2

Design Principles for Microarray Investigations

Kathleen F. Kerr

Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA.
katiek@u.washington.edu

2.1 Introduction

In the past decade, high-throughput measurement of gene expression has evolved from a tantalizing possibility to an everyday exercise, thanks to microarray technology. The initial excitement for microarrays was quickly followed, for many scientists, with apprehension about appropriately analyzing large amounts of data of sometimes questionable quality. Most scientists have now developed an appreciation for the limitations and challenges presented by the technology.

A microarray study should not be conducted without careful thought and planning, even if it is exploratory. As with any other type of scientific investigation, a successful microarray study starts with developing a well-defined project with well-defined goals. One must then develop and implement a sound experimental design based on these goals. This chapter will begin with a discussion of some of the basic issues to consider in the earliest stages of planning a microarray study. In Section 2.3, I discuss three general principles of statistical design that apply generally to scientific experimentation: *Replication*, *blocking*, and *randomization*. We will review each of these concepts in turn, and discuss each of them in the context of array experiments.

2.2 The "Pre-Planning" Stage

By the time a scientist consults with a statistician about the experimental design for a microarray study, she has probably already made some important design choices. The scientist has probably already chosen the types of mRNA to be studied. That is, she has chosen the organism and tissue type, and has decided which treatments to apply or under what conditions the mRNA will be collected. These choices are primarily made based on scientific, not statistical, considerations, although a technical consideration is whether the samples can provide a sufficient amount of mRNA for the assay.

At this stage, it is important to recognize whether a study is an *experiment* or an *observational study*. Unfortunately, microarray studies all tend to be called "experiments," but this can be a misnomer (Potter, 2003). For example, consider a study in which tissue samples are compared between patients with a particular kind of cancer and cancer-free control subjects. The investigator does not assign cancer status to the subjects, he is merely making measurements on a sample of cases and controls. This is an observational study, even though the observations happen to be measurements of gene expression for thousands of genes. The fact that the investigation is an observational study has profound implications for the interpretation of the data. For example, the investigator would *not* be automatically justified in attributing any observed differences in gene expression between the cases and controls to their cancer status because the differences could be due to a *confounding factor*. That is, the cases and controls might differ in their distributions of age, sex, environmental exposures, or what they ate for breakfast. Unfortunately, in many such observational microarray studies, data on potential confounding factors are not collected and the possible impact of such factors is ignored. Such gross oversight makes an entire study scientifically questionable (Potter, 2003).

In the early planning stage, it is important to establish realistic expectations for the array study. Because arrays produce more data than many biologists are used to, some biologists make the natural leap that they produce a vast amount of information. In a sense they do, but the information is far from complete and a successful array study will produce at least as many questions as it answers. Thus, it is important to clarify the goals of the array experiment. Dudoit et al. (2002) describe three distinct goals of microarray experiments: *Unsupervised learning* (Goal 1), *supervised learning* (Goal 2), and *class comparison* (Goal 3). I discuss each of these briefly, then focus on Goal 3 for the remainder of this chapter.

2.2.1 Goal 1: Unsupervised Learning

In very general terms, unsupervised learning attempts to organize data into groups of "similar" observations. With microarray data, this might mean using gene expression data on multiple genes to organize or "cluster" subjects into groups with similar gene expression profiles. Alternatively, one could organize genes into groups within which the expression profiles are similar across individuals. Eisen et al. (1998) presented an early and influential microarray paper that demonstrated the application of a particular flavor of unsupervised learning called hierarchical clustering. Sometimes clustering subjects and clustering genes are done simultaneously; this is especially common when hierarchical clustering is used. See Chapter 6 of this book for more information on unsupervised learning techniques. Note that unsupervised learning is also called *class discovery* and, most often in microarrays, *cluster analysis*.

Sometimes unsupervised learning is used with a specific goal in mind, for example, discovering new sub-types of cancer that have previously been hypothesized to exist. More commonly, unsupervised learning is used as a completely exploratory technique. There is an emerging consensus that unsupervised techniques are overused (Allison et al., 2006), as many studies that use these techniques would be better served supervised learning (Section 2.2.2) or class comparison (Section 2.2.3) approaches.

The literature contains little discussion of design issues for studies in which unsupervised learning will be used. Dobbin and Simon (2002) may be the only paper on the subject. However, the lack of research in this area should not be interpreted as an indication that design issues are not important in these studies. Section 2.3.3 of this chapter gives an example that illustrates how poor design can produce misleading results in cluster analysis.

2.2.2 Goal 2: Supervised Learning

Supervised learning is also know as supervised classification and discriminant analysis. An example application is a study where the goal is to develop an algorithm to make an accurate prognosis for cancer patients based on gene expression measurements on biopsy samples. An accurate prognosis could help patients and their doctors decide whether to pursue more aggressive treatment. The data include information on the eventual outcome for the subjects, and this information is used to develop (or "train") the algorithm, which is why the learning is called "supervised." See Chapter 9 for more information on supervised learning techniques.

Supervised learning is typically done with the possibility of a clinical application in mind. As such, the data used in a supervised learning analysis are invariably from an observational study, not an experiment. A truly useful classification algorithm must be able to classify new subjects, not just those in the sample. An important factor for facilitating this is to ensure that there are no obvious differences between the kinds of samples in study design. For example, suppose the biopsy samples for long-term cancer survivors tend to be older, whereas the samples for patients who died quickly tend to be fresher. Handling and storage differences could affect the array measurements, and these differences could influence the parameters of the classification algorithm. Thus, an algorithm that putatively discriminates between patients with good and poor prognoses is actually distinguishing between handling and storage differences between the RNA. Because of this design flaw, the algorithm will not perform well when tested on new samples from newly-diagnosed patients, all of whom provide fresh samples.

2.2.3 Goal 3: Class Comparison

Class comparison is probably the most common goal of gene expression studies and is the focus of the remainder of this chapter. In a typical class comparison

study, an investigator wants to identify genes that are differentially expressed between two or more classes of tissue. A class comparison investigation can be either an experiment or an observational study. For example, a comparison between laboratory mice treated with a certain drug and untreated mice is an experiment, as long as the pre-specified number of mice to receive the treatment are chosen randomly from all mice in the study. In contrast, a study that identified differentially expressed genes between patients with and without a particular malignancy is an observational study.

In class comparison studies it is important to understand that microarrays do not remove inherent limitations in determining the "cause and effect" in some system. As a measurement tool, microarrays cannot be used to make causal inferences unless the study is explicitly designed to make this possible. In the observational study comparing malignant tissue with benign controls, microarrays cannot distinguish genes whose altered expression *caused* the malignancy from genes whose expression is altered *as a result of* the malignancy. In fact, the study can only conclude that altered expression is *associated with* the malignancy, keeping in mind that such an association could be due to a confounding factor (Potter, 2003).

In the microarray experiment with the treated and untreated mice, we can justify causal inference about the effect of the drug on gene expression because of the initial randomization of the treatment. However, note that the causal inference is about the effect of the treatment. This is quite different from trying to infer the causal effect of gene expression changes.

Once these basic issues have been considered, the next step is to plan the details of the microarray study itself. We now discuss the three fundamental principles of design, replication, blocking, and randomization, focusing on their application to microarrays and in particular to microarray studies for class comparison.

2.3 Statistical Design Principles, Applied to Microarrays

2.3.1 Replication

Replication is probably the most widely-recognized principle of design. Researchers carefully plan the sample size of their studies to ensure adequate replication.

To appreciate the important role of replication, it is useful to review the general paradigm of statistics. Scientifically, we are often interested in comparing different groups or classes of individuals: Treated and untreated; diseased and non-diseased; genotypes AA, Aa, and aa (see class comparison, Section 2.2.3). In statistics, such groups are called *populations*. A population is generally either very large or infinite, so it is impossible to examine an entire population. Instead, we take a *sample* from the population. We may study the sample in excruciating detail, collecting and analyzing data. Ironically,

however, our true interest is not in the individuals in the sample. Our interest in the sample is as a means to making *inference* to the population from which it was drawn. A statistical inference is something more than a generalization or an educated guess. The theory of statistics allows us to make inferences with rigor: Using the data on a random sample, we can estimate certain characteristics of a population (for example, the mean expression of gene xyz in the population), and we can also quantify our level of certainty in the estimate (often, with a confidence interval). However, rigorous statistical inference is only possible with replication. In other words, samples of size 1 are not sufficient. Further, an adequate level of precision in inference is achieved only with an adequate amount of replication.

Understanding this fundamental statistical paradigm can help a researcher understand the appropriate level on which to replicate. In research with microarrays, it is common to differentiate between *technical replicates* and *biological replicates* (Yang and Speed, 2002). Technical replicates are typically repeated hybridizations of the same RNA to multiple arrays. Replication in early array experiments was often limited to technical replication. Technical replication allows one to make inference about the particular RNAs being studied in light of the technical error (measurement error) of the assay. However, this is usually not the desired inference. Most often, the desired inference is from the sampled individuals to the population(s) they represent. This inference is only possible with biological replication: Multiple individuals sampled from each population of interest.

Kerr (2003a) examines the relative benefits of biological and technical replication. Technical replication can be useful, but is usually unnecessary. It is usually best to use available resources to maximize biological replication and forego technical variation altogether (Simon et al., 2002; Kerr, 2003a).

2.3.2 Blocking

The term "blocking" comes from the agricultural origins of the field of statistical design. Suppose one wants to conduct a study to compare, say, the yields of different varieties of a crop. Suppose further that different blocks of land are available to use in the study. Different blocks of land will vary in many characteristics that can affect yield, e.g., the amount of sunlight or the soil composition. It would be crucial to recognize this in planning the experiment. The more variation among the blocks of land, the more important it is to explicitly address this source of variation in the experimental design. If block-to-block variability is large, an effective solution is to balance varieties with respect to blocks. For example, if there are four varieties and each block can accommodate four sub-plots, then each block should contain one of each variety (Figure 2.1). In statistical design this would be called a "complete block design." "Complete" refers to the fact that every block contains an equal number of replicates of each variety.

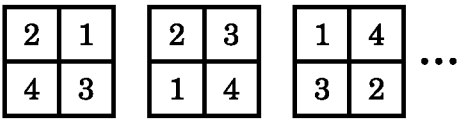

Fig. 2.1. An experiment in which the experimental units come in blocks of size 4. If there are four groups to compare, the best design is to put one of each variety in each block.

Experimentalists routinely and intuitively use the principle of blocking. For example, if an assay is known to be sensitive to humidity, then an experimentalist may make sure to conduct all assays within a short period of time when humidity is constant. Two ocular treatments might be compared by applying each of them to one eye of multiple individuals. Each pair of eyes is a "block" in such a study design. This design controls for variation between individuals by enabling the treatments to be compared "within" each individual.

In microarray studies, it can be important and useful to implement blocking as with any other kind of experiment. For example, if treatments are to be compared on mice from various litters, a litter of mice should be treated as a block. Ideally, each treatment could be applied to the same number of mice in each litter.

For two-color microarray platforms, blocking is intrinsic to the technology. This is because spot characteristics (size, density, etc.) are variable, which means a large signal could result from a high level of gene expression or from a particularly large or dense spot. However, if spot characteristics lead to a high level of signal, then the signal should be brighter in *both* channels. Therefore, the *relative* sizes of the red and green signals is used as a measure of the *relative* levels of expression in the red- and green-labeled RNAs. In other words, ratios are used because they control for spot-to-spot variation from array to array. Taking ratios (or better, log-ratios) "cancels out" uninteresting variation that is due to spot heterogeneity. This is actually a textbook example of the principle of blocking.

While the majority of analyses are based on the ratio of the red and green signals from each spot, some analytical methods start with the individual signal intensities rather than ratios. For example, see Kerr et al. (2000) and Wolfinger et al. (2001). Such methods simply handle the blocking structure of the data in a different way. In fact, the difference between intensity-based methods and ratio-based methods is somewhat more technical than substantive – see (Kerr, 2003b).

Because of spot heterogeneity, two-color arrays are used to measure relative gene expression, not absolute gene expression. A two-color array can be thought of as a comparison between the co-hybridized RNAs. When there are

multiple samples to be compared, this raises the question: Which hybridizations to perform? That is, what pairs of RNAs should be co-hybridized? Kerr and Churchill (2001) addressed this question for experiments that do not contain biological replicates. Dobbin and Simon (2002) and Kerr (2003a) update these findings for experiments with biological replicates.

When there are n replicates from two groups to be compared, an efficient and effective strategy is the multiple-dye-swap design, as seen in Figure 2.2(a). In this design, the n replicates from the two groups are randomly paired and each pair is co-hybridized to a pair of arrays, with a dye-swap to control for dye-effects. Another design, similar to those proposed by Rosa et al. (2005), is to alternate the dye-labeling between replicates (see Figure 2.2(b)). This will allow twice the number of replicates to be used for the same cost of arrays, while maintaining dye-balance. Another, popular strategy is to employ a "reference" RNA in the design; each RNA of interest is co-hybridized with the reference RNA. The reference RNA is not of interest and serves only to "connect" the other samples. In Figure 2.2(c), this strategy is employed for the two-group comparison problem, employing dye-swap. While the reference design is technically less efficient than the multiple-dye swap strategy, its efficiency disadvantage is small when biological variation is much larger than technical variation (Kerr, 2003a). It is an exceedingly simple and practical design choice for many investigations.

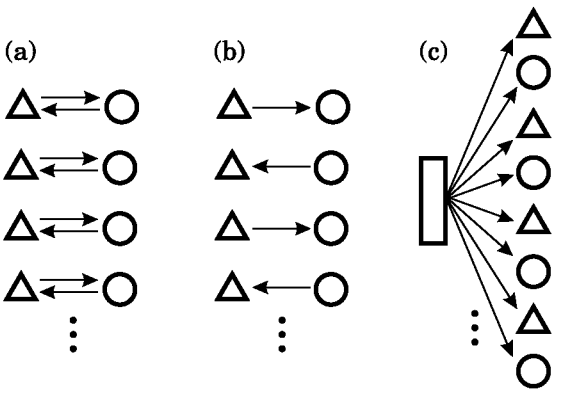

Fig. 2.2. Circles represent biological replicates from some population and triangles represent biological replicates from another population. Arrows represent two-color microarrays. An arrow between individual 1 and individual 2 indicates a hybridization with red-labeled RNA from individual 1 and green-labeled RNA from individual 2. All designs are appropriate for a two-group comparison study. (a) Multiple dye-swap design; (b) Alternating-dye pairwise design; (c) Reference design – the rectangle represents the "reference" RNA, which is not of interest.

2.3.3 Randomization

The principle of randomization says that once any blocking structure to a design is established, treatments should be applied to experimental units in random fashion. If three littermates are to be divided among treatments A, B, and C, then the mice should be randomly allocated to each treatment. "Random" here does not mean the same thing as "arbitrary." Although tedious, it is useful to assign numbers to each mouse and use a random-number generator or draw numbers out of a hat to choose the mouse for each treatment.

While blocking protects against known or anticipated biases in the data, randomization protects against unknown or unanticipated biases. For the previous example, suppose one had an unrecognized tendency to pick-up the slowest mouse out of a litter. If one assigned mice to treatments A, B, and C in sequence, treatment A mice would tend to be assigned the slowest mice and treatment C would tend to be assigned to the quickest mice. If quick mice are also healthier, the experiment would obviously be biased.

Here is a more subtle, fictionalized example from the world of microarrays that shows that randomization is important even in observational studies. An experimenter is interested in a particular human mutation and recruits 20 carriers of the mutation. The mutation is rare and non-carriers are easier to find, and she is able to recruit 40 non-carriers to serve as controls. She is interested in whether the mutation is associated with any gene expression differences in humans. The investigator is reasonably confident that there are no other variables confounding the comparison between carriers and non-carriers. Using a single-color platform, the researcher uses one array to hybridize the mRNA for every individual. There is a practical limitation of a maximum of 20 hybridizations a day, so the experiment is carried out over three days.

The researcher applies a hierarchical clustering algorithm to explore the array data. The results appear as depicted in Figure 2.3(a). To the scientist's delight, the 60 samples appear to cluster into three primary groups: The 20 samples from the carriers of the mutation, and two groups of the remaining 40 non-carriers. The natural temptation is to conclude that gene expression data can discriminate carriers of the mutation from non-carriers, and that non-carriers can further be divided into two sub-types. However, with a healthy respect for scientific skepticism, the experimenter re-examines her data. Upon closer scrutiny, she sees that the three clusters correspond exactly to the three days of hybridizations, as in Figure 2.3(b).

In detail, the schedule for the hybridizations was:

- Day1: 20 carriers
- Day 2: 20 non-carriers
- Day 3: remaining 20 non-carriers

The fatal flaw in this investigation was the lack of randomization. The day of hybridization was ignored as a factor, but it turned out to be an important source of variation. Samples should have been hybridized in random order.

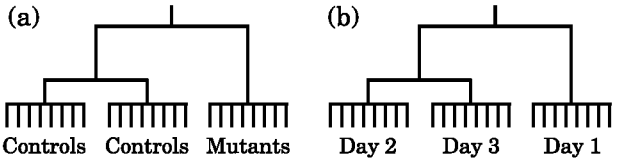

Fig. 2.3. Results of clustering samples for the example in Section 2.3.3. (a) Samples labeled by mutation status; (b) samples labeled by day of hybridization.

As is, the gene expression differences between carriers and non-carriers are hopelessly confounded with day-to-day differences in the hybridizations. There is no way to "rescue" the experiment – the confounding is complete and there is no way to separate the genetic differences of interest from the nuisance experimental artifacts.

Now that the day of hybridization is known to be an important factor, the researcher should probably "block" on the day of hybridization in future experimental plans. That is, for each group she should hybridize the same number of samples on each day.

2.4 Case Study

A plant geneticist is interested in the effects on gene expression in arabadopsis arising from infection by an agrobacterium. He plans a basic class comparison microarray study. From his initial collection of 20 plants, he randomly divides them into treatment and control groups of size 10. The treatment group is infected with the agrobacteria. The control group receives "mock" treatment, undergoing each step of infection except the introduction of the bacteria. This is to make sure that differences between the groups can properly be ascribed to infectious agent. One treated and control sample are produced every day, in random order. The RNA is extracted from each, and the treated and control RNA with same-day preparation are co-hybridized to a pair of microarrays employing dye-swap. That is, the design in Figure 2.2(a) is used, which is a very efficient design for comparing the two groups (Kerr, 2003a). This design will naturally handle any day-to-day differences in sample preparation (blocking) because day-to-day differences will cancel out in the treatment-control comparison due to the balance in the preparation schedule.

2.5 Conclusions

Replication, blocking, and randomization should all be considered in designing a microarray experiment. It usually works to consider them in the order

presented here. First, make sure there is the right kind of replication to allow the desired inferences. Replication leads directly to the question of choosing a sample size. Sample size calculations are a tricky issue with microarrays and the subject of considerable research, beyond the scope of this article. See Simon et al. (2002); Lee and Whitmore (2002); Wei et al. (2004); and Tibshirani (2005). Second, for two-color platforms the arrangement of the samples onto the arrays must be decided. For many class comparison experiments the layouts in Figure 2.2 can be adapted. See Rosa et al. (2005), for other ideas. Lastly, consider all opportunities for randomization. For example, arrays can be randomly assigned to planned hybridizations and the order of hybridizations should also be randomized.

Although microarray studies are typically exploratory, one should still be able to clearly articulate a goal for the project. A well-defined goal will inform good choices in experimental design. A seriously flawed experimental design guarantees a study will be a failure, because it produces data that cannot answer the scientific question of interest. A sound experimental design does not guarantee a study will be a rousing success, but gives it a fighting chance.

References

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nat. Rev. Gen.*, 7:55–65.

Dobbin, K. and Simon, R. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, 18(11):1438–1445.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Ass.*, 97:77–87.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sc. USA*, 95(25):14863–14868.

Kerr, M.K. (2003a). Design considerations for efficient and effective microarray studies. *Biometrics*, 59:822–828.

Kerr, M.K. (2003b). Linear models for microarray data analysis: Hidden similarities and differences. *J. Comp. Biol.*, 10:891–901.

Kerr, M.K. and Churchill, G.A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2:183–201.

Kerr, M.K., Martin, M., and Churchill, G.A. (2000). Analysis of variance for gene expression microarrays. *J. Comp. Biol.*, 7:819–837.

Lee, M.L. and Whitmore, G.A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21(1):3543–70.

Potter, J.D. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. *TRENDS in Genetics*, 19(12):690–695.

Rosa, G.J.M., Steibel, J., and Tempelman, R.J. (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics*, 6(1):123–131.

Simon, R., Radmacher, M.D., and Dobbin, K. (2002). Design of studies using DNA microarrays. *Gen. Epidem.*, 23:21–36.

Tibshirani, R. (2005). A simple method for assessing sample sizes in microarray experiments. `http://www-stat.stanford.edu/ tibs/SAM/`.

Wei, C., Li, J., and Bumgarner, R.E. (2004). Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, 5:87.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comp. Biol.*, 8:625–637.

Yang, Y.H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat. Rev.*, 3:579–588.