Biostatistics (2005), **6**, 1, *pp.* 27–38
doi: 10.1093/biostatistics/kxh015

Sample size determination in microarray experiments for class comparison and prognostic classification

KEVIN DOBBIN, RICHARD SIMON
Biometric Research Branch, National Cancer Institute, 6130 Executive Blvd., Bethesda, MD, 20892-7434, USA
dobbinke@mail.nih.gov

SUMMARY

Determining sample sizes for microarray experiments is important but the complexity of these experiments, and the large amounts of data they produce, can make the sample size issue seem daunting, and tempt researchers to use rules of thumb in place of formal calculations based on the goals of the experiment. Here we present formulae for determining sample sizes to achieve a variety of experimental goals, including class comparison and the development of prognostic markers. Results are derived which describe the impact of pooling, technical replicates and dye-swap arrays on sample size requirements. These results are shown to depend on the relative sizes of different sources of variability. A variety of common types of experimental situations and designs used with single-label and dual-label microarrays are considered. We discuss procedures for controlling the false discovery rate. Our calculations are based on relatively simple yet realistic statistical models for the data, and provide straightforward sample size calculation formulae.

Keywords: Experimental design; Gene expression; Microarrays; Sample size.

1. INTRODUCTION AND BACKGROUND

Microarray experiments are often complex, generate large amounts of data, and warrant careful planning. Here we present formulae for determining sample sizes to achieve a variety of experimental goals, including class comparison and the identification of gene expression based prognostic markers. In the microarray literature, class comparisons refer to experiments in which the goal is to compare two different classes of specimens (e.g. cancer tissue to normal tissue from the same organ, or histologically different types of cancer specimens), usually to identify genes expressed differently in the two types. We derive results describing the impact of pooling, technical replicates and dye-swap arrays on sample size requirements, and show how these calculations depend on the relative sizes of different sources of variability. We consider a variety of common types of experimental situations and designs that are used in dual-label microarray experiments, as well as the single-label case, and include discussion of procedures for controlling the false discovery rate and adjustments for small sample situations. These calculations are based on relatively simple yet realistic statistical models for the data, and provide straightforward sample size calculation formulae.

There has been relatively little work published in the microarray literature on determining the number of samples required for class comparison problems, or for the development of prognostic markers.

For interested readers, this work is reviewed as Supplementary data at `http://www.biostatistics.oupjournals.org`.

We focus here on two statistical goals: (1) class comparison, and (2) prognostic marker development, where the goal is to construct a multi-gene predictor of prognosis. These are common goals of microarray experiments in cancer research. We present equations for the sample size formulae because these give insight into the impact that variance parameters and experimental design have on sample size requirements that computational 'black box' algorithms or simulations sometimes do not. Oftentimes in practice, particularly in small sample situations, statistical software for determining sample size will be preferable to these approximate formulae because the software packages typically use computational algorithms to achieve better approximations to the true power and level, and hence better sample size estimates.

In class comparison, one is interested in identifying which genes are differentially expressed, and so powering the study in a way that focuses on inference for individual genes seems appropriate. When constructing a prognostic marker, even if the ultimate goal is to build a multivariate marker, a first step is typically to identify individual genes associated with disease outcome, so that again the study should be powered based on inference for individual genes. Multivariate approaches to sample size do not appear appropriate for these goals.

All sample size formulae presented here are based on an assumption of a normal linear model for each gene. We think this assumption is reasonably close to the truth for log-intensity data because many microarray experiments using this approach have identified biologically meaningful differential expression independently verified by other technologies. The power calculations based on a normal linear model assumption are further tested empirically in Section 9 of this paper and appear to be adequate.

2. DEFINITIONS AND NOTATION

We assume that all data have been background-corrected and normalized. In the case of Affymetrix data, we model the summary measure for a gene and not the individual PM and MM scores. We use a common notation for both single-label and dual-label microarray experiments. Y_{gadvfs} represents a fluorescent intensity; the subscript $g = 1, 2, \dots, G$ indexes the genes; the subscript $a = 1, \dots, n$ indexes the arrays or glass slides; the subscript d indexes the dye or dyes used, $d = 1$ for single-label microarrays and $d = 1, 2$ for dual-label microarrays; the subscript $v = 1, 2$ indexes the different phenotypes or varieties present, and for simplicity we will generally assume there are two; $f = 1, \dots, F$ indexes the biologically distinct samples (e.g. the different people in a human tumor experiment, or mice in a mouse model experiment) within each phenotype; we assume each phenotype is represented by an equal number of individuals; finally, $s = 1, \dots, m$ indexes the subsamples taken from the same biological source (e.g. a tissue sample from an individual may be cut into several specimens, and separate microarrays run on each specimen).

With each gene g will be associated two different levels of variation: (1) biological variation due to heterogeneity from individual to individual of gene expression within the phenotype will be denoted τ_g^2, and (2) experimental error variation due to technical inaccuracies in the microarray measurement will be denoted by σ_g^2.

In the sample size formulae, m will refer to the number of different subsamples measured from each sample (i.e. number of technical replicates run on each sample) and n will refer to the total number of microarrays used in the experiment. Also, $z_{\alpha/2}$ and z_β are the $100\alpha/2$th and 100βth percentiles of the normal distribution, respectively. Finally, δ is the distance between the class means which the study is being powered to detect.

A reference design dual-label experiment is a microarray experiment in which an aliquot from a single reference sample is applied to one channel on each microarray. A balanced block design dual-label experiment is a microarray experiment in which a single aliquot from each RNA sample is used, and samples from the different varieties are arranged in balanced block (Cochran and Cox, 1992, p. 376) layout with the individual microarrays serving as the blocks. A balanced block layout pairs a sample from each variety with a sample from every other variety together the same number of times over the microarrays.

3. Single-label microarrays

Some microarray systems, such as Affymetrix arrays, use a single label. Much work has been done on finding an adequate gene expression measure for these types of arrays (Li and Wong, 2001a; Irizarry *et al.*, 2003), and on developing methods for normalizing a set of microarrays to one another (Li and Wong, 2001b; Bolstad *et al.*, 2003; Chu *et al.*, 2002). We will not address these issues here. Instead, we will assume that a background-adjusted, normalized gene expression measurement Y_{gadvfs} is available for the genes on the arrays, which is described by the model

$$\log(Y_{gadvfs}) = G_g + GV_{gv} + (GF)_{gf(v)} + \varepsilon_{gadvfs} \tag{3.1}$$

where G_g is the average expression level of gene g in the overall population, GV_{gv} is the effect of the class or type, and GF_{gf} is the individual sample effect (from a particular biological individual), which is a random effect with variance τ_g^2 nested in the class effect, and ε_{gadvfs} represents independent, normally distributed error with gene-specific variance σ_g^2.

In the supplemental material (Section 1), we derive the formula for the number of single-label microarrays required when the variances are known, which is

$$n = 4m\left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m}\right) \tag{3.2}$$

where m is the number of technical replicates per sample, and n is the total number of microarrays required. The total number of biologically distinct samples required is

$$n/m = 4\left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m}\right). \tag{3.3}$$

An estimate of the quantity $\tau_g^2 + \frac{\sigma_g^2}{m}$ is needed. If no technical replicates are planned, so that $m = 1$, then the quantity $\tau_g^2 + \sigma_g^2$ which must be estimated is simply the variance of the log-intensities for this gene across samples of the same class. If some technical replicates are planned for each sample, so that $m > 1$, then the quantity $\tau_g^2 + \frac{\sigma_g^2}{m}$ is the variance of the average of m log-intensity measurements for this gene across samples of the same class; that is, if, for each individual, the gene expression is estimated by the average of m log-intensity measurements on arrays corresponding to that individual, then the variance of these estimates is the quantity of interest. If a lab routinely runs multiple technical replicates per sample, such an estimate may be available.

Data from a previous experiment will provide several thousand estimates of the quantity $\tau_g^2 + \frac{\sigma_g^2}{m}$, one for each gene on the array. In Section 7 we discuss several ways to construct a single estimate to be used for sample size planning.

4. DUAL-LABEL MICROARRAYS

We assume that the normalized, background corrected intensity measurements from a cDNA microarray experiment are described by the model

$$\log(Y_{gadvfs}) = G_g + GA_{ga} + GD_{gd} + GV_{gv} + (GF)_{gf(v)} + \varepsilon_{gadvfs} \tag{4.1}$$

(one spot per gene per array assumed, so that GA_{ga} specifies an unique spot on a single array). The parentheses around the GF term indicate that it is considered a random sample effect, nested in the class effect, with a distribution that is normal with zero mean and variance τ_g^2. We assume the measurement error may be gene-specific, so that ε_{gadvfs} are independent normal random variables with variance σ_g^2. The GD_{gd} dye-effect term is typically only included when it is estimable; for instance, in simple reference designs, it is absorbed into the G_g gene main effect.

This model for the dual-label microarray is equivalent to the model presented for the single-label microarray, as is shown in the supplementary material (Section 6), where a function mapping the equivalent terms in the two models is presented. Our general notation throughout is to use τ_g^2 for the biological variation and σ_g^2 for the technical error variation of the log-intensities; however, the interpretation of these population parameters will change with the context: for example, one would not expect the same technical error variation for oligonucleotide data as for cDNA data. Similarly, different experimental designs will affect the variances (Dobbin *et al.*, 2003a). Hence the context (single-label system, dual-label system, reference design, block design, paired samples design) must be kept in mind. Our goal here is not to discuss design selection issues. The reader is referred to Dobbin and Simon (2002), Dobbin *et al.* (2003a), and Kendziorski *et al.* (2003) for discussion of design selection.

Simple reference design

An approximate formula for the number of microarrays required when comparing two classes in a reference design with no technical replicates is

$$n = 4\left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_g^2 + 2\sigma_g^2\right) \tag{4.2}$$

for the total number of microarrays, with half the arrays assigned to each group. Here $\tau_g^2 + 2\sigma_g^2$ is the variance of the log-ratios for gene g within one of the types or classes of RNA samples, and will need to be estimated from previous data. See page 11 of the supplement for a bijective function relating the dual-label to the single-label model of the previous section.

If more than two classes are to be compared, then multiple comparison issues arise, and sample size determinations will depend on what type of error rates one wishes to control.

Technical replicates and dye swaps in reference designs

The formula for the number of samples required is derived in the supplement and is

$$n = 4m\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\left(\tau_g^2 + \frac{2\sigma_g^2}{m}\right). \tag{4.3}$$

The total number of biologically distinct samples required is n/m.

Unlike the sample size formula when no technical replicates are to be used, an estimate of the log-ratio variance will not be sufficient for the calculation, but separate estimates of both the biological variation

Table 1. *Effect of technical replicates on the number of arrays and samples required. Variance ratio is τ_g^2/σ_g^2. The mean and median of the variance ratios from a large human cancer data set with replicates were approximately 4 and 2 respectively (unpublished data). Settings were $\alpha = 0.001$, $\beta = 0.05$, $\delta = 1$, and $\tau_g^2/\sigma_g^2 = 0.5$. (In practice, the odd numbers would usually be rounded up so an equal number from each group taken.)*

Variance ratio	Number of technical replicates for each sample	Number of arrays required	Number of samples required
2.0	1	49	49
	2	74	37
	3	99	33
	4	124	31
4.0	1	49	49
	2	82	41
	3	114	38
	4	148	37

and the experimental error variation will be needed. This is because the relative sizes of the different sources of variation will determine the correlation between repeated measures on the same sample. Table 1 shows the impact of technical replicates both on the number of arrays required and on the number of samples required for some typical microarray design situations. The variance ratio is the ratio of biological variation to experimental error variation τ_g^2/σ_g^2, and typical estimates centering in the range of 2 to 10 have been given (Dobbin and Simon, 2002; Kendziorski *et al.*, 2003).

Similar calculations should be applicable or nearly applicable for dye-swap arrays, although some have suggested that dye swapping may improve the estimates by somewhat more than simple technical replication without dye swapping (Liang *et al.*, 2003). As can be seen from the table, there is some reduction in the number of samples required when each sample is assayed more than once, but this advantage generally comes at some cost in terms of significantly larger numbers of arrays that need to be run.

If n_m is the number of arrays required when m technical replicates of each sample are to be performed, then the relation between the required number of arrays to achieve equivalent power with one technical replicate versus m technical replicates per sample is

$$n_m = n_1 \left[\frac{m\left(\tau_g^2/\sigma_g^2\right) + 2}{\left(\tau_g^2/\sigma_g^2\right) + 2} \right] \tag{4.4}$$

(for derivation, see Section 3 of supplement). For instance, if $\frac{\tau_g^2}{\sigma_g^2} = 2$, then $n_2 = 1.5n_1$, and $n_3 = 2n_1$.

Balanced block design

Selecting a balanced block design may reduce the number of arrays required to achieve a desired power. The number of arrays required for a balanced block design is

$$n = \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} \left(\tau_{g,1}^2 + \tau_{g,2}^2 + 2\sigma_g^2\right) \tag{4.5}$$

where $\tau^2_{g,1}$ and $\tau^2_{g,2}$ are the variance in the log-intensities attributable to biological variation in gene expression in each population respectively. (By definition, the balanced block design does not use any technical replicates, so there is no m parameter in this equation.)

The variance of the log-ratios in a balanced block design will generally be larger than for a reference design, because the log-ratios manifest additional biological variation since no reference is repeated over the arrays. We have presented a conversion formula and examples for estimating the sample size required for a balanced block design from previous data from a reference design experiment (Dobbin *et al.*, 2003b).

Simple paired design and dye-swap paired design

By a paired design, we mean a design in which there is a natural pairing to the samples, such as samples from the same individual before and after some treatment, or samples from tumor and coexisting histologically normal tissue from the same organ. We do not mean the physical pairing of the cDNA samples on the same array, which is common to all dual-label microarray experiments. We discuss two paired designs: (1) by a balanced paired design we mean a design in which each sample pair is run together once on an array, and the arrays are balanced so that tumor is tagged with red dye on half the arrays and green on the other half; (2) by a dye swap paired design we mean a design in which each sample pair is run on two different arrays, and the labeling on the second array is reversed.

The number of arrays required for a balanced paired design with no dye-swap arrays is

$$n_{\text{balanced}} = \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\left(2\sigma_g^2 + \eta_g^2\right) \tag{4.6}$$

where $2\sigma_g^2+\eta_g^2$ is the variance of the log-ratios for the paired samples. We have switched notation from our earlier τ_g^2 to η_g^2 for the biological component of the variance because biological variation is conceptually much different with paired data; η_g^2 represents variation not in the expression level of the gene in the population, but in the effect the cancer has on the gene's expression level in the population. For instance, gene g may be up-regulated in tumor tissue compared to normal tissue, but the amount of up-regulation may vary from one individual to another; this variation is represented by η_g^2. If each sample is run twice, once with each dye labeling (often called a complete dye-swap), then the number of arrays required for this dye swap paired design is

$$n_{\text{dyeswap}} = \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\left(2\sigma_g^2 + \eta_g^2\right). \tag{4.7}$$

Note that, $1 \leqslant \dfrac{n_{\text{dyeswap}}}{n_{\text{balanced}}} \leqslant 2$ where n_{dyeswap} is the number of arrays required to achieve a specified power and efficiency under a dye swap design, and n_{balanced} is the number of samples required to achieve the same power and efficiency using a balanced design with no dye swaps. It makes intuitive sense that this ratio would be at most 2 because a balanced paired design requires exactly the same number of biologically distinct samples as a dye swap paired design with twice as many arrays, and one cannot lose information by measuring the same samples a second time.

5. SAMPLE SIZE FOR DEVELOPING PROGNOSTIC MARKERS

So far we have focused on sample size for the goal of class comparison, where one typically wants to identify genes expressed differentially in the classes. Another common goal in microarray studies is to develop predictors of patient prognosis from the expression of a key gene or collection of genes. In this case one does not have classes for the patients that have been predetermined, but one does typically

Table 2. *Number of arrays and samples required for various pooling levels. An independent pool is constructed for each array, so that no sample is represented on more than one array. Settings were same as Table 1:* $\alpha = 0.001$, $\beta = 0.05$, $\delta = 1$, *and* $\tau_g^2 + 2\sigma_g^2 = 0.5$. *Variance ratio is* τ_g^2/σ_g^2

Variance ratio	Number of samples pooled on	Number of arrays required	Number of samples required
2.0	1	49	49
	2	37	74
	3	33	99
	4	31	124
4.0	1	49	49
	2	33	66
	3	28	84
	4	25	100

have some continuous, right-censored clinical endpoint that has been recorded, such as time to disease recurrence or time to death. We develop formulae for predictive and prognostic markers, but not surrogate markers, which present different issues.

The number of samples required in this case (Simon *et al.*, 2002) is

$$D = \frac{(z_{\alpha/2} + z_\beta)^2}{\gamma_g \ln[h])^2}$$

where γ_g denotes the standard deviation of the log ratio for dual-label arrays, or log-intensity for single-label arrays, of the gene over the entire set of samples, h denotes the hazard ratio associated with a one-unit change in the log-ratio (or log-intensity), and ln denotes the natural logarithm. For instance, if gene expression is measured in base 2 logs, then h is the change in hazard ratio associated with a twofold change in gene expression. If we have $\gamma_g = 0.5$, as may be realistic for human data, and $\alpha = 0.001$, $\beta = 0.05$ then the sample size required to detect a change in hazard ratio of 2, 3 and 4, is 203, 81 and 51, respectively.

6. Effect of pooling on sample size requirements

Under assumptions similar to those in Kendziorski *et al.* (2003), if k independent biological samples are pooled together and applied to the arrays, then in the supplementary material we show that the sample size formula for a reference design is

$$n = 4m\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\left(\frac{\tau_g^2}{k} + \frac{2\sigma_g^2}{m}\right). \tag{6.1}$$

In order to use this formula, one needs estimates of the size of the different sources of variability.

As can be seen from Table 2, pooling can result in a smaller number of arrays required, but this comes at the cost of requiring a potentially much larger number of samples. Since it is very common in cancer research that obtaining new samples—either by completely re-running an experiment or obtaining new biological specimens (depending on the context)—is fairly expensive in terms of time and effort and cost, this table indicates that pooling will usually not result in a good tradeoff between reduced arrays and increased time and expense related to sample acquisition. Further discussion of pooling issues appears in Section 9 of the supplemental material.

7. SELECTION OF THE SIGNIFICANCE LEVEL AND POWER

In the sample size formulae we have presented, the user must provide the two parameters α and $1-\beta$. α is the probability of finding a gene is differentially expressed when, in fact, it is not. For instance, if genes are independent and there are 10 000 genes, none of which are truly differentially expressed, then setting $\alpha = 0.001$ means that on average there will be 10 false positives, i.e. 10 genes will be found to be differentially expressed. If genes are not truly independent, then the expected number of false positives will still be 10, as we show in the supplemental material (Section 2), but the distribution of the number of false positives may be different than under the independence assumption. Setting $1-\beta = 0.95$ means that on average for a gene with differential expression δ, one will have 95% probability of a true discovery, i.e. 95% of the time the gene will be found differentially expressed. For instance, in a reference design cDNA experiment, a $\delta = 1$ on the log base 2 scale would correspond to a twofold change in gene expression between the classes; if there are 20 genes with twofold differential expression (and we assume genes are independent), then the probability of identifying all 20 will be approximately 36%, and the probability of identifying 18 or more will be approximately 92%.

Setting α very small controls the expected number of false-positive genes. This type of control could be seen as inadequate for at least two reasons:

1. The resulting procedure may be so strict that in practice it will require a very large number of arrays; one may prefer to sift through a certain proportion of false-positive genes rather than take on the expense and time required in running arrays;
2. It may not be adequate to control the 'expected' number of false-positives, since this does not eliminate the possibility that one may greatly exceed this number, depending on the variation in the false-positive rate.

To address issue 1, often procedures for controlling the false discovery rate, rather than the false positive rate, are adopted. When some discoveries are made, the false discovery rate is defined as the proportion of discoveries, in this context genes identified as differentially expressed, which are, in fact, not differentially expressed. This is the proportion of mistakes in the gene list.

A false discovery is a gene that is determined (by hypothesis test) to be differentially expressed when, in fact, it is not; a true discovery is a gene that is determined to be differentially expressed when, in fact, it is. The false discovery rate can be written $FDR = E\left[\frac{\#FD}{\#FD + \#TD}\right]$ (Benjamini and Hochberg, 1995)[†] where $\#FD$ is the number of false discoveries and $\#TD$ is the number of true discoveries (and the FDR is defined to be zero when $\#FD+\#TD = 0$). Suppose that π is the proportion of genes that are differentially expressed. To simplify the calculation and keep the argument intuitive, we assume that each gene falls into one of two categories: (1) either it is not differentially expressed, or (2) it is differentially expressed by some fixed amount δ. Then the expected number of false discoveries is $E[\#FD] = \alpha(1-\pi)G$ where G is the number of genes. The expected number of true discoveries is $E[\#TD] = (1-\beta)\pi G$. This suggests an approximation based on a first-order Taylor series expansion: $E[FDR] \approx \frac{\alpha(1-\pi)}{\alpha(1-\pi) + (1-\beta)\pi} = \left\{1 + \left(\frac{1-\beta}{\alpha}\right)\left(\frac{\pi}{1-\pi}\right)\right\}^{-1}$. The estimate will have some small bias because the function is convex (Billingsley, 1995, p. 80). However, we used Monte Carlo to test the approximation (see supplementary

[†] Benjamini and Hochberg prefer the alternative definition $P(\#FD + \#TD > 0)E\left[\frac{\#FD}{\#FD + \#TD} | \#FD + \#TD > 0\right]$. However, since microarray experiments typically involve tens of thousands of genes measured simultaneously in different conditions, it seems reasonable to assume that $P(\#FD + \#TD = 0) = 0$, in which case the two FDR definitions are equivalent.

Table 3. *False discovery rate and expected false-negative rate for identifying differentially expressed genes for a number of choices of size α and power $1-\beta$, and a variety of proportions π of truly differentially expressed genes*

π	α	$1-\beta$	$\hat{E}[FDR]$	Expected number of FD in 10 000 genes	Number of truly differentially expressed genes	Expected number of truly differentially expressed genes missed
0.005	0.001	0.95	0.17	8.5	50	2.5
0.005	0.001	0.90	0.18	9	50	5
0.005	0.001	0.80	0.20	10	50	10
0.05	0.001	0.95	0.02	10	500	25
0.05	0.001	0.90	0.02	10	500	50
0.05	0.001	0.80	0.02	10	500	100
0.20	0.001	0.95	0.004	8	2000	100
0.20	0.001	0.90	0.004	8	2000	200
0.20	0.001	0.80	0.005	10	2000	400
0.005	0.01	0.95	0.68	34	50	2.5
0.005	0.01	0.90	0.69	35	50	5
0.005	0.01	0.80	0.71	36	50	10
0.05	0.01	0.95	0.17	85	500	25
0.05	0.01	0.90	0.17	170	500	50
0.05	0.01	0.80	0.19	340	500	100
0.20	0.01	0.95	0.04	20	2000	100
0.20	0.01	0.90	0.04	40	2000	200
0.20	0.01	0.80	0.05	100	2000	400
0.005	0.005	0.95	0.51	25.5	50	2.5
0.005	0.005	0.90	0.53	26.5	50	5
0.005	0.005	0.80	0.55	27.5	50	10
0.05	0.005	0.95	0.09	45	500	25
0.05	0.005	0.90	0.10	50	500	50
0.05	0.005	0.80	0.11	55	500	100
0.20	0.005	0.95	0.02	40	2000	100
0.20	0.005	0.90	0.02	40	2000	200
0.20	0.005	0.80	0.02	40	2000	400

material, Section 7), and the approximation is very good over the range of values that are likely to be used for the parameters.

Table 3 shows the approximate expected false discovery rate for a variety of values of π, α and β. As can be seen from the table, the expected FDR is most sensitive to the proportion of genes that are truly differentially expressed, π, and the significance level α; variation in the selection of the statistical power $1-\beta$ is less critical.

Using $\alpha = 0.001$ is one simple method for controlling the expected number of false discoveries. Other methods for controlling the false discovery rate have also been proposed for microarray data (Efron *et al.*, 2001; Reiner *et al.*, 2003). A step-down procedure for controlling the actual number or proportion of false discoveries has also been proposed (Korn *et al.*, 2003).

Table 4. *Simple formula Z-based sample sizes: normal approximation formula is inadequate for small sample sizes, resulting in poor power under the Monte Carlo simulation. Total of 100 000 Monte Carlo simulations used for each row. Significance level was* 0.001 *for all.* γ^2 *is the within-group variance, e.g. of the log-ratios in a dual-label reference design or log-intensities in a single-label design. Monte Carlo results are presented as: mean (sd)*

γ^2	Nominal power	N per group	Monte Carlo power
0.25	0.90	3	0.07 (0.002)
0.25	0.95	4	0.33 (0.002)
0.50	0.90	11	0.76 (0.002)
0.50	0.95	13	0.88 (0.002)
0.75	0.99	36	0.98 (0.002)

8. SELECTION OF THE VARIANCE ESTIMATE FOR THE SAMPLE SIZE CALCULATION

In all of the sample size formulae presented here, some estimate of variation is required for the calculation. What is usually required is an estimate of the variance of the log-ratios for a dual-label experiment, or of the log-intensities for a single-label experiment. The best way to determine these variance estimates is by looking at data from previous experiments that were similar to the one that is being planned. This may necessitate some thoughtful consideration of how the two experiments are the same and how they differ. Since the larger the variance, the larger the sample size requirement, the most conservative sample sizes will be those calculated from variance estimates from genes displaying the greatest variation. Typical rules of thumb would be to use the median, upper quartile or 90th percentile (Yang and Speed, 2002) of the variances from a previous similar experiment in the sample size calculations. In our experience with reference designs and cDNA arrays, we find that a median base 2 log-ratio standard deviation of 0.5 is fairly typical for human reference design data, and of 0.25 is fairly typical for inbred mice strains.

9. SMALL *n* CASES

As shown in Table 4, the sample size formulae we have presented are not appropriate for small sample sizes. This is an important point in applications because many microarray studies are limited to small sample sizes due to expense, time or logistics. The problems with the sample sizes formulae for small numbers of samples are that: (1) the formulae assume that the variance is known, but with small sample sizes even the estimated variance may be unreliable; (2) the distribution under an alternative hypothesis is non-central t, and may differ significantly from the usual normal approximation used in the explicit formulae. If the sample is large (for instance more than 30 arrays), then the sample variance will be a good estimate of the population variance and the sample size formulae may be adequate. But if the sample is small, the sample size will be inadequate to achieve the power desired. Statisticians have long recognized this problem and developed iterative procedures for obtaining adequate sample size in small sample situations. There are numerous software packages available for this purpose that can be used to correct this problem.

There are other ways to try to address the problem of poor estimation of variances when the sample size is small. In microarray studies, since many genes are being measured in parallel, trying to borrow information about variances across genes can potentially improve the estimate of the variance. Some have modeled the variation across the different genes parametrically in such a way that one can borrow information about the variance for a particular gene from other genes (Baldi and Long, 2001; Wright and Simon, 2003). The 2003 paper provides corrected degrees of freedom and sum of squares formulae that can be used to adjust the calculation of the number of arrays required.

Table 5. *Resampling study of impact of normality assumption or true power and size based on data of Rosenwald* et al. *(2002). Results using the pooled variance two-sample t-test, sample sizes rounded up to even number, minimum sample size is four (two per group), sample size determined by t iterative procedure. Observed power and level based on 1,000 resamplings for each of 7,399 genes. Results presented as Mean(SD)*

Mean shift	Fold change	Alpha level	Observed level	1−Beta power	Observed power	Mean sample size
0.5849625	1.5	0.001	0.0009 (0.001)	0.90	0.903 (0.043)	133.3
0.5849625	1.5	0.05	0.052 (0.009)	0.90	0.911 (0.021)	68.9
1	2	0.001	0.001 (0.001)	0.90	0.911 (0.027)	50.6
1	2	0.05	0.060 (0.010)	0.90	0.928 (0.022)	25.7
2	4	0.001	0.001 (0.001)	0.90	0.921 (0.033)	16.8
2	4	0.05	0.075 (0.012)	0.90	0.953 (0.026)	8.8

10. Impact of normality assumption

The sample size calculations presented here are based on assumptions of the normal linear model such as the one given in Equation (3.1). We tested the validity of the sample size calculations based on this model using the dataset Rosenwald *et al.* (2002). For each of the 7399 genes, we used Equation (4.2), with the small sample adjustment discussed in Section 8, to calculate sample size requirements for various power and significance level combinations. Then we formed samples of this size 1000 times by sampling with replacement from the cases with no missing data for that gene. Each sample was divided in half to form two groups, and to each observation in one group was added a constant δ. Then the pooled variance t-test was performed on the groups, and the power estimated from these. Computations were carried out in R and C++. Results are shown in Table 5. In general, both the observed power and the observed type I error rate tend to be somewhat higher than the nominal, and decrease towards the nominal as the sample size increases. This indicates that the sample size formulae presented in this paper should be adequate for microarray data such as this.

11. Conclusion

We have presented sample size formulae for class comparison problems for a variety of microarray platforms and experimental designs. We have also presented a sample size formula for prognostic studies. In general, determination of the number of microarrays required for an experiment does not necessitate ‘reinventing the wheel’ in the sense that much of the classical statistical sample size reasoning can be applied in this new context; on the other hand, the context of microarray experiments is novel enough that it raises interesting questions such as how the relative sources of variation will be expected to impact sample size requirements, and how design decisions such as pooling, technical replicates, and dye-swaps will impact costs associated with the arrays. We have addressed these issues in an intuitive way which we hope will be of assistance in guiding researchers in future study design.

Acknowledgments

We would like to thank a referee and the editors for helpful comments on the manuscript.

References

Baldi, P. and Long, A. G. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**, 289–300.

BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd Edition. New York: Wiley.

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

CHU, T. M., WEIR, B. AND WOLFINGER, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* **176**, 35–51.

COCHRAN, W. G. AND COX, G. M. (1992). *Experimental Designs*, 2nd Edition. New York: Wiley.

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

DOBBIN, K., SHIH, J. H. AND SIMON, R. (2003a). Statistical design of reverse dye microarrays. *Bioinformatics* **19**, 803–810.

DOBBIN, K., SHIH, J. H. AND SIMON, R. (2003b). Questions and answers on statistical design of dual-label microarray experiments for identifying differentially expressed genes. *Journal of the National Cancer Institute USA* **95**, 1362–1369.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. AND TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. AND SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

KENDZIORSKI, C. M., ZHANG, Y., LAN, H. AND ATTIE, A. D. (2003). The efficiency of pooling mRNA in microarray experiments. *Biostatistics* **4**, 465–477.

KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. AND SIMON, R. (2003). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, in press.

LI, C. AND WONG, W. H. (2001a). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences USA* **98**, 31–36.

LI, C. AND WONG, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2**, research0032.1–research0032.11.

LIANG, M., BRIGGS, A. G., RUTE, E., GREENE, A. S. AND COWLEY JR., A. W. (2003). Quantitative assessment of the importance of dye switching and biological replication in cDNA microarray studies. *Physiological Genomics* **14**, 199–207.

REINER, A., YEKUTIELI, D. AND BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. R. D., MULLER-HERMELINK, H. K., SMELAND, E. B. AND STAUDT, L. M. FOR THE LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 1937–1947.

SIMON, R., RADMACHER, M. D. AND DOBBIN, K. (2002). Design of studiesusing DNA microarrays. *Genetic Epidemiology* **23**, 21–36.

WRIGHT, G. AND SIMON, R. (2003). A random variance model for detection of differential gene detection in small microarray experiments. *Bioinformatics* **19**, 2448–2455.

YANG, Y. H. AND SPEED, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews: Genetics* **3**, 579–588.

[*Received January 13, 2003; revised May 14, 2004; accepted for publication May 17, 2004*]