CHAPTER 2

Design and Analysis of Comparative Microarray Experiments

Yee Hwa Yang and Terence P. Speed

2.1 Introduction

This chapter discusses the design and analysis of relatively simple comparative experiments involving microarrays. Some of the discussion applies to all the most widely used kinds of microarrays, that is, radiolabelled cDNA arrays on nylon membranes, two-color, fluorescently labeled cDNA or long oligonucleotide arrays on glass slides, or single color, fluorescently labeled, high-density short oligonucleotide arrays on silicon chips. The main focus, however, is on two-color complementary deoxyribonucleic acid (cDNA) or long oligonucleotide arrays on glass slides because they present more challenging design and analysis problems than the other two kinds.

As subfields of statistics, the topics of design and analysis of microarray experiments are still in their infancy. Entirely satisfactory solutions to many simple problems still elude us, and the more complex problems will provide challenges to us for some time to come. Much of what we present in this chapter could be described as first pass attempts to deal with the deluge of data arriving at our doors. Questions come in a volume and at a pace that demands answers; we simply do not have the luxury of waiting until we have final solutions to problems before we get back to the biologists. A major aim of this chapter is to stimulate other statisticians to work with their local biologists on microarray experiments and to come up with better solutions to the common problems than the ones we present here.

2.2 Experimental design

Statisticians do not need reminding that proper statistical design is essential to ensure that the effects of interest to biologists in microarray experiments are accurately and precisely measured. Much of our approach to the design (and analysis, see later) of microarray experiments, takes as its starting point the idea that we are going to measure and compare the expression levels of a single gene in two or more cell populations.
The fact that, with microarrays, we do this simultaneously for tens of thousands of genes definitely has implications for the design of these experiments, but initially we focus on a single gene.

In this section, some of the design issues that arise with the two-color cDNA or long oligonucleotide microarray experiments are discussed. Designing experiments with radiolabelled cDNA arrays on nylon membranes or for fluorescently labeled, high-density short oligonucleotide arrays is less novel. Apart from the following brief discussion about probe design, and a few other remarks in passing, we will not discuss these two platforms in detail separately.

Any microarray experiment involves two main design aspects: the design of the array itself, that is, deciding which DNA probes are to be printed on the solid substrate, be it a membrane, glass slide or silicon chip, and where they are to be printed; and the allocation of messenger ribonucleic acid (mRNA) samples to the microarrays, that is, deciding how mRNA samples should be prepared for the hybridizations, how they should be labeled, and the nature and number of the replicates to be done. We focus on the second aspect, after making a few remarks about the first.

The choice of which DNA probes to print onto the solid substrate is usually made prior to consulting a statistician; this choice is determined by the genes with expression levels that the biologist wants to measure, or by the cDNA libraries (that is, the collections of cDNA clones) available to them. With high-density short oligonucleotide arrays, these decisions are generally made by the company (e.g., Affymetrix) producing the chips, although opportunities exist for building customized arrays. Many researchers purchase pre-spotted cDNA slides or membranes in the same way as they do high-density short oligo arrays. With short (25 base pair) or long (60–75 base pair) oligonucleotide microarrays, the determination of the probe sequences to be printed is an important and specialized bioinformatic task (see Hughes et al. (2001); Rouillard et al. (2001); http://www.affymetrix.com/technology/design/index.affx for a discussion). Similarly, many issues need to be taken into account with cDNA libraries of probes, and here we refer to Kawai et al. (2001).

Advice is sometimes sought from statisticians on the use of controls: negative controls such as blank spots, spots with cDNA from very different species (e.g., bacteria when the main spots are mammalian cDNA), or spots “printed” from buffer solution, or positive controls such as so-called “housekeeping” genes that are ubiquitously expressed at more or less constant levels, and genes that are known not to be in the target samples, which are to be spiked in to it. We note that commercially produced chips (e.g., by Affymetrix) have a wide range of controls of these kinds built in. The questions typically posed to statisticians concern the nature and number of such controls, and the use to be made of signal from them in later analysis. Some controls are there to reassure the experimenter that the hybridization was a success, or indicate that it was a failure, as the case may be. Others are to facilitate special tasks such as normalization (see Chapter 1) or to permit an assessment of the quality of the experimental results. Controls used with cDNA microarrays for normalization include the so-called microarray sample pool (see Yang et al. (2002b) and the ScoreCard
system from Amersham Biosciences). At this stage we don’t have enough experience
to offer general advice or conclusions in design for controls as the full use of such
control data in cDNA arrays is still in its early stages. Yet another cDNA spot design
issue on which statisticians might be consulted is the replicating of spots on the slide,
and we discuss this next.

**Graphical representation**

First, we introduce a method for graphical representation of microarray experimental
designs. One convenient way to represent microarray experiments is to use a *multi-
digraph*, which is a directed graph with multiple edges as illustrated in Figure 2.1(a).
In such a representation, vertices or nodes (e.g., $A$, $B$) correspond to target mRNA
samples and edges or arrows correspond to hybridizations between two mRNA sam-
ples. By convention, we place the green-labeled sample at the tail and the red-labeled
sample at the head of the arrow. For example, Figure 2.1(a) depicts an experiment
consisting of replicated hybridizations. Each slide involves labeling sample $A$ with
green (e.g., Cy3) dye, sample $B$ with red (e.g., Cy5) dye and hybridizing them together
on the same slide. The number “5” on the arrow indicates the number of replicated
hybridizations in this experiment. Similar graphical representations of this nature
have been used previously in experimental design, for instance, in the context of
measurement agreement comparisons (Youden, 1969). For the rest of this chapter, we
use this representation to illustrate different microarray designs. The structure of the
graph determines which effects can be estimated and the precision of the estimates.
For example, two target samples can be compared only if an undirected path joins
the corresponding two vertices. The precision of the estimated contrast then depends
on the number of paths joining the two vertices and is inversely related to the length
of the path. In the hypothetical experiment presented in Figure 2.1(b), which consists
of three sets of hybridizations, the number of paths joining the vertices $A$ and $B$ is
2; a path of length 1 runs directly between $A$ and $B$; another path of length 2 joins
$A$ and $B$ via $C$. When we are estimating the relative abundance of target samples $A$

![Graphical representation of designs](image_url)

**Figure 2.1** Graphical representation of designs. In this representation, vertices correspond to
target mRNA samples and edges to hybridizations between two samples. By convention, we
place the green-labeled sample at the tail and the red-labeled sample at the head of the arrow.
The number 5 denotes the number of replicates of that hybridization.
and $B$, the estimate of $\log_2(A/B)$ from the path $A$ to $B$ is likely to be more precise than the estimate of $\log_2(A/B)$ by $\log_2(A/C) - \log_2(B/C)$ from the path of length 2 joining $A$ and $B$ via $C$.

The preceding discussion assumes that the spot intensities in two-color experiments are all reduced to ratios before further analysis; however, it is already the case that some authors (Jin et al., 2001) are using single-channel data and not reducing to ratios. In this case, two strata of information exist on the log scale: the usual log-ratios within hybridizations, and log-ratios between hybridizations. Because of the novelty of this analysis approach, and the absence so far of a thorough discussion of single channel normalization, we concentrate our discussion of design and analysis to what it, in effect, is within the hybridization stratum (i.e., to analyses that depend on log-ratios within slides).

**Optimal designs**

All measurements from two-color microarrays are paired comparisons, that is, measurements of relative gene expression, with microscope slide playing the role of the block of two units. We begin by discussing design choices for simple experiments comparing two samples $T$ and $C$. These include experiments comparing treated and untreated cells (e.g., drug treated and controls), cells from mutant (including knock-out or transgenic), and from wild-type organisms (Callow et al., 2000), or cells from two different tissues (e.g., tumor and normal). Suppose that we wish to compare the expression level of a single gene in the samples $T$ and $C$ of cells. We could compare them on the same slide (i.e., in the same hybridization) in which case a measure of the gene’s differential expression could be $\log_2(T/C)$, where $\log_2 T$ and $\log_2 C$ are measures of the gene’s expression in samples $T$ and $C$. We refer to this as a direct estimate of differential expression—direct because the measurements come from the same hybridization. Alternatively, $\log_2 T$ and $\log_2 C$ may be estimated in two different hybridizations, with $T$ being measured together with a third sample $R$ and $C$ together with another $R'$ sample of $R$, on two different slides. The log-ratio $\log_2(T/C)$ will in this case be replaced by the difference $\log_2(T/R) - \log_2(C/R')$, and we call this an indirect estimate of the gene’s differential expression because it is calculated with values $\log_2 T$ and $\log_2 C$ from different hybridizations.

The early microarray studies (DeRisi et al., 1996; Spellman et al., 1998; Perou et al., 1999; and many others) performed their experiments using indirect designs. These designs are also known as common reference designs in the microarray literature because each mRNA sample of interest is hybridized together with a common reference sample on the same slide. The common reference samples could be tissues from wild-type organisms, or control tissue, or it could be a pool of all the samples of interest. Common references are frequently used to provide easy means of comparing many samples against one another. More recently, several studies (Jin et al., 2001; Kerr et al., 2001; Lin et al., 2002) have performed experiments that provide direct estimates of log-ratios. In such cases, fixed or random effects linear models
and analysis of variance (ANOVA) have been used to combine data from the different hybridizations.

To date, the main work on design of two-color microarray experiments is due to Kerr and Churchill (2001), and Glonek and Solomon (2002), who have applied ideas from optimal experimental design to suggest efficient designs for some of the common cDNA microarray experiments. Kerr and Churchill (2001) based their comparisons of different designs on the $A$-optimality criterion. In addition, they introduced a novel class of designs they called loop designs, and found that under $A$-optimality, loop designs were more efficient than common reference designs.

Suppose we have a single factor experiment with $K$ levels and the goal is to compare all $K$ treatments. The $A$-optimality criterion favors designs that minimize the average variance of contrasts of interest; however, this criterion alone is often not enough to single out one design; see Designs V and VI in table of Yang and Speed (2002). Just as different microarray experiments will require different analyses, no best design class suits all experiments. Frequently, the scientific questions and physical constraints will drive the design choices.

Glonek and Solomon (2002) studied optimal designs for time course and factorial experiments. Their article introduced classes of appropriate designs based on the notion of admissibility. For the same number of hybridizations, a design is said to be admissible if there exists no other design that has a smaller variance for all contrasts of interest. Their idea is that an investigator should compare only the admissible designs and then base their design selection on scientific interest. In Glonek and Solomon (2002) and in other similar calculations in the literature, log-ratios from different experiments are regarded as statistically independent. In the “Correlation and technical replicates” section, we revisit these calculations assuming a more realistic covariance between replicates, and we examine the implications for design optimality.
Design choices

In preparing to design a cDNA microarray experiment, certain general issues need to be addressed. These can be separated into scientific and logistic (practical). The scientific issues include the aim of the experiment. It is most important to state the primary focus of the experiment, which may be identifying differentially expressed genes, searching for specific patterns, or identifying tumor subclasses. Results from previous experiments or other prior knowledge may lead us to expect only a few, or many genes differentially expressed. In addition, there may be multiple aims within a single experiment, and it is important to specify the different questions and priorities between them. Practical or logistic issues include information such as the types of the mRNA samples, the amount of material and the number of slides (chips) available. The source of mRNA (e.g., tissue samples or cell lines) will affect the amount of mRNA available, and in turn the number of replicate slides possible.

Other information to keep in mind includes the experimental process prior to hybridization such as sample isolation, mRNA extraction, amplification and labeling. These and other technical matters are discussed in Schena (2000) and Bowtell and Sambrook (2002). Keeping track of all the different aspects of the experimental process helps us better understand the different levels of variability affecting our microarray data. Finally, consideration should be given to the verification method following the experiments, such as Northern or Western blot analysis, real-time PCR, or in-situ hybridization. The amount of verification to be carried out can influence statistical methods used and the determination of sample size. All this information helps us translate an experiment’s biological goals into the corresponding statistical questions and then, following appropriate design choices, helps us obtain a ready interpretation of the results.

We begin our discussion of the design of experiments when there is just one natural design choice, when one design stands out as preferable to all others, given the nature of the experiment and the material available. For example, suppose that we wish to study mRNA from two or more populations of cells, each treated by a different drug, and that the primary comparisons of interest are those of the treated cells versus the untreated cells. In this case, the appropriate design is clear: the untreated cells become a de facto reference, and all hybridizations involve one treated set of cells and the untreated cells. Next, suppose that we have collected a large number of tumor samples from patients. If the scientific focus of the experiment is on discovering tumor subtypes (Alizadeh et al., 2000), then the design involving comparisons between all the different tumor samples and a common reference RNA is a natural choice. In both cases, the choice follows from the aim of the study, with statistical efficiency considerations playing only a small role.

The statistical principles of experimental design are randomization, replication, and local control; naturally, these all apply to two-color microarray experiments, especially the last two. We have so far found only limited opportunities for randomization, however, and the development of appropriate ways of randomizing microarray experiments would be a useful research project. In this and many similar laboratory
EXPERIMENTAL DESIGN

contexts, the challenge is to balance the requirements of uniformity (e.g., of reagents, techniques, technicians, perhaps even time of day for the experiment, which aims to reduce unnecessary variation, with the statistical need to provide valid estimates of experimental error). The situation recalls the discussion between R.A. Fisher and W.S. Gosset (“Student”) in the 1930s concerning the relative merits of random and systematic layouts for field experiments; see Pearson and Wishar (1958) and Bennett (1971–1974) for the papers. A number of issues are highly specific to experimental design in the microarray context, and we now turn to a brief discussion of some of them. Parts of what follows will also be relevant to nylon membrane and high-density oligonucleotide arrays.

**Replication**

As indicated earlier, and consistent with statistical tradition (Fisher, 1926), replication is a key aspect of comparative experimentation, its purpose being to increase precision and, more important, to provide a foundation for formal statistical inference. In the microarray context, a number of different forms of replication occur. The differences are all in the degree to which the replicate data may be regarded as independent, and in the populations that experimental samples are seen to represent. Given that replicate hybridizations are almost invariably carried out by the same person, using the same equipment and protocols, and frequently at about the same time, it is inevitable that replicate data will share many features. Most of the differences discussed next concern the target mRNA samples.

**Duplicate spots**

Many groups spot cDNA in duplicate on every slide, frequently in adjacent positions. At times, even greater within-slide replication is used, particularly with smaller customized rather than the larger general clone sets. This practice provides valuable quality information, as the degree of concordance between duplicate spot intensities or relative intensities is an excellent quality indicator; however, because replicate spots on the same slide, particularly adjacent spots, will share most if not all their experimental conditions, the data from the pairs cannot be regarded as independent. Although averaging log-ratios from duplicate spots is appropriate, their close association means that the information is less than that from pairs of truly independent measurements. Typically, the overall degree of concordance between duplicate spots is noticeably greater than that observed between the same spot across replicate slides, although exceptions exist.

**Technical replicates**

This term is used to denote replicate hybridizations where the target mRNA is from the same pool (i.e., from the same biological extraction). It has been observed that
characteristic, repeatable features of extractions exist, and this leads us to conclude that technical replicates generally involve a smaller degree of variation in measurements than the biological replicates described next. Usually, the term technical replicate includes the assumption that the mRNA sample is labeled independently for each hybridization. A more extreme form of technical replication would be when samples from the same extraction and labeling are split, and replicate hybridizations done with subsamples of this kind. We do not know of many labs now doing this, though some did initially. The section on “Correlation and technical replicates” discusses in more detail how technical replicates affect design decisions.

Replicate slides: Biological replicates—type I

This term refers to hybridizations involving mRNA from different extractions, for example, from different samples of cells from a particular cell line or from the same tissue. In most cases, this will be the most convenient form of genuine replication.

Replicate slides: Biological replicates—type II

This term is used to denote replicate slides where the target mRNA comes from the same tissue but from different individuals in the same species or inbred strain, or from different versions of a cell line. This form of biological replication is different in nature from the type I biological replicates described previously, and typically involves a much greater degree of variation in measurements. For example, experiments with inbred strains of mice have to deal with the inevitability of different mice having their hormonal and immune systems in different states, the tissues exhibiting different degrees of inflammatory activity, and so on. With non-inbred individuals, the variation will be greater still.

The type of replication to be used in a given experiment depends on the precision and on the generalizability of the experimental results sought by the experimenter. In general, an experimenter will want to use biological replicates to support the generalization of his or her conclusions, and perhaps technical replicates to reduce the variability in these conclusions. Given that several possible forms of technical and biological replication usually exist, judgment will need to be exercised on the question of how much replication of a given kind is desirable, subject to experimental and cost constraints. For example, if a conclusion applicable to all mice of a certain inbred strain is sought, experiments involving multiple mice, preferably a random sample of such mice, must be performed.

Note that we do not discuss sample size determination or power in this chapter. Despite the existence of research showing how to determine sample size for microarray experiments based on power considerations, we do not believe that this is possible. Our reasons are outlined in Yang and Speed (2002).
**EXPERIMENTAL DESIGN**

### Dye-swaps

Most two-color microarray experiments suffer from systematic differences in the red and green intensities which require correction. Details of normalization are discussed in Chapter 1. In practice, it is very unlikely that normalization can be done perfectly for every spot on every slide, leaving no residual color bias. To the extent that this occurs, not using dye-swap pairs will leave an experiment prone to a systematic color bias of an unknown extent. When possible, we recommend using dye-swap pairs. Alternatively, random dye assignments may be used, in effect including the bias in random error. A theoretical analysis of the practice of dye-swapping has yet to be presented.

### Extensibility

Often, experimenters want to compare essentially arbitrarily many sources of mRNA over long periods of time. One method is to use a common reference design for all experiments, with the common reference being a “universal” reference RNA that is derived from a combination of cell lines and tissues. Some companies provide universal reference mRNA (see e.g., [http://www.stratagene.com/gc/universal_browser.html](http://www.stratagene.com/gc/universal_browser.html)) while many individual labs create their own common reference pool. Common references provide extensibility of the series of experiments within and between laboratories. When an experimenter is forced to turn to a new reference source, it may be difficult to compare new experiments with previous ones that were performed based on a different reference source. The ideal common reference, therefore, is widely accessible, available in unlimited amounts, and provides a signal over a wide range of genes. In practice, these goals can be difficult to achieve. When a universal reference RNA is no longer available, it is necessary to carry out additional hybridizations, conducting what we term a linking experiment to connect otherwise unrelated data. More generally, linking experiments allow experimenters to connect previously unrelated experiments, with the number of additional hybridizations depending on the precision of conclusions desired. Suppose that in one series of experiments we used reference $R_1$, and that in another series we used reference $R_2$. The linking experiment compares $R_1$ and $R_2$, thus permitting comparisons between two sources, one of which has been co-hybridized with $R_1$ and the other with $R_2$; however, this ability comes at a price. Log-ratios for source $A$ co-hybridized with reference $R_1$ can be compared to ones from source $Z$ co-hybridized with reference $R_2$, only by combining $A \to R_1$ and $Z \to R_2$ together with $R_1 \to R_2$ through the identity $\log_2(A/Z) = \log_2(A/R_1) + \log_2(R_1/R_2) - \log_2(Z/R_2)$. In other words, cross-reference comparisons involve combining three log-ratios, with corresponding loss of precision, as the variance of $\log_2(A/Z)$ here is three times that of the individual log-ratios. Nevertheless, there will be times when cross-referencing is worthwhile, particularly when one notices that the variance of the linking term $\log_2(R_1/R_2)$ can be reduced to an extent thought desirable simply by replicating that experiment. The linking term in the identity would be replaced by the average of all such terms across replicates.
Robustness

Loosely speaking, we call a design robust if the efficiency with which effects of interest are estimated does not change much when small changes are made to the design, such as those following from the loss of a small number of observations, or a change in the correlation structure of the observations. It is not uncommon for hybridization reactions to fail in microarray experiments; in a context where mRNA is hard to obtain, an experiment may well have to proceed without repeating failed hybridizations. In such cases, robustness is a highly desirable property of the design. A situation to be avoided is one where key comparisons of interest are estimable only when a particular hybridization is successful. It follows that heavy reliance on direct comparisons is not so desirable. Preferable is the situation where all quantities of interest are estimated by a mix of direct and indirect comparisons, that is, where many different paths connect samples in the design graph.

A nonstandard way to improve the robustness of a design is to give careful thought to the order in which the different hybridizations in the experiment are carried out. More critical hybridizations could be done earlier, and full sets of hybridizations completed before further replicates are run, leaving the greatest opportunity for revising the design in the case of failed hybridizations. Note that this practice is contrary to the generally desirable practice of randomizing the order in which the parts of an experiment (here hybridizations) are carried out. In this context, such randomization is frequently achievable, but it is not popular with experimenters because it will often require a greater number of preparatory steps and so an increased risk of failure.

As we will see later, the use of technical replicates introduces correlation between measured intensities and relative intensities. The precise extent of this correlation is typically difficult to measure. A design where the performance is more stable across varying technical replicate correlations would usually be preferred to one that is more efficient for one range of values of the unknown correlations, but less efficient for another range of values. This is a different form of robustness.

Pooling

An issue arising frequently in microarray experiments is the pooling of mRNA from different samples. At times pooling is necessary to obtain sufficient mRNA to carry out a single hybridization. At other times, biologists wonder whether pooling improves precision even when it is not necessary. Is this a good idea?

To sharpen the question, suppose that we wish to compare mRNA from source A with that from source B, using three hybridizations. We could carry out three separate extractions and labellings from each of the sources, arbitrarily pair A and B samples, and do three competitive hybridizations, each a single A sample versus a single B sample. We would then average the results; see Section 2.3 ("Two-sample comparisons") for further discussion. Alternatively, we could pool the labelled mRNA samples, one pool for the A and one for the B samples, then subdivide each pool into
three technical replicates, and carry out three replicate competitive hybridizations of pooled A versus pooled B. Again, the results of the 3 comparisons would be averaged. Which is better? An analogous question can be posed with the single-color hybridizations (high-density short-oligonucleotide arrays, and nylon membranes). An *a priori* argument can be made for either approach. Pooling may well improve precision, that is, reduce the variance of comparisons of interest. But does it do so at the price of permitting one sample (or a few) to dominate the outcome, and so give misleading conclusions overall? These are hard questions to answer.

We know of no experiment with two-color microarrays aimed at answering these questions, but we have seen the results of such an experiment with the Affymetrix technology (Han et al., 2002). There we saw that averaging pooled samples and then comparing across the types described previously was slightly more precise than doing the same thing with averaged results from single samples, and in that case there seemed to be no obvious biases from individual samples. Our conclusion was that the gain in precision arising from pooling probably does not justify the risks, and that it is probably better to be able to see the between-sample variation, rather than lose the ability to do so. It remains to be seen whether this conclusion will stand up over time, and whether it applies to two-color microarrays.

*Our design focus*

With most experiments, a number of designs can be devised that appear suitable for use, and we need some principles for choosing one from the set of possibilities. The remainder of this chapter focuses on the question of identifying differentially expressed genes, and discusses design in this context. The identification of differentially expressed genes is a question that arises in a broad range of microarray experiments (Callow et al., 2000; Friddle et al., 2000; Galitski et al., 1999; Golub et al., 1999; Spellman et al., 1998). The types of experiments include: single-factor cDNA microarray experiments, in which one compares transcript abundance (i.e., expression levels) in two or more types of mRNA samples hybridized to different slides. Time-course experiments, in which transcript abundance is monitored over time for processes such as the cell cycle, can be viewed as a special type of single-factor experiment with time being the sole factor. We discuss them briefly from this perspective. Factorial experiments, where two or more factors are varied across the mRNA are also of interest, and we discuss their design and analysis as well.

### 2.3 Two-sample comparisons

The simplest type of microarray experiment is the two sample or binary comparison, where we seek to identify genes that are differentially expressed between two sources of RNA. Such comparisons might be between knock-out and wild-type, tumor and normal, or treated and control cells. With “single color” systems, such as the nylon membranes and high-density short oligonucleotide arrays, the comparisons can be
COMPARATIVE MICROARRAY EXPERIMENTS

between results from two arrays, or two sets of arrays. With the two-color cDNA or long oligonucleotide arrays, the comparison can be within a single slide, across each of a set of replicate slides involving direct comparisons, or involving indirect comparisons between slides.

Case Studies I and II

We illustrate the ideas of this section with two sets of two sample comparisons. These studies both aim to identify differentially expression between a mutant and a wild-type organism, but they do it differently. Both studies are with two-color cDNA microarrays. Case Study I involves replicates of direct comparisons made within a slide. By contrast, Case Study II involves indirect comparisons between samples co-hybridized to a common reference mRNA.

In order to identify and remove systematic sources of variation in the measured expression levels and allow between-slide comparisons, the data for experiments I and II (as well as experiments III and IV introduced next) were normalized using the within-slide spatial and intensity dependent normalization methods described in Yang et al. (2002b). Normalization methods are discussed in more detail in Chapter 1, and we make no further mention of the topic, apart from noting that it is a critical preprocessing step with almost any microarray experiment.

Case Study I: Swirl zebrafish experiment

The results from the swirl zebrafish experiment were given to us by Katrin Wuennenberg-Stapleton from the Ngai Lab at University of California, Berkeley, while the swirl embryos themselves were generously provided by David Kimelman and David Raible from the University of Washington in Seattle. The experiment was carried out using zebrafish to study early development in vertebrates. Swirl is a point mutation in the BMP2 gene that causes defects in the organization of the developing embryo along its dorsal-ventral body axis. This results in a reduction of cells showing ventral cell fates (i.e., cell types that are normally formed only within the ventral aspect of the embryo), such as blood cells are reduced, whereas dorsal structures such as somites and notochord are expanded. A goal of this swirl experiment was to identify genes with altered expression in the swirl mutant compared to the wild-type zebrafish. The data are from four replicate slides: two sets of dye-swap pairs. For each of these slides, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild-type mutant is labeled using the other dye. Figure 2.3 is the graphical representation of this experiment.

In this case study, target cDNA was hybridized to microarrays containing 8848 cDNA probes. The microarrays were printed using $4 \times 4$ print-tips and are thus partitioned into a $4 \times 4$ grid matrix. Each grid consists of a $22 \times 24$ spot matrix that was printed with a single print-tip. In this and the other studies discussed next, we call the spotted
Case Study I: The swirl experiments provided by Katrin Wuiennberg-Stapleton from the Ngai Lab at the University of California, Berkeley. This experiment consists of two sets of dye swap experiments comparing gene expression between the mutant swirl and wild-type (wt) zebrafish. The number on the arrow represents the number of replicated experiments.

cDNA sequences “genes,” whether or not they are actual genes, ESTs (expressed sequence tags), or cDNA sequences from other sources.

Case Study II: Scavenger receptor BI mouse experiment

The scavenger receptor BI (SR-BI) experiment was carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice, Callow et al. (2000). The SR-BI gene is known to play a pivotal role in high-density lipoprotein (HDL) metabolism. Transgenic mice with the SR-BI gene overexpressed have very low HDL cholesterol levels, and the goal of the experiment was to identify genes with altered expression in the livers of these mutant mice compared to “normal” FVB mice. The treatment group consisted of eight SR-BI transgenic mice, and the control group consisted of eight normal FVB mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and labeled using the red-fluorescent dye Cy5. The reference sample used in all 16 hybridizations was prepared by pooling cDNA from the eight control mice and was labeled with the green-fluorescent dye Cy3. The design would have been better if the reference sample had come from a different set of control mice. In this experiment, target cDNA was hybridized to microarrays containing 5548 cDNA probes, including 200 related to lipid metabolism. These microarrays were printed in a $4 \times 4$ matrix of sub-arrays, with each sub-array consisting of a $19 \times 21$ array of spots. The data are available.

Case Study II: The SR–BI experiments provided by Matt Callow from the Lawrence Berkeley National Laboratory. This experiment consists of eight slides comparing gene expression between the transgenic SR–BI mice and the pooled control (WT*). Another eight slides comparing gene expression between normal FVB (WT) mouse and pooled control. The number on the arrow represents the number of replicated experiments.
Comparative microarray experiments


Single-slide methods

A number of methods have been suggested for the identification of differentially expressed genes in single-slide, two-color microarray experiments. In such experiments, the data for each gene (spot) consist of two fluorescence intensity measurements, \((R, G)\), representing the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples, respectively. (The most commonly used dyes are the cyanine dyes, Cy3 and Cy5, however, other dyes such as fluorescein and X-rhodamine may be used as well). We distinguish two main types of single-slide methods: those which are based solely on the value of the expression ratio \(R/G\) and those that also take into account overall transcript abundance measured by the product \(RG\).

Early analyses of microarray data (DeRisi et al., 1996; Schena et al., 1995, 1996) relied on fold increase/decrease cutoffs to identify differentially expressed genes. For example, in their study of gene expression in the model plant Arabidopsis thaliana, Schena et al. (1995) use spiked controls in the mRNA samples to normalize the signals for the two fluorescent dyes (fluorescein and lissamine) and declare a gene differentially expressed if its expression level differs by more than a factor of 5 in the two mRNA samples. DeRisi et al. (1996) identify differentially expressed genes using a \(\pm 3\) cutoff for the log-ratios of the fluorescence intensities, standardized with respect to the mean and standard deviation of the log ratios for a panel of 90 “housekeeping” genes (i.e., genes believed not to be differentially expressed between the two cell types of interest).

More recent methods have been based on statistical modeling of the \((R, G)\) pairs and differ mainly in the distributional assumptions they make for \((R, G)\) in order to derive a rule for deciding whether a particular gene is differentially expressed. Chen et al. (1997) propose a data dependent rule for choosing cutoffs for the red and green intensity ratio \(R/G\). The rule is based on a number of distributional assumptions for the intensities \((R, G)\), including normality and constant coefficient of variation. Sapir and Churchill (2000) suggest identifying differentially expressed genes using posterior probabilities of change under a mixture model for the log expression ratio \(\log R/G\) (after a type of background correction, the orthogonal residuals from the robust regression of \(\log R\) versus \(\log G\) are essentially normalized log expression ratios). A limitation of these two methods is that they both ignore the information contained in the product \(RG\). Recognizing this problem, Newton et al. (2001) consider a hierarchical model (Gamma–Gamma–Bernoulli model) for \((R, G)\) and suggest identifying differentially expressed genes based on the posterior odds of change under this hierarchical model. The odds are functions of \(R + G\) and \(RG\) and thus produce a rule which takes into account overall transcript abundance. The approach of Hughes et al. (2000b) (supplement) is based on assuming that \(R\) and \(G\) are approximately independently and normally distributed, with variance depending on the mean. It thus also produces a rule which takes into account overall transcript abundance.
TWO-SAMPLE COMPARISONS

Figure 2.5 (See color insert following page xxx.) Single-slide methods: An MA-plot showing the contours for the methods of Newton et al. (2001) (orange, odds of change of 1:1, 10:1, and 100:1), Chen et al. (1997) (purple, 95% and 99% “confidence”), and Sapir and Churchill (2000) (cyan, 90%, 95%, and 99% posterior probability of differential expression). The points corresponding to genes with adjusted p-value less than 0.05 (based on data from 16 slides) are colored in green (negative t-statistic) and red (positive t-statistic). The data are from transgenic mouse 8.

As a result, each of these methods produces a model dependent rule which amounts to drawing two curves in the \((\log R, \log G)\)-plane and calling a gene differentially expressed if its \((\log R, \log G)\)-falls outside the region between the two curves. We apply Chen et al. (1997), Newton et al. (2001), and Churchill (2000)* single-slide methods to one slide from the SR–BI experiment. The different methods are used to identify genes with differential expression in mRNA samples from individual treatment mice compared to pooled mRNA samples from control mice. Using an MA-plot, Figure 2.5 displays the contours for the posterior odds of change in the Newton et al. (2001) method, the upper and lower limits of the Chen et al. (1997) 95% and 99% “confidence intervals” for \(M\), and the contours for Sapir and Churchill’s 90%, 95%, and 99% posterior probabilities of differential expression. The regions between the contours for the Newton et al. (2001) method are wider for low and high intensities \(A\), this being a property of the Gamma distribution which is used in the hierarchical model. The genes identified as having differential expression between the SR–BI

* Note that we are not performing the orthogonal regression for the log transformed intensities (Part I of the poster). The orthogonal residuals of Sapir and Churchill are essentially normalized log expression ratios. We have simply implemented Part II of the poster and are applying the mixture model to our already normalized log-ratios.
transgenic and the wild-type on the basis of all 16 slides (Callow et al., 2000) are highlighted in green (down-regulated) and red (up-regulated) in the figure. None of the methods satisfactorily identify all 13 genes found using all of the data on this slide, and the nature of their failure strongly suggests that these methods should not be relied upon in general. In our view, the statistical assumptions the different methods make are just too strong, and inconsistent with the data, being unlikely to capture the systematic and random variability inherent these data. Furthermore, it is hard to see how a within-slide error model can capture between-slide variation, and error probabilities relating to detection of differentially expressed genes should relate to repeated hybridizations.

No single-slide or single-chip comparison exists for radiolabeled target hybridized to cDNA spots on nylon membranes or for high density short oligo arrays (i.e., for the single color systems). In order to compare mRNA from two cell populations in these cases, we need at least two nylon membranes or two chips, including at least one with target mRNA from each of the populations of interest. Assuming that we have exactly one membrane or chip from each of the cell samples, the problem becomes formally quite similar to the single-slide, two-color problem just discussed, though the details differ in important ways between nylon membranes and for high-density short oligo chips. With exactly two nylon membranes the situation really is quite similar to a single two-color slide, in that we have no more to go on that the two log intensities for each spot. Thus, determining differentially expressed genes can be no more than drawing lines in the associated plane, and the previous discussion applies, though not all the methods mentioned have been advocated in the nylon filter context. With high-density short oligo arrays, the situation is better. Approximately 11–20 probes for each gene or EST, and so there is information that permits us to estimate a standard error for each estimated log-ratio, or to carry out a significance test. Details of the Affymetrix methods for comparing two chips can be found at the following Web site: http://www.affymetrix.com/products/software/specific/mas.affx. This approach works reasonably well in practice, though it is not clear that the \( p \)-values can be given their usual interpretation.

**Replicate slides: design**

Before considering methods for identifying differentially expressed genes involving replicates slides, let us briefly discuss the design question for the simple treatment-control comparison with two-color arrays.

Consider the two designs described in Figure 2.2. The goal of both designs is to compare two target samples \( T \) and \( C \), and identify differentially expressed genes between them. Suppose that we plan on doing two hybridizations, and that quantity of RNA is not a limiting factor. For a typical gene on a slide, we denote the intensity value for the two target samples by \( T \) and \( C \). The log base 2 transformation of these values will be written \( \log_2 T \) and \( \log_2 C \), respectively, and when reference samples \( R \) and \( R' \) are used, we will write \( \log_2 R \) and \( \log_2 R' \). In addition, we denote the means of
the log-intensities across slides for a typical gene by $\alpha = \mathbb{E} \log_2 T$ and $\beta = \mathbb{E} \log_2 C$, respectively. Then, for the gene under study, $\phi = \alpha - \beta$ is the parameter representing the differential expression between samples $T$ and $C$, which we want to estimate. The variances and covariances of the log-intensities for a typical gene across slides will be assumed to be the same for all samples, that is, we suppose that differential gene expression is exhibited only through mean expression levels, and we always view this on the log scale. In addition, we assume for the moment that the replicate measurements on different slides are independent. For any particular gene, let us assume that $\sigma^2$ is associated with the variance for one such measurement. (This may vary from gene to gene.) It follows that the direct estimate of the differential expression and its corresponding variance are:

$$\hat{\phi}_D = \frac{1}{2} (\log_2(T/C) + \log_2(T'/C'))$$

and 

$$\text{var}(\hat{\phi}_D) = \sigma^2/2$$

respectively. Alternatively, if we make use of a common reference $R$ say, then from our two hybridizations, the indirect estimate of log-ratio and its variance are:

$$\hat{\phi}_I = \log_2(T/R) - \log_2(C/R')$$

and 

$$\text{var}(\hat{\phi}_I) = 2\sigma^2$$

The resulting relative efficiency of the indirect versus the direct design for estimating $\alpha - \beta$ is thus 4. This is the key difference between direct and indirect comparisons, and the reason why we recommend under many circumstances that direct comparisons are to be preferred. The factor 4 depends critically on our independence assumption, but we will see shortly that under very general assumptions, the direct comparison is never less precise than the indirect one.

**Replicate slides: direct comparisons**

A number of approaches can be used here, and we briefly discuss each of them.

**Classical**

Suppose that we have $n$ replicate hybridizations between mRNA samples $A$ and $B$. For each gene, we can compute the average $\bar{M}$ and the associated variance $s^2$ of the $n$ log-ratios $M = \log_2 A/B$. In line with the early work summarized above, it would be natural to identify differentially expressed genes by taking those whose values $|\bar{M}|$ exceed some threshold, perhaps one determined by the spread of $\bar{M}$ values observed in related self-self hybridizations. It would be equally natural to statisticians to calculate the $t$-statistic $t = \sqrt{n} \bar{M} / s$, and make decisions on differential expression on the basis of $|t|$. Both strategies are reasonable, the first implicitly assigning equal variability to every gene, the second explicitly permitting gene-specific variances across slides; however, neither strategy is entirely satisfactory on its own. Large values of $\bar{M}$ can be driven by outliers, as the value of $n$ is typically quite small (in our experience $\approx 2 - 8$), and the technology is quite noisy. On the other hand, with tens of thousands of $|t|$ statistics, it is always the case that some are quite large in comparison with
the others because their denominators $s$ are very small, even though their numerators $\sqrt{n}|M|$ may also be quite small, perhaps almost zero.

**Empirical Bayes**

Several more or less equivalent solutions are available to the problem of very small variances giving rise to large $t$-statistics, ones which lead to a compromise between solely using $t$ and solely using $\bar{M}$. One solution is to discount genes with a small $\bar{M}$ whose standard errors are in the bottom 1%, for example. This leaves open the choice of cutoffs on $\bar{M}$ and the standard error. More sophisticated solutions in effect standardizes $\bar{M}$ by something midway between a common and a gene-specific standard error. For example, Efron et al. (2000) slightly tune the $t$-statistic by adding a suitable constant to each standard deviation, using

$$t^* = \frac{\sqrt{n}\bar{M}}{a + s}$$

One choice for $a$ is the 90th percentile of standard deviations, while another minimizes the coefficient of variation (see Efron et al., 2000; Tusher et al., 2001). This solution recalls the empirical Bayes (EB) approach to inference, which is natural in the microarray context where thousands of genes exist. A more explicitly EB approach, which is almost equivalent to the preceding one, apart from the choice of $a$, is presented in Lönnstedt and Speed (2001); we illustrate it next. In addition to these two just cited, there are other EB formulations of the problem of identifying differentially expressed genes (see e.g., Efron et al., 2001; Long et al., 2001; Baldi and Long, 2001).

In Lönnstedt and Speed (2001), data from all the genes in a replicate set of experiments are combined into estimates of parameters of a prior distribution. These parameter estimates are then combined at the gene level with means and standard deviations to form a statistic $B$, which is a Bayes log posterior odds for differential expression. $B$ can then be used to determine if differential expression has occurred. It avoids the problems of the average $\bar{M}$ and the $t$-statistic just mentioned. In the same article, a comparison is conducted between the $B$-statistic, the previous statistics $t^*$, and truncation of small standard errors. The differences are not great.

Note that the preceding analysis, and others like it, treats $M$ values from different genes as conditionally independent, given the shared parameters, which is very far from the case, although it may not matter much at this point. Any attempt to provide a semiformal analysis of replicated log-ratios may require this assumption, or something very similar to it. By contrast, the permutation-based analyses described next do not require this assumption, but they do not apply here.

**Robustness**

The EB solution to the problems outlined in the classical approach focus on smoothing the empirical variances of the genes, thereby avoiding the situation where tiny variances can create large $t$-statistics. A different approach is to replace $\bar{M}$ in the
numerator, in effect avoiding the problems which result from outliers coupled with small sample sizes. The obvious solution is to use a robust method of estimation of the parameter $\phi$ wherever possible (Huber, 1981; Hampel et al., 1986; Marazzi, 1993). Alternatively, use could be made of a nonparametric testing procedure, for example one based on ranks. The problem here is that the sample sizes are small, not infrequently as low as two or three.

Let us note in passing that the use of robust estimates of location with sample sizes as small as two or three is not without its problems. In such cases, estimates of standard errors can hardly be relied upon, and instability in the associated “$t$-statistics” is not uncommon. Users of standard robust procedures such as $r1m$ in R and SPlus should take care and not simply rely upon default parameter settings in the algorithms. It is to be hoped that the greatly increased use of robust methods stimulated by microarray data will lead to further research on this topic.

**Mixture models**

Many approaches to identifying differentially expressed genes in this context use mixture models, including the EB ones just discussed. Lee et al. (2000) use a two-component normal mixture model for log-ratios in two-color arrays, one component for differentially expressed genes and another for the remainder of the genes. They estimate the parameters of their model by maximum likelihood and compute posterior probabilities using the estimated parameters. In a sense, this is another EB model, but not one with gene-specific variances, and so quite different in character from those discussed previously. Efron et al. (2001) also have a two-component mixture model, but it is for Affymetrix GeneChip data, and it is more general that the previous one. Their mixture model is for the statistic $t^*$ instead of for log-ratios, and they make no parametric assumptions about their mixture components. More recently, Pan (2002) discussed a multicomponent normal mixture model, extending the of analysis of Lee et al. (2000) toward that of Efron et al. (2001).

**Fixed and random effects linear models**

One approach to the analysis of two-color microarray data and the determination of differentially expressed genes makes use of linear models and the analysis of variance (Kerr et al., 2001; Wolfinger et al., 2001; Jin et al., 2001). These authors model un-normalized log intensities with linear models which include terms for slide, dye, gene and treatment, a subset of the interactions between these effects, and a random error term. Important differences exist between the approaches of Kerr et al. (2001) and the other two papers in the way in which normalization is incorporated, in whether terms are fixed or random, and in assumptions about the error variances. We illustrate the approach of Kerr et al. (2001) next, and we begin by presenting their model for our Case Study I. We label array effects (A) by $i$, dye effects (D) by $j$, treatment effects (V) by $k$ and gene effects (G) by $g$. Their model for the log of the intensity $y_{ijkg}$ is:

$$\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg}$$
where \( i = 1, \ldots, 4 \), \( j = 1, 2 \), \( k = 1, 2 \), and \( g = 1, \ldots, 8848 \). In Case Study I, which consists of a two sample comparison, the treatment effect \( V_k \) represents the mutant \((V_2)\) or wild-type \((V_1)\) samples. In this model, all terms are fixed apart from the random error terms \( \epsilon \). The term of interest here is \((VG)_{kg}\) and the value \((VG)_{2g} - (VG)_{1g}\) estimates the level of differential expression between the mutant and the wild type samples for gene \( g \). This model can easily be extended to cover multiple samples and factorial designs, although we do it differently in the section 3 on “Linear model analyses”. Further discussion of fixed and random effects linear model is given there.

Error models

Several groups have used more fully developed error models for measurements on microarrays, and sought to identify differentially expressed genes by making use of their error model. These include Roberts et al. (2000), Ideker et al. (2000), Rocke and Durbin (2001), Theilhaber et al. (2001), and Baggerly et al. (2001). Ideker et al. (2000) has made their software publicly available, so we illustrate this approach in Case Study I. Their error model is for pairs \( T \) and \( C \) of unnormalized un-logged intensities for the same spot, namely,

\[
T = \mu_T + \mu_T \epsilon + \delta \\
C = \mu_C + \mu_C \epsilon' + \delta'
\]

where \((\epsilon, \epsilon')\) and \((\delta, \delta')\) are independently bivariate normally distributed across spots with means \((0, 0)\) and separate, general covariance matrices that are common to all spots on the array. Thus, six parameters exist for the error model, in addition to parameters determining the expected values. This model is close, but not quite identical to a similar model for \((\log T, \log C)\), one that permits the two components to be correlated, with a correlation which is common to all the spots. Intensities from distinct spots on the same slide, or spots on different slides are taken to be independent. Roberts et al. (2000) and Rocke and Durbin (2001) use a similar model, but do not allow the different channel intensities to be correlated.

In order to identify differentially expressed genes, Ideker et al. (2000) fit the model to two or more slides by maximum likelihood. Differential expression is then determined by carrying out a likelihood ratio test of the null hypothesis \( \mu_T = \mu_C \) for each gene separately, resulting in a likelihood ratio test statistic \( \lambda \) for each gene.

Other approaches

Working with Affymetrix GeneChip data, Thomas et al. (2001) use a variant on the simple two-sample \( t \)-statistic which starts from a nonlinear model including sample-specific additive and multiplicative terms. In the same GeneChip context, Theilhaber et al. (2001) present a fully Bayesian analysis, building on a detailed error model for such data. We refer readers to these articles for further details.
Replicate slides: indirect comparisons

When the comparisons of log-ratios in a two-color experiment are all indirect, the procedures just described need to be modified slightly. Typically, we would have some number, \( n_T \), of slides on which the gene expression of sample \( T \) is compared to that of a reference sample \( R \), leading to \( n_T \) log-ratios \( M = \log_2 T/R \), and a similar set of \( n_C \) log-ratios \( \log_2 C/R' \) from sample \( C \). The analogues of the average and \( t \)-statistics here are the differences between the means \( \bar{M}_T - \bar{M}_C \) and the two-sample \( t \)-statistic is

\[
t = \frac{\bar{M}_T - \bar{M}_C}{s_p \sqrt{1/n_T + 1/n_C}}
\]

where \( s_p \) is the pooled standard deviation.

The problems with these two statistics are completely analogous to those described in the previous section with \( M \) and \( t \), and the solutions are similar: modify \( s_p \) or use robust variants of \( \bar{M}_T - \bar{M}_C \) and \( t \).

We note in closing this brief review that some non-statisticians addressing these two-sample problems in the microarray literature have devised novel approaches. Galitski et al. (1999) and Golub et al. (1999) sought to identify single differentially expressed genes by computing for each gene the correlation of its expression profile with a reference expression profile, such as a vector of indicators for class membership. In the case of two classes, this correlation coefficient is a type of \( t \)-statistic. Genes were then ranked according to their correlation coefficients, with a cutoff derived from a permutation distribution.

Illustrations using our case studies

We now describe the results of applying some of the methods just discussed to Case Studies I and II. It should be understood that we are not attempting to present thorough analyses of these data sets, in part because of lack of space, and in part because more effort always goes into the determination of differential expression than the application of a single statistical analysis. Nevertheless, we hope that what follows gives an indication of the potential of the methods we illustrate. We have certainly found them useful in similar contexts, particularly the graphical displays.

Plots: averages, SDs, \( t \)-statistics and overall expression levels

Important features of the genes which might be differentially expressed can be found by examining plots of \( t \)-statistics, their numerators \( M \) and denominators \( s \), and the corresponding overall expression levels. The overall expression level for a particular gene is conveniently measured by the quantity \( A \), the average of \( A = \log_2 \sqrt{RG} \) over all the slides in the experiment.
Let us begin with the swirl experiment. In Figure 2.6b, 2.6c and 2.6d we have plotted the average log ratio $\bar{M}$, the $t$-statistic and the log-likelihood ratio statistic $\lambda$ from Ideker et al. (2000) against $\bar{A}$. We began by defining three groups of 250 spots each, being those having the largest values of $|M|$, $|t|$ and $B$, respectively, and then plotted the points according to the color code depicted in Figure 2.6a. Thus, points corresponding to spots in all three groups are colored heavy black, while the very light black spots, the overwhelming majority, are in none of the groups. Clearly, the heavy black spots are the main candidates for differential expression. Spots belonging to large $|t|$ group only are green, while those in the large $|M|$ and large $B$ and not large $|t|$ groups are pink. The absence of points colored yellow is noteworthy: any spots with large $|t|$ and large $|M|$ already have large $B$. This is not uncommon, and shows that $B$ is, in general, a useful compromise between $t$ and $M$.

It is clear from these plots, and is not infrequently the case, that some points are well separated from the cloud. The genes corresponding to these points are likely to be differentially expressed, and we recommend this informal approach identifying such genes. In many cases, this evidence is as solid as any we are able to obtain for differential expression. Note the broad agreement on the black points in all three of the panels. Figure 2.7a shows the nature of $B$ rather clearly, while Figure 2.7b shows the role of the standard deviation (SD) in determining whether a given $\bar{M}$ ends up
having a large $|t|$ as well, with it being clear how large $B$ related to the other two. The green points are those with smaller SDs, in comparison with the solid black, pink, and light blue spots. Of course, much of this is dependent on the cutoffs defining our groups, which were determined after looking at the plots, but the message of these plots is general.

Turning to the SR–BI experiment, Figure 2.8a shows a plot of $\bar{M}_T - \bar{M}_C$ against $\bar{A}$, while Figure 2.8b shows $t$ against $\bar{A}$. Here, we have given no analogue of $B$, though one could be developed. In this case it is the yellow spots that should attract our attention, and it should be clear that our choice here of 250 in the two groups is too large. Looking at Figure 2.9a and 2.9b, we see that in general a large $|\bar{M}_T - \bar{M}_C|$ is more likely to go with a large $|t|$ if the SD is not too large and not too small, which
58 COMPARATIVE MICROARRAY EXPERIMENTS

Figure 2.9 (See color insert following page xxx.) Case Study II: SR–BI experiment. (a) \( t \) vs. \( \log_2(SD) \); (b) \( \log_2(SD) \) vs. \( \bar{A} \)

is roughly equivalent to having a \( \bar{A} \) value that is not too small and not among the largest ones. We see clearly from Figure 2.9b that larger \( \bar{A} \) values go with smaller values of the SD, and we remark that many plots of this kind would show a much greater increase in SD for lower values of \( \bar{A} \). That is often quite evident in the \( M\bar{A} \)-plot (Figure 2.8a) as a ballooning in the low intensity ranges, a phenomenon we have prevented by using a smaller and less variable background adjustment (see Yang et al., 2002a).

How would we use these statistics to identify differentially expressed genes? It is tempting to answer in the following way: Simply rank the genes on the basis of \( |M|, t \) or \( B \), and determine a cutoff in some sensible way. This is in effect what we have done in producing the plots, without being careful about cutoffs; however, the plots also tell us that a ranking based on just one of these statistics is not necessarily the best we can do. The reason is this: A spot’s overall intensity value \( \bar{A} \) can be a useful indicator of the importance that can be attached to its \( M, t, \) or \( B \) value. Typically, considerably fewer genes exist with large \( \bar{A} \) values, and a spot can stand out from the cloud in the larger \( \bar{A} \) range with a smaller \( |M|, t, \) or \( B \) value than than would be necessary to stand out in the low \( \bar{A} \) region. At times, the difference can be striking, although this is not the case with our illustrative examples. Most methods of identifying differentially expressed genes in the present context do not make explicit use of the overall intensity or any related values, although approaches using fully specified error models in effect do so. One problem with full error models is that they invariably assume that observations on different slides are independent, which can be very far from the case, and this vitiates their exact probability calculations.

In summary, we regard the determination of differentially expressed genes on the basis of a small number of replicates as a problem for which more research is needed. It is clear that the values of \( M, t, B \) (or their analogues) are highly relevant, but the values of \( \bar{A} \) and the SD should not be ignored. For the present moment, we feel that determining cutoffs is best done informally, following visual inspection of plots like the ones we have shown. Naturally, scientific considerations such as the expected
number of differentially expressed genes and the number of follow-up experiments that might be feasible, are also relevant in determining cutoffs, and we feel that $\bar{A}$ should also play a role.

**Ranking genes by sets of statistics**

Whether we use of $|\bar{M}|, |t|, B, \lambda$ or some similar statistic to provide a ranking of genes corresponding to the strength of evidence of differential expression, it is intuitively clear that our preferred genes should rank highly on all these criteria. Hero and Fleury (2002) describe a valuable method of selecting genes which are highly ranked in a suitable multidimensional sense. They describe what they term *Pareto fronts* in *multicriterion scattergrams*, which are points that are maximal in the componentwise ordering (Pareto optimal) in the $P$-dimensional scatterplots of a desired set of $P$ criteria. As well as giving some modified versions of Pareto optimal genes, they demonstrate that their ideas are applicable to a wide range of gene filtering tasks, not just that of detecting differentially expressed genes. We refer to the paper Hero and Fleury (2002) and also Fleury et al. (2002b) for a fuller discussion of the method.

**Assessing significance**

After ranking the genes based on a statistic, a natural next step is to choose suitable cutoff values defining the genes that might be considered as significant or differentially expressed. In this section, we consider the extent to which this can be done informally. We focus on two types of plots.

**Quantile–quantile plots**

A simple graphical approach is to examine the *Quantile–Quantile plots* (Q–Q plots) for certain statistics. Q–Q plots are a useful way to display the $\bar{M}$ or $t$-statistics for the thousands of genes being studied in a typical microarray experiment. The standard normal distribution is the natural reference, and the more replicates that enter into an average or $t$-like statistic, the more we can expect the majority of the statistics to look like a sample from a normal distribution. With just four replicates, which is equivalently eight log intensities entering into our averages, we cannot expect and do not get a very straight line against the standard normal; however, the plot can have value in indicating the extent to which the extreme $t$-statistics diverge from the majority. Q–Q plots informally correct for the large number of comparisons, and the points which deviate markedly from an otherwise linear relationship are likely to correspond to those genes whose expression levels differ between the two groups. At times, we can tell where the outliers end and the bulk of the statistics begin (see e.g., Dudoit et al. (2001b), but unfortunately this is not the case with either of the present examples.
Figure 2.10  Quantile-quantile (Q–Q) plot for (a) one-sample $t$-statistics from the swirl experiment; (b) two-sample $t$-statistics from the SR–BI experiment.

Figure 2.10a is typical of many Q–Q plots which are not very helpful. The tails of the distributions of these $t$-statistics are far from those of a normal (or $t$) distribution. The best we can hope for in such cases is that some points are obviously outliers with respect to the nonnormal distribution, and this is true to some extent here, especially for large negative $t$s.

The picture in Figure 2.10b is clearer. This is for the SR–BI experiment, and is perhaps what we would expect, as there are 16 observations in each of these $t$-statistics. We see about a dozen genes with “unusual” $t$-statistics, and these are obviously good candidates for genes exhibiting differential expression (both up and down-regulation). Some of these genes were verified to have the observed behaviour in follow-up experiments.

$p$-value vs. average $M$ (volcano) plot

Another plot which allows outliers to reveal themselves among thousands of statistics is the so-called volcano plot (Wolfinger et al., 2001), Figure 2.11. We have already seen one plot like this in Figure 2.7a, where the log-odds corresponding to a given value of the statistic $B$ was plotted against $M$. More commonly, people plot the logs of raw (i.e., model-based, unadjusted) $p$-values against the estimated fold change on a log scale, $\hat{M}$ (Wolfinger et al., 2001). Whether the $p$-values are calculated assuming a $t$ or a normal distribution is not so important here. The color code indicates how these plots capture some aspects of the plots we presented earlier in what is perhaps a more convenient form; see especially the solid black points in Figure 2.11a and the yellow points in Figure 2.11b.

So far, the approaches we have offered for identifying differential expression are all informal. In many cases, this will be adequate. The number of genes that are selected as possibly differentially expressed will in general depend on many things: the aims
of the experiment, the number of genes expected to be differentially expressed, the number of replicates, and the nature and extent of follow-up (validation). We cannot expect a formal procedure to take all these factors into account.

In order to examine more formally whether the extreme $t$-statistics do indeed reflect real differences between the control and transgenic mice we turn to adjusted $p$-values. These cannot always be calculated in a reliable way, but when they can, we feel they are worthwhile.

**Multiple comparisons**

A more formal approach to testing the null hypothesis of constant expression can be obtained by calculating $p$-values (or posterior probabilities) under some model. With a typical microarray dataset comprising thousands of genes, however, there are at least two major impediments to doing this correctly. One immediate concern for the validity of any single gene unadjusted $p$-values or posterior probabilities is the underlying statistical model. As we suggested, in our discussion of single-slide methods for identifying differentially expressed genes, parametric statistical models generally have difficulty adequately capturing all the details of microarray data; this is worst where it matters most for testing, namely, in the tails of the distributions of statistics. For example, collections of single-slide log-ratios $M$ are frequently approximately normally distributed, but this breaks down in the tails. We might hope for averages to behave more like normal random variables by virtue of the central limit theorem, but this will only be true when all systematic effects have been removed and the individual terms really do exhibit a high degree of independence. In certain
cases, permutation-based methods can be used, but this is not as frequent as we might like. As we will see, use of permutation methods can raise yet other problems.

The other major concern is the multiplicity issue: Carrying out thousands of significance tests (or computing thousands of posterior probabilities) brings with it the need to recognize and deal with the issues that arise when so many comparisons are made.

Two different approaches to the multiple comparisons problem have emerged in the microarray literature. One makes use of the traditional literature on the problem, seeking to control family-wise type I error rates (see Dudoit et al., 2002b), who build on the work of Westfall and Young, 1993), while the other develops and extends the notion of false discovery rate of Benjamini and Hochberg (1995) (see e.g., Efron et al., 2000; Tusher et al., 2001). We summarize and illustrate both Approaches; refer to Dudoit et al. (2002a) and Storey et al. (2002) for fuller details and further references. We also refer to Manduchi et al. (2000) for a contribution to this problem which falls outside the approaches just mentioned.

Consider the problem of simultaneously testing $m$ null hypotheses $H_j, j = 1, \ldots, m$, and denote by $R$ the number of rejected hypotheses. In the frequentist setting, the situation can be summarized by the following table (Benjamini and Hochberg, 1995). The $m$ hypotheses are assumed to be known in advance, while the numbers $m_0$ and $m_1$ of true and false null hypotheses are unknown parameters, $R$ is an observable random variable, and $S$, $T$, $U$, and $V$ are unobservable random variables. In the microarray context, a null hypotheses $H_j$ exists for each gene $j$, and rejection of $H_j$ corresponds to declaring that gene $j$ is differentially expressed. In general, we want to minimize the number $V$ of false positives, or type I errors, and the number $T$ of false negatives, or type II errors. The standard approach is to prespecify an acceptable type I error rate $\alpha$ and seek tests that minimize the type II error rate (i.e., maximize power), within the class of tests with type I error rate $\alpha$.

In terms of these random variables, we can define the main rates used in the present context. The per-comparison error rate (PCER) is defined as the expected value of the number of Type I errors over the number of hypotheses (i.e., \( PCER = E(V)/m \)). The family-wise error rate (FWER) is the probability of at least one Type I error, \( FWER = \text{pr}(V \geq 1) \). The false discovery rate (FDR) is the expected proportion of Type I errors
among rejected hypotheses, \( FDR = \mathbb{E}(V/R; R > 0) = \mathbb{E}(V/R|R > 0)\Pr(R > 0) \), in Benjamini and Hochberg (1995), while the positive false discovery rate (pfDR) in Storey (2002) is the rate that discoveries are false, \( pfDR = \mathbb{E}(V/R|R > 0) \).

It is important to note that the preceding expectations and probabilities are conditional on which hypotheses are true or false (i.e., on which genes are differentially expressed). We distinguish between strong and weak control of the Type I error rate. Strong control refers to control of the Type I error rate under any combination of true and false hypotheses (i.e., for any combination of differentially and constantly expressed genes). In contrast, weak control refers to control of the Type I error rate only when none of the genes are differentially expressed (i.e., under the complete null hypothesis \( H_0^C \) that all the null hypotheses are true). In general, weak control without any other safeguards is unsatisfactory. In the microarray setting, where it is very unlikely that none of the genes are differentially expressed, it seems particularly important to have strong control of the Type I error rate.

As we will see, a wide variety of multiple testing procedures are used. How should we choose which to use? No simple answers are available here, but a procedure might be judged according to a number of criteria. One criterion is interpretation: Does the procedure answer a question that is relevant to the investigation? Another is type of control: strong or weak? We have already suggested that strong control is highly desirable in the microarray context. An important criterion is validity: Are the assumptions under which the procedure applies clearly true, or perhaps plausibly true, or are they unclear, or most probably not true? A fourth is computability: Are the procedure’s calculations straightforward to carry out accurately? Or, is there perhaps numerical or simulation error, or discreteness, which might cast doubt on the exactness of the result?

Adjusted p-values

To account for multiple hypothesis testing, one may calculate adjusted \( p \)-values (Westfall and Young, 1993). Given a test procedure, the adjusted \( p \)-value corresponding to the test of a single hypothesis \( H_j \) can be defined as the level of the entire test procedure at which \( H_j \) would just be rejected, given the values of all test statistics involved. We can distinguish three ways of adjusting \( p \)-values: the single-step, step-down, and step-up procedures. In single-step procedures, equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the test statistics or unadjusted \( p \)-values. Improvement in power, while preserving Type I error rate control, may be achieved by stepwise procedures, in which rejection of a particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. Step-down procedures order the unadjusted \( p \)-values (or test statistics) starting with the most significant, while step-up procedures start with the least significant.

For strong control of the FWER at level \( \alpha \), the Bonferroni procedure rejects any hypothesis \( H_j \) with \( p \)-value less than or equal to \( \alpha/m \). The corresponding Bonferroni
single-step adjusted $p$-values are thus given by $\tilde{p}_j = \min(mp_j, 1)$. While single-step adjusted $p$-values are simple to calculate, they tend to be very conservative. Let the ordered unadjusted $p$-values be denoted by $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$. Then, the Holm step-down adjusted $p$-values are given by

$$\tilde{p}_{r_j} = \max_{k=1, \ldots, j} \left\{ \min\left((m-k+1)p_{r_k}, 1\right) \right\}$$

Holm’s procedure is less conservative than the standard Bonferroni procedure which multiplies the $p$-values by $m$ at each step. However, neither Holm’s method nor Bonferroni (nor other single-step methods) take into account the dependence between the test statistics, which can be quite strong for co-regulated genes. Westfall and Young (1993) propose adjusted $p$-values which take into account quite general dependence, and which are less conservative. Their step-down minP adjusted $p$-values are defined by

$$\tilde{p}_{r_j} = \max_{k=1, \ldots, j} \left\{ pr\left(\min_{l \in \{r_k, \ldots, r_m\}} P_l \leq p_{r_k} \mid H_0^C \right) \right\}$$

while their step-down maxT adjusted $p$-values are defined by

$$\tilde{p}_{s_j} = \max_{k=1, \ldots, j} \left\{ pr\left(\max_{l \in \{s_k, \ldots, s_m\}} |T_l| \geq |t_{s_k}| \mid H_0^C \right) \right\}$$

where $|t_{s_1}| \geq |t_{s_2}| \geq \ldots \geq |t_{s_m}|$ denote the ordered test statistics.

These adjusted $p$-values lead to procedures which guarantee weak control the FWER in all cases, and strong control under the additional assumption of subset pivotality, which applies in the example described next. See Dudoit et al. (2002a) for fuller details, including details of permutation-based calculation of the $p$-values and their step-down adjustments.

Turning now to a different kind of adjusted $p$-value, the following formula of Benjamini and Hochberg (1995) gives a step-up adjustment which leads to strong control of the FDR under the additional assumption of independence of the test statistics. Even though this is not a realistic assumption with microarray data, we nevertheless offer the formula

$$\tilde{p}_{r_i} = \min_{k=1, \ldots, m} \left\{ \min(m p_{r_k} / k, 1) \right\}$$

Recently, Benjamini and Yekutieli (2001a) showed that the preceding adjustment is applicable when a particular form of dependence they term positive regression dependency obtains between the test statistics, and gave a conservative modification applicable quite generally. The latter formula is equivalent, for large $m$, to replacing the factor $m$ multiplying the unadjusted $p$-values by $m \log m$.

The final multiple comparison method we discuss briefly is the $pFDR$ of Storey (2002). This novel approach does not seek to control the Type I error or the false-discovery rate, nor does it provide adjusted $p$-values. Instead, Storey (2002) takes the view that conservatively estimating the FDR or $pFDR$ for rejection regions defined beforehand or by the actual statistics is a more appropriate task. In a separate paper, Storey (2001) introduces the notion of $q$-value, which loosely speaking, is the
minimum \( pFDR \) that can occur when rejecting a statistic equal to the observed one for a nested set of rejection regions. Space does not permit us to present full definitions or explanations of this highly interesting theory, so we refer the reader to the article by Storey (2002, 2001). These articles are mainly concerned with the case where the test statistics are independent and identically distributed, while Storey and Ribshirani (2001) requires only identically distributed test statistics and an ergodicity condition.

\textit{Comparison of multiple testing procedures.}

We now apply the procedures to Case Study II, the SR–BI experiment, and display the results in Figure 2.12. All \( p \)-values depicted there are for the two-sample \( t \)-statistics introduced in the previous section for this example, and are calculated using all 12,870 assignments of 16 mice to two groups of 8, called “treatment” and “control.” For example, the unadjusted (two-sided) \( p \)-value \( p_i \) for gene \( i \) will be 2/12,870 when none of the assignments of the 16 mice to two groups of 8, other than the true one, leads to a \( t \)-statistic larger in absolute value than \( t_i \), the value observed with the true assignment.

Figure 2.12 presents a variety of adjusted \( p \)-values for the genes with the 100 smallest unadjusted \( p \)-values. As the solid line indicates, all of the latter are well below 0.01. The Holm (and hence the Bonferroni) adjustment (dashes) is already far too conservative, suggesting that none of the genes are differentially expressed, something we know
from follow-up studies to be false; but with over 6,000 genes and a smallest possible unadjusted \( p \)-value of roughly \( 1/6,000 \), this is not surprising. The \( \text{minP} \) adjustment (short dash, long dash) is likewise too conservative, and the explanation is the same. Even an observed \( t \)-statistic whose magnitude is not exceeded by that of any of the other (false) assignments of treatment and control status to the mice, leads to a \( \text{minP} \) adjusted \( p \)-value of 0.53, a value which is determined by the number of permutations and not by the magnitude of the test statistic. The adjustment of Benjamini and Yekutieli (2001b) (dot-dash) is similarly conservative. The one adjustment that does lead to fairly small values is the \( \text{maxT} \) adjustment (curved dash), and here we must point out that the permutation distributions of the 6000 \( t \)-statistics are not all identical, and consequently not all tests contribute equally to the \( \text{maxT} \) adjusted \( p \)-values, Westfall and Young (1993). In this example, the adjustment of Benjamini and Hochberg (1995) (which assume independence) turns out to be very close to the \( q \)-values of Storey (2001) (dots). While not giving values that are conventionally small (e.g., less than 0.05), these adjustments do seem less conservative than the others described previously.

In closing this brief discussion of multiple testing, we draw some tentative conclusions. The well-established and generally applicable \( p \)-value adjustments such as Bonferroni’s, Holm’s, and Westfall and Young’s \( \text{minP} \) do not seem to be helpful in the microarray context. Although 12,870 permutations might seem rather few in some contexts, 8 replicates of a treatment and 8 of a control are large numbers for a microarray experiment: 2, 3, and occasionally 5 are more common numbers of replicates. Also, it is not easy to argue that there should be more, simply to permit the use of this multiple testing methodology. In many, perhaps most cases, 3–5 replicates will be enough to reveal many differentially expressed genes. Thus, we are unlikely to see many situations in which permutation based (and hence distribution-free) adjusted \( p \)-values can be used. Westfall and Young’s \( \text{maxT} \) appears somewhat more useful. Procedures based on controlling or estimating the FDR (or \( \text{pFDR} \)) appear to be the most promising alternatives. When thousands of tests are being carried out, concern about making one or more false positive decisions does seem misplaced. Instead, permitting a small percentage of false positives among the rejected hypotheses seems more reasonable. However, the theory for FDR and \( \text{pFDR} \) needs to be developed more before it is generally applicable to microarray data. We need procedures controlling FDR under fairly general dependence assumptions between the test statistics that are less conservative than the one presented earlier, while the theory for \( \text{pFDR} \) needs to be applicable beyond independent or identically distributed test statistics to a framework that definitely covers microarray data.

Correlation and technical replicates

As explained in the section on “Replicate slides: design,” an important decision with two-color microarray experiments when choice is available is whether to use direct or indirect comparisons, that is, whether to measure expression differences within slides or between slides. Most statisticians would immediately answer that direct
comparisons are better, and they are, but the situation is not quite as simple as it might look upon first examination.

In the “Replication” section, we explained how biologists conducting simple gene expression comparisons, such as between treated and control cells, will typically carry out replicate experiments on different slides. These will frequently involve what we have called technical replicates, where the target mRNA in each hybridization is from the same RNA extraction, but is labeled independently for each hybridization. We have noticed that estimates of differential gene expression based on technical replicates tend to be positively correlated, whereas the same estimates based on replicates involving different RNA extractions and labelings tend to be uncorrelated. When the more extreme form of technical replication is used, that is, when a labeled sample is split and used in replicate hybridizations, the correlation can be very strong. For scatter plots illustrating these assertions, see Figure of Bowtell and Sambrook (2002). These observations suggest that we should reexamine the independence assumption underlying the experimental design calculations presented in the section on “Design choices.” When variances are calculated for linear combinations log-ratios across replicate slides, it appears desirable to use the most realistic covariance model for the measurements. The discussion that follows is based on Speed and Yang (2002) to which we refer for fuller details.

A more general covariance model

Let us reexamine comparisons between two target samples $T$ and $C$. Following the notation from the “Design choices” section, for a typical gene on the slide, we denote the means of the log-signals across slides by $\alpha = \mathbb{E} \log_2 T$ and $\beta = \mathbb{E} \log_2 C$, respectively. The variances and covariances of the log signals across slides will be assumed to be the same for all samples. Our dispersion parameters are a common variance $\tau^2$, a covariance $\gamma_1$ between measurements on samples from the same hybridization, a covariance $\gamma_2$ between measurements on technical replicate samples from different hybridizations, and a covariance $\gamma_3$ between measurements on samples which are neither technical replicates nor in the same hybridization. These parameters will in general be different for different genes, but we suppress this dependence in our notation. Later, we attempt to estimate typical values for them.

Consider again the two different designs illustrated in Figure 2.2, Design I illustrates an indirect comparison, where $T$ and $C$ are each hybridized with a common reference sample $R$. Design II involves two direct comparisons, where the samples $T$ and $C$ are hybridized together on the same slide. We denote technical replicates of $T$, $C$, and $R$ by $T'$, $C'$, and $R'$, respectively. Note that both designs involve two hybridizations, and we emphasize that in both cases, our aim is to estimate the expression difference $\phi = \alpha - \beta$ on the log scale. We now calculate the variances of the obvious estimates of this quantity from each experiment. For Design I, this is

$$v_1 = \text{var}(a - r - b + r') = 4(\tau^2 - \gamma_1) - 2(\gamma_2 - \gamma_3)$$

while for Design II, this is one-half of $y = a - b + a' - b'$, and we have
Yang (2002), we presented an analysis based on the data from the swirl experiment. The direct design is evidently never less precise than the indirect one, and the extent of its advantage depends on the values of $\gamma_1$, $\gamma_2$, and $\gamma_3$. To see this, consider the covariance matrix for the four log intensities from Design II:

\[
\text{cov} \begin{pmatrix} a \\ b \\ a' \\ b' \end{pmatrix} = \begin{bmatrix} \tau^2 & \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \tau^2 & \gamma_2 & \gamma_3 \\ \gamma_2 & \gamma_3 & \tau^2 & \gamma_1 \\ \gamma_3 & \gamma_2 & \gamma_1 & \tau^2 \end{bmatrix}
\]

It is easy to check that the eigenvalues of this matrix are $\lambda_1 = \tau^2 + \gamma_1 + \gamma_2 + \gamma_3$, $\lambda_2 = \tau^2 + \gamma_1 - \gamma_2 - \gamma_3$, $\lambda_3 = \tau^2 - \gamma_1 + \gamma_2 - \gamma_3$, and $\lambda_4 = \tau^2 - \gamma_1 - \gamma_2 + \gamma_3$, corresponding to the eigenvectors $(1, 1, 1, 1)'$, $(1, 1, -1, -1)'$, $(1, -1, 1, -1)'$ and $(1, -1, -1, 1)'$, respectively. In terms of these eigenvalues, we see that $v_1 = \lambda_3 + 3\lambda_4$ and that $v_2 = \lambda_3$. Thus, the relative efficiency of the indirect versus the direct design for estimating $\alpha - \beta$ is

\[
\frac{v_1}{v_2} = 1 + \frac{3\lambda_4}{\lambda_3}
\]

The direct design is evidently never less precise than the indirect one, and the extent of its advantage depends on the values of $\tau^2$, $\gamma_1$, $\gamma_2$, and $\gamma_3$. Notice that when $\lambda_4 = 0$ (equivalently, $\tau^2 = \gamma_1 = \gamma_2 = \gamma_3$), we see that $v_1 = v_2$. This shows that under our more general model, the reference design could, in theory, be as efficient as the direct design. This is very unlikely in practice, as these conditions are equivalent to the variance $\text{var}(a - b)$ of a log-ratio coinciding with the covariance $\text{cov}(a - b, a' - b')$ between two log-ratios derived from technical replicate samples. At the other extreme, when $\gamma_2 = \gamma_3$, that is when the covariance between measurements on technical replicates coincides with that between any two unrelated samples, we have $v_1 = 4v_2$. This is the conclusion which is obtained when log-ratios from different experiments are supposed independent.

The preceding discussion focused on a single gene. It is not an easy task to obtain estimates of these eigenvalues or of $v_1$ and $v_2$ for single genes; however, in Speed and Yang (2002), we presented an analysis based on the data from the swirl experiment, which sought to obtain estimates of the average eigenvalues, and corresponding estimates of averages for $v_1$ and $v_2$. We found there that the relative efficiency $v_2/v_1$ of the indirect to direct designs for estimating $\log_2(\text{swirl}/\text{wt})$ was 4 for dye-swap set 1, consistent with independence, and 2.5 for dye-swap 2, suggesting a measure of dependence.

When we average log-ratios, as we do in Design I, we want the terms to be as independent as possible to minimize the relevant variance. In this case, it would be best if we could avoid using technical replicates, and use truly independent samples. On the other hand, when we take differences, as we do in Design II, we want the technical replicate terms ($R$ and $R'$) to be as dependent as possible. This could be achieved by using the same extraction and the same labeling (extreme technical replication) for the common reference mRNA. Further, the results we have just described suggest that, in some cases, the covariance due to technical replication needs to be considered.

We will revisit this discussion with a three-level factor in the next section.
2.4 Single-factor experiments with more than two levels

A simple extension to two-sample comparisons is single-factor experiments where we wish to compare the effects of $K$ treatments, $T_0, \ldots, T_{K-1}$, on gene expression. Examples of such experiments include comparing multiple drug treatments to a particular type of cell (Bedalov et al., 2001) or comparing different spatial regions of the retina (Diaz et al., 2002b).

Case Study III: mouse olfactory bulb

To provide an illustration, we consider an experiment where comparisons were made between different spatial regions of mouse olfactory bulb to screen for possible region specific developmental cues (Lin et al., 2002). The target cDNA was hybridized to glass microarrays containing 18,000 cDNA probes obtained from the RIKEN consortium. The olfactory bulb is an oblong spherical structure, so in order to make a three-dimensional representation using binary comparisons, the bulb was dissected into three sections along the three orthogonal axes, leading to six samples termed anterior (front), posterior (back), medial (close to the axis of bilateral symmetry), lateral (away from the axis of symmetry), dorsal (top), and ventral (bottom). In what follows, these cell samples will be denoted by $A, P, M, L, D$, and $V$, respectively. Initially, comparisons were to be made between regions that were maximally separated ($A – P, M – L, D – V$), but later it was decided to make all possible comparisons. Figure 2.13 is a graphical representation of some selected arrays of this experiment.

Design

Let us begin our discussion on design with an example comparing three types of mRNA samples ($K = 3$) and let us suppose that all pairwise comparisons are of
equal interest. An example of this type of experiment would arise in investigating the differences in expression between different regions \((T_0, T_1, T_2)\) of the brain. The scientific aim of this experiment would be to identify genes that are differentially expressed between different regions of the brain; that is, to identify genes with differential expression between \(T_0\) and \(T_1\), or between \(T_1\) and \(T_2\), or between \(T_2\) and \(T_0\). Figure 2.14 depicts two designs for such a three-level, single-factor experiment, where \(R\) denotes a common reference source of mRNA.

For comparing the efficiency of different designs, we fit a linear model and examine the variance associated with the least squares estimates for the parameters of interest. For example, let us consider the design depicted in Figure 2.14b. For any particular gene, we denote the means of the log base 2 intensities across slides by \(\alpha_0 = E \log_2 T_0\), \(\alpha_1 = E \log_2 T_1\) and \(\alpha_2 = E \log_2 T_2\). As all measurements are paired comparisons, only the differences between the effects are estimable and the contrasts of interest are thus the pairwise differences. For estimation purposes, we can treat \(T_0\) as a “pseudo” common reference. It follows that our parameters of interest are \(\phi_1 = \alpha_1 - \alpha_0\) and \(\phi_2 = \alpha_2 - \alpha_0\). In another context, we might be more interested in \(\alpha_2 - \alpha_1\), and so wish to make that one of our parameters.

We fit the following linear model \(y = X\phi + \epsilon\) to the vector \(y\) of log-ratios
\[
y_1 = \log_2(T_1/T_0), \quad y_2 = \log_2(T_2/T_1) \quad \text{and} \quad y_3 = \log_2(T_0'/T_2)
\]
from different slides:
\[
X = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}; \quad \phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \text{cov}(y) = \sigma^2 I
\]

We have used primes (e.g., \(R'\)) to denote technical replicate material, as this would almost always be the case in experiments like these. Nevertheless, we will begin by assuming that the different log-ratios are independent. The least squares estimates of the parameters are \((X'X)^{-1}X'y\) and the corresponding variances of estimates are given by the diagonal elements of the matrix \(\sigma^2(X'X)^{-1}\). Table 2.1 provides comparisons for a few design choices, where for presentation, \(\sigma^2\) is set to 1. The value in each cell of the table is the average variance associated with the three pairwise comparisons of interest: \(\text{var}(\hat{\phi}_2) + \text{var}(\hat{\phi}_1) + \text{var}(\hat{\phi}_2 - \hat{\phi}_1)\). This is the same as the \(A\)-optimality criterion defined in Kerr and Churchill (2001). This criterion is suitable...
SINGLE-FACTOR EXPERIMENTS WITH MORE THAN TWO LEVELS

Table 2.1 Single factor experiment—variance of estimated effects for the three different designs.\textsuperscript{a}

<table>
<thead>
<tr>
<th>Design choices</th>
<th>Indirect design</th>
<th>Direct design</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(a)</td>
<td>$T_0 \rightarrow T_1 \rightarrow T_2$</td>
<td>$T_0 \rightarrow T_1 \rightarrow T_2$</td>
</tr>
<tr>
<td>Number of slides</td>
<td>$N = 3$</td>
<td>$N = 6$</td>
</tr>
<tr>
<td>Units of material</td>
<td>$T_0 = T_1 = T_2 = 1$</td>
<td>$T_0 = T_1 = T_2 = 2$</td>
</tr>
<tr>
<td>Average efficiency</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>II</td>
<td>$T_0 \rightarrow T_1$</td>
<td>$T_0 = T_1 = T_2 = 2$</td>
</tr>
<tr>
<td>N = 3</td>
<td></td>
<td>0.67</td>
</tr>
</tbody>
</table>

\textsuperscript{a}For presentation, $\sigma^2$ was set to 1.

here because all pairwise comparisons are of equal interest and the main scientific constraint is that they are estimable.

Different design choices will be made depending on different physical constraints. For example, if the experimenters have unlimited amounts of reference material, but only one sample of mRNA from each of $T_0$, $T_1$ and $T_2$, then Design I(a) is the only possible choice out of these three. However, if the experimenters have two samples of mRNA from each of the $T_0$, $T_1$, and $T_2$ regions, then both Designs I(b) and II are feasible, with one using twice as many slides as the other. Direct comparison, Design II, will lead to more precise comparisons between the different regions, and in addition, save on the number of slides required.

In general, extending the designs in Table 2.1 to include $K$ target samples, we considered the following two classes of designs:

**Indirect design.** Perform each $R \rightarrow T_k$ hybridization as well as the corresponding dye-swaps, $T_k \rightarrow R$, $k = 0, \ldots, K - 1$, for a total of $2K$ hybridizations.

**Direct design.** Perform each $T_k \rightarrow T_l$ hybridization, $k, l = 0, \ldots, K - 1$, $k \neq l$, for a total of $K(K - 1)$ hybridizations.

For direct design, the comparison of two samples $T_k$ and $T_l$ can be done within slides, while for indirect design, the contrast $\alpha_k - \alpha_l$ are not estimated directly, but instead across slides through the common reference $R$. It can be shown, using the linear model that estimates of contrasts $\alpha_k - \alpha_l$ have variance $\sigma^2$ for indirect design and variance $\sigma^2/K$ for direct design. Adjusting for the different number of slides, the relative efficiency ratio between indirect versus direct design (i.e., ratio of variances for estimates of $\alpha_k - \alpha_l$ using both designs is

\[
\frac{v_{(i)}}{v_{(ii)}} = \frac{2K}{K - 1}
\]  

(2.1)
Therefore, for an equivalent number of slides, the limiting efficiency ratio, as the number of treatments $K$ increases, is 2, clearly illustrating the point that direct comparisons within slides yield more precise estimates than indirect ones between slides. Whenever comparisons are made through a common reference $R$, which alone is not of primary interest, every hybridization involving $R$ is in some sense “wasted,” roughly doubling the required number of hybridizations for a given level of precision. As observed before, the relative efficiency is largest for $K = 2$ treatments: where using direct comparisons gives a variance of $1/4$ of that obtainable from an indirect design involving the equivalent number of slides.

Moreover, to compare the two design classes for an equivalent number of target samples, the relative efficiency of the indirect versus the direct design for estimating $\alpha_k - \alpha_l$ is

$$\frac{v_{(i)}}{v_{(ii)}} = \frac{K}{K-1} \tag{2.2}$$

As the number of treatments $K$ increases, the limiting efficiency ratio is 1, and the number of slides used in the indirect design will be twice as large as that used with the direct design. Comparing this result with Eq. 2.1, we see that the differences in efficiency between the two designs are due largely to the amount of material involved in the experiments.

Considering Table 2.1, it is evident that, when $K$ becomes larger, the situation becomes more complex. The analogues of Designs I(a) and I(b) are clear, they are the so-called reference designs. The analogue of Design II—which we call all-pairs design—is unlikely to be feasible or desirable for a large number of comparisons. For example, with 6 sources of mRNA, there are 15 pairwise comparisons requiring 5 units of each target mRNA; for 7 there are 21, requiring 6 units, and so on. Alternative classes of designs that involve far fewer slides include the loop designs of Kerr and Churchill (2001), but these designs can suffer from having long path lengths between some of the comparisons. In Figure 2.15, we offer two alternative designs for six sources of target mRNA and six hybridizations. For Design, I, the least squares estimates of contrasts $\alpha_k - \alpha_l$ have variance $2\sigma^2$ for any pairwise comparisons, and hence an average variance of $2\sigma^2$. In contrast, the average efficiency for all pairwise least squares estimates of contrasts in Design II is $1.8\sigma^2$, with some contrasts being relatively more precise (e.g., $\text{var}(\hat{\alpha}_1 - \hat{\alpha}_2) = \sigma^2$) and others that are relatively less precise (e.g., $\text{var}(\hat{\alpha}_1 - \hat{\alpha}_4) = 3\sigma^2$). It should be clear from this example that, instead of regarding the problem of choosing a design as a decision between classes of designs (reference, loop, all-pairs), a more productive approach is to ask which comparisons are of greatest interest and which are of lesser interest, and seek a design that gives higher precision to the former and lower precision to the latter. Such a design then will involve a mix of direct and indirect comparisons, tailored to the needs and constraints of the particular context, including the robustness discussed in the section on “Robustness.”
SINGLE-FACTOR EXPERIMENTS WITH MORE THAN TWO LEVELS 73

Figure 2.15 Two possible designs comparing the gene expression between six target samples $T_0, \ldots, T_5$ of cells. (a) Indirect comparison: this design compares the expression levels of samples $T_k$ and $T_l$ to the reference sample separately on two different slides, and estimates the log-ratio $\log_2(T_k/T_l)$ by the difference $\log_2(T_k/R) - \log_2(T_l/R')$, where samples $R$ and $R'$ are reference samples. (b) Direct comparison: loop design. This design compares the gene's expression between samples $T_k$ and $T_l$ directly on the same slide.

A more general covariance model

The preceding comparisons do not take into account the different types of replication. Now, let us consider how the efficiencies of our designs change when the covariances arising from the use of technical replicates are included in our analysis. We revisit the designs for a three-level, single-factor experiment depicted in Table 2.1 using Design II first. As before, we denote by $y = (y_1, y_2, y_3)$ the log-ratios observed from slides 1, 2, and 3. To estimate the parameters of interest $\beta = (\phi_1, \phi_2)$ where $\phi_1 = \alpha_1 - \alpha_0$ and $\phi_2 = \alpha_2 - \alpha_0$, we once again fit the linear model by least squares, but this time using generalized least squares. The covariance matrix of the observations is given by:

$$
cov \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & -\rho & -\rho \\ -\rho & 1 & -\rho \\ -\rho & -\rho & 1 \end{pmatrix}
$$

Here, $\sigma^2$ is the variance of a single log-ratio, which in the notation of the “Correlation and technical replicates” section is $2(\tau^2 - \gamma_1)$, while the covariance between two log-ratios involving a common technical replicate labeled with the same dye in the opposite position is $-\sigma^2 \rho = -(\gamma_2 - \gamma_3)$. Analogous linear models and covariance matrices are used for the indirect Designs I(a) and I(b) of Table 2.1. In these cases, we not only have terms in the covariance matrix corresponding to log-ratios sharing a single technical replicate, in Design I(b) we also have terms $2\sigma^2 \rho$ corresponding to the covariances between log-ratios sharing two technical replicates. We omit the straightforward details.

With these covariances we can calculate the average variances of the generalized least-squares estimates of the contrasts of interest under the more general model. Because
the designs are symmetric, these will be the same functions of \( \rho \) as the variance \( \text{var}(\hat{\phi}_1) \) of the generalized least squares estimator \( \hat{\phi}_1 \) of \( \phi_1 \). Table 2.1 has the results in the case where \( \rho = 0 \) (i.e. there is zero covariance between log-ratios derived from technical replicate samples). Under the more general covariance model, the variance of the indirect estimate of \( \hat{\phi}_1 \) from Design I(b) of Table 2.1 is just \( \sigma^2 \), independent of \( \rho \). By contrast, the estimator from Design II, which is a mix of direct and indirect information, increases from \( \frac{2}{3} \) to 1 as \( \rho \) increases from 0 to 0.5. In fact, in this case the generalized least squares estimator of \( \phi_1 \) coincides with its ordinary least squares estimator. We conclude that the variance of the indirect estimate (Design I(b)) is 1.5 times that of the mix of direct and indirect (Design II) at \( \rho = 0 \), but comes to equal it at the other extreme \( \rho = 0.5 \). These observations and the results for Design I(a) can be found in Figure 2.16.

### Linear model analyses

#### General remarks

The approach we adopt to combining gene expression data from replicate experiments to estimate gene expression differences between two types of samples was introduced informally in the “Replicate slides: design” and “single-factor experiments with more than two levels” sections. In its simplest form, the combination is carried out by averaging observations from individual experiments (“treatment” and “control”) and taking differences, to produce a combined estimate of the gene expression effect that is a result of the treatment. This straightforward idea may be extended to experiments such as Design II in Figure 2.14b by employing fixed effects linear models, where we
estimate certain quantities of interest (e.g., the $T_1 - T_0$ difference) for each gene on our slide. Further, the idea applies much more generally, and will serve our purpose here, to provide least squares estimates of all estimable contrasts such as the anterior-lateral difference in the bulb experiment of Figure 2.13. As with all our earlier work, we focus on parameters that are, in effect, expectations of log-ratios (e.g., of $\log(A/L)$).

Some small but important differences occur between our use of linear models here and those published in the microarray literature. Others, most notably Kerr et al. (2000), use linear models and the analysis of variance (ANOVA) to estimate expression differences and assess the variability of their estimates. As explained in the section on “Replicate slides: direct comparisons”, these writers typically assume a fixed effects linear model for the logged intensities, with terms for dye, slide, treatment, and gene main effects, as well as selected interactions between these factors, and error terms having common variance across genes. Differentially expressed genes are those that exhibit significant treatment $\times$ gene interaction, while normalization is effected by the inclusion of dye terms in the linear model. By contrast, our linear models are for log-ratios from experiments in which normalization has been carried out separately for each slide, typically using a decidedly nonlinear adjustment which could not be captured in a linear model. Indeed, artifacts removed by our nonlinear normalization frequently correspond to interactions assumed absent in the linear approach. We do not include different genes in the same linear model, and so we do not assume a common error variance, and the only terms we do include in the model relate to the mRNA samples and their treatments. We find that our approach deals more satisfactorily with dye differences that are nonlinearily intensity-dependent and spatially dependent in varying ways across a set of slides. Also, our designs are usually not orthogonal, and so would not lead to a unique analysis of variance.

In more recent research, Jin et al. (2001) and Wolfinger et al. (2001) use linear models on a gene by gene basis, with a separate error for each gene, but still including normalization as part of their linear modelling. Their linear models for normalized data also include random effects for arrays. In a sense, these studies are quite ambitious, as the authors took the bold step of treating the signals from the two channels of their cDNA experiment as two separate sources of data, not passing to ratios or log-ratios, but keeping both for their analysis. They have some effects estimated within hybridizations (age), which are therefore based on ratios within hybridizations, and others (sex and strain) that are not, but result from comparisons across hybridizations not expressible as functions of within-hybridization ratios. In our view, this approach can only be adopted following a very thorough multi-slide normalization of all of the single channels, because many systematic nonadditive spatial and intensity-dependent hybridization biases disappear when ratios are taken, but can remain otherwise. When this is done, it would appear that the appropriate analysis should be a two-stratum one, distinguishing the within-hybridization and the between-hybridization strata, for in our experience, the variances in these two strata are likely to be quite different. In brief, the appropriate analysis in this two-stratum context is likely to be significantly more complex than the present one, and for this reason we do not pursue it further here.
To define the notation used in our linear model, for a typical gene \(i\), let us denote the gene’s observed intensity value corresponding to the six different regions of the bulb by \(A_i, P_i, D_i, V_i, M_i,\) and \(L_i\). Further, define the average values across hybridizations of the log transformation of these quantities to be \(a_i = \mathbb{E} \log_2 A_i, p_i = \mathbb{E} \log_2 P_i, d_i = \mathbb{E} \log_2 D_i, v_i = \mathbb{E} \log_2 V_i, m_i = \mathbb{E} \log_2 M_i,\) and \(l_i = \mathbb{E} \log_2 L_i\). To estimate the spatial gene expression for gene \(i\), we fit the following linear model:

\[
y_i = X\beta_i + \epsilon_i
\]

where \(y_i\) is the vector of log-ratios from all the slides; \(X\) is the design matrix; \(\epsilon_i\) is a vector of the disturbance term; and \(\beta_i\) is a vector of parameters. The five estimable parameters we choose are given by \(a_i = a_i - l_i, p_i = p_i - l_i, d_i = d_i - l_i, v_i = v_i - l_i\) and \(m_i = m_i - l_i\). In addition, we assume that the error terms associated with the different slides are independent and identically distributed with \(\mathbb{E}(\epsilon_i) = 0\) and \(\text{cov}(\epsilon_i) = \sigma^2 I\), where \(I\) is the identity matrix. Note that in this experiment, some but relatively little use was made of technical replicates. For simplicity, we ignore possible correlations in our discussion. The design matrix and parameters corresponding to the data of Case Study III is given next.

\[
\begin{pmatrix}
y_{1i} \\
y_{2i} \\
y_{3i} \\
y_{4i} \\
y_{5i} \\
y_{6i} \\
y_{7i} \\
y_{8i} \\
y_{9i} \\
y_{10i} \\
y_{11i} \\
y_{12i} \\
y_{13i} \\
y_{14i} \\
y_{15i}
\end{pmatrix} = 
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 \\
-1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
-1 & 1 & 0 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 \\
-1 & 0 & 1 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 \\
0 & 0 & 1 & 0 & -1
\end{pmatrix} \begin{pmatrix}
a_i \\
p_i \\
d_i \\
v_i \\
m_i
\end{pmatrix} + \epsilon_i
\]

To estimate the parameter \(\beta_i\), we can use \(\hat{\beta}_i = (X'X)^{-1}X'y_i\). In practice, we used a robust linear model, so that our estimates are less affected by outliers. The data are fitted to the linear model described previously by iteratively re-weighted least squares (IWLS) procedure using the function \texttt{rlm} provided in the library MASS in the statistical software package R. Details of the theory and implementation of the “robust linear model” can be found in Venables and Ripley (1999).

Although we did not carry out the bulb study with a common reference mRNA source, this was initially contemplated, and we would have used “whole bulb” mRNA, \(W\). Had we done so, we would have had seven mRNA samples and could then have
SINGLE-FACTOR EXPERIMENTS WITH MORE THAN TWO LEVELS

estimated six parameters. The natural parameterization would then have been $a_i^{(w)} = a_i - w_i$, $p_i^{(w)} = p_i - w_i$, $d_i^{(w)} = d_i - w_i$, $v_i^{(w)} = v_i - w_i$, $m_i^{(w)} = m_i - w_i$, and $l_i^{(w)} = l_i - w_i$, where $w_i = E \log_2 W_i$, and the analysis would have begun with robustly fitting a suitable linear model to estimate these parameters.

Returning to the parametrization we did use, what came next? It is clear from the design matrix that all 15 pairwise comparisons of the form $a_i - l_i$ are estimable, and so our first summary description of a gene’s spatial pattern was the profile consisting of the set of 15 parameter estimates of this form. Not surprisingly, profiles of this kind were not as easy to interpret as we would have liked, having a high degree of redundancy. Accordingly, we switched to the more economical profile consisting of the estimates of the six parameters $a_i$, $p_i$, $d_i$, $v_i$, $m_i$, and $l_i$, subject to the zero sum constraint $a_i + p_i + d_i + v_i + m_i + l_i = 0$, that is, to estimates of the six parameters $a_i - \frac{1}{6}(a_i + p_i + d_i + v_i + m_i + l_i)$, . . . . These profiles were much easier to visualize, and so we worked with them in the subsequent analysis. We refer to Lin et al. (2002) for the rest of the story, noting here that our main focus following the preceding linear model analysis was the clustering of the spatial profiles. We restricted ourselves to profiles which seemed to be real, that is, not noise, and we used Mahalanobis distance between profiles based on (robust variant of) the variance-covariance matrix $\sigma^2(X'X)^{-1}$ of the parameter estimates.

Remarks on assessing parameter estimates

Most of the issues we discussed in Section 2.2 (“Two-sample comparisons”) have analogues in this more general setting involving linear models for expectations of log-ratios. This includes the use of estimated standard deviations for estimates of parameters in the linear models, $t$-statistics (which are standardized parameter estimates here), modified $t$-statistics, robust estimates, empirical Bayes statistics, and the use of Q–Q plots and more formal tests of significance. Once we leave replicated treatment versus control or multiple-treatment experiments, few opportunities exist for permutation-based, distribution-free methods, but the bootstrap appears to be a promising alternative. The techniques just mentioned are still relatively unexplored in this context, no doubt because the use of linear models with microarray data has so far been limited. As the use of linear models increases, we can expect all the analogues of techniques effective in treatment versus control experiments to find their place in this more general setting.

Time-course experiments

This class of microarray experiments is among the most widely analyzed and reanalyzed in the literature, with a relatively large number of papers discussing one or more of the cell-cycle data sets (Cho et al., 1998; Chu et al., 1998; Spellman et al., 1998). We will not attempt to review all this literature, but instead comment briefly on the general approaches and give references. A major drawback of this methods-based
summary is the absence of any careful discussion of the biological motivation of the methods, present in some but by no means all of the articles we mention.

Microarray time-course data differs from traditional time series data in that they tend to be either short and irregularly spaced (e.g., 0, 1, 4, 12, and 24 hours after some intervention), or longer and equally spaced, but still with relatively few time points, typically only 10–20. The cell cycle data involved about 16 time points and included just two full cycles, which is quite short by normal time series analysis standards. A further common feature of such data (e.g., developmental time-courses) is that there are few if any \textit{a priori} expectations concerning the patterns of responses of genes, apart from vague notions such as “may come on some time, stay on for a while, and then go off,” or will be on “early” or “late” in the time period observed. Depending on the time increments and the context, smoothness of a gene’s temporal response may or may not be a reasonable expectation. All this suggests that the analysis of microarray time-course data will offer many new challenges to statisticians, and that it will frequently be more exploratory than model-based.

Bearing these remarks in mind, it is not surprising that initial approaches to analyzing time-course data relied on simple mathematical modeling and cluster analysis (see Chu et al., 1998; Cho et al., 1998; Spellman et al., 1998; and Chapter 4 of this book). Indeed, clustering is still the most widely used technique with time-course data, although regression and other model-based statistical methods are becoming more common.

\textit{Design}

A short time-course experiment may be viewed as a single factor experiment with \textit{time} being the factor. The additional information in this context is that there is a natural ordering between the different target samples, in contrast with the experiment described in Section 2.4 (“Single-factor experiments with more than two levels”), where no ordering existed between the different samples. The ordering of the “levels” of the factor time will single out certain comparisons (e.g., each time with baseline, or between consecutive time points) and linear contrasts (e.g., concerning linearity, concavity, or monotonicity) of interest to the researcher, and so the design choice will definitely depend on the comparisons of particular interest. Further, the best design can depend critically on the number of time points.

In a small study (e.g., with four time points), finding the design that is optimal in some suitable sense can be done simply by enumerating all possibilities (see table 2 in Yang and Speed (2002) for some simple examples). It is not feasible to enumerate all possible designs for problems with a much larger number of times points, such as the cell-cycle experiment. Developing algorithms for finding near-optimal designs is a research topic of interest here.
Clustering time-course data

Noteworthy early research on large-scale gene expression was conducted by Wen et al. (1998), an RT-PCR study of 112 genes, each measured on cervical spinal cord tissue in triplicate at nine different time points during the development of the rat central nervous system. These authors clustered the temporal responses of the genes, and found five main groups. Their approach is typical of that many other authors have taken since then with microarray data. Tamayo et al. (1999) introduced self-organizing maps into the microarray literature in the context of time-course data, illustrating them on the yeast cell-cycle data and another dataset concerning hematopoietic differentiation in four activated cell lines, each sampled at 4–5 times. Other research adopting a similar approach include Saban et al. (2001) and Burton et al. (2002), both of which have just 5 time points, with 4 and 2 replicates, respectively, at each time. Langmead et al. (2002) provides a critique of the Fourier approach to short time-course data, and offers a novel analysis incorporating autocorrelation and a new metric for clustering. Their methods were applied to simulated data and the cell-cycle data sets.

Principal component, singular value decompositions, and related methods

It is worth separating this category of methods from the clustering, although in most cases the aims are very similar. Chu et al. (1998) and Raychaudhuri et al. (2000) are early examples of the application of principal component analysis (PCA) to microarray time series data. Alter et al. (2000) address similar issues using singular value decompositions (SVD), but go much further in creatively displaying the results. Ghosh (2002) uses resampling methods to estimate the variability in SVD, and applies his methods to the Cho et al. (1998) data. A dynamic model of gene expression is presented in Holter et al. (2001), who estimate the time-independent translation matrix in a difference equation model for the “characteristic modes” (i.e., the eigenvectors of the SVD of published time-course data sets), (Spellman et al., 1998; Chu et al., 1998; Iyer et al., 1999), using 12, 7, and 13 time points, respectively.

Regression and related model-based approaches

In a series of articles, (Zhao et al., 2001; Thomas et al., 2001; Xu et al., 2002), L. P. Zhao and colleagues promote the use of appropriately defined regression models to analyze microarray data. The first and third of these articles discuss time-course data (Zhao et al., 2001) and revisit the cell-cycle data. Perhaps the most interesting of these papers is Xu et al. (2002), which analyzes parallel eight-time-point series from both transgenic and control mice. This article presents an approach to the joint analysis of related series which can be compared with the one discussed briefly in “Design and analysis of factorial experiments.”

Two articles that make very explicit use of the temporal pattern expected of (or sought from) genes are Kato et al. (2000) and Sasik et al. (2002). Both invoke differential equation models for mRNA concentrations, and solve them under special assumptions.
to arrive at regression equations, which are then fitted to the data. Kato et al. (2000) is the more ambitious of the two, attempting to infer genetic networks from the data using multiple regression. They apply their method to an early yeast data set from Derisi et al. (1997) involving seven different times. By contrast, Sasik et al. (2002) model genes in isolation. Their data were sampled every 2 hours for 24 hours during the developmental program of the slime mold *Dictyostelium discoideum*. These authors begin their analysis by fitting to the data on each gene a simple nonlinear kinetic model based on a first-order differential equation. The equation embodies a threshold model for transcription, and focusses on genes which come on “sharply.” They then assess the fit of this model to each gene, and go on to examine the temporal patterns of the genes which fit the model satisfactorily.

Another noteworthy article essentially adopting a regression approach is Shedden and Cooper (2002), which presents a critical analysis of the Cho et al. (1998) cell-cycle data. In fact, their reanalysis suggests that “all apparent cyclicities in the expression measurements may arise form chance fluctuations,” and that “there is an uncontrol- lable source of experimental variation that is stronger than the innate variation of gene expression in cells over time.” In the light of the large number of reanalyses of the cell-cycle data, this article is well worth reading.

The use of Bayesian networks is most appropriately noted under this heading (Friedman et al., 2000). This is an ambitious attempt to infer gene interactions from the cell-cycle data of Spellman et al. (1998), one in which the role of time in the analysis is by no means straightforward.

Aach and Church (2001) show how to align two time series when the time scales might have become “warped” due to different time sampling procedures or the different rates at which common biological processes evolve in related experiments. This notion is also present in Bar-Joseph et al. (2002), who present a rather general model-based approach to clustering, supposing that each gene’s time-series can be viewed as having gene-specific and class-specific parameters multiplying spline basis functions, with Gaussian noise added. The estimation is by maximum likelihood and a byproduct of this is an assignment of genes to classes. Both of these articles illustrate their methods on the yeast cell-cycle data of Spellman et al. (1998), while Bar-Joseph et al. (2002) also analyze a later yeast data set (Zhu et al., 2000).

Articles by Butte et al. (2002) and Filkov et al. (2002) focus on analytical issues underlying the identification of pairs of genes that are co-regulated in time-course experiments, and both illustrate their approach on the cell cycle data sets Cho et al. (1998) and Spellman et al. (1998). Our final reference in this category is to Klevezl (2000) and Klevecz and Dowse (2000) who also study the cell cycle data, this time using wavelets. Their interesting analyses suggest some quite novel conclusions, including the possible regulation of the yeast cell cycle by “an attractor whose fundamental period is an emergent property of dynamic interactions within the yeast transcriptome.”
Contrasts

A simple and potentially powerful approach to extracting information from time-course data makes use of linear combinations (or functionals) of the expression values across times. Following standard statistical usage, we will call these contrasts, which usually but do not always require that the coefficients sum to zero. The coefficients in contrasts can be tailored to the aims of the analysis, and they can be fixed a priori, or determined by the data. An example of a data-driven contrast would be one defined by the expression profile of a prespecified gene, something we might use in seeking other genes with temporal patterns of expression match those of that gene. A fixed set of coefficients might be one that estimates a linear trend. Another example is the discrete Fourier transform, which can be viewed as a fixed set of constraints (corresponding to different frequencies) evaluated on an input series. Contrasts thus generalize many familiar calculations, and as we shall see shortly, they also generalize single variable regression modeling. When we have more than one contrast, it is not uncommon to orthogonalize them sequentially, in the hope that, with approximately normally distributed input data, the resulting estimated contrasts are approximately independent. This is not necessary, however.

When \( c_1, c_2, \ldots \) is a set of contrast coefficients corresponding to times 1, 2, \ldots, and \( E_1, E_2, \ldots \) are the corresponding expression time series for a given gene, the sum \( <c, E> = \sum_t c_t E_t \) is the value of the contrast for that gene. If we normalized this quantity by \( |c|^2 = \sum_t c^2_t \), supposing for simplicity that \( \sum_t c_t = 0 \), then \( <c, E> / |c|^2 \) would be the regression coefficient corresponding to the fit of expression values to the temporal pattern \( (c_t) \). As this normalizing quantity is the same for all genes, there is no real need to include it in the fitting and assessing procedure. We note here that the expression values \( E_t \) may be “absolute,” as with Affymetrix chip data, or “relative” as with two-color cDNA or long oligonucleotide data, and these will usually be on the log scale.

Coefficients have many natural candidates \( (c_t) \), and we simply illustrate with a few that we have found helpful. In a time-course experiment (Lönnstedt et al., 2002), where mRNA from a cell line was sampled at 0.5, 1, 4, and 24 hours following stimulation with a growth factor, we sought “early” and “late” responding genes. An ad hoc but apparently useful working definition of these genes was as follows: those genes with large values of \( <c, E> \) with \( c_t = (t - 24.5)^2 \) were termed early, and those with \( c_t = t^2 \) (i.e., with \( c = (0.25, 1, 16, 576) \)) were termed late, respectively. It was an easy matter to obtain candidates for these genes: we simply did a Q–Q plot of the values \( <c, E> \) and somewhat arbitrarily selected cutoffs determining genes with unusually high or low values. More formal tests of significance are clearly possible, just as they were with similar contrasts defining gene expression differences.

The theme of contrasts in the analysis of time-course data is prominent in Fleury et al. (2002b), Fleury et al. (2002a), and Hero and Fleury (2002). In these articles, many examples of the use of contrasts are given, and the procedure for selecting genes from the various Pareto fronts is illustrated (refer to the section on “Illustrations using our case studies”). In essence, Pareto fronts and the variants presented in these articles
all seek to identify genes that have large values for all of a set of contrasts of interest. This ends our brief survey of models, methods, and the literature on the analysis of time-course experiments.

2.5 Factorial experiments

The previous examples have all been single-factor or one-way designs, where the factor has two, three, or more levels. A more complex class of designs arises when two or more factors are considered jointly, each factor having two or more levels. These factorial experiments are used to study both the expression differences caused by single factors alone, as well as those resulting from the joint effect of two or more factors, especially when it differs from what might be predicted in the basis of the factors separately, the phenomenon known as interaction. Factorial experiments were introduced by R.A. Fisher in 1926, and studied extensively by his collaborator Yates (1937). More recent references are Cox (1958) and Box et al. (1978).

Case Study IV: The weaver mouse mutant

This is a case study examining the development of certain neurons in wild-type (wt) and weaver mutant (wv) mice, Diaz et al. (2002a). The weaver mutation affects cerebellar granule neurons, the most numerous cell-type in the central nervous system. In the near-absence of these cells, the mice have a weaving gait. The wv mutant mouse strains were purchased from the Jackson Laboratory, and their genotypes were determined by a restriction site-generating PCR protocol. In the mutant mice, granule cells proliferate in the external granule cell layer, but terminally differentiated cells die before they migrate to the internal granule cell layer. As a result, the weaver mutants have greatly reduced numbers of cells in the internal granule cell layer, in comparison with the wt mice of the same strain. Consequently, the expression of genes which are specific to mature granule cells or expressed in response to granule cell-derived signals is greatly reduced.

Figure 2.17 is a graphical representation of four selected slides from Diaz et al. (2002a) in a form convenient to illustrate the parametrization of factorial experiments that we will use.

Design and analysis of factorial experiments

In this $2 \times 2$ factorial experiment, gene expressions levels are compared between the two strains wt and wv at the two postnatal times $P_{11}$ and $P_{21}$ (days). The four possible mRNA samples are $wtP_{11}$, wild-type at postnatal day 11, $wvP_{11}$, weaver at postnatal day 11, $wtP_{21}$, wild-type at postnatal day 21, and $wvP_{21}$, weaver at postnatal day 21.
Figure 2.17 Case Study IV: A portion of the weaver experiment provided by Elva Diaz from the Ngai Lab at the University of California, Berkeley. One parametrization for this $2 \times 2$ factorial experiment is indicated, see text for more explanation.

Figure 2.18 shows one parametrization for this experiment. Here, $\mu$ is the expectation of the log intensity $\log wtP_{11}$, $\mu + t$ is the expectation of $\log wtP_{21}$, etc. for a generic gene. Note that our parametrization is different from the symmetric ANOVA parametrization used in most statistics text books. We do it this way because we find the terms are more readily interpretable to biologists. Thus, the main mutant effect $v$ is the expectation $E \log_2 (wvP_{11}/wtP_{11})$ of the gene expression difference (on the log scale) between $wv$ and $wt$ at $P_{11}$. Similarly, the main time effect $t = E \log_2 (wtP_{21}/wtP_{11})$ is the expected gene expression difference (on the log scale) between days 21 and 11 for $wt$ mouse. The mutant by time interaction is defined by

$$v.t = E \log_2 \frac{wtP_{21}/wtP_{11}}{wvP_{21}/wvP_{11}}$$

Genes with a nonzero interaction term can be interpreted as genes for which the gene expression difference between $wtP_{11}$ and $wtP_{21}$ is different, on average, from that between $wvP_{11}$ and $wvP_{21}$; equivalently, genes whose expression difference

Figure 2.18 Experimental Design II of Table 2.2, with individual hybridizations numbered.
between *wtP*11 and *wwP*11 are different, on average, from those between *wtP*21 and *wwP*21.

How do we design $2 \times 2$ factorial experiments in this context? Four possible designs for the weaver mutant case study, each involving six hybridizations, are represented in Table 2.2, where here we follow Glonek and Solomon (2002). The designs head the columns, and the table entries are the corresponding variances of the ordinary least squares estimates of the main effect parameters $v$ and $t$ and the interaction parameter $v.t$. Suppose that our main goal is to identify genes with large interaction with six hybridizations. Designs II and IV give the smallest variance for the interaction term, but the main effect for $t$ is not even estimable in Design IV. Design I, which many biologists would use instinctively (perhaps without the dye-swaps), is by far the worst for precision in estimating the interaction, but it uses less mRNA: two units from each source, compared to three in all other designs. The preferred design will depend on the level of interest in the main effects in relation to the interaction, assuming all physical constraints are satisfied. In general, Design II (or its complement, with dye-swaps horizontally instead of vertically) will probably be the design of choice, offering good precision for all comparisons, though more for one main effect than the other.

The preceding design options help us to make an important general design point, namely, that in addition to experimental constraints, design decisions will be driven by an awareness of which effects are of greater interest to the investigator, and which are of lesser interest. More fully, the effects for which the greatest precision is required should be estimated *within slides* to the greatest extent possible, while effects of lesser interest can be measured less precisely, *between slides*. In extreme cases, where there is no interest at all in quantifying an effect, it need not even be estimable within slides (refer to Design IV in Table 2.2). Similar points were made by Kerr and Churchill (2001), who recommended greater use of loop designs.

We now revisit the design problem just discussed when correlations between technical

---

**Table 2.2** $2 \times 2$ factorial experiment—variance of estimated effects for the four different designs.\(^a\)

<table>
<thead>
<tr>
<th>Design choices</th>
<th>Indirect design</th>
<th>A balance of direct and indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>$w^*\ P_{11}$</td>
<td>$w^*\ P_{21}$</td>
<td>$w^*\ P_{11}$</td>
</tr>
<tr>
<td>$w^*\ P_{11}$</td>
<td>$w^*\ P_{21}$</td>
<td>$w^*\ P_{11}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effect $t$</td>
<td>0.5</td>
<td>0.67</td>
<td>0.5</td>
<td>—</td>
</tr>
<tr>
<td>Main effect $v$</td>
<td>0.5</td>
<td>0.43</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Interaction $v.t$</td>
<td>1.5</td>
<td>0.67</td>
<td>1</td>
<td>0.67</td>
</tr>
</tbody>
</table>

\(^a\)For presentation, $\sigma^2$ was set to 1.
replicate data are included. Figure 2.18 provides a representation of this $2 \times 2$ factorial experiment (Design II in Table 2.2) and the number next to the arrows in the diagram is the slide number. Here, the parameters of interest are the main effect $v$, main effect $t$, and the interaction effect $v \cdot t$.

For the observed log-ratios $y = (y_1, \ldots, y_6)$, we fit the following linear model:

$$
\mathbb{E} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & -1 & -1 \\ -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ t \\ v \cdot t \end{pmatrix}
$$

and

$$
\Sigma = \text{cov} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & \rho & -2\rho & 0 \\ \rho & 1 & -\rho & 0 & -\rho & \rho \\ 0 & -\rho & 1 & -\rho & 0 & -2\rho \\ \rho & 0 & -\rho & 1 & -\rho & \rho \\ -2\rho & -\rho & 0 & -\rho & 1 & 0 \\ 0 & \rho & -2\rho & \rho & 0 & 1 \end{pmatrix}
$$

Here, as in the section on “Design”, we denote the covariance between two log-ratios that share a technical replicate term in the same position (numerator or denominator) by $\sigma^2 \rho$, with the signs being reversed if the shared term is in the opposite position, while the covariance is $2\sigma^2 \rho$ or $-2\sigma^2 \rho$ between log-ratios with two technical replicate terms in common. The diagonal elements of the matrix $(X'\Sigma^{-1}X)^{-1}$ provide the estimates for $\text{var}(v)$, $\text{var}(t)$ and $\text{var}(v \cdot t)$, and similarly we can estimate these variances for all four designs shown in Table 2.2.

Figure 2.19 Comparing variances for a series of $\rho$ values for parameter estimates from four designs for a $2 \times 2$ factorial experiment (see Table 2.2). Plotted are scaled variances for (a) main mutant effect $v$; (b) main time effect $t$; and (c) interaction effect $v \cdot t$, for $\rho$ ranging from 0 to 0.5. The vertical dotted line in all three panels denotes the estimated $\rho$ from the second swirl experiment described in the “Case Studies I and II” section.
Figure 2.19 compares the variances of the generalized least squares parameter estimates for different values of $\rho$. As we have already remarked, the main effect $t$ is not estimable for Design IV. In general, the variances increase as the correlation among technical replicates increases, because we do not gain as much independent information while averaging, although this is not the case with the estimate of the interaction parameter in Design I. Perhaps the most striking conclusion from this figure is that for high enough correlation between technical replicate Design I becomes competitive for estimating the interaction parameter. Note also that the designs which lead to estimates of parameters with equal variance when $\rho = 0$ (refer Table 2.2), specifically Designs I and III for the main effects $t$ and $v$, Designs II and IV for the interaction parameter, also lead to estimates with equal variances for these parameters for all $\rho$.

It is worth pointing out that by showing how to take covariances between target samples into account, our aim here is simply to help make more informed design decisions. We do not advocate the use of technical replicates in place of biological replicates. As explained in the section on “Replication”, the type of replication to be used in a given experiment depends on the precision and on the generalizability of the experimental results sought by the experimenter. In general, an experimenter will want to use biological replicates to validate generalizations of conclusions. Often technical replicates are the result of physical constraints on the amount of mRNA samples an investigator can obtain. For example, in the Case Study III, the investigators were comparing small regions of the brain from newborn mice. In these experiments, the material is rare and an amplification technique is used to generate more material. This technique will inevitably introduce correlation between subsamples of the amplified material. Thus, we see that at times the consideration of covariances between technical replicates is a form of physical constraint that we need to recognize in making design decisions. In summary, when making designs choices, investigators should consider the pattern and approximate level of correlation between technical replicates, in addition to identifying which effects are of greater and which are of lesser scientific interest.

**Analysis of factorial experiments**

We have already discussed the estimation of parameters in linear models for log-ratios (see “Linear model analyses”); it all applies to our main effects and interaction parameters here. Figure 2.20 shows a plot of the estimates of the interaction parameter $v \cdot t$ (vertical axis) against the average values $A$ (horizontal axis) for the genes involved in Case Study IV. We have retained the coloring convention of Figure 2.6, and we observe that there appears to be a number of genes with quite large interactions by all three of the measures (estimated effect size, standardized estimate of effect size, and the Bayesian compromise between these two). We could conduct other analyses, including Q–Q plots, but we are content with just this one illustrative plot.

Most of what we have discussed for $2 \times 2$ factorial experiments extends straightforwardly to $2 \times 3$, $2 \times 4$, $3 \times 3$, and more general two-way factorial experiments, as well as to $2 \times 2 \times 2$ and other higher-way factorial experiments. It should all be
Figure 2.20 (See color insert following page xxx.) Estimated effect v.t vs. A plot for the interaction v.t between time and mutant.

quite clear how to proceed, but we need a special comment. We continue to regard nonstandard parametrizations analogous to the one used previously as preferable to the usual ANOVA-type ones from the viewpoint of the biological interpretability of the parameters.

Comparison of time series as factorial experiments

One area in which we have made considerable use of factorial experiments is worth discussing further: comparing tissue from two (or more) different mouse strains (e.g., wt and mutant) over several times (i.e., in extensions of studies such as IV above to more than two times). In the microarray literature, as in many other places, it is common to regard time as a covariate instead of a factor, as we are suggesting here, and use a more traditional regression approach (see the section “Time-course experiments”. We have found that for a small number of times (e.g., 2-5), where there is no special spacing or a priori expectation of particular patterns, particulary in developmental studies, it is advantageous to treat time as a factor. When doing so, we find that particular interest attaches to interactions of other factors with time, frequently focusing on just the points of biological interest to the experimenter. The interaction in Case Study IV is one such example.
2.6 Some topics for further research

A large number of problems occur in the design and analysis of microarray data, therefore, further research is required. Some problems are implicit in the discussion we presented earlier (e.g., on design, the analysis of comparative and time-course experiments, where the incompleteness of our knowledge is abundantly clear). In this short section, we touch on a few problems that were not discussed elsewhere in this book.

Involving $\hat{A}$ in ranking genes for differential expression

In the section on “Illustrations using our case studies,” we discussed why we think it desirable to take notice of a gene’s average intensity value $\hat{A}$ in determining whether it is differentially expressed. At present, none of the measures $\hat{M}, t,$ or $B$ that we have described, or the many variants on them, do this in any direct way. To the extent that variability is frequently lower at the higher end of the $\hat{A}$-range, standardizing by SD goes part of the way toward this end, but it does not fully incorporate the information we see in the plots of $\hat{M}$ (or its analogues such as estimated linear model parameters) against $\hat{A}$.

A first step that we have used with some satisfaction is the following simple approach. For example, we block the $\hat{A}$-range into 20 intervals; within each, we calculate upper and lower limits containing 95% of the values, and then smoothly join the values across the full range. Ideally, we would like this for a number of percentages, such as 90%, 95%, 99%, 99.9%, etc., and then one could determine (to some extent) the total number of points outside the range. What appears to be lacking is a formal approach to this question, one that recognizes and deals with the failure of the log transformation to be homoscedastic over the whole range of $\hat{A}$, as well as the changing density of points over this range.

Our last comment raises the question of transformations other than log-ratios of the intensities measuring gene expression. Not surprisingly, a number of statisticians have considered this topic, including Rocke and Durbin (2001), Durbin et al. (2002) and Huber et al. (2002). In our view, it is highly desirable to work with log-ratios if at all possible, as this is a simple and intuitive transformation for biologists, especially when compared with differences of $\arcsinh$. Although there considerable heteroscedacity always exists in the distribution of untransformed intensities, it is usually greatly reduced although not completely eliminated after taking logs. The authors just cited use a mean-variance relationship which includes a dominant additive component of variation at low intensities and a dominant multiplicative component at high intensities. One consequence of this is that the transformations they derive behave like log at high intensities, despite a body of evidence that this is too severe a transformation in that range. Further, we believe that the need for an additive component at low intensities can be reduced if not completely removed by better background adjustment, see Yang et al. (2002a). In brief, we are not sure that the problem discussed here can be solved by using a transformation of intensities other than the log-ratio.
Multiple gene analyses

It is not hard to see that, for the most part, the approaches to design and analysis offered so far in this chapter can be described as a single gene approach. We summarize them as follows:

1. Find appropriate ways of carrying out inference for single genes, and do this with each gene separately.
2. Deal with the special problems that arise (e.g., multiple testing, empirical Bayes) because we have tens of thousands of genes.

In brief, we have outlined a one-gene-at-a-time approach. On the other hand, all the many forms of cluster analysis used with microarray data (and discussed in Chapter 4) are explicitly all (or most or many) genes at a time, and for the most part do not focus on single genes. It is clear to us that room exists for approaches to statistical inference, that is, to answering biologists’ questions. These questions are, in a suitable sense, midway between single gene and cluster analyses.

Here is a line of reasoning that leads naturally to multiple gene analyses. Suppose that we seek pairs of genes that are jointly differentially expressed in a treatment-control comparison. In many cases, these genes would be differentially expressed separately, and therefore should show up in single gene analyses, but there may well be gene pairs that are significantly affected jointly, but not much so separately. For example, the association between two (or more) genes may be different under one condition to that under another. Can we find the gene pairs so affected? One approach we have tried, which failed miserably, was to go beyond the single gene one- or two-sample t-tests outlined in the section “Two-sample comparison” to one or two-sample bivariate Hotelling $T^2$-tests. With an experiment similar to Case Study II having eight treatment and eight control slides, we found nothing useful: the millions of pairs of genes led to $T^2$-statistics that were hardly distinguishable from noise. It became clear that a “head-on” approach was doomed to failure because of the sheer number of pairs of genes.

The following question logically precedes the search for pairs of genes that are jointly differentially expressed: Can we find pairs (more generally sets) of genes with transcriptional levels that are associated in an experiment? Questions like this have been the subject of a series of articles by S. Kim, E. Dougherty, and colleagues (Kim et al., 2000a; Kim et al., 2000b; Dougherty et al., 2000). In these articles, two-channel microarray data are reduced to ternary form (−1, 0, and +1 for downregulated, invariant, and upregulated, respectively) and nonlinear systems, such as neural networks, are used to determine when one gene’s transcriptional response can be predicted by that of other genes. With thousands of genes, and thus millions of sets of predictors, this is clearly a formidable task, and yet it is but a small beginning down the road of true multiple gene analyses. We refer readers to these articles and commend the general problem to them.

A final example of multiple gene analyses is the synthesis of classification and clustering that leads to a more formal statistical treatment of clustering than is usually
the case. Here, we simply refer the reader to Fridlyand and Dudoit (2001) and references therein.

**Significance testing**

We have already indicated that the theory for multiple testing in the microarray context is not satisfactorily complete, even for simple problems such as identifying differentially expressed genes. A number of other contexts exist that involve microarray data in which biologists would like to make statements concerning statistical significance and where the multiple testing issue arises. Two such contexts are described next.

In the package GenMAPP (<http://www.genmapp.org/about.asp>) genes belonging to a given biochemical pathway can be colored red or green or yellow according to whether they are up- or down-regulated or unchanged, in some treatment mRNA relative to mRNA from a reference condition. If the genes in a pathway are in truth unaffected by the treatment, one would expect the pattern of red and green coloring to be “random” in some suitable sense, whereas if that pathway is indeed affected by the treatment intervention, one might expect more genes changing in one direction or another than might occur “by chance.” Can these notions be made rigorous and a test of significance developed for the null hypothesis that a specified biochemical pathway is not affected by treatment intervention? For a first attempt at addressing this question, see Doniger et al. (2002). What about the same question when we do not specify the pathway in advance, but search over pathways after the analysis?

A very similar question arises when biologists make use of the Gene Ontology (Go) (<http://www.geneontology.org/>, which is a framework for assigning genes a molecular function, a biological process and a cellular location. Naturally, in most cases, the assignments should be regarded simply as principal ones, for many perhaps most genes will not have unique assignments to categories within these headings. Nevertheless, the GO is extremely valuable to biologists seeking to interpret the results of microarray experiments, and in this context another significance question arises. Frequently, the immediate outcome of an analysis of a microarray experiment is a long list of genes that have been doing something in the experiment; let us say that they have large interaction in a \(2 \times 2\) factorial experiment. The experimenter then compares each gene on the list to the GO and find that 15 of the genes are transcription factors, or are concerned with locomotory behavior, or are located in the nucleus. Is this number unusual, or could it have readily occurred “by chance?” A reference set could be all genes on the microarray. Biologists are interested in the answers to questions like this because they are looking for clues concerning the cellular processes affected by the experimental intervention.

It may be that questions like the two just mentioned cannot be well-posed, but it would be good to have a thorough discussion of the issues involved here.
Combining other data with microarray data

A class of problems of ever-increasing importance concern the combining of microarray data with clinical data. It is easy to envisage the time when an individual’s cancer diagnosis, for example, will involve a wide variety of clinical observations—the traditional diagnostic indicators—together with the absolute or relative expression levels of tens of thousands of genes. Further, we can expect such data together clinical outcomes, such as survival data on patients. Early articles on this topic include Alizadeh et al. (2000), Bittner et al. (2000), and Sorlie et al. (2001). It is hard to be specific here, but the challenges and opportunities appear great.

One of the common outcomes of a microarray experiment is a number of lists (perhaps clusters) of genes which are thought to be co-regulated, that is, to be acting in concert in the biological processes underlying the experiment. A very natural question for a biologist with access to the genome sequence of the organism under study is: Can we find short regulatory sequences in the genome, upstream of any these sets of genes, that might be responsible for their being co-regulated? In the jargon of molecular biology, such regulatory sequences are termed cis-acting, in order to distinguish them from trans-acting sequences, which are much more difficult to identify computationally. For some early research on this problem, see Bussemaker et al. (2000), Hughes et al. (2000a), Cohen et al. (2000), Bussemaker et al. (2001), Pilpel et al. (2001), Chiang et al. (2001), and Keles et al. (2002). A related problem is discussed in Liu et al. (2002). These articles, and others like them, are just the beginning of what will surely grown into a significant body of research that is likely to require considerable statistical input: the addressing of questions that involve both gene expression and genome sequence data. This appear to be an appropriate theme on which to close our brief summary of the design and analysis of gene expression experiments.