using the computed standard errors, the desired false positive rate, and approximate values for the expected contrasts, compute power based on the noncentral t-distribution.

This method can be applied for any suitable experimental design over a grid of values for the variance components, false-positive rates, and the contrasts.

ACKNOWLEDGMENTS

We gratefully acknowledge helpful comments from Ash Alizadeh, Leping Li, Shyamal Peddada, John Sall, Priya Sudarsanam, David Umbach, and Fred Winston.

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L, Marti, G.E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Straudt, L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Chen, Y., Dougherty, E.R., and Bittner, M.I. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2(4), 364–374.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705.

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

Cochran, W.G., and Cox, G.M. 1957. *Experimental Designs*, Wiley, New York.

Dudoit S., Fridlyand, J., and Speed, T.P. 2001a. Comparison of methods for the classification of tumors using gene expression data. Submitted to *JASA*, *www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html*.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. 2001b. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Submitted to *JASA, www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html*.

Eisen, M., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863–14868.

Ewing, R.M., Kahla, A.B., Poirot O., Lopez F., Audic S., and Claverie J.-M. 1999. Large-scale statistical analysis of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9, 950–959.

Federer, W.T., and Wolfinger, R.D. 1998. SAS Code for recovering intereffect information in experiments with incomplete block and lattice rectangle designs. *Agronomy J.* 90, 545–551.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P.O., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A.A., Staudt, L., and Botstein D. 2000. Gene shaving: A new class of clustering methods for expression arrays. *www-stat.stanford.edu/~hastie/Papers/.*

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126.

Kerr, M.K., and Churchill, G.A. 2001a. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Submitted manuscript, *www.jax.org/research/churchill/pubs/index.html.*

Kerr, M.K., and Churchill G.A. 2001b. Experimental design for gene expression microarrays. *Biostatistics*. To appear in *www.jax.org/research/churchill/pubs/index.html.*

Kerr, M.K., Martin, M., and Churchill, G.A. 2001. Analysis of variance for gene expression microarray data. To appear in *J. Comp. Biol.*, *www.jax.org/research/churchill/pubs/index.html.*

Klus, G.T., Bittner, M.L., Chen, Y., Wahde, M., and Szallasi, Z. 2000. Use of overall quantitative features of cDNA microarray measurements in cancer research. *www.usuhs.mil/pha/faculty/zoltan.shtml.*

Lazaridis, E., Gieser, P., and Yanev, G. 2000. Statistical methods in modern molecular biology: RNA and protein analysis, *in Joint Statistical Meetings Continuing Education Series*, American Statistical Association, Washington, DC.

Lazzeroni, L., and Owen, A.B. 2000. Plaid models for gene expression data. Submitted manuscript, *www-stat.stanford.edu/~owen/reports/.*

Lee, M.-L.T., Kuo, F.C., Whitmore, G.A., Sklar, J. 2000. The importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proc. Natl. Acad. Sci.* 97, 9834–9839.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. 1996. *SAS System for Mixed Models,* SAS Institute Inc., Cary, NC.

Muller, K.E., LaVange, L.M., Ramey, S.L., and Ramey, C.T. 1992. Power calculations for general linear multivariate models including repeated measures applications. *J. Am. Stat. Assoc.* 87(420), 1209–1226.

Netwon, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Submitted manuscript.

Rocke, D.M., and Durbin, B. 2001. A model for measurement error for gene expression arrays. *http://handel.cipic.ucdavis.edu/~dmrocke/.*

Sapir, M., and Churchill, G.A. 2000. Estimating the posterior probability of differential gene expression from microarray data. Poster, *www.jax.org/research/churchill/pubs/index.html.*

SAS Institute Inc. 2000. *SAS/STAT Software Version 8*, SAS Institute Inc., Cary, NC.

Searle, S., Casella, G., and McCulloch C. 1993. *Variance Components*, Wiley, New York

Snedecor, G.W., and Cochran, W.G. 1980. *Statistical Methods.* Iowa State University Press, Ames.

Steel, R.G.D., Torrie, J.H., and Dickey, D. 1997. *Principles and Procedures of Statistics: A Biometrical Approach*, 3rd ed., McGraw–Hill, New York.

Sudarsanam, P., Vishwanath, R.I., Brown, P.O., and Winston, F. 2000. Whole-genome expression analysis of snf/swi mutants in *Saccaromyces cerevisiae. Proc. Natl. Acad. Sci.* 97, 3364–3369.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Tanaka, T.S., Jaradat, S.A., Lim, M.K., Kargul, G.J., Wang, X., Grahovac, M.J., Pantano, S., Sano, Y., Piao, Y., Nagaraja, R., Doi, H., Wood III, W.H., Becker, K.G., and Ko, M.S.H. 2000. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl. Acad. Sci.* 97, 9127–9132.

Tibshirani, R., Hastie, T. Eisen, M., Ross, D., Botstein, D., and Brown, P.O. 1999. Clustering methods for the analysis of DNA microarray data. Submitted manuscript, *www.stat.stanford.edu/~tibs/lab/publications.html.*

Westfall, P.H., Zaykin, D.V., and Young, S.S. 2001. Multiple tests for genetic effects in association studies. To appear as a chapter in *Methods in Molecular Biology: Biostatistical Methods*, Stephen Looney, ed., Humana Press, Totowa, NJ.

Wittes, J., and Friedman, H.P. 1999. Searching for evidence of altered gene expression: A comment on statistical analysis of microarray data. *J. Natl. Cancer Inst.* 91, 400–401.

Address correspondence to:
Russell D. Wolfinger
Director of Genomics
SAS Institute Inc.
Cary, NC 27513

E-mail: russ.wolfinger@sas.com