In this chapter, we review the design and low-level analysis of microarray experiments. Microarray experiments are widely used to quantify and compare gene expression on a large scale. Such experiments can be costly in terms of equipment, consumables, and time. For this reason, careful design is particularly important if the resulting experiment is to be maximally informative, given the effort and the resources. A number of issues must be addressed when designing a microarray experiment: what will have the most impact on the accuracy and precision of the resulting measurements? How should the different components
of the experimental design be balanced to reach a decision? For example, should we replicate, and if so, how? Which samples should be hybridized to which slides? Should samples be pooled? If the design is inadequate, the experimenter will be left with a less than effective use of resources and the resulting conclusions might be biased. The scientific question of interest may even be left unanswerable with the collected data.

Low-level analysis is carried out between the image analysis phase and interrogation of gene expression data. The goal of low-level analysis is to take raw data from the scanner, without any biological interpretation, and process it to produce cleaner and ultimately more meaningful gene expression measures. This is in contrast to higher level analysis in which questions of a more biological nature are addressed. Such high-level questions include detecting differential expression in treatment and control tissues, gene function, pathway analysis, and changes in gene expression over time, among others. We shall not further consider such high-level analysis here. However, improved low-level analysis ultimately aids the downstream data investigations and for this reason is very important to consider. In this chapter, we consider the following low-level topics: normalization, in which the goal is to reduce or remove sources of nonbiological variability (for both complementary DNA [cDNA] and Affymetrix arrays), summarization, in which one combines multiple probes to produce a gene expression measure (a topic of great importance for Affymetrix-like arrays), data quality, as may be assessed by image analysis for cDNA arrays and from probe-level modeling of Affymetrix arrays, and the detection of absolute expression.

II. Design of Experiments

A. Design of Experiments Using Affymetrix Arrays

Comparative experiments with an Affymetrix chip share many similarities with comparative experiments more generally. In contrast with two-color cDNA microarray experiments, which we discuss next, we can make immediate use of the extensive statistical literature on experimental design (Box et al., 1978; Cobb, 1998; Cox, 1992; Montgomery, 2000). These books present excellent discussions of the general principles of randomization, replication, and local control within the context of agricultural, industrial, and scientific experimentation. In this section, we briefly summarize the particular aspects of these issues relevant to comparative experiments involving Affymetrix and other high-density short oligonucleotide chips and refer the reader to the books for more general considerations.

Randomization is not widely used and not popular in biomedical laboratory experiments. In our view, it would be a step forward if this attitude changed. One
of the most striking aspects of Affymetrix chip experiments is the extent to which the conditions of an experiment—the messenger RNA (mRNA) extraction and processing, the reagents, the operator, the hybridization conditions, the scanner, and so on—can leave a “global signature” in the resulting expression data. If an experiment involving a number of chips is carried out over a number of days, more often than not, chips processed on the same day are readily identifiable as such, for example, by cluster analysis or some other similar global analysis. If different parts of an experiment are carried out on different equipment, at very different times, or with different personnel, these global differences can be striking, raising the possibility that they could obscure or confound the very differences the experiment seeks to measure. The normalization methods we discuss can reduce the impact of some of these differences, but if they cannot be controlled by good design, randomization is the best way to deal with them. We illustrate the idea next.

Replication is the foundation stone of statistical inference. No one should expect to be able to reach conclusions about variable populations based on a sample size of one. Appropriate replication is the key to generalizing from a sample to a population, but the term appropriate is important here. Taking an RNA sample from, say, the liver of one mouse from an inbred strain, treating it correctly, and hybridizing portions of the result on to two Affymetrix chips leads to what is known as technical replicate data: two sets of measurements on one mouse. This is to be contrasted with taking RNA from the livers of two mice from the same strain and hybridizing RNA from the different mice onto different chips. These are called biological replicates: one set of measurements on each of two mice. Clearly biological replication leads to data that are better for reaching conclusions that might apply more generally, and using more than two mice is better still. Equally clearly, technical replication leads to data better for reaching conclusions about that particular mouse, under the conditions of that particular experiment. The term appropriate depends on the aim of the experiment, where this term includes the level of generalization sought for any conclusions.

In most experiments involving animals or humans, biological replicates are more relevant to the aims of an experimenter than technical replicates. Indeed, it is arguable that in an experiment in which variable material such as animal or human tissue is involved, technical replicates are never preferable to biological replicates. Chips are expensive, and a technical replicate will (in a well-conducted experiment) usually give results very similar to the original hybridization, whereas even inbred animals can lead to quite different results. Biological replication can also be expensive, but in general it will be better to spend limited resources getting more data at the level exhibiting higher variation, across animals, say, than at the level exhibiting lower variation, within animals. Averaging results can reduce the impact of chance variation on summary statistics, but if there are chance features common to all results being averaged, as in the case of technical replicates, averaging will not reduce their impact. If the ability to obtain
technical replicates is limited, but resources for experimentation are not, then carrying technical replicates is a good way to improve the quality of the data you have, but only about the samples measured. In a way, you get better information about a more limited universe, which is reasonable if it is not possible to expand that universe, but in general the expanded universe leads to more robust scientific conclusions.

There are many intermediate levels of replication between what one might call pure technical and pure biological replicates, and in general a more efficient use of resources results from carrying out replication at the level at which the greatest variability is to be found.

Local control is the general term statisticians use for arranging experimental material, in the present case, RNA samples to be treated and hybridized to chips, in relation to extraneous sources of variability. By extraneous we mean variability in the measurement process, not variability in the experimental material, which is what is relevant to replication. For example, if we propose to compare gene expression between wild-type and particular mutant animals and plan to use six animals in each group, it would be most unwise to carry out the processing of all six animals in one group on one day and then process the six animals from the other group on a second day. If this was done, day-to-day differences, which are sometimes difficult to avoid, could be wrongly seen as genetic differences. Equally, it would be unwise to carry out the processing of the six wild-type animals first, spread out over days, and then follow with the six mutant animals. A more appropriate approach in this example would be to proceed as follows. First decide how many samples can be processed in a single day, and hence the number of days over which the processing will be spread. Suppose that six per day for 2 days is deemed feasible. Then an appropriate design would be to process three wild-type and three mutant animals on each day, each set of three chosen at random from the available six, and within a day, process the mix of three wild-type and three mutants in a random order. A more complex design might seek to avoid processing all three mutants or all three wild-type animals on any given half-day, although here it is clearly impossible to arrange for equal numbers in each group to be processed in each half-day. The principles being illustrated should be clear here: group the material so that any unique chance features of a day are shared equally by the two groups and thus are not potentially confounded with group differences, and to the extent that this is not possible, randomize over the remaining features, such as time within day. Similar considerations apply whenever differences will occur across other factors known to contribute extraneous variability, such as chip batch, operator, reagents, scanners, and so on. Arrange those varying factors that can be controlled in a manner similar to that just described, and randomize across the remainder. This is the general idea, and more details of its implementation can be found from the books listed. For the design of a complex experiment, it is advisable to seek statistical advice.
We are frequently asked about the advantages or disadvantages of pooling RNA samples before hybridization. For example, in the small wild-type versus mutant experiment just discussed, RNA from the six wild-type animals could be pooled, and the resulting mix hybridized to six (or fewer) chips, with the mutant animals being dealt with likewise. Many people feel that pooling provides a form of “biological averaging” and should make it possible to get more precise results with fewer chips than hybridizing RNA from individual animals to separate chips. Our view is that although this may well be true, compelling evidence is not yet present in the literature. On the other hand, we have seen evidence that pooling could mislead, but perhaps more importantly, could mislead without this being apparent. Suppose that one animal contributing to the pool is very different from the remainder, for example, perhaps one mouse has been injured by fighting with other mice, and that its immune system is in a quite different state. Depending on the tissue under study, pooling this animal’s RNA with that of the other mice gives it the ability to exert a large influence on the measured expression values for many genes, whereas running it on a separate chip leaves open the possibility of identifying this animal as an outlier. If the level of a particular mRNA transcript is 50 times the average of the other animals, biological averaging is not taking place, rather this outlier animal is biasing conclusions reached for the group. At this point it is probably best to say that if pooling is envisaged in an experiment that could be carried out at the same cost without pooling, it is better not to pool. If pooling is carried out to save cost, the possible drawback just mentioned should be borne in mind. Frequently, pooling is seen to be necessary to get sufficient mRNA from the tissue in question, in which case the possible drawbacks must inevitably be accepted or at least weighed against the possible drawbacks of the alternative, which is usually amplification.

B. Design of Experiments Using cDNA Arrays

Because the two-color microarray system is inherently comparative, the major design issue for cDNA microarrays is which samples should be cohybridized. For any proposed design to be desirable, it needs to satisfy two types of constraints: physical and scientific. Physical constraints include the quantity of RNA available, the number of slides, and other cost considerations. Scientific constraints should motivate the design in that more important questions should be able to be answered more precisely than less important ones.

The primary decision with cDNA experiments is which samples to hybridize together. This often becomes a question of whether to use a direct or an indirect comparison. Direct comparisons are between two samples hybridized to the same slide. Indirect comparisons are those between samples that can be compared only via multiple slides. In other words, should we be comparing within or between
slides? Figure 1 demonstrates some of the basic components of cDNA experiments using a simple graphical representation. The squares represent target mRNA samples and the arrows represent hybridization between the two samples. By convention, the sample at the head of the arrow will be labeled with Cy5 (red) and the tail with Cy3 (green).

Sometimes there is an obvious design choice, given the available materials and the goal of the experiment. For example, suppose we have a series of cells each receiving treatment from a different drug and the aim is to compare them with untreated cells. An appropriate design in this case would be one in which the untreated cells become a de facto reference and each one of the treated sets of cells is hybridized with the untreated cells (as in Fig. 1B).

The key difference between direct and indirect comparisons can be illustrated by examining the variance of the estimated log fold change in each comparison. Consider a direct comparison of the form shown in Fig. 1A. We will assume that the variance of the log ratio \( \log(B/A) \) is given by \( \sigma^2 \). With two direct comparisons, we would take the average of the two independent observed ratios, yielding an estimate with variance \( \sigma^2 / 2 \). If we instead make an indirect comparison, as in Fig. 1B, then an estimate of the log ratio is given by \( \log(B/A) = \log(B/R) - \log(R/A) \). Assuming independence, this would have a variance of \( 2\sigma^2 \). In other words, two direct comparisons have one fourth the variance of an indirect comparison. Note that in practice this might not be observed, because the independence assumption is not always valid.
A dye-swap experiment (Fig. 1C) is one in which each hybridization is done twice, with the dye assignments reversed in the second hybridization. This is useful for removing systematic bias. Systematic dye biases are commonly observed with cDNA microarrays, and although normalization (as will be discussed later) can partially correct for this effect, it is still possible that individual spots have a residual color bias. By swapping the dyes and averaging across the two hybridizations, it is hoped that this bias will be reduced.

Figure 1D represents a simple loop design. In this example, a comparison between any two pairs can be made directly (e.g., A–B) and indirectly (through C). Maintaining our previous assumptions, the variance of such a comparison would be $2\sigma^2/3$. With more than three sources of mRNA, the number of hybridizations needed to allow direct comparisons between every pair of sources increases rapidly. In practice, the comparisons of greatest interest would be done directly and those of lesser interest done more indirectly, with those of least interest the most distant.

In a time-course experiment, the design choice depends on the comparison of interest. For instance, if comparisons to the initial time are of interest, then each of the subsequent time points should be hybridized with a sample from the initial time. If changes between time points are of particular interest instead, then a more sequential design would be desirable, with samples from consecutive time points hybridized together. Often, a combination of the two might be desirable, for example, a direct comparison between the first and final time points and direct comparisons between each of the intermediate time points.

The second important decision in cDNA microarray experiments is how replication should be carried out. Replication is important because it reduces variability in summary statistics and allows data to be analyzed using more formal statistical methods. With cDNA microarrays, individual genes can be quite variable between hybridizations. By replicating, then averaging, less variable estimates are obtained.

A common form of replication is to place replicates of the same spot (cDNA probe) on each slide. However, because such spots will typically share systematic effects from printing, general hybridization, and scanning, the lack of independence between these measurements reduces their value for more sophisticated statistical inference. If duplicate spots are to be used, it is recommended that they are distributed across the slide because this will give a better reflection of intraslide variability.

There are two methods by which between-slide replicates can be created: technical replication and biological replication. Technical replication is when mRNA from the same extraction is hybridized to multiple slides. Because RNA extraction typically has characteristic repeatable elements, technical replicates usually have smaller variability than biological replicates. In addition, these shared systematic features will remain even after averaging. Biological replication
refers to hybridizations that use mRNA from different extractions from, for example, the same cell line or tissue. Ideally, sample labeling is also carried out separately for each extraction. In the context of a microarray experiment, this leads more closely to independent experimental results. Biological replication should therefore, be favored as the primary method of replication.

*Biological replication* also refers to the situation in which target mRNA is extracted from different individuals or perhaps from different versions of a cell line. This form will be more variable than the biological variation discussed earlier. Although this might make it more difficult to detect real expression differences, the conclusions made from such an experiment might be more generalizable. For example, multiple extractions from a single individual cannot be treated as being representative of all individuals with the same condition, so any conclusions made using the former might be flawed.

A mixture of biological and technical replication is often desirable. Biological replication will allow more generalization of conclusions. Technical replication will reduce variability. Just how much and which types of replication can be done are determined by the physical constraints of the experiment.

### III. Sample Size Considerations

#### A. A Classic Power Calculation

Suppose that we use a microarray to measure the expression of one gene in a class of cells on unmatched samples from cases and controls. (The approach in what follows applies equally to matched samples, but the actual figures will change.) We suppose that the measurements are given in the log₂ scale. They will be subject to measurement error, which we view on the same log₂ scale. Our aim in this section is to explore power issues in a context in which the aim of the study is to identify differentially expressed genes based on gene expression microarray data.

The test we discuss here is the standard two-sided, two-sample *t* test, with type I error (i.e., false-positive) rate of 5%. Such a test seeks to identify differential expression between the case and control groups on the log₂ scale, in other words, relative expression, because a difference of logs of means equals a log ratio of means. For example, a mean difference of 1 between the two groups on this log₂ scale is equivalent to a twofold difference, on average, on the original (concentration) scale. Microarray gene expression measurements are subject to measurement error, and in our analysis, we make a range of realistic assumptions concerning the standard deviation (SD) of this error.
Of primary interest in power studies is the probability of detecting a mean difference of a given magnitude $\Delta$ between the two groups, given a sample size $n$ and value $SD$ for the error standard deviation, and our test procedure (with given type I error). Put another way, of interest is the power of our test to reject the null hypothesis of no mean difference in the gene’s expression between the two groups, given that the true mean difference has a magnitude $\Delta$, the sample size is $n$, and the measurement error has standard deviation $SD$.

Realistic values of $\Delta$ in this context are 0.5, 1.0, and 2.0, corresponding to fold changes of 1.4, 2.0, and 4.0. Similarly, realistic values of $SD$ (on the log$_2$ scale) are 0.5, 1.0, and 2.0, while we consider sample sizes (for each of the case and control groups) of 10, 20, and 30. We will see shortly that the main points we make are not dependent on the precise values of these numbers. As foreshadowed previously, we calculate and display later the power to reject the null hypothesis for all 27 combinations of these input variables, using the abbreviations LO and HI for values of the power close to 0 or 1, respectively.

Each of the three tables later in this chapter correspond to one value for the SD, the first to the value 0.5, the second to 1.0, and the third to 2.0. Rows of the tables are labeled by the value for $\Delta$, and columns by the value of $n$. All the entries of the tables are power values, that is the probability of (correctly) rejecting the null of no mean group difference, given the values for the parameters and the fact that we are using a two-sided two-sample $t$ test, with type I error 5%. The power values are taken from Table 10 of Pearson and Hartley (1962).

<table>
<thead>
<tr>
<th>$\Delta$ = 0.5</th>
<th>.60</th>
<th>.8</th>
<th>.95</th>
<th>LO</th>
<th>LO</th>
<th>.5</th>
<th>LO</th>
<th>LO</th>
<th>LO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$ = 1</td>
<td>.98</td>
<td>HI</td>
<td>HI</td>
<td>.6</td>
<td>.8</td>
<td>.95</td>
<td>LO</td>
<td>LO</td>
<td>.5</td>
</tr>
<tr>
<td>$\Delta$ = 2</td>
<td>HI</td>
<td>HI</td>
<td>HI</td>
<td>.98</td>
<td>HI</td>
<td>HI</td>
<td>.6</td>
<td>.8</td>
<td>.95</td>
</tr>
</tbody>
</table>

SD = .5

SD = 1

SD = 2

What can we learn from these values? The most striking thing is that the actual power varies from LO to HI through the values 0.5, 0.6, 0.8, and 0.95 as we vary the postulated mean group difference $\Delta$, the SD, and the sample size through plausible values in the microarray context. Put another way, depending on the group mean difference $\Delta$ and the single gene expression measurement SD, we can have a power from HI to middling to LO, as the sample size varies. For large $\Delta$ and small SD, we can get by with small samples and feel sure that group mean differences of this magnitude will be detected, whereas for smaller values of $\Delta$, larger SDs, or both, detection of such group mean differences may be difficult with even the highest sample size we currently contemplate.
B. SOME OBSERVATIONS ABOUT POWER CALCULATIONS IN THE MICROARRAY CONTEXT

We now make some important observations about power calculations in this microarray context. Together, they suggest that the previous power calculations are of little direct value in the microarray context.

First, in general we do not necessarily know in advance the genes whose between case control mean differential expression we wish to ascertain. Some of these will be probes corresponding to unknown ESTs, and our finding them to be differentially expressed between cases and controls might be the first interesting fact found out about them. Often the main reason for doing a microarray experiment is to find genes that are differentially expressed between cases and controls.

Second, even if we do know the names of the genes of interest, and there are often many likely candidates that could be specified in advance, we will not necessarily be able to state in advance the magnitude $\Delta$ of the relative mean expression that is of interest. Of course, we might nominate a fold change such as 1.4 and claim interest in detecting all genes with fold changes of 1.4 or greater. However, there are undoubtedly genes for which smaller fold changes are still biologically significant and that might be detected with greater replication. But we will not know in advance which these are.

Third, even if we did know the gene or genes of interest and we were able to specify to the magnitude of the difference $\Delta$ of interest, we are extremely unlikely to know in advance the SD of a single microarray expression measurement on that gene. Again, we might nominate a (perhaps conservative) value for the SD and hope that in so doing, we avoid this objection. But again, we will not know in advance which genes have SDs smaller than this conservative value, and which do not. And SDs of genes can vary enormously.

Fourth, the analysis underlying the power calculation just given assumes that all cases have the same normally distributed gene expression measurements (on the log2 scale) and similarly for the controls. It ignores the well-known tendency for cases in particular to be heterogeneous in gene expression, and to a lesser extent, this also applies to controls. Heterogeneity compromises the power calculation, and unknown heterogeneity does so to an unknown extent.

Summarizing these four points, a conventional power analysis is of limited value in the context of analyzing microarray gene expression values, because we will only rarely be able to nominate in advance the gene, the magnitude $\Delta$ of interest, or the magnitude of the measurement error SD, or be sure about homogeneity of response. Both the SD and the biologically meaningful change $\Delta$ are notorious variables across genes in expression microarray studies. Nevertheless, it might still be argued that we should be able to determine the power we have with a given sample to detect all genes whose log fold change is some value
or greater, and whose measurement SD is no more than a given (conservative) value, even if we will not know in advance which genes these are.

However, there is a much greater obstacle to making use of conventional power analyses in this context than these four points. In microarray studies, or other similar "genome-wide" studies, we are not measuring just one gene’s expression; in a typical experiment, we measure the gene expression in 20,000 genes. When we do this, another important issue comes to the fore: the issue of multiple testing. Put simply, the multiple testing issue is that even when no genes are differentially expressed, on average, between cases and controls, many will appear to be so, by chance, according to conventional analyses. In other words, even though the power analysis presented earlier is appropriate for examining whether a single (named) gene is differentially expressed between cases and controls, that is not the context in which we operate with gene expression microarrays. In this context, we are screening tens of thousands of genes for differential expression, and the conventional power analysis just presented is quite inappropriate. To identify differentially expressed genes in the context of tens of thousands of genes, we would never use the conventional two-sided two-sample t tests with a conventional cutoff appropriate to a (single test) type I error rate 5%, or we would risk obtaining hundreds or thousands of type I errors. Indeed, we might not even use the t statistic at all.

C. WHAT CAN BE DONE? PART I

The next question is naturally, which power analyses are appropriate in this multiple testing context? Sadly, but almost inevitably, the answer is that with our present understanding of the statistical analysis of microarray gene expression data (Speed, 2003), there are no analyses strictly similar to conventional power analyses. The reason is not difficult to see. As already noted, with tens of thousands of genes, the observed log2 fold changes vary greatly, and the variability of single gene expression measurements vary greatly. If the analysis is to be global, that is, if it is to use all genes with probes on the array, then many genes will be genuinely differentially expressed at levels that would be of interest to an experimenter if he or she knew, but these will not be identified, because they will be masked by the thousands of genes quite probably not differentially expressed, but simply varying greatly. These statements can be (and have been) supported by theoretical and simulation analyses. Put crudely, for a gene to be noticed against a background of tens of thousands of other genes, it must stand out, that is, it must have a level of differential expression substantially greater than the others, and in particular, substantially greater than we would require in conventional one-gene-at-a-time testing.
Statisticians have developed a number of ways of dealing with the generic testing issues in this context, with names like family-wise (type I) error rate (FWER), false-discovery rate (FDR), and positive FDR (pFDR). These are all multiple testing analogs of the simple type I error for a single test, with different assumptions and different properties (Ge et al., 2003). Their natures differ greatly: FWER refers to single genes, and FDR and pFDR only to sets of genes. The ability to use them differs greatly, with FDR techniques more broadly applicable and FWER approaches less so. And their values can differ greatly. The story is not so simple, but the main point is we cannot and should not speak of the power to reject the null hypothesis of no difference in the mean expression for single genes, with microarray experiments. If we want to use the advantage of the microarray experiment in permitting us to screen tens of thousands of genes simultaneously, we have to appreciate that decisions on individual genes in a microarray experiment are always made in the context of thousands of similar decisions on other genes. This is both a boon for biologists and a drawback for them, because their thinking concerning power (and a few other issues) needs to be modified.

D. IS THERE A MULTIPLE TESTING POWER ANALYSIS?

The nearest analog to a power analysis in the large-scale multiple testing context is embodied in the following idea. We could conduct a computational study in which expression levels for cases and controls for all genes in a large set are simulated, where we include a specified subset of genes as having predetermined levels of differential expression between the groups, and we give all genes predetermined standard deviations for their measurement errors. Then for a specified test statistic and multiple testing procedure, we could examine how many of the genes “known” to be differentially expressed from the design of the study are correctly identified as differentially expressed. This could be done for many procedures at the same time, and the performances compared (Lönnstedt and Speed, 2002).

The catch is that we need to simulate data that we hope look like our actual data. All power studies assume that the actual data are very close in their statistical properties to the underlying theoretical model, and frequently this is not such a bad assumption. In other words, power studies frequently turn out to be useful. However, with gene expression measurements on 20,000 genes, nobody knows how to describe a model (and hence simulate data) that leads to data closely approximating the data we observe. The microarray process is simply too complex for this to be achievable right now, and perhaps for the foreseeable future. We can hope that our simulations capture features of the data relevant to our analysis, and there is some evidence—it is early days yet for this—that our hopes are fulfilled.
Such simulation studies have been carried out, and one of the clear conclusions is that, just as one might expect, when there are a lot of genes differentially expressed between two conditions and these differences in average level vary from high to low, there is no chance of identifying all of them. False-positive rates are controllable, but at the price of what might appear at first glance to be disturbingly high false-negative rates, that is, very low power, in the traditional (one-gene-at-a-time) thinking. When there are just a few genes differentially expressed and the differences in average level are not small, then with a large enough number of replicates, it is reasonable to expect them all to be identified. As is so often the case with microarrays, the catch is that we do not generally know in advance which case we are in, although we frequently suspect that it is the first, and more difficult, rather than the second, relatively easier case.

E. WHAT CAN BE DONE? PART II

Where are we now? If we think that our average gene expression differences between cases and controls are not likely to be large, then we are simultaneously faced with the prospect of possibly missing many genes whose differential expression might be of interest to us and having little that we can do in advance to influence this. It remains true that more independent replicates always helps our aims, as long as the heterogeneity does not increase at the same rate and the analyses are appropriate. What we cannot do is say how many replicates are enough. Even answering the question “enough for what?” is not easy.

Two simple conclusions follow from this discussion. First, we should aim to get as many case and control samples as we can, bearing in mind the important requirement of homogeneity. And secondly, we need to improve the quality of our statistical analysis, knowing that we cannot increase our sample size beyond a certain point. In brief we need a “smarter analysis” to try and overcome the limitations of modest sample size. We need to do better than using the standard $t$ test (or its analogs) and instead use appropriate calibrated moderated $t$ statistics. We need to depart from the conventional “context-free,” search for differentially expressed genes that get embodied in power calculations and their multiple testing analogs, and more fully integrate biological knowledge with statistical analysis. Recent efforts for connecting pathways, the Gene Ontology classification and related tools are along these lines (Mootha et al., 2003). Even when we do these things, multiple testing issues will remain, so we have to extend that theory to apply to our stronger analysis. In brief, we are still in the early days of the statistical analysis of microarray data. Much traditional thinking must be extended and strengthened.
IV. Normalization

Normalization is a process performed to compensate for systematic technical differences both between and within arrays. The process of normalization should reduce or remove this variation while leaving the more scientifically interesting biological differences that may exist.

Systematic nonbiological differences between samples become apparent in several common ways. For instance, it is often observed that one array is brighter overall than another. With cDNA arrays, a systematic difference in the intensity of signals from different dyes is a frequently observed source of variation.

There are many possible causes of systematic nonbiological variation. Differing amounts of RNA, scanner settings, and differing hybridization or experimental conditions are all commonly observed contributors to this sort of variation. In two-color arrays, dye biases are commonly observed. These biases could be due to various factors such as physical properties of the dyes (light and heat sensitivity), how efficiently dye is incorporated, or experimental variation in the labeling process.

It should be remembered that all normalization methods require some level of assumption about the underlying data. The most common assumption is that most genes are not changing across conditions. A second common assumption is that the number of upregulated genes is roughly equal to the number of downregulated genes across conditions.

A. Normalization for cDNA Arrays

The MA-plot (Dudoit et al., 2002a) is a very useful tool for normalizing two-color microarray data. For each spot on the array, we have a (R, G) fluorescence intensity pair (where R = red, for Cy5, and G = green, for Cy3). An MA-plot is used to represent these (R, G) data pairs, where we define $M = \log_2 \frac{R}{G}$ and $A = \log_2 \sqrt{RG}$. MA-plots help to identify spot artifacts and to detect intensity-dependent patterns in the log ratios. Note that the MA-plot is a rotation and rescaling of a plot of R versus G.

Because equal amounts of RNA are generally hybridized to both channels, a difference in brightness between channels must be due to different uptakes of the labeling dye. The simplest adjustment one can make to two-color array data is to scale the data so that both channels have equal total intensity. We can think of this as relating the two channels by a constant so that $R = kG$. Note that this is equivalent to subtracting a constant from the log ratio. Thus, the transformation is
We would usually choose a constant that centers the distribution of \( M = \log_2 R/G \) around 0, by setting \( c = \log_2 k \) to be the mean (or median) of the \( M \) values. However, this method does not adequately deal with nonlinear intensity-dependent differences in the dye bias.

A commonly used intensity-based adjustment (Yang et al., 2002b) is a loess scatterplot smoother (Cleveland and Devlin, 1988) fitted to the MA-plot. A loess smoother is a locally robust linear fit. It will not typically be affected by the small fraction of differential genes, which would appear as outliers on an MA-plot. The loess adjustment is given by

\[
\log_2 (R/G) \rightarrow \log_2 (R/G) - c \rightarrow \log_2 (R/kG),
\]

where \( c(A) \) is the loess fit to the MA-plot. A span of \( f = 0.4 \) is typically used for the loess curve.

Although the global loess adjustment deals effectively with intensity dependent differences in the dye bias, we sometimes observe differences resulting from spatial location on the array. This is most clearly illustrated in Fig. 2 where print-tip–specific effects are still evident after global loess normalization has been applied. An improvement to the global loess method is to use print-tip–specific loess smoothers (Yang et al., 2002b). The adjustment is now given by

\[
\log_2 (R/G) \rightarrow \log_2 (R/G) - c_i(A) \rightarrow \log_2 (R/k_i(A)G),
\]

where the \( c_i(A) \) are loess smoothers fitted individually to data from each print tip. This method is recommended for routine use because it deals with both intensity-dependent effects and subarray variation.

There are many other features of the data that could be used in the normalization process. However, further normalization should be applied only when there is clear evidence from diagnostic plots indicating the need for such normalization. Unnecessary estimation of such effects and trend removal may add noise to the data. One variation, after using print-tip loess normalization, is to further standardize the \( M \) values from each print tip to have the same scale. In particular, it is assumed that the variance of the \( M \) from each print-tip group is given by \( \sigma^2 \).

We can robustly estimate \( \hat{a}_i = \text{MAD}_i / \sqrt{\prod_{i=1}^{l} \text{MAD}_i} \), where \( \text{MAD} \) is the median absolute deviation. The scale-normalized values for grid \( i \) are then given by \( M_i = M_i / \hat{a}_i \). This normalization is typically not required except in cases in which the arrays are extremely noisy. An example in which such normalization might be required is shown in Fig. 3, where we see that the variability of \( M \) from the fourth row of grids is significantly larger than that for the other grids. Applying the scale normalization removes this difference.

Usually all the spots on the array are used in the normalization methods described previously, because this provides the most stability in terms of the
number of spots and the flexibility to operate in a print-tip–specific manner. However, sometimes the expression profiles in the biological samples are more divergent than has been assumed in the cases mentioned. The previous strategies

Fig. 2. An MA-plot after global loess normalization (top) shows print tip effects, which are eliminated using the print-tip specific normalization (bottom).
can be employed if a suitable set of control spots that are known to be not differentially expressed are printed on the array. Ideally these would span the range of possible concentrations. One such method is to use a microarray sample pool (MSP) titration series in which the entire clone library is pooled and then titrated at different concentrations. Because, in theory, all labeled cDNA sequences should hybridize to this series, it should not be subject to sample-specific biases. Differential genes should not bias a loess curve through the control spots.

Sometimes it is useful to combine the MSP normalization and the print-tip–specific normalization. It is suggested (Yang et al., 2002) that one take a weighted average of the print-tip–specific adjustment and the MSP normalization, in which the weights are dependent on the intensity. Define

$$
\epsilon_i(A) = w(A)\hat{g}(A) + (1 - w(A))\hat{f}_i(A),
$$

where $\hat{g}(A)$ is a loess curve fitted to spots from the MSP series, $\hat{f}_i(A)$ is a loess for print-tip group $i$, and $w(A)$ is usually defined as the proportion of spots less than intensity $A$. The adjustment is then done as before. The idea is to increasingly use the MSP curve at higher intensities, where there are fewer spots and the print-tip–specific curves may be more unreliable.
We have discussed normalization within slides, but sometimes there are large differences in scale when comparing data between slides. The advised procedure is to first normalize within slides using the methods previously discussed, and then consider scaling of $M$ between slides, as described previously. This adjustment is needed so that the relative expression levels from one slide do not dominate the expression levels from others when averaging across replicate slides. It should be noted that there is a tradeoff between the gains achieved by scale normalization and any variability that may be introduced. Often this normalization will not be required.

Software implementing these normalization methods for cDNA data may be found in the SMA package (Dudoit et al., 2002b) and downloaded from CRAN (http://cran.r-project.org/).

B. NORMALIZATION FOR AFFYMETRIX ARRAYS

There are two main approaches to normalization of Affymetrix GeneChip data. A recent paper (Bolstad et al., 2003) categorizes these into methods that use a baseline array and methods that are complete data methods. A complete data method does not use a baseline array, instead using data from all the chips to form the normalization.

Examining box plots of raw probe intensities by array can often show the need for normalization. Such a plot is shown in Fig. 4A, for five arrays from part of a dilution series dataset (Gene Logic, 2001). The only difference between the arrays is the scanner that was used, yet the box plot shows quite different levels of expression for each array.

A number of normalization methods have been proposed. The simplest approach, scaling, is to scale each array so that all arrays in a dataset have the same mean intensity. Trimmed means are often used instead of means, and this is the method used by Affymetrix in the MAS 5.0 software (Affymetrix, 2001). If $\bar{X}_i$ is the mean (trimmed) intensity for array $i$ and $K$ is the target mean intensity, then array $i$ is normalized by multiplying by $K/\bar{X}_i$. The target intensity is often chosen to be the mean of one of the arrays. We would, thus, classify this method as a baseline method. Figure 4B shows the five arrays after scaling normalization. The scaling approach can be applied in a time-efficient manner, but it does not adequately deal with possible nonlinear trends between arrays, as shown in Fig. 5. The Affymetrix HG-U113A chip has 100 normalization control probe sets that may be used for normalization in this context. These probe sets have been chosen because of their stability of expression across a wide range of tissues.

Another approach is to choose a baseline array, then fit nonlinear relationships between the baseline array and each of the other arrays (in this context, we call these the treatment arrays). Such an approach fitting splines was suggested by
Schadt et al. (2001) and used with a running median line (Li and Wong, 2001a,b). A rank invariant set of probes is chosen between the baseline and the treatment array. These probes are then used to fit the nonlinear relation. The curve is then used to map from the treatment array to the baseline array and defines the normalization.

Several complete data adaptations of the MA-plot loess method for cDNA arrays have been proposed for normalizing Affymetrix arrays. The first is the cyclic loess method, in which arrays are normalized against each other in a pairwise fashion using a loess fit to an MA-plot. Unfortunately this requires $O(N^2)$ MA-plot normalizations and so it is quite time consuming.

A second adaptation of the MA-plot loess method is to transform the data using an orthonormal basis to give a set of contrasts (Åstrand, 2003). The normalization is applied to the transformed data. The data are then transformed back to the original basis. This method requires only $O(N)$ MA-plot normalizations and is, therefore, faster than the cyclic loess method. However, loess normalizations are slow for probe-intensity data. Typical implementations use only a subset of the probes to improve the processing time.

Another complete data method is the quantile normalization method, in which the goal is to normalize arrays so that each array has a common intensity.

Fig. 4. Box plots of log-scale PM intensities for five arrays from different scanners when (A) unnormalized, (B) normalized by scaling, and (C) quantile normalized.
distribution. This method uses a simple non-parametric algorithm to quickly normalize a batch of arrays. In particular, averaging the quantiles of all the arrays in the set forms the reference distribution. Each array is then assigned the reference intensity distribution. The quantile normalization method is a specific case of the transformation $x'_i = F^{-1} G(x_i)$, where we estimate $G$ by the empirical distribution of each array and $F$ using the empirical distribution of the averaged sample quantiles. Extensions of the method could be implemented where $F^{-1}$ and $G$ are more smoothly estimated. However, we have found the current method to perform satisfactorily in practice.

Figure 4C shows the arrays after quantile normalization. Figure 5 demonstrates that the quantile normalization deals adequately with nonlinear relationships in the data. In practice, this normalization can be carried out in a very time-efficient manner.

These methods were compared in a recent paper (Bolstad et al., 2003), where it was demonstrated that the scaling method was least effective at reducing variability. Figure 6 illustrates this result using the RMA expression measure. These graphs show the ratios of the variance of the expression measure across five arrays plotted against mean expression for two different normalization

![Figure 5. MA plots comparing PM probes from array 1 with array 5 when (A) unnormalized, (B) normalized by scaling, and (C) quantile normalized.](image-url)
methods. We see that scaling and quantile normalization both reduce the variability, with the greater reduction achieved by the quantile normalization method. The complete data methods were favored in Bolstad et al. (2003) because baseline methods can introduce peculiarities of the baseline array into the data for the treatment arrays. It was found that the quantile method was the fastest, with acceptable reductions in variance and little change in bias. Software implementing these normalization methods may be found in the affy package (Gautier et al., 2003), which is part of the Bioconductor project (see www.bioconductor.org).

V. Expression Summaries for GeneChip Data

Figure 7 shows that background-corrected probe intensities follow an additive model on the log scale. For each probe set, we can write the following model:

\[ \log_2(B(PM_{ij})) = a_i + m_j + \varepsilon_{ij} \]  

for \( i = 1, \ldots, I \) and \( j = 1, \ldots, J \). The quantity \( B(PM_{ij}) \) is the background-adjusted, normalized PM intensity, \( a_i \) is a probe affinity effect, \( m_j \) is a quantity
proportional to the amount of transcript on array \( j \), and \( \varepsilon_{ij} \) is an independent identically distributed error term with mean 0. For identifiability of the parameters, we assume that the sum of the \( a_i \) is 0 for each gene. Notice that this assumption translates to assuming that the Affymetrix technology has probes with expected intensities that, on average, are representative of the associated gene expression.

Under this model an unbiased estimate of \( m_j \) for each array \( j \) could be obtained using the average of the \( \log_2(B(\text{PM}_{ij})) \) across the \( i = 1, \cdots, I \) probes. This average can be used to estimate a simple expression measure. We can demonstrate empirically that this expression measure works well. If the errors are normally distributed, this estimate is according to various statistical criteria. However, many researchers (Li and Wong, 2001b) have observed that outliers (observations too extreme to occur under the normality assumption) are relatively common. For some arrays, the proportion of outliers is as high as 15%. This suggests that the aforementioned model should be fit using robust procedures.

Median polish is a simple ad hoc procedure for fitting such a model robustly (Holder et al., 2002; Tukey, 1977). Irizarry et al. (2003a) demonstrate that the expression measure obtained using median polish provides estimates with comparable accuracy to and much better precision than the two leading expression measures, namely those obtained from MAS 5.0 and from dChip MBEI (Li and Wong, 2001a,b). Irizarry et al. (2003a,b) call this procedure the Robust Multi-Array Analysis (RMA).
The additive model lends itself to various practical extensions. For example, if we are comparing two populations of RNA species for which we have many technical replicates that we assume have the same expected RNA expression, we can write

\[
\log_2(B(\text{PM}_{ijk})) = a_i + m_j + \varepsilon_{ijk}
\]

for \(i = 1, \cdots, I\), \(j = 1, \cdots, J\), and \(k = 1, \cdots, K\). The estimate of \(m_j\) would then be based on \(K\) times more data than RMA. If we had technical replicates instead of biological replicates, we could add a \(\zeta_{ij}\) term to the model, representing a random effect (Chu et al., 2002).

VI. Quality Assessment

A. Quality Assessment for Affymetrix GeneChip Expression Data

Producing gene expression data using microarray technology is an elaborate process with many potential sources of variability. To maximize the scientific value of gene expression information derived from microarrays, we must make rigorous quality assessments throughout the process. Standard sample preparation protocols include a number of qualitative assessments meant to ensure that good quality RNA is used in the hybridization experiments. After hybridization and image processing, each microarray provides a wealth of information that can be used to assess the quality of the data. Recommended post-hybridization quality assessments include general image quality assessment and analysis of intensity measures of specialized probes (Affymetrix, 2001).

In this section, we suggest some methods to assess data quality based on the analysis of residuals from the models fitted to estimate gene expression. Departures from quality standards may be attributable to various sources: RNA preparation, hybridization, chip scan, wash, image processing, or faulty chips. The effects of departures from quality may be localized to a small area on a chip or may be uniformly distributed over an entire array, possibly affecting numerous arrays. In most cases, departures from quality standards attributable to processing failures will be reflected by inflated residuals from fits to models such as Eq. 1. Residuals are, therefore, expected to provide useful information for data quality assessment.

Quality assessment can be focused at different levels: at the level of individual probes, of probe set summaries, of probe sets, or of chips. Fitting the probe level models robustly will automatically reduce the effect of malfunctioning probes (cross-hybridizing or non-responding probes) on the estimated expression values, so diagnosis of dysfunctional probes is not required to obtain good expression...
summaries in this context. It may still be useful to identify dysfunctional probes (by means of residual analysis) for other purposes, for example, when seeking cross hybridizing probes or genes with alternative splicing.

At the probe set summary level, residuals can be combined to produce estimated standard errors of probe set summaries. These can be used to derive weights for individual probe set summaries for downstream analysis. Careful analysis is required to ensure that these weights are beneficial to the downstream analysis.

At the probe set level, residuals can be used to estimate the scale of the residual variance for each probe set or to produce a goodness-of-fit measure for the models fitted to each probe set. These goodness-of-fit measures can be used to derive appropriate weights for combining expression measures for different probe sets.

Our focus in this section is on obtaining an overall chip data quality index, which can be used to distinguish among chips of varying quality. We also suggest a way to visualize the distribution of residuals on a chip to help diagnose the source of departures from quality. Finally, we suggest some chip data quality assessment based on analysis of relative log expression.

To illustrate the methodology we use a set of 19 cel files from the Affymetrix HG-U95A Spike-In Experiment, the 2353 series. The cel files and corresponding chips are identified by the letters A through T (note that the C experiment is missing from this series). Differential concentrations of 14 human transcripts were spiked in a common pool of pancreas mRNA. The behavior of the 14 spike-in probe sets does not play a role in overall chip data quality assessment. For the remainder of the probe sets, the arrays in this experiment constitute a set of technical replicates. The data are available from www.affymetrix.com/analysis/download_center2.a2v.x.

1. Summarizing Residuals from Fits

A simple way to summarize residuals for an entire chip is by means of their empirical distribution. Box plots provide a useful way to compare distributions for a large data set. The top panel of Fig. 8 shows box plots of residuals for each chip. In these, we note a slightly inflated variability in residuals for experiments A and P of the series. Note that the box plots of residuals will be centered close to zero (exactly zero for a least-squares fit), and that their distribution is approximately symmetrical about zero, so the differences between chips could effectively be summarized by the 75th percentile of the chip residual distribution.

Because our biggest concern is the effect of low-quality probe data on expression summaries, it makes sense to combine residuals into estimated standard errors of expression estimates and summarize these at the chip level. To derive the standard errors, we assume that the models were fitted robustly by iteratively re-weighted least squares (IRLS). This fitting procedure can be used to obtain the various $M$ estimators (Holland and Welsch, 1977) as well as the
maximum likelihood fit assuming $t$ error distributions (Lange et al., 1989). IRLS estimates of parameters are obtained as weighted least-squares estimates. The weights are updated at each step by applying a transformation to the residuals from the previous fit. The choice of weight function depends on the particular $M$ estimator desired (Huber, 1972, 1981).

Applying the $M$-estimation techniques to the model specified in Eq. 1 we get

$$\hat{m}_j = \sum_{i=1}^{I} w_{ij} \log_2(B(\text{PM}_{ij}))$$

$$\text{SE}(\hat{m}_j) = \hat{\sigma} / \sqrt{\sum_{i=1}^{I} w_{ij}}$$

For each chip, indexed by $j$, we thus get both an expression value and a standard error estimate. These vectors can be summarized at the chip level to obtain an index of quality for each chip.
The standard errors of estimated expression within a chip form a heterogeneous set by virtue of the fact that the value of $\hat{\sigma}$ varies from probe set to probe set. We can remove this source of heterogeneity by using unscaled standard errors to assess the precision of the estimated expressions. Removing the $\hat{\sigma}$ factor from the standard error does not affect the assessed relative precision of estimated expressions across chips, which is our main interest.

There still remains some heterogeneity in the unscaled standard errors across probe sets, because the effective number of probes used in estimating the expression for chip $j$ may vary from probe set to probe set. To remove this source of heterogeneity, we can normalize the unscaled standard errors by dividing by the average or median standard error across a set of chips.

The bottom panel of Fig. 8 shows box plots of normalized unscaled standard errors (NUSE) of probe set summaries for each chip. In these, we see that the NUSE of probe set summaries are quite sensitive to deviations in assessed expression variability, with experiments A and P clearly standing out from the rest.

We can summarize the batches of NUSEs for each chip by the median, for example, and this value can be used as a chip data quality index. NUSE values fluctuate around 1.0. A median value of 1.05 for a chip may be interpreted as a 5% average loss in precision. The question that naturally arises is what is a good range for this quality index? This is a difficult question and may not have a single answer. The answer depends on the specific application and the various costs involved. For a specific application, one could judge at what level of quality including a chip in an analysis becomes disruptive, by a “leave-one-out” comparison, for example. For a carefully performed analysis, such as one that combines expression measures robustly, the answer may be that including a small number of lower quality chips in an analysis will be harmless in most cases. This does not mean that detecting departures from quality standards that have small effects on downstream analysis is not useful. For example, in a large-scale production environment, having a sensitive tool to monitor quality may help detect and correct problems before they have an impact on expression measures and critical results of downstream analysis.

Affymetrix recommends a number of quality checks to be performed after the analysis of the raw data by the Affymetrix MAS 5.0 software (Affymetrix, 2001). Some are qualitative and involve judging the overall quality of the chip image by visual inspection, whereas others are quantitative. Of the quantitative assessments, some involve examining the expression level of special-purpose probe sets—the hybridization controls, poly(A) controls, and housekeeping genes. Other quantitative assessments are based on a more comprehensive summary of expression and signal level on a chip. Figure 9 examines the relationship between the chip data quality index derived from the residuals (the median NUSE) and three of the quantitative quality assessment measures recommended by Affymetrix: Scaling Factor (target = 500), RawQ, and Percent Present calls.
Other recommended measures that summarize the probe intensities are highly correlated with these and do not provide much additional information.

Figure 9 demonstrates that the median NUSE is highly sensitive to departures from quality standards. In sets of chips varying over a wide range of quality levels, we find that the index of quality based on assessed variability of expression, the median NUSE, is highly correlated with some of the recommended quality assessment measures. We believe that assessed variability provides a better basis for making decisions to rerun an experiment or exclude a chip from an analysis set, whereas other measures are potentially more useful at identifying the source of a problem.

Analyzing expression levels of specialized spike-ins or housekeeping genes for quality assessment purposes poses a special challenge. Because there are only a few spike-in probe sets, measures derived from them tend to be noisy, requiring substantial departures from quality standards for a problem to be detectible. These measures may nonetheless be useful for tracking the source of departures from quality standards that are more easily detectible by other means.

2. Spatial Analysis of Residuals

Residuals can be imaged in a manner similar to the way probe cell intensities are typically imaged. It is common practice to assess chip quality by visually inspecting probe-intensity images. Artifacts like bright or dim spots, scratches, or uneven brightness can be identified this way. Because cell intensities within a chip vary over a wide range and most of this variation comes from the fixed part of the model (Eq. 1), the imaged residuals are expected to provide increased resolution for visually detecting image artifacts.
Spatial patterns of residuals can be profitably examined when seeking an explanation for elevated standard errors of expression estimates on a chip. Spatial patterns may provide evidence of SAPE residue caused by poor wash, uneven hybridization, bubbles, or other local artifacts. A uniform distribution of elevated residuals is another possibility, indicating a different kind of problem with the assay. Note that spatial patterns of residuals may sometimes detect artifacts that are not detectible at the level of gene expression variability. Such artifacts would probably not play a role in accepting or rejecting a chip for analysis but may be valuable in monitoring a chip production process.

Spatial patterns of residuals themselves have proven difficult to visualize. The challenge is to capture spatial patterns of a dense scatter of numbers having both sign and amplitude. Each of these features, sign and amplitude, are readily visualized separately. The weights used in the IRLS fit can be imaged to capture the magnitude of the residuals, highlighting residuals that deviate substantially from an overall estimated scale. The sign of residuals can also be imaged and such images add to the pseudo-images of the weights by telling us whether a region of outlying residuals corresponds to a bright or a dim region on the chip. In addition, the image of the sign of residuals will capture small effects that are not detectible in the weights, which are insensitive to small deviations of the residuals from their expected value of zero.

In Fig. 10 the log intensities (top row), probe weights (middle row), and the residuals (bottom row) are imaged for three chips: two with elevated assessed variability of expression, A and P, and one with average assessed variability of expression, H. Low probe weights, corresponding to residuals with high absolute values, appear as the intense green spots on the chip pseudo-image of the weights (middle row). Clusters of probes with high absolute residuals are clearly visible for chips A and P. The patterns are also discernible in the pseudo-images of log intensities, but not nearly as clearly. Clusters of positive residuals corresponding to bright areas on the chip are clearly visible in the images of the sign of the residuals (bottom row). Determining the source of variability accounting for specific patterns of high absolute residuals—local versus global, and possible trends—is an open question. Software for producing these images is available as part of the affyPLM library from www.bioconductor.org.

3. Quality Assessment Based on Relative Expression

The standard error estimates provide a measure of expression summary variability that is independent of expression level. We can also gauge variability of expression measures by summarizing the distribution of relative log expressions. To compute relative log expression values, we use a virtual median chip constructed by taking, for each probe set, the median log expression from a reference set of chips. We can summarize a vector of relative expression by a measure of bias: median(RE), a measure of variability: IQR(RE), or total error:
The summaries are sensitive to technical sources of variability that are large compared to biological variation. This assessment will be highly correlated with an assessment based on estimated standard errors of probe set expressions, but it has the advantage of being derived from the expression estimates alone (as opposed to probe-level residuals).

Figure 11 shows box plots of relative log expressions for the 2353 series. We can readily see the elevated variability in chips A and P, as was assessed by the residual analysis. In addition, we note a downward bias in the expressions for chip P. As the chips being compared here were hybridized with a common source of RNA, the relative log expression should be zero for all non spike-in probe sets, and the differences in variability between chips can therefore be attributed to technical or processing variability. When comparing chips with different sources of RNA, the variability in relative log expression will be inflated by real biological variability. This is not seen as a serious handicap in the use of relative log expression to assess data quality, because the technical variability that we are interested in is typically greater than the biological variability.
B. QUALITY ASSESSMENT FOR CDNA MICROARRAY EXPERIMENTS

The quality of the expression data derived from cDNA microarray experiments depends on experimental and production factors similar to those affecting oligonucleotide microarrays. The extraction of gene expression information from a scanned array requires a complicated image analysis process. This process is an additional source of potential variability. Yang et al. (2002a) discuss image analysis for spotted arrays in detail. As a by-product of the image analysis step, a number of spot characteristics are generated: spot size and shape, spot intensity, and background intensity. These can be used as quality indicators (Wang et al., 2001). When some clones are spotted at several locations on the array, the repeated measurements for clones can be combined to obtain some assessment of the reproducibility of the measurements, just as probes within probe sets are used to measure reproducibility with the oligonucleotide microarrays. Jenssen et al. (2002) and Tseng et al. (2001) discuss the use of multiply spotted clones in quality assessment. Ritchie et al. (2003) demonstrate that spot-quality measures are correlated to spot reproducibility for the multiply spotted clones and suggest that this relationship could be exploited to derive spot weights to be used in gene-wise regressions.

VII. Detection of Absolute Gene Expression

The problem of classifying genes as present or absent in a given sample has been largely overlooked in the literature. The only widely used detection call for oligonucleotide microarrays is the one implemented in the MAS software developed by Affymetrix (2001). Although the detection of absolute expression is
not generally regarded as important as that of differential expression, it has
definite biological relevance in some circumstances. For example, a biologist
studying gene expression in neural stem cells may want to know which genes
go from being absent to being present at a particular time, and vice versa.

A. THE AFFYMETRIX PRESENCE/ABSENCE ALGORITHM

The Affymetrix MAS 5.0 software makes a detection call for each probe set
by defining a discrimination score

\[ R_i = \frac{PM_i - MM_i}{PM_i + MM_i} \]

where \( PM_i \) is the perfect match intensity of the \( i \)th probe in the probe set, and
\( MM_i \) is the corresponding mismatch intensity. This is done for the nonsaturated
probe pairs. A one-sided Wilcoxon signed-rank test is then used to test

\[ H_0: \text{median}(R_i) = \tau \]
\[ H_1: \text{median}(R_i) > \tau \]

where \( H_0 \) is the null hypothesis and \( H_1 \) is the alternate. \( \tau \) is a small positive
number, tunable by the user, and set to a default of 0.015. Affymetrix has
determined this value as being one that minimizes the number of incorrect calls
without sacrificing sensitivity.

The \( p \) value from the signed-rank test is used as a determinant of gene presence
or absence. MAS 5.0 actually uses two user-configurable significance levels \( \alpha_1 \)
and \( \alpha_2 \), such that \( 0 < \alpha_1 < \alpha_2 < 0.5 \). Probe sets are called present if \( p \leq \alpha_1 \), absent
if \( p \geq \alpha_2 \), and marginal (no call) if \( \alpha_1 \leq p < \alpha_2 \). The defaults in MAS 5.0 are
\( \alpha_1 = 0.04 \), \( \alpha_2 = 0.06 \). These are found to be optimal (based on analyses of spike-in
data) for the default value of \( \tau \). More details about the Affymetrix presence/
absence methodology can be found in Liu et al. (2001) and Liu et al. (2002).

B. ALTERNATIVE METHODS

Zhou and Abagyan (2002) have developed an algorithm to calculate expression
summaries that use only the PM intensities. As a side effect of their procedure,
they perform a detection call. The 5% lowest intensity probe sets are designated as
background, and the empirical cumulative distribution of the background inten-
sities on the linear scale, \( B(I) \), is then calculated. For each probe set, they calculate
the empirical cumulative distribution of the probe signals, \( S_k(I) \), and compare this
to \( B(I) \). The authors’ claim is that genes that are absent will tend to have integral
distributions that are close to the background distribution. They, therefore, compare
each \( S_k \) to \( B \) using a Kolmogorov-Smirnov test. There is no recommendation
for an appropriate threshold on the \( p \) value from the K-S test for calling presence or absence. Instead, the authors state that “those signal sets that can be easily explained by noise are assigned a log10\( p \) value closer to zero.”

Rubinstein and Speed (2003) have approached the problem of transcript detection using several novel methods. They define three broad classes of detection algorithms: thresholding rank sums of probe-pair summaries, thresholding robust averages of probe-pair summaries, and thresholding expression-level estimates. Possible probe-pair summaries include \( \log(\text{PM}_i/\text{MM}_i) \) and \( (\text{PM}_i - \text{MM}_i)/(\text{PM}_i + \text{MM}_i) \). The latter is the summary used by the MAS software. The PM and MM values may or may not be background corrected and normalized across chips. The authors have developed a framework for evaluating different detection algorithms, using the ROC (Receiver Operating Characteristic) Convex Hull method. Under this scheme, the cost of misclassification is defined as follows:

\[
P(p) \cdot (1 - TPR) \cdot C(N, p) + P(n) \cdot FPR \cdot C(P, n),
\]

**Fig. 12.** Receiver Operating Characteristic (ROC) curves for some of the detection algorithms discussed in the text. NRALR is the normalized robust average of log ratios, the NRSLR is the normalized rank sum of log ratios, and the RMA is the expression-level estimate obtained using RMA. The gray polygon is the convex hull of the ROC curves and represents the best possible classifier.
where $P(p)$ is the prior probability of an example being positive and $P(n)$ of being negative, TPR and FPR are the true and false positive rates, and $C(P, n)$ and $C(N, p)$ are the costs of false negatives and false positives, respectively. Requiring that this cost be minimized allows one to evaluate the optimality of detection algorithms over a particular range of false-negative and false-positive costs, given a set of ROC curves for those algorithms. Rubinstein and Speed find that their normalized robust average of log ratios (NRALR) and normalized rank sum of log ratios (NRSLR) outperform the MAS 5.0 algorithm for a wide range of costs, while thresholding on either expression level derived from RMA (Irizarry et al., 2003a) or the MAS 5.0 signal estimate do not perform as well. ROC curves for these five representative algorithms are shown in Figure 12.

References


