at the perfect match probe minus the intensity at the mismatched paired probe may be a better estimate of the intensity due to hybridization to the true target transcript.

Current GeneChips™ use 11 to16 probe pairs for each target gene but the lengths of the probes are smaller than for cDNA arrays. The differences in perfect-match minus mismatch intensities are averaged across the probe pairs to give an estimate of intensity of hybridization to the target transcript; see Section 4.3.

2.6 Other Microarray Platforms

Several companies such as Protogene (Menlo Park, CA) and Agilent Technologies (Palo Alto, CA) in collaboration with Rosetta Inpharmatics (Kirkland, WA) have developed methods of in situ synthesis of oligonucleotides on glass arrays using ink-jet technology that does not require photolithography. The ink-jet technology of Agilent can also be used to attach pre synthesized DNA probes to glass slides.

Another class of DNA microarrays utilizes cDNA probes printed on a nylon membrane, and radioactive labeling of the sample. The radioactive label provides a stronger signal than fluorescent dye. This is useful when the amount of mRNA available for labeling is limited, but the wide scattering of label limits the density of probes that can be printed on the array, and larger format arrays are necessary. Although most of the principles of experimental design and analysis apply equally to arrays using radioactively labeled samples as to arrays using fluorescent labels, we generally talk in terms of the latter.

Design and Analysis of Microarray Investigations

Ch. 3 Simon et al. 2004

3 Design of DNA Microarray Experiments

3.1 Introduction

Microarray based experiments, like all experiments, should be carefully planned. Careful planning begins with a clear objective. The objective drives the selection of specimens and the specification of an appropriate analysis strategy. It is a common misconception that microarray experiments do not require planning or objectives; in this view, expression profiles are placed in a pattern recognition blackbox and discoveries emerge. Although pattern recognition algorithms have a role for some objectives involving microarrays, most successful microarray-based experiments have a definite focus.

There is substantial confusion about the role of "hypothesis testing" in studies using microarrays. It is true that microarray-based research is generally not based on a mechanistic biological hypothesis focused on specific genes. Other technologies are more suitable for testing hypotheses about specific genes. Nevertheless, most good microarray experiments are based on a hypothesis. For example, the hypothesis might be that there are genes whose expression is up-regulated or down-regulated in a tumor compared to normal tissue of the same tissue type. Or, the hypothesis might be that different tumors of the same tissue type and the same stage are not homogeneous with regard to gene expression profiles. Clearly identifying the general hypothesis of the study is important for ensuring that the type and number of specimens collected are appropriate. Clarity on the general hypotheses is also important for selecting methods of data analysis. A DNA microarray is just a highly parallel assay. It does not herald an era in which good practices of carefully thinking about the objectives of the experiment and of carefully planning the experiment and its analysis are obsolete.

Because DNA microarray investigations are not focused on a prespecified gene-specific hypothesis, there is much more opportunity for spurious findings than with more traditional types of investigations. Although the contexts in which microarrays are used are exploratory, strong claims are often made about which genes are differentially expressed under specified condi-

tions, which are disregulated in diseased tissue, and which are predictive of response to treatment. The serious multiplicity problems inherent in examining expression profiles of tens of thousands of genes mandate careful planning and special forms of analysis in order to avoid being swamped by spurious associations.

Design issues can be divided into those relating to the design of the DNA microarray assay itself and issues involving the selection, labeling, and arraying of the specimens to be assayed. In this chapter, we focus on the latter issues. Section 3.2 describes the importance of defining the study objectives for designing a microarray study. Section 3.3 discusses the difficulties in satisfying study objectives when only two RNA samples are compared. The sources of variation and the levels of replication of the experiment, discussed in Section 3.4, are important to consider when designing a study. Section 3.5 discusses the possibility of pooling samples and assaying the pooled sample with a microarray. With dual-label microarrays, the different ways of pairing and labeling the samples are discussed in Sections 3.6 and 3.7, respectively. The chapter ends with a discussion of the sample sizes required to meet the study objectives.

3.2 Study Objectives

DNA microarrays are useful in a wide variety of investigations with a wide variety of objectives. Many of these objectives fall into the following categories.

3.2.1 Class Comparison

Class comparison focuses on determining whether gene expression profiles differ among samples selected from predefined classes and identifying which genes are differentially expressed among the classes. For example, the classes may represent different tissue types, the same tissue under different experimental conditions, or the same tissue type for different classes of individuals. In cancer studies, the classes often represent distinct categories of tumors differing with regard to stage, primary site, genetic mutations present, or with regard to response to therapy; the specimens may represent tissue taken before or after treatment or experimental intervention. There are many study objectives that can be identified as class comparison. The defining characteristic of class comparison is that the classes are predefined independently of the expression profiles. Many studies are performed to compare gene expression for several types of class definition. For example, two genotypes of mice may be studied under two different experimental conditions. One analysis may address differences in gene expression for the two types of animals under the same experimental condition and the other analysis may address the effect of the experimental intervention on gene expression for a given genotype.

3.2.2 Class Prediction

Class prediction is similar to class comparison except that the emphasis is on developing a statistical model that can predict to which class a new specimen belongs based on its expression profile. This usually requires identifying which genes are informative for distinguishing the predefined classes, using these genes to develop a statistical prediction model, and estimating the accuracy of the predictor. Class prediction is important for medical problems of diagnostic classification, prognostic prediction, and treatment selection.

3.2.3 Class Discovery

Another type of microarray study involves the identification of novel subtypes of specimens within a population. This objective is based on the idea that important biological differences among specimens that are clinically and morphologically similar may be discernible at the molecular level. For example, many microarray studies in cancer have the objective of developing a taxonomy of cancers that originate in a given organ site in order to identify subclasses of tumors that are biologically homogeneous and whose expression profiles either reflect different cells of origin or other differences in disease pathogenesis (Alizadeh et al. 2000; Bittner et al. 2000). These studies may uncover biological features of the disease that pave the way for development of improved treatments by identification of molecular targets for therapy.

3.2.4 Pathway Analysis

The objective of some studies is the identification of genes that are coregulated or which occur in the same biochemical pathway. One widely noted example is the identification of cell cycle genes in yeast (Spellman et al. 1998). Pathway analysis is often based on performing an experimental intervention and comparing expression profiles of specimens collected before and at various time intervals after the experimental intervention. In some cases, however, pathway analysis may involve comparing the wild type organism to genetically altered variants.

3.3 Comparing Two RNA Samples

The initial cDNA microarray studies involved the cohybridization of one mRNA sample labeled with one fluorescent dye and a second mRNA sample labeled with a second fluorescent dye on a single microarray (DeRisi et al. 1996). This type of study, and the high cost of microarrays, left many investigators hoping and believing that no replication was needed. It also led to the publication of a variety of statistical methods for comparing the expression levels in the two channels at each gene on a single microarray. Even today,

Affymetrix software is designed to compare gene expression on just two arrays (one sample on each array) and to compare two classes of specimens one must compare the specimens two at a time (Affymetrix 2002).

The main problems with drawing conclusions based on comparing two RNA samples apply to both dual-label and Affymetrix arrays. First, the relative intensity for a given gene in the two specimens can reflect an experimental artifact in tissue handling, cell culture conditions, RNA extraction, labeling, or hybridization to the arrays that is not removed by the normalization process. The analysis of two RNA samples each arrayed once provides very little evidence that if the same two samples were rearrayed the results would be similar.

. Even more important, the conclusions derived from comparing two RNA samples, even if they are arrayed on replicate arrays, apply only to those two samples and not to the tissues or experimental conditions from which they were derived. For example, in comparing two RNA samples, none of the biological variability is represented. In comparing expression profiles of tumors of one type to tumors of another type, there is generally substantial variation among tumors of the same class (e.g., Hedenfalk et al. 2001). There may even be substantial variation in expression within a single tumor. Hence, comparison of one RNA sample from one tumor of the first type to one RNA sample from one tumor of the second type is not adequate. In comparing tissue from inbred strains of mice, the biological variability is generally less than for human tissue but some biological replication is still necessary. Even for comparing expression of a cell line under two conditions, there is biological variability resulting from variation in experimental conditions, growth and harvest of the cells, and extraction of the RNA. Hence some replication of the entire experiment is important. This is discussed further in the next section.

3.4 Sources of Variation and Levels of Replication

Some important sources of variation in microarray studies can be categorized as

- between individuals within the same "class" or between complete replication of tissue culture experiments under the same experimental conditions;
- between specimens from the same individual or same experiment;
- between RNA samples from the same specimen;
- between arrays for the same RNA sample;
- between replicate spots on the same array.

Replicate arrays made from the same sample of RNA are often called *technical replicates*, in contrast to *biological replicates* made from RNA from biologically independent samples (Yang and Speed 2002b). There are, however, several levels of biological replicates.

Suppose we wish to determine gene expression differences between breast tumors with a mutated BRCA1 gene and tumors without a mutation. If we performed array experiments on one breast tumor with a BRCA1 mutation and one without a mutation we would not be able to draw any valid conclusions about the relationship of BRCA1 mutations to gene expression because we have no information about the natural variation within the two populations being studied. The situation would not improve even if the tumors under investigation were large enough for us to be able to perform multiple mRNA extractions and run independent array hybridizations on each extraction. Sets of tumors representative of the BRCA1 mutated population and the non-BRCA1 mutated population are necessary to draw valid conclusions about the relationship of BRCA1 mutations to gene expression.

There is sometimes confusion with regard to the level of replication appropriate for microarray studies. For example, in comparing expression profiles of BRCA1 mutated tumors to expression profiles of non-BRCA1mutated tumors, it is not necessary to have replicate arrays of a single RNA sample extracted from a single biopsy of a single tumor. Having such replication may provide protection from having to exclude the tumor if the one array available is of poor quality, but such replications are merely assay replicates and do not satisfy the crucial need for studying multiple tumors of each type. Often the biological variation between individuals will be much larger than the assay variation and it will be inefficient to perform replicate arrays using specimens from a small number of individuals rather than performing single arrays using a larger number of individuals.

In comparing expression profiles between two cell lines, or for a given cell line under different conditions, the concept of "individual" may be unclear. Suppose, for example, we wish to compare the expression profile of a cell line before treatment to the expression profile after treatment. Cell lines change their expression profiles depending on the culture conditions. Growing the cells and harvesting the RNA under "fixed conditions" will result in variable expression profiles because of differences in important factors such as the confluence state of the culture at the time of cell harvesting. Consequently, it is important to have independent biological replicates of the complete experiment under each of the conditions being compared. The degree of variation between independent biological samples may be less for experiments involving cell lines or inbred strains of model species compared to those involving human tissue samples, and this will influence the number of biological samples required as described in Section 3.8.

In some cases it is useful to obtain two specimens from the same individual. For example, if you are attempting to discover a new taxonomy of a disease based on an expression profile, it is useful to establish that the classification is robust to sampling variation within the same individual. For many studies of human tissue, however, the tissue samples will not be large enough to provide multiple specimens for independent processing. It is important to note that there is a distinction between multiple specimens from the same

individual and multiple independently labeled aliquots of one RNA sample. The latter will show less variability than the former, especially when the tissue is heterogeneous. However, even without tissue heterogeneity, variation may be observed among expression profiles of multiple specimens taken from the same individual because of differences in tissue handling and RNA extraction.

Performing technical replicate arrays with independently labeled aliquots of the same RNA provides information about the reproducibility of the microarray assay, that is, the reproducibility of the labeling, hybridization, and quantification procedures. It is useful to know that the reagents, protocols and procedures used provide reproducible results on aliquots of the same RNA sample. Generally, it will be sufficient to obtain such technical replicates on just a few RNA samples. Serious attention should be devoted to reduce technical variability in a study. If possible, RNA extraction, labeling, and hybridization of all arrays in an experiment should be performed by the same individual using the same reagents. If spotted arrays are used, it is desirable to use arrays from the same print set and certainly the same batch of internal reference RNA. If samples become available at different times in a long-term study, it is best to save frozen specimens so that all of the array assays can be done at approximately the same time.

When technical replicate arrays of the same RNA samples are obtained, they can be averaged to improve precision of the estimate of the expression profile for a given RNA sample. If reproducibility is poor, however, it may be preferable to discard technically inferior arrays rather than average replicates. Although averaging of replicates may seem ad hoc, analysis of variance methods also average replicates although they account for the differences in precision available for different samples based on their possibly varying number of replicates. Replicate arrays of the same RNA samples are also sometimes used in dye-swap experimental designs described later in this chapter.

3.5 Pooling of Samples

Some investigators pool samples in the hope that through pooling they can reduce the number of microarrays needed. For example, in comparing two tissue types, a pool of one type of tissue is compared to a pool of the other tissue type. Replicate arrays might be performed on each pooled sample. Although the pooled sample approach may be applicable for preliminary screening, the approach does not provide a valid basis for biological conclusions about the types of tissues being compared. If only one array of each pooled sample is prepared, then even the two pools cannot be validly statistically compared because there is no estimate of the variability associated with independently labeling and hybridizing the same pool onto different arrays. Even if the two pools are hybridized to replicate arrays, one cannot assess the variability among pools of the same type and so one doesn't know how adequate a pool of that number of RNA specimens is in reflecting the population of that

tissue type. Unless multiple biologically independent pools (of distinct specimens) of each type are arrayed, only the pooled samples themselves can be compared, not the populations from which they were derived. Biological replication is necessary. It can be achieved either by assaying individual samples, or by assaying independent pools of distinct samples. Studying independent pools of samples would be necessary in studying small model species where it may be necessary to pool in order to obtain enough RNA for assay (Jin et al. 2001).

3.6 Pairing Samples on Dual-Label Microarrays

With Affymetrix GeneChips™, single samples are labeled and hybridized to individual arrays. Spotted cDNA arrays, however, generally use a dual-label system in which two RNA samples are separately labeled, mixed, and hybridized together to each array. When using dual-label arrays one must decide on a design for pairing and labeling samples.

3.6.1 The Reference Design

The most commonly used design, called the reference design, uses an aliquot of a reference RNA as one of the samples hybridized to each array. This serves as an internal standard so that the intensity of hybridization to a probe for a sample of interest is measured relative to the intensity of hybridization to the same probe on the same array for the reference sample. This relative hybridization intensity produces a value that is standardized against variation in size and shape of corresponding spots on different arrays. Relative intensity is also automatically standardized with regard to variation in sample distribution across each array inasmuch as the two samples are mixed and therefore distributed similarly. The measure of relative hybridization generally used is the logarithm of the ratio of intensities of the two labeled specimens at the probe. Figure 3.1 is taken from Brody et al. (2002) who cohybridized labeled RNA from C2C12 myoblast cells and from 10T1/2 fibroblasts on an array that contained 100 spots for the glycerol-3-phosphate dehydrogenase gene. The figure shows the vast range of intensities among spots printed with the same clone on the same array. The ratio of intensities for the two samples, however, has little variation as evidenced by the tight linear association.

The reference design is illustrated in Figure 3.2. Generally, the reference is labeled with the same dye on each array. Any gene specific dye bias not removed by normalization affects all arrays similarly and does not bias class comparisons. Using a reference design, any subset of samples can be compared to any other subset of samples. Hence the design is not dependent on the specification of a single type of class comparison. For example, in studying BRCA1 mutated and BRCA1 nonmutated tumors, one might be interested in comparing samples based on their mutation status, comparing samples based

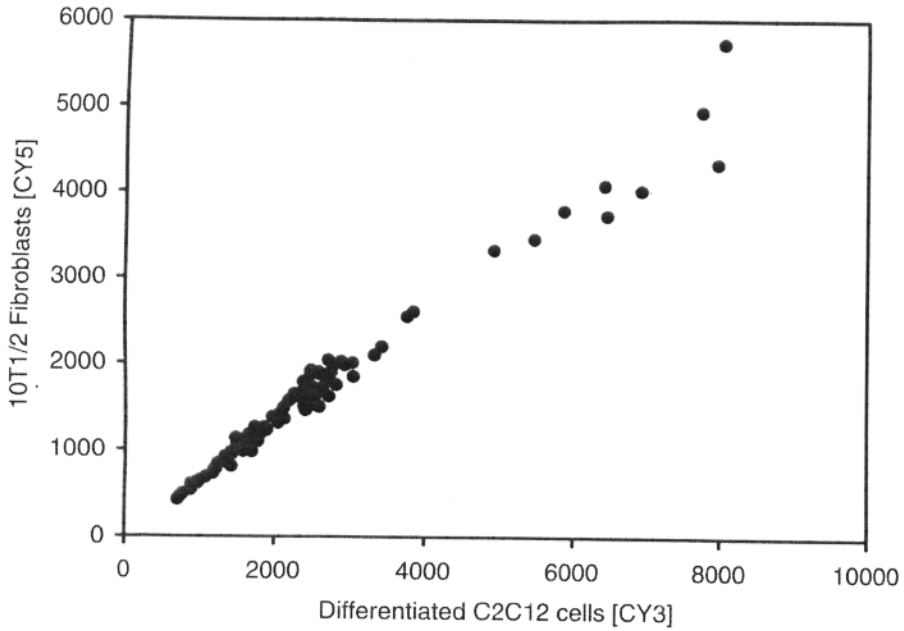

Fig. 3.1. Intensities of labeled RNA from C2C12 myoblast cells and labeled RNA from 10T1/2 fibroblasts hybridized to one array containing 100 spots of the glycerol-3-phosphate dehydrogenase gene. The figure shows the vast range of intensities among spots printed with the same clone on the same array. The ratio of intensities for the two samples, however, has little variation. From Brody et al. (2002).

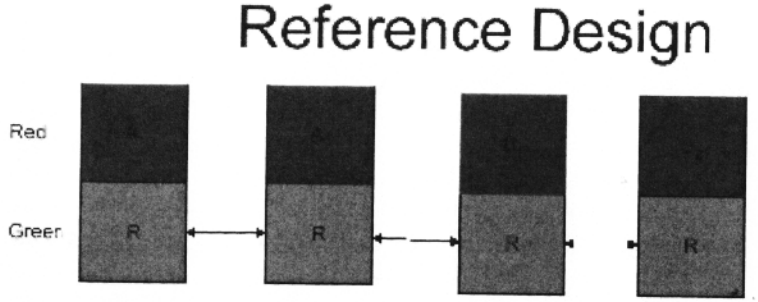

Fig. 3.2. Reference design. Aliquot of reference sample is labeled with the same label and used on each array.

on their estrogen receptor status, or comparing samples based on the stage of disease of the patient. The reference design is also convenient for class discovery using cluster analysis because the relative expression measurements are consistently measured with respect to the same reference sample.

If a laboratory uses reference designs with the same reference sample for all of their arrays, even those for different experiments, then all of their expression profiles can be directly compared. Consequently, expression signatures of different tissues studied in different experiments can be compared. This latter advantage can even extend to comparisons of expression profiles made by different laboratories using reference designs with the same reference sample.

There is sometimes confusion about the role of the reference sample. Some investigators erroneously believe that analysis is always based on combining single array determinations of whether the Cy5 (red) label is differentially expressed compared to the Cy3 (green) label for a given spot on a given array. Therefore they assume that the reference sample must be biologically relevant for comparison to the nonreference samples. In fact, the reference sample does not need to have any biological relevance. The analysis will usually involve quantitative comparisons of the average logarithm of intensity ratios for one set of arrays to average log ratios for another set of arrays.

It is desirable that most of the genes be expressed in the reference sample but not expressed at so high a level as to saturate the intensity detection system. Often, the reference sample consists of a mixture of cell lines so that nearly all genes will be expressed to some level. It is also important that a single batch of reference RNA is used for all arrays in a reference design. Different batches of reference RNA may have quite different expression profiles. When assaying samples collected over a long period of time, it is generally best to freeze the RNA samples and to perform the microarray assays at one time when all reagents can be standardized.

3.6.2 The Balanced Block Design

A disadvantage of the reference design is that half of the hybridizations are used for the reference sample, which may be of no real interest. Balanced block designs (Dobbin and Simon 2002) are alternatives that can be used in simple situations. For example, suppose one wished to compare BRCA1 mutated breast tumors to BRCA1 nonmutated breast tumors, that equal numbers of each tumor were available and that no other comparisons or other analyses were of interest. One could hybridize on each array one BRCA1 mutated tumor sample with one nonmutated sample. On half of the arrays the BRCA1 mutated tumors should be labeled with the red dye and on the other half the nonmutated tumors should be labeled with the red dye. This block design is illustrated in Figure 3.3. The analysis of data for the block design is discussed in Section 7.9. In its simplest form, a paired value t-test or Wilcoxon signed-rank test is performed for each gene, pairing the samples cohybridized to the same array. The block design can accommodate n samples of each type using

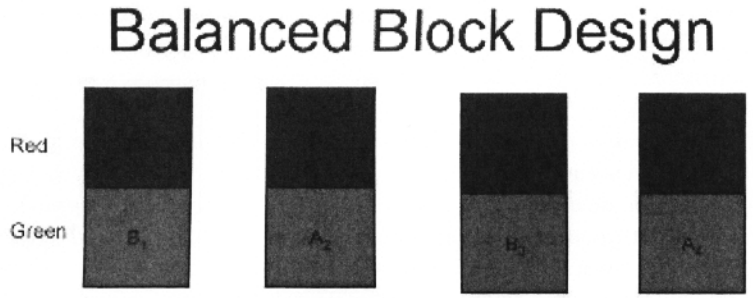

Fig. 3.3. Loop design for comparing two classes of samples. Each biologically independent sample is subaliquoted and hybridized to two arrays, once with the Cy3 label and once with the Cy5 label. Each array contains a sample from each class.

only n microarrays. No reference RNA is used at all. The reference design would require $2n$ arrays to accommodate n nonreference samples from each of the two classes.

The balanced block design is very efficient in the use of arrays, but it has major limitations. For one, cluster analysis of the expression profiles cannot be performed effectively. Without a common reference, any comparisons between expression profiles of samples on different arrays will be subject to noise resulting from variation in size and shape of corresponding spots on different arrays and variation in sample distribution patterns on individual arrays (Dobbin and Simon 2002).

Another important limitation of the balanced block design is that it is based on a single specified two-class comparison. It does not easily accommodate analyzing the data in different ways for contrasting different groups of samples. Because it may be difficult to pair the samples simultaneously with regard to all of the class comparisons of interest, the block design is most effective when there is a single type of class comparison. The block design is also not effective for developing class predictors as described in Chapter 8.

In addition, the balanced block design also requires an arbitrary pairing of samples from the two classes and is less effective than the reference design when there is large intersample variability or when the number of samples, rather than the number of arrays, is limiting (Dobbin and Simon 2002).

3.6.3 The Loop Design

Loop designs (Kerr and Churchill 2001a) are another alternative to reference designs. When cluster analysis is planned, two aliquots of each sample must be arrayed for the loop design (Figure 3.4). For example, the first array would consist of one aliquot of the first sample labeled red and an aliquot of the second sample labeled green. The second array would consist of a second aliquot of the second sample, labeled red this time, and an aliquot of a third sample

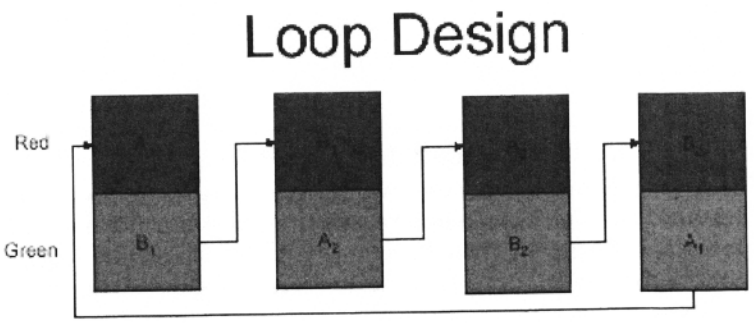

Fig. 3.4. Balanced block design for comparing two classes of samples. Each array contains a biologically independent sample from each class. Each class is labeled on half the arrays with one label and on the other half of the arrays with the other label. Each biologically independent sample is hybridized to a single array.

labeled green. The third array would consist of a second aliquot of the third sample labeled red this time, and an aliquot of a fourth sample labeled green. This loop continues and concludes with the nth and final array which consists of a second aliquot of the final sample n labeled red and hybridized with a second aliquot of the first sample, labeled green this time. This uses n arrays to study n samples, using two aliquots of each sample. The loops permit all pairs of samples to be contrasted in a manner that controls for variation in spot size and sample distribution patterns using a statistical model. Contrasting two samples far apart in the loop, however, involves modeling many indirect effects corresponding to the arrays linking the two arrays of interest and this adds substantial variance to many of these contrasts (Dobbin and Simon 2002). Consequently loop designs are not effective for cluster analysis. Loop designs can be used for class comparisons, but are less efficient than balanced block designs and require more complex methods of analysis than do common reference designs. Loop designs are less robust against the presence of bad quality arrays; two bad arrays break the loop. Loop designs also require enough RNA to be available for each sample for at least two hybridizations. Because of these limitations, loop designs are not generally recommended.

3.7 Reverse Labeling (Dye Swap)

Some investigators believe that all arrays should be performed both forward- and reverse- labeled. That is, for an array with sample A labeled with Cy3 and sample B labeled with Cy5, there should be another array with sample A labeled with Cy5 and sample B labeled with Cy3. In general, this is unnecessary and wasteful of resources (Dobbin et al. 2003a,b). Balanced labeling, as

described in Section 3.6.2 is in general much more efficient than replicating hybridizations of the same specimens with swapped dye labeling. We discuss here, however, one circumstance where some reverse-labeling of samples is appropriate.

Dye swap or dye balance issues arise because the relative labeling intensity of the Cy3 and Cy5 may be different for different genes. Although the normalization process may remove average dye bias, gene-specific dye bias may remain. This is not important for comparing classes of nonreference samples using a reference design when the reference is consistently assigned the same label. Suppose, however, that we wanted to compare tumor tissue to matched normal tissue from the same patient using dual-label microarrays. As discussed in Section 3.6.2, one effective design would be to pair tumor and normal tissues from the same patient for cohybridization on the same array, with half of these arrays having the tumor labeled with Cy3 and the other half having the tumor labeled with Cy5 (Figure 3.3). Because the dye assignments are balanced, it is not necessary to perform any reverse-labeled replicate arrays of the tissues from the same patient (Dobbin et al. 2003a,b). For a fixed total number of arrays, it is best to use the available arrays to assay tissue from new patients, using the balanced block design described, rather than to perform replicate reverse-labeled arrays for single patients. The balanced block design is also best when there are n tumor tissues and n normal tissues even though the tissues are not from the same patients, or for comparing any two classes of samples. In these cases, the samples may be randomly paired, or paired based on balance with regard to potentially confounding variables such as the age of the specimens.

In some cases a reference design is used in which the primary objective is comparison of classes of the nonreference samples but comparison to the internal reference is a secondary objective. For example, there may be several types of transgenic mouse breast tumors for comparison and the internal reference may be a pool of normal mouse breast epithelium. Because the primary interest is comparison among multiple tumors models, a reference design may be chosen. The use of a pool of normal breast epithelium as the internal reference, rather than a mixture of cell lines, reflects some interest in comparison of expression profiles in tumors relative to normal breast epithelium. Comparison to a pool of normal breast epithelium is somewhat problematic, however, for reasons described previously in Section 3.5. The conclusions derived from comparison of the tumor samples to the internal reference will apply to that pool of normal epithelium, but it will not be possible to evaluate how representative that pool is. Nevertheless, the comparison may be of interest.

In order to ensure that the comparison of tumor expression to that of the reference is not distorted by gene-specific dye bias when using a reference design, some reverse-labeled arrays are needed. One can then fit a statistical analysis of variance model to the logarithms of the intensities for each channel as described in Section 7.9. Not all arrays need to be reverse-labeled; 5 to10 reverse-labeled pairs of arrays will generally be adequate. Except for this purpose of comparison of experimental samples to the common reference in a reference design, however, Dobbin et al. (2003a,b) recommend against reverse-labeling of the same two RNA samples.

3.8 Number of Biological Replicates Needed

As indicated in Section 3.3, it is not generally meaningful to compare expression profiles in two RNA samples without biological replication. The number of independent biological samples needed depends on the objectives of the experiment. We describe here a relatively straightforward method for planning sample size for testing whether a particular gene is differentially expressed between two predefined classes. Such a test can be applied to each gene if we adjust for the number of comparisons involved (Simon et al. 2002).

This approach to sample size planning may be used for dual-label arrays using reference designs or for single-label oligonucleotide arrays. For dual-label arrays the expression level for a gene is the log ratio of intensity relative to the reference sample; for Affymetrix GeneChip^{TM} arrays it is usually the log signal, discussed in Chapter 4. The approach to sample size planning described here is based on the assumption that the expression measurements are approximately normally distributed among samples of the same class. Let σ denote the standard deviation of the expression level for a gene among samples within the same class and suppose that the means of the two classes differ by δ for that gene. For example, with base 2 logarithms, a value of $\delta = 1$ corresponds to a twofold difference between classes. We assume that the two classes will be compared with regard to the level of expression of each gene and that a statistically significant difference will be declared if the null hypothesis can be rejected at a significance level α. The *significance level* is the probability of concluding that the gene is differentially expressed between the two classes when in fact the means are the same ($\delta = 0$). The significance level α will be set stringently in order to limit the number of false positive findings inasmuch as thousands of genes will be analyzed. The desired statistical power will be denoted $1 - \beta$. *Statistical power* is the probability of obtaining statistical significance in comparing gene expression between the two classes when the true difference in mean expression levels between the classes is δ. Statistical power is one minus the false negativity rate (β).

Under these conditions, the total number of samples required from different individuals or different replications of the experiment approximately satisfies the equation:

$$n = \frac{4\left(t_{\alpha/2} + t_{\beta}\right)^2}{(\delta/\sigma)^2}, \tag{3.1}$$

where $t_{\alpha/2}$ and t_{β} denote the $(100)\alpha/2$ and 100β percentiles of the t distribution with $n - 2$ degrees of freedom. Because the t percentiles depend on n, however, the equation can only be solved iteratively. When the number of

samples n is sufficiently large, Equation (3.1) can be adequately approximated by

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2}{(\delta/\sigma)^2}, \tag{3.2}$$

where $z_{\alpha/2}$ and z_β denote the corresponding percentiles of the standard normal distribution (Desu et al. 1990). The normal percentiles do not depend on n, and hence equation (3.2) can be solved directly for n. For example, for $\alpha = 0.001$ and $\beta = 0.05$ as recommended below, the standard normal percentiles are $z_{\alpha/2} = -3.29$ and $z_\beta = -1.645$, respectively. Expressions (3.1) and (3.2) give the total number of biologically independent samples needed for comparing the two classes; $n/2$ should be selected from each class.

The fact that expression levels for many genes will be examined indicates that the size of α should be much smaller than 0.05. The 0.05 value is only appropriate for experiments where the focus is on a single endpoint or single test. If $\alpha = 0.05$ is used for testing the differential expression of 10,000 genes between two classes, then even if none of the genes is truly differentially expressed, one would expect 500 false discoveries; that is, 500 false claims of statistical significance. The expected number of false discoveries is α times the number of genes that are nondifferentially expressed. This is true regardless of the correlation pattern among the genes.

In order to keep the number of false discoveries manageable with thousands of genes analyzed, $\alpha = 0.001$ is often appropriate. For example, using $\alpha = 0.001$ with 10,000 genes gives 10 expected false discoveries. This is much less conservative than the multi-test adjustment procedures used for clinical trials where the probability of even one false discovery is limited to 5%. We recommend $\beta = 0.05$ in order to have good statistical power for identifying genes that really are differentially expressed. If the ratio of sample sizes in the two groups is k:1 instead of 1:1, then the total sample size increases by a factor of $(k+1)^2/4k$ compared to formula (3.2).

The parameter σ can usually be estimated based on data showing the degree of variation of expression values among similar biological tissue samples. σ will vary among genes. For log ratio expression levels, we have seen the median values of σ of approximately 0.5 (using base 2 logarithms) for human tissue samples and similar values for Affymetrix GeneChips™. The parameter δ represents the size of the difference between the two classes we wish to be able to detect. For $\log_2$ ratios or $\log_2$ signals, $\delta = 1$ is corresponds to a twofold difference in expression level between classes. This value of δ is reasonable because differences of less than twofold are difficult to measure reproducibility with microarrays. Using $\alpha = 0.001$, $\beta = 0.05$, $\delta = 1$ and $\sigma = 0.50$ in (3.2) gives a required sample size of approximately 26 total samples, or 13 in each of the two classes. The more accurate formula (3.1) gives a requirement of 30 total samples, or 15 in each of the two classes.

The within-class variability depends somewhat on the type of specimens; human tissue samples have greater variability than inbred strains of mice or

than cell lines. In experiments studying microarrays of kidney tissue for inbred strains of mice, the median standard deviation of log ratios for a normal kidney was approximately 0.25, with little variation among genes. For cell line data on Affymetrix GeneChips™, we have seen similar standard deviations for $\log_2$ signals. Using $\alpha = 0.001$, $\beta = 0.05$, $\delta = 1$ and $\sigma = 0.25$ in formula (3.1) gives a required sample size of 11 total samples. Because we cannot have 5.5 samples per class, we should round up to 6 samples per class. If this were a time-series experiment with more than two time points, then one should plan for 6 animals per timepoint in order to enable expression profiles to be compared for all pairs of time points.

The discussion above applies either to dual-label arrays using a reference design and log ratio as the measure of relative expression, or to single-label arrays such as the Affymetrix GeneChip™ arrays using log transformed signals or another measure of expression. When dual-label arrays are used with the block design to compare either naturally paired or independent samples from two classes, then the same formulas apply but the definition of σ changes. For the block design, σ represents the standard deviation of variation across arrays of the log ratio computed with one sample from each class (Dobbin et al. 2003). Preliminary data are generally needed to estimate σ.

Many of the considerations for comparing predefined classes also apply to identifying genes that are significantly associated with patient outcome (Simon et al. 2002). When the outcome is survival and not all patients are followed until death, the analogue of expression (3.2) is

$$E = \frac{(z_{\alpha/2} + z_\beta)^2}{(\tau \ln(\delta))^2}. \tag{3.3}$$

E denotes the number of events (e.g., deaths) that need to be observed in order to achieve the targeted statistical power. For survival comparisons, the statistical power often depends on the number of events, rather than the number of patients. For a given number of patients accrued, the number of events will increase as the duration of followup increases. There is a tradeoff between number of patients accrued and duration of followup in order to achieve a targeted number of events. In expression (3.3), τ denotes the standard deviation of the log ratio or log signal for the gene over the entire set of samples. δ represents the hazard ratio associated with a one-unit change in the log ratio or log signal and ln denotes the natural logarithm. Note that we are assuming that the log ratio or log signal values are based on logarithms to the base 2, so a one-unit change in the expression level represents a twofold change.

If $\tau = 0.5$ and $\delta = 2$, then 203 events are required for a two-sided significance level of 0.001 and power of 0.95. This makes for a large study in most cases because to observe 203 events in a group of patients with a 50% event rate requires 406 patients. The large number of events results from assuming that a doubling of hazard rate requires a two standard deviation change in log ratio. Hence most patients would have expression levels that had very limited effects on survival. Therefore it may be more reasonable to size the study for

detecting statistically significant differences in only the more variable genes, for example, $\tau = 1$ and $\delta = 2$, which results in 51 required events. Genes that have small standard deviations across the entire set of samples are difficult to use for prognostic prediction in clinical situations.

The multivariate permutation tests described in Chapter 7 are a more powerful method for finding differentially expressed genes than the univariate parametric test that is the basis for formulas (3.1) and (3.2). Nevertheless, the sample size formulas given here are useful for planning purposes and provide control of the number of false discoveries in a reasonable manner. Other methods have been described by Black and Doerge (2002), Lee and Whitmore, (2002), and Pan et al. (2002). Adequate methods for determining the number of samples required for gene expression studies whose objectives are class prediction or class discovery have not yet been developed. Hwang et al. (2002) provide a method of planning sample size to test the hypothesis that the classes are completely equivalent with regard to expression profile. The sample size formulas given above provide reasonable minimum sample sizes for class prediction studies. Often, however, developing multivariate class predictors or survival predictors involves extensive analyses beyond determining the genes that are informative univariately. Consequently, larger sample sizes are generally needed for class prediction studies (Rosenwald et al. 2002).

In class prediction studies it is important to estimate the misclassification rate of the identified multivariate predictor. There is a problem using the same data to develop a prediction model and to estimate the accuracy of the model, particularly when the number of candidate predictors is orders of magnitude larger than the number of cases (Simon et al. 2003). Consequently, special methods of analysis must be used to provide unbiased estimates of prediction accuracy. One approach is to split the data into a training set and a validation set (Rosenwald et al. 2002). Other approaches involve more sophisticated methods such as cross-validation or bootstrap resampling (discussed in Chapter 8). Cross-validation and bootstrap resampling can be used when the derivation of the prediction rule can be clearly defined as an algorithm with no subjective elements. In many cases, the derivation of the prediction rule is more complex and involves numerous analyses that cannot be easily specified in a manner that can automatically be applied to resampled datasets. In these cases, it is necessary to use the split sample approach to obtain an unbiased estimate of the accuracy of the class prediction rule. Initially the data are divided into a training set and a validation set. This division may be done randomly or may be balanced by factors such as an institution contributing the specimens. The validation set is put aside and not analyzed at all until a completely specified prediction rule is defined based on analyses conducted using the training set. The prediction rule resulting from the analysis should be completely specified, including the estimation of parameters and the establishment of threshold values for class prediction. After the analysis of the training set is completed, the validation set is unlocked and the completely specified prediction rule is applied to the cases in the validation set. The prediction accuracy is determined based on the performance of the predictor on the validation set. Usually about one third to one half of the total dataset is reserved for the validation set. Therefore in planning the size of a class prediction study that will use the split sample approach, it should be recognized that perhaps only half of the data will be available for development of the multivariate predictor.

Often investigators use validation sets that are far too small to provide meaningful validation. Table 3.1 shows the upper 95% confidence limit for the misclassification rate as a function of the observed proportion of misclassified specimens and the number of specimens in the validation set. Suppose that the true misclassification rate is 10% and you are lucky enough to observe no misclassifications in the validation set. With only 10 specimens in the validation set you would be able to bound the true misclassification rate to be no greater than 26%. But it is just as likely that you would obtain 20% of the validation set misclassified. In this case, even with 50 samples in the validation set, you would only be able to bound the true misclassification rate to be no greater than 32%. Consequently, a substantial validation set is needed in order to adequately estimate the true misclassification rate.

Table 3.1.

Number of Validation Samples	Upper 95% Confidence Limit for Misclassification Rate	
	None Misclassified in Validation Set (%)	20% Misclassified in Validation Set (%)
5	45	66
10	26	51
20	14	40
50	6	32