Some Considerations for the Design of Microarray Experiments

John H. Maindonald, Yvonne E. Pittelkow and Susan R. Wilson

Abstract

Issues relevant for the design of gene expression experiments using spotted cDNA microarrays and gene chip microarrays are overviewed. Emphasis is placed on the uses of replication, and on the importance of identifying major sources of variation.

Keywords: microarrays; oligonucleotide; design of experiments; variability; replication; gene expression

1 Introduction

Microarrays are new and evolving technologies that enable large numbers of genes, up to the order of tens of thousands, to be evaluated simultaneously. Our aim is to give a brief overview of principles of experimental design, and to comment on their application to microarray experiments. A major theme is that, for purposes of design, the different sources of variation in gene expression are not well understood.

The objective of a microarray experiment might be to investigate genes which are differentially up or down regulated in cells between, say, a control group and cells which have undergone some treatment, or between cells of animals of different genetic background (*e.g.*, control mice compared to knockout mice) or between cells in healthy tissue and diseased tissues, or between cells at different time points (*e.g.*, developmental biology). Many studies search for genes that have similar expression profiles, often in an attempt to determine genes involved in biological pathways, or in development, or genes involved in regulatory functions. The focus would then be on the analysis of dependency structure. Time course experiments may investigate how the pattern of expression or relative expression changes over the cycle of cell division, or following administration of a drug. Finally, interest may be in estimation of gene expression levels.

The primary goal of the experiment should be clear, as this gives focus to the investigation, desirable even if a major part of the analysis will be a general search for interesting patterns of expression. Many experiments have multiple aims; these must be prioritized. Both in its scale and in the processes that are under investigation, the biology has a large element of novelty, with implications for statistical design and analysis. Vingron [52], commenting on the "big science" issues that such large-scale technologies raise, draws attention to "a major upcoming challenge for the bioinformatics community to adopt a more statistical way of thinking and to interact more closely with statisticians." Bioinformaticians need to educate themselves in statistics. "Not so much with the goal of mastering all of statistics but with the goal of sufficiently educating ourselves in order to pull in the statisticians."

Our focus here is on design issues for comparative studies for two types of array platform – two-channel cDNA spotted microarrays [17, 20, 24], and high density oligonucleotide microarray chips produced by Affymetrix [1] for expression analysis, which we refer to as gene chip microarrays. For both types of array, DNA sequences are laid out in a grid on a solid substrate. Occasionally we refer to the spotted microarrays as *slides*, recognising however that glass is just one of several possible substrates, and we refer to Affymetrix oligonucleotide microarrays as *chips*. Much of our discussion of spotted cDNA microarrays applies also to oligonucleotide spotted microarrays (distinct from Affymetrix oligonucleotide arrays, which are produced by photolithography rather than spotting), which we do not explicitly discuss. We note that gene chip microarrays can in principle, with suitable calibration, yield *absolute* expression measures. Each individual spotted microarray slide is by contrast used to yield *relative* expression measures, for example between a treatment and a reference, or between one treatment and another. We note also that, perhaps inevitably for technology that is rapidly changing and developing, there is no single established nomenclature that distinguishes clearly between the different types of arrays. A feature that distinguishes microarray experiments from more conventional experiments described in the biostatistical literature is the very large number of parallel measurements on typically only a few cases. Summary measurements are typically provided for each of a large number of genes or of Expressed Sequence Tags (ESTs), which are partial gene sequences. The small number of cases is, in part, a function of the (initial) high costs of the microarrays, especially chips, limitation of available sample, and the (apparent) failure to involve scientists with statistical training in the early stages of the development of microarrays.

The processing of microarray data raises a variety of statistical, mathematical and computational issues, see for example [12, 19, 45, 47, 49]; some of these are alluded to in passing.

The remainder of the paper is organized as follows: Section 2 gives examples of experiments, Section 3 considers outcome measures, Section 4 notes experimental design principles and discusses their application to microarray experiments, Section 5 considers sources of variation, Section 6 discusses the design of microarray slides and chips, and Section 7 summarizes the discussion.

2 Examples of Experiments

2.1 Spotted Microarrays

In a typical spotted microarray experiment, samples from a treatment and from a reference are combined in equal proportions and hybridized to cDNA probes that have been spotted on a slide. A key question is whether the comparisons that are of interest will be made directly or indirectly. In an indirect comparison, each treatment that is of interest is compared with a reference sample, and the responses of the treatments relative to this reference sample are then compared. In a direct comparison, treatments are directly compared with each other.

For example, Callow *et al.* [6] used the indirect comparison approach to search for genes that were differentially expressed between liver tissue from apolipoprotein apoAI-knockout (test) mice and liver tissue from C57B1/6 (control) mice. Each of 8 test mice was compared with the reference sample, and each of 8 control mice was also compared with the reference sample. For a reference sample, material from the same eight control mice was pooled.

For each of the 16 mice, cDNA, labeled to reflect the source of the mRNA, was prepared by reverse transcription of mRNA. The experiment we describe used Cy5 ("red") and Cy3 ("green") dyes, with Cy5 for individual mice and Cy3 for the reference. The cDNA from each mouse was combined with the cDNA from the reference sample and hybridized to a slide. This experiment resulted in 8 comparisons between control mice and reference, and 8 comparisons between test mice and reference.

Preparation of a spotted microarray slide involves choosing and fixing a large number of spots on a slide, with each spot containing a number of strands of DNA or cDNA that are intended to uniquely hybridize, or bind, to the corresponding gene in the labeled cDNA sample. In this experiment around 6000 spots, one or two per gene, were laid down (spotted) on each of 16 microarray slides (one per "treatment"). After separate labeling, the mixed sample was hybridized to the slide in specially humidified chambers. Laser-induced fluorescence imaging was then used to detect dye intensities. This gave two images of the slide, one for the treatment (test or control) and one for the reference. Image analysis software, together with some post-processing, was then used to derive a background-corrected relative intensity measure for each spot.

Results, for each spot on each of the 16 slides, were expressed as the logarithm of a ratio of the intensity value for each mouse to the intensity value for the pooled reference. Two-sample t-tests, with an adjustment for the large number of comparisons made, were then used to compare the log-ratios from the test mice and the control mice. The study identified eight spots, corresponding to four genes, that were under-expressed in test mice relative to controls.

2.2 Gene chip expression microarrays

In a typical gene chip microarray experiment, prepared cRNA sample is hybridized to the probes on a chip. The chip is then scanned to obtain fluorescence intensity readings of stains incorporated during the laboratory procedures. Image processing software is then used to compute intensity values for each probe.

In contrast to typical spotted microarray experiments, only one sample is hybridized to a chip, allowing, in principle, the estimation of absolute expression values. Because of the high cost of these chips, efficient use is important.

The main characteristics of gene chip microarrays are:

1. Thousands of short oligonucleotide probes (commonly 25-mer, *i.e.*, 25 bases in length) are synthesized *in situ* on a glass substrate, using photolithographic techniques. Multiple paired sets of probes (commonly 11, 16 or 20) are used for each gene or EST. The probe sequences are chosen according to specific criteria described in Lockhart *et al.* [35].

2. One probe in a pair has the exact sequence from the gene or EST, while in the other member of the pair the middle base is changed to its complement. The mismatched probes (*MM*) provide a probe-specific control or nonspecific hybridisation control. The collection of perfect match (*PM*) probes and mismatched probes (*MM*) corresponding to one gene or EST makes up a probe set.

3. User control over the choice and layout of probes requires the construction of custom arrays, whose cost is beyond the resources of many laboratories.

We note that probes are not chosen at random, nor are they independent, although some analyses make this assumption.

In an experiment described by Efron *et al.* [14], the aim was to study transcriptional responses to ionising radiation in the context that some cancer patients have severe life-threatening reactions to radiation treatment. It is important to understand the genetic basis of this sensitivity so that patients with high rates of sensitivity can be identified before being allocated treatment. The design was a factorial experiment with two levels each of two factors, namely (i) RNA was taken from two wild-type human lymphoblastoid cell lines; (ii) the growing state was either irradiated or unirradiated; in addition RNA samples were labeled and divided into two identical aliquots for independent hybridizations. Each microarray provided expression estimates for 6810 genes/ESTs.

Another type of gene chip microarray experiment is described by Golub *et al.* [23]. Their aims were essentially class prediction (assigning tumours to known classes) and class discovery (identifying new cancer classes). They analysed leukemia data of 38 bone marrow samples obtained at time of diagnosis: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML).

3 Issues Concerning Outcome Measures

As noted, spotted microarrays typically yield two intensity measurements for each spot, which are combined into a single ratio or logratio. Gene chip microarrays yield one intensity measurement for each probe. The information from each probe set is generally combined into a single expression index for the probe set. The outcome measure is, in either case, essentially multivariate.

Evidence for the form of the link between expression summary measures and mRNA concentration (or number of molecules) is sparse; however see [8, 25, 28, 32] for gene chip microarrays. When an antibody amplification step is employed, the link is more tenuous, due to nonlinearity in its action. It is important to note that even with replicate slides or chips that use different subsamples from the same sample, and where laboratory procedures have been carried out as similarly as possible, the scanned images can show considerable differences. The normalization or scaling techniques that attempt to make intensity measures comparable between slides or chips are different for the two technologies; see [28] for chips, [55] for slides.

Saturation effects, *i.e.* intensity readings close to or above the upper detection limit of the scanner, are an extreme form of nonlinearity. At high mRNA concentration or high laser power, all intensity measurements may be inaccurate due to saturation. Where one of two estimates being compared is affected by saturation, the estimated difference is attenuated. If both are affected by saturation, the difference will be meaningless [26]. Due to the large number of genes or probes, each with a potentially different saturation level, global avoidance of all such regions may not be feasible, and detection strategies are required.

For both technologies, negative controls (*i.e.* spots or probe sets that should never show a signal) or positive controls (*i.e.* should always show a signal), can be useful checks.

3.1 Spotted Microarrays

Each slide may be used either for a comparison between treatment and reference, or for a comparison between two treatments. In either case, there is one intensity ratio or log-ratio for each spot.

There are typically separate background corrections for the red and the green signals. Both foreground and background signals will differ, depending on the scanner settings and on the image analysis software used [54]. Important considerations are the identification of the spot boundary, the choice of the region used to estimate background and the form of the background adjustment. Negative intensity estimates that can result from background subtraction are a nuisance for later data processing, and should be avoided.

Ramdas *et al.* [41] noted that signal quenching associated with excessive dye concentrations led to nonlinearity in signal intensities. Spot size and morphology can affect intensity measurements. Thus, the routine use of the intensity ratio or logarithm of the intensity ratio as the comparative expression measure is open to question. If, for example, the intensity measurements were changing additively, then differences could be used. On the other hand, if the intensity measurements were changing proportionately then differences in the log values would be used. Currently this is the scale that is widely chosen. If there are three (or more) treatments, then an experiment that has all pairwise comparisons allows us in principle to check that the chosen scale is appropriate. It is prudent to check, to the extent that this is possible, that measurements are in a range where response is linear.

3.2 Gene chip microarrays

In statistical terms, the data from each chip is a single multivariate response vector, with complex dependencies inherent from the biology and the technology. As mentioned earlier, generally a summary measure or estimate of expression is computed from the multiple probes in each probe set, following suitable background estimation and chip normalization (calibration). A number of different summary measures or expression indices are in use. Some are based on differences between the probe intensity (*PM*) and its nonspecific hybridization (*MM*) control; examples include the Affymetrix trimmed average difference (AvDiff, [1]), the model-based expression indices of Li and Wong [33], and the average median filtered differences of Alon *et al.* [2]. Since as many as a third of the *MM* control probes can have intensity readings higher than their paired *PM* probe, truncation, filtering or transformation are often used to accommodate the negative values of $PM - MM$ differences. Some measures do not use the nonspecific hybridization control probes except to calculate a background estimate [28, 34, 39]. Other possibilities include the log of the ratio of the *PM* probe to *MM* probe [1, 32, 39], the robust multi-array average (RMA) approach [28], and empirical Bayes estimation [14]. Other summary measures are also found in the biological literature (*e.g.* [21]).

4 Experimental Design

This section is organized as follows: An introductory subsection discusses aims and principles of experimental design, then bias and replication are discussed in more detail; 4.1 discusses pooling, which is an issue for both types of array; finally 4.2 discusses special issues for spotted microarrays, including the choice between direct and indirect comparison, and dye bias. There are many excellent texts and papers that discuss general principles of experimental design, including [5, 9, 10, 15, 42, 36]. Here we discuss these in the context of microarrays.

Design questions relevant to the aim of the experiment that should be clear before proceeding include:

1. What are the "treatments"?

2. What are the experimental units?

3. What are the experimental measurements?

4. What is measured, and what do the measurements mean?

5. What comparisons are of interest? (Note that interactions are a form of comparison.)

For microarray experiments, "treatments" refer not only to defined procedures, for example treatment by a drug, but also to qualitatively different units, such as tissues from healthy and unhealthy organs, or tissues from wild type model organisms and genetically modified organisms.

For example, in the Callow *et al.* [6] experiment the comparison was between test (knockout) mice and control mice. In the Efron *et al.* [14] experiment, the main interest was in the comparison between irradiated and unirradiated cells, allowing for a possible difference in effect between cell lines, *i.e.*, for a possible interaction between the irradiation effect and cell line.

Cox and Reid [10, p. 4] define an experimental unit as the "smallest subdivision of the experimental material such that any two different experimental units might receive different treatments". The sample may be from a single organism, or it may be a pooled sample of material from several organisms.

In the Callow *et al.* [6] experiment, it is convenient to regard the separate red and green labeled samples that are mixed and hybridized onto a slide as a pair of experimental units, yielding separate intensity information that will (usually), for analysis, be combined into a single log intensity ratio. In Efron *et al.* [14], the experimental units are, strictly, the four separate mRNA samples, each of which is repeated.

A broad over-riding aim of experimental design is to use resources in the manner that will best achieve the intended purpose and produce conclusions that are widely valid (*i.e.*, that are not restricted to too specific a set of conditions). However, this needs to be balanced against the need for simplicity and robustness of design. We begin with a list of broad aims and principles of statistical experimental design, using experiments with spotted microarrays for illustrative purposes, followed by further discussion of some of the issues. Later, we consider special issues for the design of spotted microarray experiments.

Broadly, the aims are to find designs that:

1. *Allow generalization* of results to the relevant wider population;

2. *Avoid bias*, or systematic error;

3. *Minimize the effects of random error*, for a given cost;

4. Allow an *assessment of the accuracy of estimates* of effects that are of interest;

5. *Are robust*, in the sense that they will still give useful results even if there are occasional failures in the experimental protocol, or if some assumptions that motivated the design prove to be false.

Basic devices that are available to achieve these aims are:

1. *Controlling for all "fixed" effects* for which this is possible. For example, the expression of genes in some tissues will be different depending on whether the tissue is from a male or female;

2. *Blocking*, or local control, to allow an accurate assessment of effects under varying experimental conditions. In two-channel spotted microarray experiments, each pair of samples is a block. In general, it is desirable to match the treatment and control samples as closely as possible;

3. *Randomisation* of treatment allocations with respect to factors that cannot be controlled. For example, in a two channel spotted microarray experiment, it is inherently desirable to randomise the allocation of dyes to treatments, in such a way that each treatment occurs equally often with each dye;

4. *Replication* of experimental units, at least to an extent that an estimate of accuracy is possible. In principle, replication may be further increased to achieve a pre-specified accuracy. Additionally, by reducing the opportunity for one unsatisfactory replicate to damage results, replication makes experiments more robust;

5. *The use of repeats*, *e.g.*, repeated spots, within experimental units, where this makes a useful contribution to reducing variability between experimental units. As with replication of experimental units, this has the additional effect that experiments are more robust;

6. Giving first priority in use of experimental resources to *controlling the effects that have the largest implications* for results. For example, once appropriate forms of correction have been applied, the dye effect may, for the present spotted microarray technology, be inconsequential; *i.e.*, any remaining bias from this source may be dwarfed by other sources of variability.

Avoiding Bias

The best way to deal with bias is to modify instrumentation or experimental procedures to avoid it. Where a bias is associated with instrumentation, it may be possible to find an analytical adjustment that verifiably removes or reduces the bias. If neither of these approaches is completely successful, and the necessary information is available, one of devices 1–3 above can be used.

A major difficulty in discussing methods of avoiding bias in microarray experiments is that there is insufficient systematic information available about the biases involved. At present, the exception for spotted microarrays is the bias arising from differences between the dyes used to label the different samples [13]. There is some evidence of day effects, *i.e.* changes in response from one day to another, for both types of microarray. Concerning other sources of bias, until appropriate experiments are performed it might be prudent to make the laboratory situations as uniform as possible during the course of an experiment and to randomise treatment allocation over any potential sources of bias that are not otherwise controlled.

Replication

A discussion of replication and decisions on the optimal level of replication are intimately linked with understanding the sources of error, which we address in a later section. In the context of replication, it is useful to consider a hierarchy of corresponding variation, as in Yang and Speed [56], with the following levels:

1. Separate slides/chips to (separately) obtain measurements on samples from distinct biological sources – biological replicates;
2. Separate slides/chips to probe each of several replicate preparations of RNA from the same biological source (sometimes, and rather misleadingly, also referred to as biological replicates);
3. Technical replicates that use distinct slides/chips to obtain measurements on different target samples of RNA from the same preparation;
4. For spotted microarrays, replicate spots on the slide.

Biological replication is essential when the intention is to make claims about a broader population of patients, plants or animals. Since biological organisms can vary substantially, such replication would be necessary even if the measurement device gave exactly reproducible results when repeated on an individual. Note in this context the broad distinction between technical reproducibility and biological reproducibility. Note also that in the above hierarchy, variation at any lower level contributes to variation at all higher levels.

Since the reasons for replication are not transparent to all, we repeat them here in the microarray context: (i) to allow generalization to the wider biological population (and replication at the biological level is essential for this); (ii) to provide information that will make it possible to do a better experiment next time; (iii) to reduce variation (and increased replication at the biological level will certainly do this, but may be an unnecessarily expensive method if a similar improvement could be achieved by increased replication further down the hierarchy); (iv) to allow identification of major sources of variability, in the hope that something might be done about some of them (and in this context we might want to consider crossed, *i.e.* nonhierarchical, sources of variation); (v) to allow identification of outliers, at levels where that may be important; (vi) to make experiments more robust.

The calculation of the number of replicates required to be able to detect a difference of a given size (power calculations) is challenging in microarray experiments, not only because the newness of the field means that even rough guides to variance estimates for given probe sequences are unknown but also because estimates will change between probe sequences.

Above, we distinguish "technical replicates" from biological replicates. When replication is used to reduce variance (because analysis can be based on the mean or other summary measure) it is important that the replicates be as independent as possible. For example, using different sample preparation hybridized to chips/slides is probably preferable here to using duplicate chips/slide but the same mRNA sample.

At least for spotted microarrays, a further level of replication is possible, namely replicate spots on the same slide, as recommended in Tseng *et al.* [51]. However, the placement of these duplicate spots needs to be carefully considered to avoid potential systematic bias; see Yang and Speed [56]. Removal of one apparently contaminated spot may enable remaining spots to be used in further analysis [51].

For gene chip microarrays, limited available sample material and the relatively high cost of chips often limit the number of biological or technical replicates. While noting that there are no firm standards on the number of replicates required in a microarray chip experiment, Novak *et al.* [40] mention that they commonly design their initial experiments to include three replicates for each biological state, including control. Li and Wong [33] recommend 10 replicates for estimating standard errors used for detecting outliers in gene chip microarray studies. Glynne at at [22] recommend between two and five replicates.

The value of replication in a spotted microarray experiment was shown by Lee *et al.* [31] who, limiting their attention to the red signal, carried out an experiment in which 32 out of 288 genes were expected to be strongly expressed, while the remaining genes should not have been expressed. They used a mixture model to identify genes that were expressed. Although the assumptions required for their analysis can be questioned, their qualitative conclusion holds, in particular that results from individual replicates are unreliable, and of unknown accuracy. With two replicates, there is some indication of the extent of irreproducibility; however, Lee *et al.* recommend doing at least three replicates.

In general, and depending on the tissue, experiments with human tissue are likely to require more extensive replication than experiments with tissue from highly inbred strains of laboratory animals.

Multiple independent estimates of treatment effects

Designs that allow multiple independent estimates of treatment effects may allow reduced replication, or even no replication. For example, for spotted microarrays consider the "all possible pairs" experimental design with three treatments A, B and C. There are two estimates of the contrast between A and B: one that is obtained directly by comparing B with A, and the other that is obtained by subtracting the A versus C effect from the B versus C effect. Thus, if each pairwise comparison is made only once, there is one degree of freedom that can be used for the estimation of "noise"; we prefer this term to the commonly used term "error". If the design has two replicates of each of the three two-way comparisons, there are four degrees of freedom for estimation of noise.

With four or more treatments, there are several alternatives to designs in which all comparisons are with a reference. The design that has each of the six possible comparisons between four treatments has three degrees of freedom for estimation of noise for evaluating each treatment comparison. An alternative is the loop design [30] that compares A with B, B with C, C with D, and D with A. This design has one degree of freedom for estimation of noise. The comparisons that must be made indirectly, between A and C and between B and D, are on average less precise than the comparisons that can be made directly. Where there are many treatments, some comparisons in a loop design will involve many links, with a consequent loss of precision. Modification of loop designs to add comparisons that avoid many connecting links is therefore desirable.

Considerations that will affect the choice between the different designs include: the number of slides that are required; the precision of the comparisons that are of chief interest; the amount of available mRNA, for treatments and where relevant for the reference; the robustness of the design; and the ease of carrying out the analysis.

Factorial designs

Following the structuring of comparisons in terms of main effects and interactions of factors, it may be possible to incorporate into the noise term high order interactions that are not statistically significant, thus increasing the available degrees of freedom for estimating the relevant noise variance. This should be considered at the design stage, although often it is left to the analysis stage.

For example, Efron *et al.* [14] used an initial exploratory analysis to satisfy themselves that the effect of radiation was similar for both levels of cell line, for both aliquots. Hence, they felt able to assume that the three interactions involving irradiation were zero, giving three degrees of freedom for estimating the relevant noise variance. This does, however, ignore the implications for variance structure of the nesting that arises from the way that aliquots were formed in this experiment, namely by splitting samples in two.

For a general discussion of factorial design issues, see Cox [9, pp.94–96] and Cox and Reid [10, pp.99–101].

4.1 Pooling – an issue for both technologies

If there is insufficient RNA from the tissues under investigation from one individual, then it is common practice to prepare RNA from, say, several individuals from a pure (inbred) line, kept as far as possible in a common environment. Other reasons for pooling include provision of adequate quantities of a standard that can be maintained consistently over time, and to "reduce" variation. An alternative to pooling is amplification. Depending on how it is done, however, amplification can bias abundance relationships [4, 29]. At the same time amplification can, for spotted microarrays, lead to results that are more consistent between slides.

A concern is that pooling might increase or modify potential masking effects that may arise from the hybridization of RNA to itself or to other strands of RNA. Self-hybridization is an aspect of secondary structure as described in Zuker [57]. Consistently with comments in Yang and Speed [56], we have been unable to find direct experimental evidence on this point. If masking is not a serious problem and pooling is indeed a form of averaging, then it should be used wherever possible, for treatments as well as for any control. Replication will then require the use of replicate pooled samples, with different individuals used for the different pooled samples. Or is pooling perhaps more problematic for treatment samples than for reference samples, *e.g.*, for knockout or transgenic organisms? There is a clear demand for better knowledge of effects at this level.

For gene chip microarray experiments, Novak *et al.*. [40] suggested that pooling to reduce biological variation is of limited value. On the other hand, Bakay *et al.* [3] concluded that pooling is of value. Such conflicting claims are due, in part, to the different methods used to examine variability, but the issue is clearly unresolved.

4.2 Some special issues for spotted microarrays

The issues that we discuss here are special to spotted microarrays because each slide gives comparative information – either between two treatments, or between a treatment and a reference.

The design used by Callow *et al.* [6], described above, is analogous to the conventional completely randomised design. Note that the use of a common reference sample creates a correlation between the two sets of comparisons with the reference. Additionally, for this experiment one of the comparisons is between the reference and individual mouse samples that are correlated with the reference. An alternative is a design in which each slide gives a direct comparison between a test mouse and a control mouse. Such a direct comparison will, with 8 slides, be more precise than the indirect comparison that used 16 slides, while requiring less mRNA from each control mouse and the same amount of mRNA from each test mouse. Often, though not in the Callow *et al.*. experiment, the comparison with reference will have intrinsic interest. The choice is then between the design that has all pairwise comparisons, and the design that has only the comparisons between treatments and reference.

We have noted that a direct paired comparison of the two treatments should be more precise than the indirect comparison (see also Dudoit *et al.* [13]; Yang and Speed [56]; Kerr and Churchill [30]). Applying such a design to the Callow *et al.* experiment, each slide compares a test mouse with a control mouse. A consequence of the correlations alluded to above is that, as demonstrated in [50], the improvement in precision is not as great as a naive analysis might suggest. Paired comparison designs are a simple type of block design, with each pair of samples (mice) that are compared forming a block. Readers who are familiar with classical experimental design will recognise this as a "paired comparison" experiment, though now with many such comparisons made using a single slide. Fisher [15] discusses such experiments. They are the subject of David's [11] book; see also Cox [9]. These designs have been widely used in food tasting and other sensory evaluation experiments [18]. They are a special case of more general balanced incomplete block designs. For technical details, see Yang and Speed [56] who also discuss and compare many different experimental designs.

The precision of the comparisons that are of interest is not the only consideration. Depending on the experimental context and aim, the experiment in which all comparisons are with a baseline has the following merits: assuming that dye bias affects all comparisons with the reference equally, though perhaps differently for different probe sequences, the swapping of dyes is unnecessary; the comparison between treatments and reference may have an intrinsic interest of its own; limitations in the amount of available mRNA, for one or all of the treatments, may require the use of a design that compares treatments with a reference [56]; use of a reference that is common over different experiments allows treatment effect comparisons across those experiments.

Dye bias

It is now well known that the dye bias varies nonlinearly with the average intensity of the signals [13]. The loess correction, which is one of several corrections that Dudoit *et al.* [13] discuss, seems to work well, but like other such corrections can at best ensure that the bias over all spots is on average reduced to zero. It is in principle possible that the strength of the binding may vary with the sequence of bases to which the dye binds, thus leading to variation between different differentially expressed genes. A cautious approach therefore requires the routine use of dye flips, *i.e.*, each dye occurs equally often with each treatment. This allows an analysis that averages out any bias that remains after the correction.

5 Sources of Variation

The following scheme, adapted from Cox and Reid [10, p. 10], gives a framework for discussion of sources of variation in microarray experiments. Inevitably, it cannot capture the complex ways in which sources of variation may interact:

1. Intrinsic or baseline noise (or "error"), *i.e.*, variation that is inherent in the subjects of the experiment
   (a) Errors associated with the biological, genetic/environmental sources (*e.g.* SNP or different animals or cultures)
   (b) Errors associated with hybridization process (which may be probe dependent);

2. Intermediate noise, *i.e.*, variation associated with the process that leads from treatment to response
   (a) Laboratory (RNA extraction, amplification and labeling)
   (b) Biological sample sources (tissue, homogeneity, contamination);
3. Measurement error, *i.e.*, error associated with the instrumentation
   (a) Chip/slide manufacture (including for spotted microarrays the size and shape of spots)
   (b) Scanning
   (c) Algorithms, including the image processing and scaling procedure used
   (d) Defects arising in the manufacturing process, or in the subsequent handling of slides or chips.

References addressing these sources of variation include [25, 28, 32, 37, 38, 39, 40, 46, 56].

A hierarchy of levels of variation can be envisaged, as detailed in Yang and Speed [56], and might be formalized in a multi-level model, with components of variance attached to each level of the hierarchy. Such models provide a useful framework for thinking about sources of noise, and in addition have a role in the examination of the effects of individual genes. They allow us, *e.g.*, to compare the improvement in precision that arises from the use of multiple spots for the one probe sequence with the improvement from increased technical or biological replication, a point that is demonstrated in the next section. We note that from its beginning, the analysis of variance has been multi-level; see Speed [48]. Many of the models that Fisher [15] analysed had multiple levels of variation.

From a design perspective, we require an estimate of technical variability because we wish to know the contribution that it makes to the variability of biological measurements. Where technical variability is a substantial component, it will be necessary to break it down further, so that we can identify the major sources of noise and take whatever steps are possible to reduce their effect. For a variety of biological and technical measurement reasons, the relative contributions of different noise sources may vary between probe sequences.

Note that:

1. There are several different components of the experimental procedure. If one of these components is, relative to the others, a major component of the variation, attempts should be made to identify it;
2. Comparisons made within individuals, *e.g.*, a cell line from an individual versus a knockout cell line created from the same individual, can be more precise than when the sample and the knockout sample are from different individuals. Experimental procedure becomes more than ever important for controlling the variation that remains;

3. If interest is in getting an accurate estimate of variation, for purposes of generalizing (*e.g.*, to mice generally of a particular strain), then the demand is for repeat results from several individuals, *i.e.*, for genuine biological replication. Then although the standard errors of treatment comparisons can be estimated, it will not be possible to distinguish between variation that arises from experimental procedure and the effects of variation between individuals. The distinction between these two sources of variation may be useful in deciding whether effort on the improvement of laboratory procedure is justified.

6 The Design of Microarray Chips and Slides

There are two aspects of microarray experiment design – the design of the array/chip, and the allocation of the mRNA samples to the array/chip. Because the fabrication of a custom gene chip is expensive, most users accept one of a set of standard gene chip microarray designs. By contrast, users of spotted microarrays do often design their own slides. They then face important issues that include the choice of genes (or ESTs), the number of repeats of each probe sequence, and the relative positioning of repeats. In addition, each gene may be represented by more than one probe sequence. A major advantage of fabricated oligonucleotide sequences, for spotted arrays as well as for chips, is in the opportunities that they offer for selecting and testing probe sequences. This is an important ongoing research area, which is however beyond the scope of this paper; we refer the reader to Rouillard *et al.* [44].

The remaining discussion will comment on the number and possible prioritization of genes represented on the slide or chip, and the use of repeats. Our comments have direct relevance to cDNA microarray slides, where there is ordinarily one probe sequence, perhaps repeated, for each gene or EST, but the principles are general.

Many probes, or few probes

It is tempting to include as many probes for genes as possible on a slide. However, as the number of different genes represented on the slide increases, so also does the potential for false positives when, say, analysing a comparative experiment. To avoid this situation, the criteria for establishing differential expression becomes more stringent for statistical tests as the number of tests are increased. For example a t critical value that equals 2.1 for a single t-test (for a single gene) may, depending on the adjustment used and on the choice of reference distribution, increase to 4.5 when there are 5000 such tests.

An attractive design option can be to divide probes into two groups – a smaller "likely" group, and a much larger "possible" group. Statistical comparisons can then be done separately for the two groups, with a much less stringent criterion for establishing differential expression used for probes in the smaller group. The highest priority for the use of repeated spots will be given to the smaller group of genes chosen for careful scrutiny. Such a classification of genes into two groups builds in prior knowledge, with implications for the subsequent statistical inference.

Repeated spots

What is the effect on precision from repeating probes multiple times on a single slide, by comparison with repeating slides?

Writing m_b for the between array mean square, and m_w for the within array mean square, and with k spots per probe sequence, and assuming a simple form of multi-level model where the between spots (within array) component of variance is σ^2, while the between array component of variance is σ_b^2, it follows that:

$$\begin{aligned} \mathrm{E}[m_b] &= k\sigma_b^2 + \sigma^2 \\ \mathrm{E}[m_w] &= \sigma^2. \end{aligned}$$

Thus $\mathrm{E}[m_b]/\mathrm{E}[m_w]$ equals 1 if $\sigma_b^2 = 0$, and is otherwise greater than one.

The variance of the mean $\bar{x}$ over all k spots on each of n slides is

$$\mathrm{var}[\bar{x}] = \frac{\sigma_b^2}{n} + \frac{\sigma^2}{kn}.$$

If $\sigma_b^2 = 0$, then $\mathrm{var}[\bar{x}] = \frac{\sigma^2}{kn}$, and the repeating of spots is just as effective, for increasing precision, as the repeating of slides.

The Callow *et al.* [6] data are interesting in this connection. Out of 5544 non-blank spots, 175 were duplicates of the same probe sequence, while 6 were triplicates. For each of these probe sequences, we can thus use an analysis of variance calculation to determine both a within array (between spot) mean square, and a between array mean square.

Individual sample ratios are too inaccurate and variable, ranging from 0.11 to 11.2, to give useful indications for experimental design. We can however use a quantile-quantile plot (Figure 1) to study the pattern of change of the ratio over many different genes, and assess the extent to which these ratios behave like independent ratios from an F-distribution with 14 and 16 d.f.

The smallest 156 values are consistent with the assumption that the ratios follow the theoretical F-distribution corresponding to $\sigma_b^2 = 0$, independently between probe sequences. Included among these 156 probe sequences are the only two out of the 181 that were identified as differentially expressed.

Thus for these duplicated or replicated data, for the majority of probe sequences, increasing the number of spots on a slide gives the same improvement in precision as increasing the number of slides by the same factor. There is no way to know whether the same would be true for the probe sequences that were not repeated. Data from a less homogeneous biological population, *e.g.*, tissues from distinct human sources, are inherently likely to show stronger evidence of biological variation. In some types of

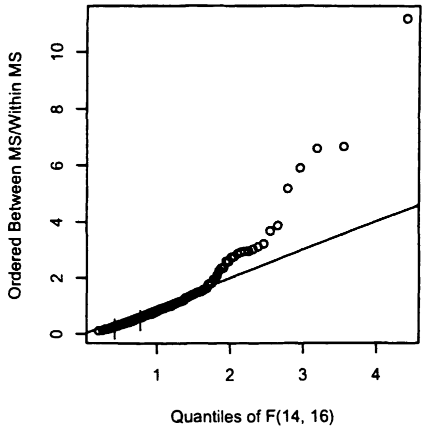

Figure 1: Quantile-quantile plot that compares ordered ratios of between to within slide mean squares, for 181 probe sequences that appear more than once, to quantiles of the F-distribution with 14 and 16 d.f. The line $y = x$ is superimposed on the plot. The two points that correspond to genes identified as differentially expressed are marked with a vertical bar (|).

study, for some probe sequences, increasing the number of spots per gene may be a highly effective way to improve precision.

Note that in more traditional applications of multi-level models, the relevant variances are rarely known with sufficient accuracy that they give a secure basis for use in setting priorities in the future use of experimental resources. For microarray experiments, the combining of information across large numbers of probe sequences can provide such a secure basis. This is an area that requires further investigation.

Published information on mean square ratios such as just given, for a range of different experimental conditions and probe sequence sets, would greatly assist the design of future experiments.

Some special issues for gene chip microarrays

Important chip design issues that require further investigation include the following:

1. There is some evidence that the use of mismatch probes in expression indices reduces precision; see [28, 39]. Further research is required on the optimal assessment of nonspecific hybridization and background.

2. Some probes appear consistently unresponsive, arguing for their removal or replacement.

3. The inclusion of control probe sets can assist quality control and calibration.

In designing experiments, consideration should be given to the inclusion of "spikes" of known concentration in the sample, to allow for more accurate normalization between chips.

7 Discussion

While statistical methodology is now seen as an important part of microarray experiments, its penetration into this area remains, in many respects, superficial. This is especially true for experimental design. Effort at the design phase of a microarray experiment will often save considerable effort and frustration at the analysis stage; see Yang and Speed [56] for further discussion. Good experimentation can be seen as a sequential learning process in that what has been learned from one experiment can contribute to the design of the next experiment.

This paper outlines many of the issues that require consideration when designing a microarray experiment. There has been emphasis on replication and sources of error because of their pivotal role in analysis and subsequently inference. For example, in a comparative experiment researchers should consider that an observed difference is 'real' only if it is greater than what could be expected by chance. The estimate of the size of that difference is a function of all the noise that has contributed to the difference, and is obtained from replicates. Too often, the need for replication has been overlooked in microarray experiments. Yet recall Fisher's [16] comment over seventy years ago concerning plant experimentation:

> No one would now dream of testing the response to a treatment by comparing two plots, one treated and the other untreated.

It is unusual when measuring with, say, a tape measure, to make replicate measurements on the same object. The accuracy of the instrument is commonly high relative to the variability of the object that is measured. Hopefully, technological improvements will lead to arrays with correspondingly high levels of technical reproducibility. In the meantime, there are large potential gains that may come from a better understanding both of the technology and of quantitative aspects of gene expression. Experiments that will assist in an understanding of the technical characteristics of this methodology and the sources of variation and bias should be a priority.

Combining information from the different platforms and laboratories also is important (see, for example, Glynne *et al.* [22]). As yet, we are not aware of studies that directly investigate the extent to which results from a microarray experiment can be reproduced by other workers in other laboratories. If, however, results from some microarray studies point in one direction and some in another, it may be necessary to undertake a statistical overview analysis, or meta-analysis, such as is done in clinical medicine (see for example, Chalmers and Altman [7]). In a related context, Ionnidis *et al.* [27] examined the extent to which genetic association studies stand up when repeated by other researchers, and found that results from the first study often suggest a stronger effect than is found in later studies, and show poor correlation with subsequent research on the same association. This observation may be in part a manifestation of the so-called "file drawer problem" [43], that positive results are more likely to be published than negative results. Epistatic effects such as are discussed in Wilson [53] provide another likely explanation.

The challenges that arise from the massively parallel measurement of gene expression are new. At the analysis stage, what choice of designs will ease the task of interpreting and summarizing the potentially huge number of individual results? This is clearly an area for further research. Meanwhile, we recommend the use of designs that are both reasonably robust against unexpected behavior, and that are also capable of revealing effects that have not been anticipated.

Acknowledgements

We thank Dr. Aude Fahrer, Dr. Matthew Wakefield and the anonymous referee for helpful comments, which led to several improvements.

John H. Maindonald, Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, `john.maindonald@anu.edu.au`

Yvonne Pittelkow, Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, `yvonne.pittelkow@anu.edu.au`

Susan Wilson, Centre for Mathematics and its Applications, Mathematical Sciences Institute and Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, `sue.wilson@anu.edu.au`

References

[1] Affymetrix. *Affymetrix Microarray Suite User Guide, Version 4 edition.* Affymetrix, Santa Clara, CA.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, D. Mack S. Ybarra, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by microarray chips. *Proceedings of the National Academy of Sciences, USA*, 96:6745–6750, 1999.

[3] M. Bakay, Y.-W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. P. Hoffman. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics*, 3:4, 2002.

[4] L. R. Baugh, A. A. Hill, E. L. Brown, and C. P. Hunter. Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Research*, 29(5):e29, 2001.

[5] G. Box, W. Hunter, and S. Hunter. *Statistics for Experimenters*. Wiley, New York, 1978.

[6] M. J. Callow, S. Dudoit, E. L. Gong, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 10:2022–2029, 2000.

[7] I. Chalmers and D. G. Altman. *Systematic Reviews*. BMJ Publishing Group, London, 1995.

[8] E. Chudin, R. Walker, A. Kosaka, S. X. Wu, D. Rabert, T. K. Chang, and D. E. Kreder. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip® arrays. *Genome Biology*, 3(1):research0005, 2001.

[9] D. R. Cox. *Planning of Experiments*. Wiley, New York, 1958.

[10] D. R. Cox and N. Reid. *Theory of the Design of Experiments*. Chapman and Hall, London, 2000.

[11] H. A. David. *The Method of Paired Comparisons*. Oxford University Press, New York, 1988.

[12] S. Dudoit, Y. H. Yang, and B. Bolstad. Using R for the analysis of DNA microarray data. *R News*, 2:24–32, 2002. `http://cran.R-project.org/doc/Rnews`.

[13] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–140, 2002.

[14] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[15] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935; 7th edition 1960.

[16] R. A. Fisher and J. Wishart. The arrangement of field experiments and the statistical reduction of the results. *Imperial Bureau of Soil Science (London). Technical Communication*, 10:1–23, 1930.

[17] S. H. Friend and R. B. Stoughton. The magic of microarrays. *Scientific American*, 286:34–39, 2002.

[18] M. D. Gacula and J. Singh. *Statistical Methods in Food and Consumer Research*. Academic Press, Orlando, FL, 1984.

[19] R. Gentleman and V. Carey. Bioconductor. Open source bioinformatics using R. *R News*, 2:11–17, 2002. `http://cran.R-project.org/doc/Rnews`.

[20] G. Gibson and S. V. Muse. *A Primer of Genome Science*. Sinauer Associates, Madison, WI, 2001.

[21] R. J. Glynne, S. Akkaraju, J. I. Healy, J. Rayner, C. C. Goodnow, and D. H. Mack. How self-tolerance and the immuno-suppressive drug FK506 prevent B-cell mitogenesis. *Nature*, 403:672–676, 2000.

[22] R. J. Glynne, G. Ghandour, and C. C. Goodnow. Genomic-scale expression analysis of lymphocyte growth, tolerance and malignancy. *Current Opinion in Immunology*, 12:210–214, 2000.

[23] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[24] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cDNA microarray analysis. *Biotechniques*, 29:548–562, 2000.

[25] A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim. Evaluation of normalization procedures for oligonucleotide microarray data based on spiked cRNA controls. *Genome Biology*, 2(12):research0055, 2001.

[26] L.-L. Hsaio, R. V. Jensen, T. Yoshida, K. E. Clark, J. E. Blumenstock, and S. R. Gullans. Short technical report: Correcting for signal saturation errors in the analysis of microarray data. *Biotechniques*, 32:330–336, 2002.

[27] J. P. A. Ioannidis, E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nature Genetics*, 29:306–309, 2001.

[28] R. A. Irizzary, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2002. In press.

[29] N. N. Iscove, M. Barbara, M. Gu, M. Gibson, C. Modi, and N. Winegarden. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnology*, 20:940–943, 2002.

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

[31] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences, USA*, 97:9834–9839, 2000.

[32] W. T. Lemon, J. T. Palatini, R. Krahe, and F. A. Wright. Theoretical and experimental comparisons of gene expression estimators for oligonucleotide arrays. *Bioinformatics*, 18:1470–1476, 2002.

[33] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences, USA*, 98:31–36, 2001.

[34] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology*, 2:1–11, 2001.

[35] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[36] J. H. Maindonald. Statistical design, analysis and presentation issues. *New Zealand Journal of Agricultural Research*, 35:121–141, 1992.

[37] J. Mar and S. Grimmond. A review of image analysis software for spotted microarrays. Technical Report, 2002.

[38] J. C. Mills and J. I. Gordon. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Research*, 29(15):e72–2, 2001.

[39] F. Naef, D. A. Lim, N. Patil, and M. O. Magnasco. From features to expression: High density oligonucleotide microarray analysis revisited. LANL e-print physics/0102010. To appear in the Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data, 2001.

[40] J. P. Novak, R. Sladek, and T. J. Hudson. Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics*, 79:104–113, 2002.

[41] L. Ramdas, K. R. Coombes, K. Baggerly, L. Abruzzo, W. E. Highsmith, T. Krogmann, S. R. Hamilton, and W. Zhang. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*, 2(11):research0047, 2001.

[42] G. K. Robinson. *Practical Strategies for Experimenting.* Wiley, New York, 2000.

[43] R. Rosenthal. The “file drawer problem” and tolerance for null results. *Psychological Bulletin,* 86:638–641, 1979.

[44] J.-M. Rouillard, C. J. Herbert, and M. Zuker. Oligarray: genome-scale oligonucleotide design for microarrays. *Bioinformatics,* 18:486–487, 2001.

[45] G. Sawitzki. Quality control and early diagnostics for cDNA microarrays. *R News,* 2:6–9, 2002. `http://cran.R-project.org/doc/Rnews.`

[46] E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry,* 80:192–202, 2000.

[47] G. K. Smyth, Y. H. Yang, and T. P. Speed. Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols.* Humana Press, 2002.

[48] T. P. Speed. What is an analysis of variance? *Annals of Statistics,* 15:885–910, 1987.

[49] T. P. Speed. *Statistical Analysis of Gene Expression Microarray Data.* CRC Press, 2002. In press.

[50] T. P. Speed and Y. H. Yang. Direct versus indirect designs for cDNA microarray experiments. Technical Report # 616, 2002.

[51] G. C. Tseng, M.-K. Oh, L. Rohlin, J.-C. Liao, and W.-H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research,* 29:2549–2557, 2001.

[52] M. Vingron. Editorial. Bioinformatics needs to adopt statistical thinking. *Bioinformatics,* 17:389–390, 2001.

[53] S. R. Wilson. Epistasis. In *Encyclopedia of the Human Genome.* Macmillan, 2003. In press.

[54] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Statistical Graphics,* 11:108–136, 2002.

[55] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* 30(4):e15, 2002.

[56] Y. H. Yang and T. P. Speed. Design issues for cDNA microarray expression experiments. *Nature Reviews,* 3:579–588, 2002.

[57] M. Zuker. Calculating nucleic acid secondary structure. *Current opinion in structural biology*, 10:303–310, 2000.