BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1620–1627
DOI: 10.1093/bioinformatics/btg227

The effect of replication on gene expression microarray experiments

Paul Pavlidis[1,], Qinghong Li[2] and William Stafford Noble[3,†]*

[1]Columbia Genome Center, Columbia University, 1150 St. Nicholas Avenue, New York, NY 10032, USA, [2]Department of Computer Science, Columbia University, 1214 Amsterdam Avenue, New York, NY 10027, USA and [3]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA

Received on 27 November 2002; revised on 28 February 2003; accepted on 28 March 2003

ABSTRACT

Motivation: We examine the effect of replication on the detection of apparently differentially expressed genes in gene expression microarray experiments. Our analysis is based on a random sampling approach using real data sets from 16 published studies. We consider both the ability to find genes that meet particular statistical criteria as well as the stability of the results in the face of changing levels of replication.

Results: While dependent on the data source, our findings suggest that stable results are typically not obtained until at least five biological replicates have been used. Conversely, for most studies, 10–15 replicates yield results that are quite stable, and there is less improvement in stability as the number of replicates is further increased. Our methods will be of use in evaluating existing data sets and in helping to design new studies.

Contact: pp175@columbia.edu

Supplementary information: http://microarray.cpmc.columbia.edu/pavlidis/pub/gxrep

INTRODUCTION

Replication is a straightforward method for improving the quality of inferences made from experimental studies. However, replication increases the cost of experiments and, typically, the amount of material needed. In general, it makes sense to do as much replication as is necessary to achieve a desired level of sensitivity and specificity, but not much more. This trade-off between cost and statistical power arises frequently in gene expression microarray experiments. Replication is clearly necessary in this domain (Lee *et al*., 2000; Novak *et al*., 2002), but microarray experiments are costly and involve RNA samples that are often difficult to obtain. We therefore need techniques for estimating in advance how many replicates should be performed in a given study.

A standard approach to the problem of estimating the statistical properties of a planned set of data is 'power analysis'. Power analysis estimates the probability of correctly rejecting the null hypothesis in favor of a specific alternative while maintaining a particular Type I error rate. For the situations we consider here, the alternative hypothesis is usually expressed in terms of 'effect size', the actual difference in the group means (relative to the variance) that is desired to be detected. A mathematical model of the data is then used to estimate how many replicates are needed to achieve the desired Type I and Type II error rates. Certain parameters for the modeled data (most critically, the expected variability) are often estimated from real data, perhaps from a pilot study.

Although clearly a useful tool, power analysis comes with some caveats. First, the estimated variability is critically dependent on the assumptions of the model and the quality of the input parameter estimates. A second set of assumptions enters into the statistical test that is used to evalute the null hypothesis. In addition, for gene expression studies, power analysis is potentially extremely complex, with a separate set of parameters for each gene, not to mention the need to account for complex interactions among genes. To our knowledge such a complete power calculation has not been attempted, though some papers have used simpler power analyses to study microarray expression data (Zien *et al*., 2002; Hwang *et al*., 2002; Pan *et al*., 2002).

In this paper, we study the effect of increasing (or decreasing) replication on the detection of differentially expressed genes in real data sets, avoiding the assumptions required to simulate data. However, because in real data sets we do not know which genes truly show differential expression, we cannot directly assess power. Instead, we examine aspects of the results which are of interest to biologists and which complement traditional power analyses. We make our findings as general as possible by analyzing many data sets.

We consider a simple general type of experiment, the goal of which is to identify genes that are differentially expressed between two experimental groups (for example, tumor and normal tissue). The two groups each contain a number of

*To whom correspondence should be addressed.
† Formerly William Noble Grundy: see www.gs.washington.edu/~noble/name-change.html

Published by Oxford University Press

replicate samples. These replicates are derived from different biological sources, as opposed to so-called 'technical replicates', in which the same biological sample is tested multiple times. Differentially expressed genes are identified by a statistical test for group comparison (such as a t-test), where the null hypothesis is equality of the group means. A p-value threshold is applied following the test to establish a desired Type I error rate. The final result obtained from this hypothetical experiment is a list of genes that are differentially expressed at a particular level of statistical confidence.

To study various levels of replication, we use a random sampling approach. Given a real data set, we simulate smaller data sets of various sizes by randomly selecting samples from it. For example, if we start with a data set containing at least 12 replicates in each group, then we can make data sets of any level of replication (up to 12) by randomly selecting from the real samples (Fig. 1). We then examine properties of each of these sampled data sets with methods described below. We repeat this procedure on many data sets, for every possible level of replication, for many random samples, to generate a large set of statistics on the properties of data sets of various sizes.

We consider two qualities of each sampled data set. The first and most important is the ability to obtain any results at all, that is, to find genes that meet our statistical criteria. We refer to this property as 'apparent power' to distinguish it from power in the strict sense. Because increasing sample size will essentially always increase power, it might be reasonable for an experimenter to choose a level of replication that is sufficient to yield 'enough' high-confidence candidates, where 'enough' must be defined by the needs of the experiment.

The second quality that we consider is the stability of the results. Note that stability is only meaningful if some genes have met our statistical criteria. We define stability as the tendency for the results to remain the same as the replication level is changed. We define two metrics of stability, which differ in their stringency. First, we consider the stability of the identities of the genes that meet the statistical criteria. Second, we consider the rank order of those genes. Details of our metrics are provided in the methods section, below.

Our goal is to identify, for each data set, a level of replication that yields good performance according to our metrics, but without requiring an unreasonably large number of replicates. We wish to ask, 'Can we find useful results with only a few replicates?' and at the other extreme, 'Do we need 30 replicates?' Although the experimental design used here is simple—identifying differentially expressed genes across two conditions—the techniques that we describe could be applied to a wide range of situations.

Our results suggest that while statistical power is a critical consideration in experimental design, researchers should also consider the stability of the results they obtain. While the specific findings are data dependent, we found that good apparent power and stability can usually be obtained with fewer than 15 replicates, and often with fewer than 10. On the other hand, using fewer than five replicates almost always results in poor apparent power and low stability. The methods we present can be used in study design, by researchers who have pilot data and wish to estimate the benefits of performing more experiments, and for evaluating the reliability of existing data.

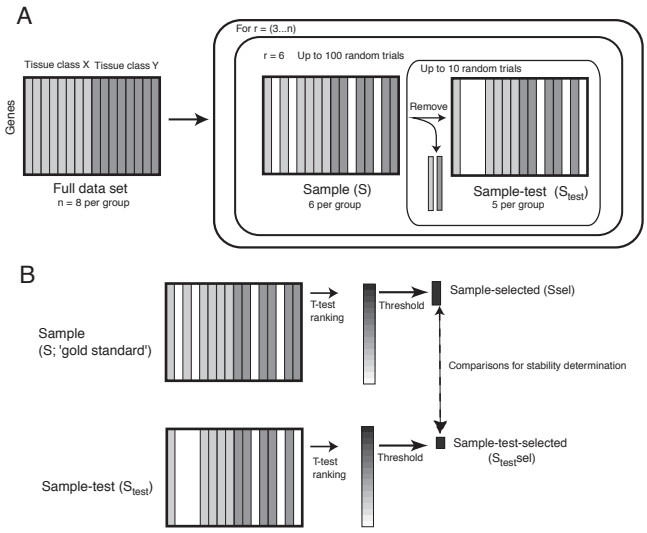

Fig. 1. Outline of methods. A schematic of the methods used in this study. The data sets are depicted as collections of data from individual microarrays (indicated by grey bars). In this simple experiment, there are two types of tissues, X (light grey) and Y (dark grey), which were tested with the goal of identifying genes that are differentially expressed between groups X and Y. (**A**) The sampling procedure. The ovals indicate procedures that are repeated during the experiment. To create a sample S of size r, arrays are removed (indicated by blanks) at random from the full data set. The properties of this data set are tested by comparison to a test sample ($Sample_{\text{test}}$ or S_{test}), generated by removing one additional sample at random from each group. Up to 10 S_{test}s are made randomly from each S (innermost loop) to establish stability measures for S. (**B**) Outline of the statistical testing procedure, starting with the output of a single sampling trial as shown in (A). First, a t-test is used to generate a ranking of the genes in the two data sets. Then a statistical threshold is applied to select genes. As indicated by the dashed line, the results from S are compared to S_{test}. See text and Figure 2 for further details.

METHODS

Data sets

We used microarray gene expression data sets from publically available sources, as summarized in Table 1. If the original data set included more than two types of samples, we generally chose the two groups with the largest number of replicates for further study. If necessary, for some cDNA data sets we imputed missing data points as the mean of the values for the gene, and genes that were missing more than 20% of the data were not used. We did not attempt to replicate precisely the

Table 1. Data sets used in this study

Reference	Replicates	Genes	Type	Description
Allander *et al*. (2001)	6	1987	cDNA	Sarcoma (GIST versus spindle cell carcinoma)
Hedenfalk *et al*. (2001)	7	3226	cDNA	Breast cancer (BRCA1 versus BRCA2)
Callow *et al*. (2000)	8	6384	cDNA	Knockout mice (apoIA-/- versus control)
Huang *et al*. (2001)	8	12 558	oligo	Thyroid cancer (papillary tumor versus normal)
Luo *et al*. (2001)	9	2303	cDNA	Prostate (cancer versus BPH)
Ramaswamy *et al*. (2001)	10	16 063	oligo	Cancer (breast versus prostate adenocarcinoma)
Eaves *et al*. (2002)	12	39 114	oligo	Mouse (spleen versus thymus)
Shipp *et al*. (2002)	19	7129	oligo	Lymphoma (DLBCL versus FL)
Khan *et al*. (2001)	20	2303	cDNA	Sarcoma (EWS versus RMS)
Armstrong *et al*. (2002)	20	12 582	oligo	Leukemia (MLL versus ALL)
Alon *et al*. (1999)	22	2000	oligo	Colon cancer (tumor versus normal)
Golub *et al*. (1999)	25	7129	oligo	Leukemia (ALL-Bcell versus AML)
Yeoh *et al*. (2002)	27	12 625	oligo	Leukemia (E2APBX versus TEL AML).
Gruvberger *et al*. (2001)	28	3389	cDNA	Breast cancer (ER+ versus ER−)
Garber *et al*. (2001)	32	22 115	cDNA	Lung cancer (adenocarcinoma versus other)
Singh *et al*. (2002)	50	12 600	oligo	Prostate (tumor versus normal)

Summary of the 16 data sets used to study the effect of replication. The table lists the largest number of replicates we used, which is not necessarily the complete published data set. The number of genes (or, more precisely, the number of array elements) is also indicated. The 'Type' column refers to the type of array which was used in the study, either 'cDNA' for data that was collected using two-color cDNA microarrays, 'oligo' for Affymetrix-type oligonucleotide arrays. The last column description of the experiment and the comparison we studied. For details, see the web supplement. Abbreviations: apoIA, apolipoprotein IA; E2APBx, GIST, gastrointestinal stromal tumor; ER, estrogen receptor; AML, acute myeloid leukemia; BPH, benign prostate hyperplasia; DLBC, diffuse large B cell lymphoma; FL, follicular lymphoma; EWS, Ewing's sarcoma; RMS, rhabdomyosarcoma; MLL, mixed lineage leukemia; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia.

analysis of interest in the original study. For example, in one case (Eaves *et al*., 2002), we examine a distinction (spleen versus thymus) that was not examined in the original publication. Therefore, our study should not taken as an evaluation of the quality of the original work. Details on the data sets as used in our study are available on our web site.

Gene evaluation

To evaluate the significance of a gene, we use the Student's t-test. This test is performed on each gene in the data set, in each case testing the null hypothesis that the mean expression level for the gene is equal in the two groups of samples. Our use of the t-test implies assumptions about the distribution of the data, namely normality and homoscedasticity, which may or may not be valid in various cases. The t-test is a reasonable choice because it is simple to perform, and commonly used in published microarray studies. Our basic findings are unlikely to be dramatically affected by the exact choice of test, and this is supported by preliminary results with the Mann–Whitney 'U' test (see the web supplement).

Multiple test correction

To set appropriate statistical thresholds for each trial, accounting for the multiplicity of testing thousands of genes in each data set, we use a method that controls the false discovery rate [FDR; Benjamini and Hochberg (1995)]. The false discovery rate is the number of null hypotheses that can be expected to be falsely rejected (that is, false positives) expressed as a fraction of the total number of genes selected. For a desired FDR (e.g. 0.05), we identify a t-test p-value threshold using the method of Benjamini and Hochberg (1995).

Experimental design

Our testing procedure for each data set is illustrated in Figure 1. This entire procedure is repeated for each data set listed in Table 1. Each data set originally contains two groups of samples with size n and n' ($n \leq n'$). The largest data set we consider in a simulation is of size $2n$. From the full data set, a number of replicates r ($3 \leq r \leq n$) is chosen. By randomly choosing r samples (microarrays) from each experimental group, a new data set called a *Sample* (abbreviated S) is created. Up to 100 such randomly generated pseudo data sets are created for each value of r, though only the possible distinct subsamples of the data are studied if this is fewer than 100. For each S, multiple test data sets S_{test} are created by randomly removing one sample from each group (Fig. 1A). Thus, if S contains four per group, then S_{test} contains three replicates per group. We test up to 10 randomly created S_{test}s for each S (when $r = 3$, only nine different S_{test}s are possible). The stability statistics (see below) for the 10 S_{test}s are averaged to yield measures of stability for S. This procedure is repeated for each value of r. For $r = n$ in data sets where $n = n'$, there is only one possible S; thus, in this situation we instead create up to 100 random S_{test} data sets instead of just 10 to help ensure that we collect sufficient samples to get good statistics. This procedure of removing a single replicate is similar in character to a jackknife sampling (Efron and Tibshirani, 1998). However, unlike the jackknife, we do not

attempt to assess the effect of all possible sample removals because usually these are too numerous. Another difference from a jackknife sample is that, in order to keep our experiments simple, for each trial, we remove one sample from each group—two samples in total—instead of a single sample.

For S and S_{test}, a Student's t-test is performed on each gene. This generates a ranking of the genes in each data set, where the highest ranked genes are most likely to exhibit changes in expression. A statistical threshold based on FDR (described earlier) is then applied to both ranked lists. The ranking and the selected genes in each S_{test} are compared to the ranking and selected genes in S using the metrics described above. The median and interquartile range of the stability and power metrics for all psuedo data sets is determined for each r.

Measuring stability and power

As outlined above, we developed three metrics for assessing the apparent power and stability of each data set S. We refer to the genes selected in S at a given FDR as $Ssel$ (for 'genes Selected from Sample'). The 'apparent power' is the size of $Ssel$. This metric indicates how many replicates are needed before genes begin to meet a particular statistical threshold. Apparent power is expressed in units of genes selected, unlike power in the strict sense, which is expressed as a value between 0 and 1, because we do not know how many genes actually show changed expression.

The two stability metrics involve comparing the ranked list of genes selected from S to the ranking obtained when one replicate is removed (S_{test}). A simple example illustrating how both stability metrics operate is shown in Figure 2. The first metric is the fraction of genes in $S_{\text{test}}sel$ that are also in $Ssel$. We refer to this metric as 'recovery'; it ranges from 0.0 (none of the genes in $S_{\text{test}}sel$ are in $Ssel$) to 1.0 (all are in $Ssel$). For example, a value of 0.5 means that when one replicate is removed (yielding S_{test}), some genes still meet the statistical criterion, but only half of them met the criterion before (in S). Thus, this measure captures an important aspect of data stability: in a data set with a high recovery score, the identities of the genes that are selected by the statistical test would likely be similar if an additional replicate experiment were performed.

The second stability metric measures the degree to which the ordering of genes selected from S and S_{test} is preserved. The test statistic is the Spearman rank correlation of the rankings for genes that occur in both $S_{\text{test}}sel$ and $Ssel$. This 'order' metric varies from −1.0 (exactly reversed order) to 1.0 (exactly the same order). The 'order' measure captures a more subtle but still important aspect of data stability than 'recovery': in a data set with a high order score, the relative ranking of the selected genes would not change dramatically if an additional replicate experiment were performed. Note that neither stability metric is intended to measure the correctness of the results.

When apparent power is zero, the stability metrics cannot be computed because $S_{\text{test}}sel$ contains no genes. We report recovery and order statistics for a given p-value threshold only if the number of 'successful' tests was at least 10, sufficient to collect reasonable statistics. In addition, when only small numbers of genes are selected, both 'order' and 'recovery' can only attain a very restricted set of values, and will be highly variable. To help limit these effects, we only show results for a given FDR if at least two genes were selected on average for that setting.

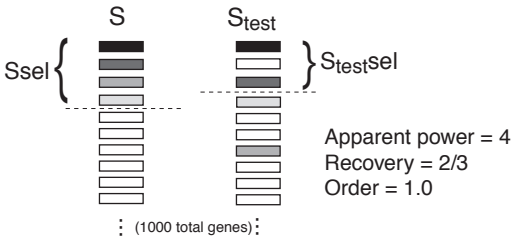

Fig. 2. Metrics. A toy example illustrating the metrics used in this study. Genes are indicated by bars, and the vertical order indicates the statistical ranking. In this example, there are 1000 genes, of which only the top ranked are shown. In S, four genes are selected (indicated by four differently shaded bars) at a given threshold (indicated by the dotted line). In S_{test} (which contains one fewer replicate than S), only three genes are selected at the same threshold; the genes that were selected in S now have the locations shown. The values of the three metrics we used are shown on the right. The 'apparent power' is simply the number of genes above the threshold in S. The 'recovery' score is 2/3, because of the three genes selected in S_{test}, two of them were selected in S. The 'order' score is 1.0, because the two genes that appear in both selected sets occur in the correct relative order; that is, the top gene is still ranked higher than other selected genes. See text for details.

RESULTS

We study the effect of replication in 16 published data sets. We present only a portion of our results in detail here; full results for all data sets are available on our web site (microarray.cpmc.columbia.edu/pavlidis/pub/gxrep). We focus on results at one false discovery rate setting, 0.05, which is the third-most stringent we used. Figure 3 summarizes the main results for all 16 data sets. Figure 4 shows more detailed graphs for selected data sets. In our presentation of the results, we first consider apparent power and then each stability measure in turn.

Apparent power

As expected, when the number of replicates is small, the apparent power is low. In some data sets containing many replicates, many genes are assigned p-values of 10^{-10} or smaller. However, when the number of replicates is reduced to a low level, small p-values become rare. Using multiple testing correction based on FDR, often few or no genes are selected from data sets containing five or fewer replicates (Fig. 3A). This is true even at the most lax FDR we tested (0.1; see web site).

By definition, increasing the number of replicates increases statistical power (in the strict sense). Similarly, if we examine

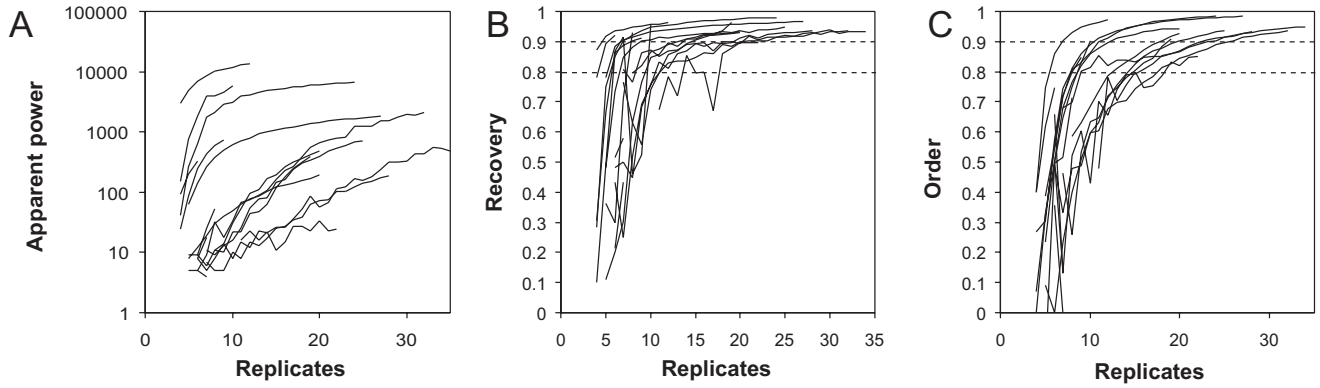

Fig. 3. Summary of results. Each line represents results for one data set shown in Table 1, at an FDR of 0.05. Not all of the 16 data sets are illustrated on these graphs, because some failed to meet criteria at this FDR (see our web site for more results). The plots are of the median values for all trials. Error bars are omitted for clarity. The dashed lines in (B) and (C) indicate the 0.8 and 0.9 levels. (**A**) Plot of the number of genes selected (apparent power, the size of *Ssel*). Note that the scale is logarithmic. (**B**) Recovery stability. (**C**) Order stability. Values below zero are not shown. Larger versions of this and the other figures are available as supplementary data.

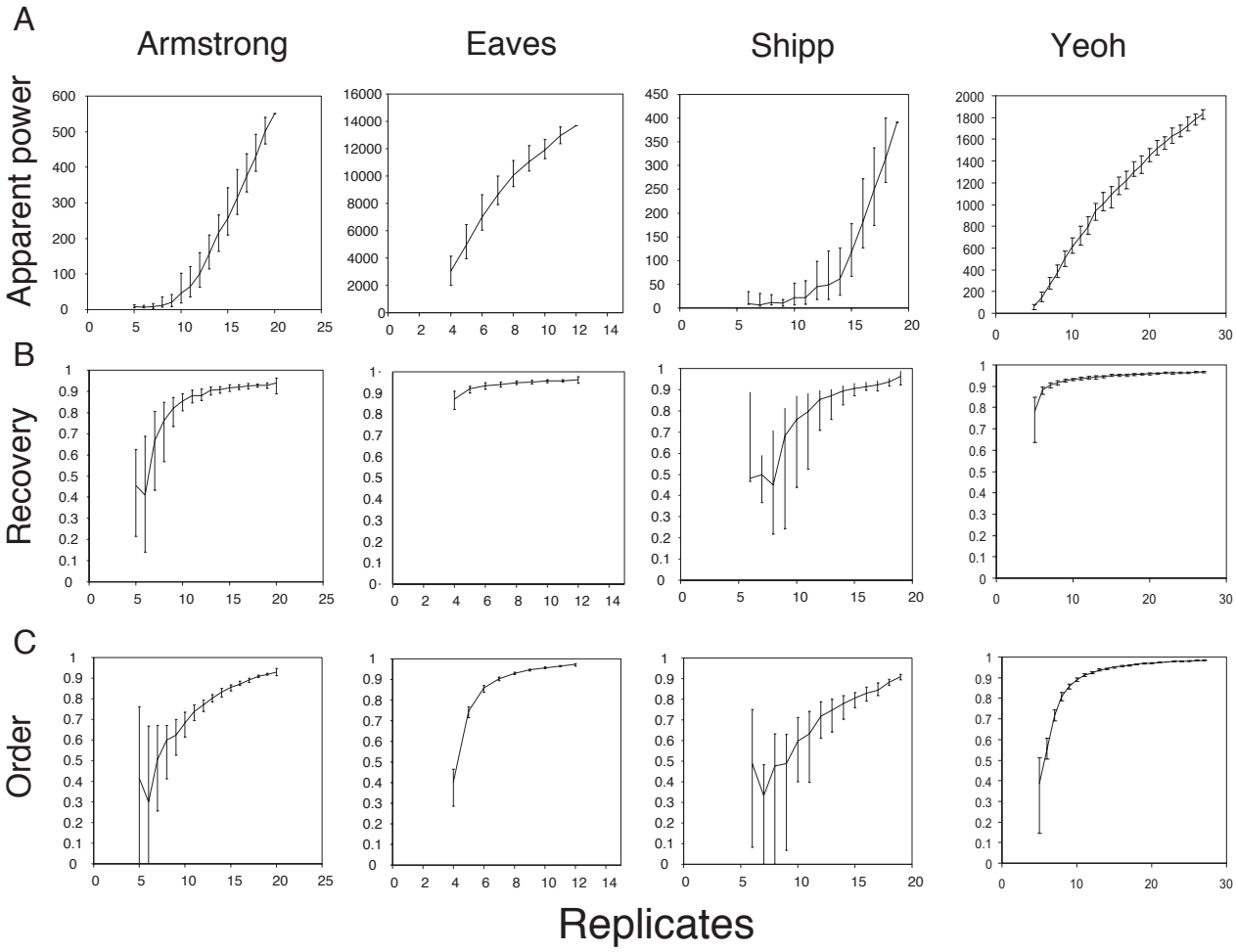

Fig. 4. Detailed results for four data sets. The plots show detailed results for four of the data sets illustrated in Figure 3. The FDR is 0.05. The error bars indicate the interquartile range of the scores for all trials at each level of replication. (**A**) Apparent power. (**B**) Recovery stability. (**C**) Order stability.

how many genes are *actually selected* as replicates increase (apparent power), then we observe a steady increase, with little leveling off (Figs. 3A and 4A; note the logarithmic scale). This effect illustrates that, for the available data, there is no obvious point at which all differentially expressed genes have been statistically detected.

The apparent power is expected to be inversely proportional to the subtlety of the biological effect under investigation as well as inversely proportional to the amount of experimental noise. Not surprisingly, given the variety of data sets we analyzed, we observe a wide range in apparent power. Table 2 summarizes this range, listing the apparent power for each data set when using all available replicates. For some data sets, apparent power at the threshold we focus on (0.05) is as much as 35% of the genes on the array (Table 2). This is not only an effect of the different maximal levels of replication among the data sets. Even if we look at a single level of replication across multiple data sets, a similarly wide range of numbers of genes are selected (Table 3; see also Fig. 3A).

Our stability results (below) must be viewed in the context of the effect of replication on apparent power. In fact, for most data sets, at very low levels of replication and at higher p-value thresholds, the stability metrics cannot be used because no genes are found. This result suggests that when very few replicates are available, power is too low for simple statistical criteria to be of much use in detecting differential expression.

Table 2. Number of genes selected at each FDR (apparent power), when using all replicates

Data set	0.0125	0.025	0.05	0.1	Fract.
Callow	—	—	4	10	0.0006
Huang	15	32	53	127	0.0042
Hedenfalk	—	3	18	91	0.0056
Alon	12	14	24	68	0.0120
Armstrong	222	309	478	820	0.0380
Gruvberger	80	123	184	298	0.0543
Shipp	104	190	391	755	0.0548
Khan	101	139	192	283	0.0834
Garber	833	1309	2112	3403	0.0955
Golub	300	460	708	1092	0.0993
Yeoh	1342	1551	1833	2252	0.1452
Allander	137	213	328	472	0.1651
Singh	619	954	2091	4402	0.1660
Luo	319	488	741	1169	0.3218
Eaves	10 183	11 738	13 666	16 286	0.3494
Ramaswamy	1764	3598	5797	8041	0.3609

The fraction of the genes on the array that are selected at a FDR of 0.05 is shown in the last column. See Table 1 for the microarray sizes. A '—' indicates that too few genes met criteria for inclusion in the study (< 2 on average). A plot of the data in this table is given in Figure 3A.

Table 3. Number of genes selected at each FDR (apparent power), when using 10 replicates

Data set	0.0125	0.025	0.05	0.1	Fract.
Alon	—	—	—	8	—
Singh	—	6	8	20	0.0006
Garber	11	29	34	166	0.0015
Golub	8	8	16	40	0.0022
Armstrong	12	15	31	87	0.0025
Gruvberger	5	6	10	13	0.0030
Shipp	7	12	22	54	0.0031
Khan	25	33	49	82	0.0213
Yeoh	327	450	615	876	0.0487
Eaves	8556	10 095	11 929	14 491	0.3050
Ramaswamy	1764	3598	5797	8041	0.3609

The fraction of the genes on the array that are selected at a FDR of 0.05 is shown in the last column. See Table 1 for the microarray sizes. A '—' indicates that too few genes met the threshold for inclusion (< 2 genes on average). Only data sets containing at least 10 replicates could be included in this table.

Stability

The two stability metrics, recovery and order, are illustrated in Figures 3 and 4, C and D, respectively. Unlike apparent power, both stability metrics tend to level off past a certain number of replicates. Thus, increasing the number of replicates beyond a certain value yields a relatively small increase in stability. For most data sets, the change in recovery score slows at approximately 8–12 replicates; some reach this level with as few as 6 replicates (Fig. 3C; Fig. 4C, Eaves). In general, the recovery metric levels off with scores of 0.8–0.95. Increases in the more stringent order metric continue until at least 10 or 15 replicates are used (Fig. 4D), although as few as eight replicates are necessary for some data sets. Maximal order stability scores are typically 0.8–0.9.

For some of the smaller data sets (less than 10 replicates available), we are unable to observe any leveling off of the stability metrics. We consider the results for these data sets (in particular, the Huang, Callow and Hedenfalk data sets) to be inconclusive, because we do not know how many replicates would be required to obtain stability.

Details of four data sets

As mentioned, we observe a range of behaviors for individual data sets, and it is informative to examine the results in detail. Here we discuss the four data sets shown in Figure 4, which cover a wide range of the behaviors we observed. They were selected for display in part because they have many replicates, so we can make relatively confident conclusions about the results.

The Eaves data set is somewhat unusual, first because unlike most of our data sets it comes from mouse, and second because of the large difference between the tissues we compared (spleen and thymus). We reiterate that this comparison was not the one of primary interest to Eaves *et al.* (2002) who were interested in the (relatively very small) differences between defined groups of samples within each tissue type. The number of replicates for those comparisons was only four, so we did not attempt to study them. Using all 12 replicates, the

apparent power represents thousand of genes (out of ≈40 000 on the array). This data set also shows high stability at low numbers of replicates compared to many of the cancer data sets: even the order stability metric reaches 0.8 with only six replicates.

The Shipp, Armstrong and Yeoh data sets all involve comparing different tumor types. The Armstrong and Shipp data sets have lower apparent power than Yeoh, even when many replicates are used, suggesting that the biological distinction for those data sets is relatively subtle. For the Yeoh data set, we compared E2APBx with TEL AML (acute myeloid leukemia), which differ in characteristic chromosomal rearrangements, though both are B-cell lineage leukemias (Yeoh *et al*., 2002). Using all 27 replicates, nearly 2000 genes show significant differences in expression at a FDR of 0.05. Recovery stability was high even with only five or six replicates; order stability is over 0.8 once nine replicates are used. The full data set shows very high stability, among the highest we measured in this study. Several other cancer data sets seemed similar in character, including those of Khan, Golub, Luo and Allander, though not all showed as strikingly high apparent power.

Like Yeoh *et al*. (2002), Armstrong *et al*. (2002) studied leukemia, and compared acute lymphoblastic leukemia (ALL) with a subtype that involve particular chromosome translocations (mixed-lineage leukemia, MLL). It required somewhat more replicates than Yeoh to reach our stability threshold of 0.8. The Garber and Ramaswamy data sets showed similar behavior.

Shipp *et al*. (2002) compared diffuse large B-cell lymphoma with follicular lymphoma. This data set required more replicates to reach stability (10–15) than the Yeoh and Armstrong data sets. The Alon and Singh data sets showed similar behavior. The samples used for these data sets may have greater heterogeneity than the Yeoh data set, for example.

DISCUSSION

Our results show that in most cases, using fewer than five replicates results in rather poor results in a statistical analysis, both in terms of apparent power and in stability. For most data sets, near-maximal levels of stability are obtained between eight and 15 replicates, with most of the improvement occurring by ten replicates. This is also the range of replicate levels that typically result in the detection of differential expression at quite high levels of statistical confidence (FDR 0.05 or lower). Even with the most stringent measure of stability (order), using more than 15 replicates has a diminishing effect for most data sets where we could test this. These numbers are naturally quite data-dependent. Therefore, planning a future experiment would require looking at apparent power and stability scores for data sets that are likely to have similar properties to the proposed study.

We developed these methods to complement power analysis by more familiar methods. While in our studies we are unable to directly address questions of power, we base our observations entirely on realistic situations. Power analysis, on the other hand, directly estimates power, but its ability to do so is contigent on the realism of the model used. Interestingly, these two different approaches yield results that are generally in agreement. For example, Zien *et al*. (2002) suggest that 15 or more replicate samples are needed to be able to detect fairly large changes in expression (3-fold) with good power (>0.8), using models based on five data sets. In another study, (Hwang *et al*., 2002) suggest that eight replicates should be sufficient to detect an effect of size 2 for one particular data set (that of Golub *et al*., 1999) at 0.95 power and 0.95 confidence (Hwang *et al*., 2002). Pan *et al*. (2002) found that eight replicates were needed to detect an effect of size 3 with power 0.8 in a rat radiolabeled microarray data set.

If the above results are accepted as broad guidelines, then it would seem that most published studies probably have very little power. In fact, even doing three replicates has only recently become common, and other than cancer studies, we are aware of very few studies that do as many as five. However, a brief review of the literature reveals dozens of papers that use hardly any replication if any, yet appear to yield at least some results of value. Why do researchers find poorly replicated experiments useful? The simple explanation is that high power is not always necessary to yield some useful results from a microarray study. The usefulness of a microarray study is often gauged by *how many* high quality differentially expressed genes are obtained, not by the *fraction* of all such genes that are detected (which is generally unknowable anyway). The latter requires high power; the former only requires that some of the expression changes be robust enough to be reliable. Even with no replication, some of the most striking findings are likely to be 'real'.

Given that a statistical approach breaks down in the presence of few replicates, a 'fold change' or other heuristic method must be adopted to select genes from such data. The cost of such an approach is that very stringent criteria need to be applied, and the results must be confirmed with an alternative method. This approach is unlikely to be effective if the expected expression changes are subtle and are restricted to a small number of genes—a common situation, as shown in Table 2. In particular, specificity would likely be very low. In many cases, the 'fold change' method will yield so many erroneous results as to be of highly questionable use. We note that our methods can easily be extended to an examination of 'fold change', a topic we leave for further study.

Our methods can be readily applied to new data sets to assess their reliability. First, assessing apparent power obviously does not require a sampling approach: lack of apparent power will be obvious if no genes meet reasonable statistical criteria. For the recovery and order metrics, the situation is equivalent to the trials in our experiments when all replicates are being used. The random sampling would be applied (or, preferably and if computationally feasible, a full jackknife

sampling) to generate pseudo data sets containing one fewer replicate, and our stability metrics applied.

ACKNOWLEDGEMENTS

We thank Walter L. Ruzzo, Kathleen Kerr and Manisha Desai for comments on the manuscript. This work is funded in part by a Sloan Foundation Research Fellowship to WSN and by National Science Foundation grant ISI-0093302.

REFERENCES

Allander,S.V., Nupponen,N.N., Ringner,M., Hostetter,G., Maher,G.W., Goldberger,N., Chen,Y., Carpten,J., Elkahloun,A.G. and Meltzer,P.S. (2001) Gastrointestinal stromal tumors with kit mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Res.*, **61**, 8624–8628.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Callow,M.J., Dudoit,S., Gong,E.L., Speed,T.P. and Rubin,E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**, 2022–2029.

Eaves,I.A., Wicker,L.S., Ghandour,G., Lyons,P.A., Peterson,L.B., Todd,J.A. and Glynne,R.J. (2002) Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the nod model of type 1 diabetes. *Genome Res.*, **12**, 232–243.

Efron,B. and Tibshirani,R. (1998) *An Introduction to the Bootstrap*. Chapman and Hall, London.

Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Gengelbach,M., van de Rijn,M., Rosen,G.D., Perou,C.M., Whyte,R.I. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.

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.

Gruvberger,S., Ringner,M., Chen,Y., Panavally,S., Saal,L.H., Borg,A., Ferno,M., Peterson,C. and Meltzer,P.S. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.*, **61**, 5979–5984.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O.P. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.

Huang,Y., Prasad,M., Lemon,W.J., Hampel,H., Wright,F.A., Kornacker,K., LiVolsi,V., Frankel,W., Kloos,R.T., Eng,C., Pellegata,N.S. and de la Chapelle,A. (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl Acad. Sci. USA*, **98**, 15044–15049.

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

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Lee,M.L., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Luo,J., Duggan,D.J., Chen,Y., Sauvageot,J., Ewing,C.M., Bittner,M.L., Trent,J.M. and Isaacs,W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.*, **61**, 4683–4688.

Novak,J., Sladek,R. and Hudson,T. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, **79**, 104–113.

Pan,W., Lin,J. and Le,C.T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.*, **3**, R22.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Shipp,M.A., Ross,K.N., Tamayo,P., Weng,A.P., Kutok,J.L., Aguiar,R.C., Gaasenbeek,M., Angelo,M., Reich,M., Pinkus,G.S. *et al.* (2002)Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Zien,A., Fluck,J., Zimmer,R. and Lengauer,T. (2002) Microarrays: How many do you need? In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, pp. 321–330.