

## DESIGN ISSUES FOR cDNA MICROARRAY EXPERIMENTS

*Yee Hwa Yang<sup>\*‡</sup> and Terry Speed<sup>\*§</sup>*

Microarray experiments are used to quantify and compare gene expression on a large scale. As with all large-scale experiments, they can be costly in terms of equipment, consumables and time. Therefore, careful design is particularly important if the resulting experiment is to be maximally informative, given the effort and the resources. What then are the issues that need to be addressed when planning microarray experiments? Which features of an experiment have the most impact on the accuracy and precision of the resulting measurements? How do we balance the different components of experimental design to reach a decision? For example, should we replicate, and if so, how?

The ever-increasing rate at which genomes are being sequenced is attracting attention to functional genomics — an area of genome research that is concerned with assigning biological function to DNA sequences. More precisely, the completion of the sequencing of any given genome immediately raises the essential and formidable task of defining the role of each gene, and of understanding the interactions between sets of genes in that genome. These tasks can be carried out in various ways, including protein prediction, homology searching and expression analysis. We are interested in DNA microarrays, which are part of a new class of technology that allows simultaneous monitoring of the expression levels of numerous genes. This technology is being more and more widely applied in biological and medical research to address a wide range of questions (see, for example, REFS 1–5). Microarray experiments generate large and complex multivariate data sets, and some of the greatest challenges lie not in generating these data but in the development of computational and statistics tools to analyse the large amounts of data. However, an important ancillary task is to design the experiments so that the efficiency and reliability of the obtained data can be improved.

High-density oligonucleotide microarray experiments provide direct information about the expression levels in a mRNA sample of the 200,000–500,000 probed gene fragments<sup>6</sup>. By contrast, cDNA microarray experiments typically involve hybridizing two mRNA

samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000–20,000 cDNA probes. Data from such experiments provide information on the relative expression of the sample genes, which correspond to the probes (BOX 1). Our discussion is mainly relevant to these two-colour experiments, in which the main design issue is which samples should be co-hybridized. However, several of the points we make (in particular, see later sections on Multifactorial designs, Variability and replication, and Power and sample-size determination) are also relevant to single-label experiments.

What are the dangers of not paying adequate attention to design issues? At one extreme, an experiment that is not carefully designed might be entirely satisfactory, but possibly less efficient in its use of the available material than it could have been otherwise, thereby sacrificing a potential gain in efficiency. At the other extreme, a badly designed experiment might leave an experimenter unable to answer a question of interest with the data that has been collected, or perhaps leave a potential bias in the data that might compromise the interpretation of the results. In most cases, neither of these extremes will apply, but careful attention to experimental design will ensure that good use is made of the available resources, obvious biases will be avoided and it will be possible to answer the primary questions of interest<sup>7</sup>.

*\*Department of Statistics and Program in Biostatistics, 367 Evans Hall, 3860, University of California, Berkeley, California 94720-3860, USA.*

*‡Department of Epidemiology and Biostatistics, Box # 0560, University of California, San Francisco, California 94143-0560, USA.*

*§Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Post Office, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia.*

*Correspondence to T.S.*

*e-mails: terry@stat.berkeley.edu; terry@wehi.edu.au*

*doi:10.1038/nrg863*

Box 1 | **What are cDNA microarray experiments?**

cDNA microarrays consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide by a robotic arrayer. The relative abundance of the spotted DNA sequences in two DNA or RNA samples can be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples or targets are reverse transcribed into cDNA, labelled using different fluorescent dyes (usually a red-fluorescent dye, Cyanine 5 (Cy5), and a green-fluorescent dye, Cyanine 3 (Cy3)), then mixed in equal proportions and hybridized with the arrayed DNA sequences or probes (we follow the definition of probe and target adopted in REF. 29). After this competitive hybridization, the slides are imaged using a scanner, and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. See REF. 29 for a more detailed introduction to the biology and technology of cDNA microarrays and oligonucleotide chips.

**COMPETITIVE HYBRIDIZATION**  
A mixture of differently labelled target cDNA fragments that are hybridized together in the presence of a common probe or collection of probes.

**LOG RATIO**  
The logarithm, usually to the base 2, of the ratio of the measured signal intensities in the two channels of a two-colour microarray experiment. If we denote these two signals by R (red channel) and G (green channel), then their log ratio is  $\log_2(R/G)$ .

This review describes the experimental design and related issues that are important for carrying out cDNA microarray experiments. In addition, we hope to facilitate discussion and understanding between biologists who do the experiments and statisticians or others who do the analyses. We first describe the objectives of experimental design in the context of microarray experiments. The next section introduces the reader to a display that summarizes the hybridizations that are carried out in an experiment. Furthermore, we discuss how scientific aims affect the choice of design, and how practical issues constrain our design options. Finally, we use three case studies to illustrate the ways in which scientific and physical constraints can be used to choose a design.

**Why experimental design?**

The objective of experimental design is to make the analysis of the data and the interpretation of the results as simple and as powerful as possible, given the purpose of the experiment and the constraints of the experimental material. As described in BOX 1, the underlying idea of

a cDNA microarray experiment is a competitive hybridization between a sample that is labelled with the red-fluorescent dye Cyanine 5 (Cy5) and a sample that is labelled with the green-fluorescent dye Cyanine 3 (Cy3). Unlike gene-expression data from nylon membranes (filter) or GeneChip (Affymetrix), cDNA microarray data are inherently comparative. This is because the filter or Affymetrix data measure gene-expression levels for each sample separately, whereas, in the case of cDNA experiments, the pairing of target samples for hybridization leads to relative expression values and constrains the types of design that can be considered. So, each cDNA microarray experiment gives us the relative abundance of two sets of mRNA.

The principles of design for comparative experiments of this kind are not new. They first arose in agricultural research many years ago with Ronald A. Fisher<sup>8</sup>, who studied yields from different plant varieties (see also REFS 9,10). The varieties to be compared were grown on the same land, as the variation between plots was substantial. This planting arrangement is conceptually equivalent to using **COMPETITIVE HYBRIDIZATION** to compensate for the variation between glass-slide microarrays. This similarity means that the design and analysis of comparative experiments can be accommodated in a classical statistical framework, an important point to which we return in later sections. In cDNA microarray experiments, we see more variation between slides than within slides (for further discussion, see the section on Variability and replication), and so the most important design issue is to determine which mRNAs are to be labelled with which fluor, and which are to be hybridized together on the same slide. In addition, there can be constraints on the number of slides, the amount of RNA available, or other cost considerations, all of which will affect the experimental design.

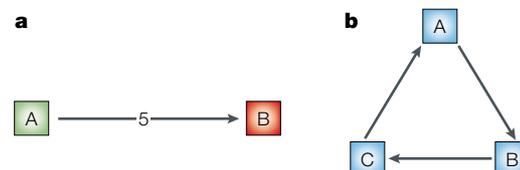
**Graphical representation of designs**

Before discussing design choice, we introduce a graphical representation of microarray experimental designs (BOX 2). In the rest of this review, we use this way of illustrating microarray designs. The structure of the graph determines which gene-expression differences can be estimated and the precision of these estimates. For example, gene-expression levels in two target samples can be compared only if there is a 'path' (that is, sequence of hybridizations) that joins the corresponding two vertices (mRNA samples). The precision of the estimates of relative expression, then, depends on the number of paths that join the two vertices, and is inversely related to the length of these paths.

The sample experiment depicted in BOX 2, panel b, consists of three sets of hybridizations (A, B and C). There are two paths that join the vertices A and B: a path of length 1 joins A and B directly and a path of length 2 joins A and B through C. When we are estimating the relative abundance of mRNA between target samples A and B, the estimate of **LOG RATIO**  $\log(A/B)$  from the direct path A to B — that is, the experiment in which A and B are co-hybridized — is likely to be more precise than the indirect estimate of  $\log(A/B) = \log(A/C) - \log(C/B)$

Box 2 | **Graphical representation of experimental designs**

One way to represent microarray experiments is to use 'multi-digraphs' — directed graphs (that is, one that contains vertices or nodes, and edges; see figure), possibly with multiple edges. In this representation, vertices or nodes (for example, A and B) correspond to target mRNA samples, and edges or arrows correspond to hybridizations between two mRNA samples. By convention, we place the green-labelled sample at the tail and the red-labelled sample at the head of the arrow. For example, panel a shows an experiment that consists of replicate hybridizations. Each slide involves labelling sample A with green (Cy3) dye and sample B with red (Cy5) dye and hybridizing the samples together on the same slides. Number 5 on the arrow indicates that there are five replicate hybridizations in this comparison. Panel b depicts the simplest loop design: three samples — A, B and C — are hybridized together in consecutive pairs, each labelled once in red and once in green. Graphical representations of this nature have been used previously in experimental design in the context of paired comparisons in precision measurement<sup>30</sup>.



## Box 3 | Issues that affect the design of array experiments

## Scientific

- Aim of the experiment.
- Specific questions to be answered and how they are prioritized.
- How will the experiments answer the posed questions?

## Practical (logistic)

- Types of mRNA samples: reference, control, treatment 1 (T1), and so on.
- Amount of material available: count the amount of mRNA involved in one CHANNEL of one hybridization as one unit.
- Number of slides available for the experiment.

## Other factors

- The experimental process before hybridization: sample isolation, mRNA extraction, amplification and labelling.
- Controls planned: positive, negative, ratio, and so on.
- Verification method: northern blot, reverse transcriptase (RT)-PCR, *in situ* hybridization, and so on.

from the path of length 2 that joins A and B through C (that is, from the two experiments in which A is co-hybridized with C, and B is co-hybridized with C, separately).

## Scientific aims and design choice

BOX 3 contains a list of general issues that need to be addressed when designing a cDNA microarray experiment. We hope this list helps to translate biological questions into appropriate statistical questions. Most importantly, the primary focus of the experiments needs to be stated, whether it is to identify differentially expressed genes, to search for specific gene-expression patterns or to identify tumour subclasses. It is important to bear in mind that results from previous experiments or other information might lead us to expect only a few or many genes to be differentially expressed or to have specific expression patterns.

An independent verification method should also be considered, such as northern or western blot analyses, reverse transcriptase (RT)-PCR or *in situ* hybridization, as a follow up to the experimental results. The amount of verification that is carried out can influence the choice of statistical methods and the sample size (see below for further discussion). The source of mRNA (for example, tissue samples or cell lines) will affect the amount of mRNA available and, in turn, the number of replicate slides possible. Details of sample isolation, mRNA extraction and labelling also affect the number of times the experiment has to be repeated and how this needs to be done. Controls can be used simply for checking that the experiment went well, or they might be useful in data analysis — for example, in normalizing the experimental results.

**One design choice.** We begin by looking at experimental design with an example of an experiment in which there is just one obvious design choice. In this case, one design stands out as preferable to all others, given the nature of the experiment and the material available. Let us

imagine that we wish to study mRNA from populations of cells, each of which has been treated with a different drug, and that the main goal is to compare the treated with the untreated cells. In this case, the appropriate design would dictate that the untreated cells become a *de facto* reference, and that all hybridizations involve one treated set of cells and the untreated cells.

To take a different example, suppose that we have collected numerous tumour samples from patients. If the scientific focus of the experiment is to discover tumour subtypes<sup>11</sup>, microarray experiments in which tumour samples are compared with a common reference RNA are an obvious choice. Here, the design choice is dictated by the aims of the study, although considerations of statistical efficiency also affect the choice of design. BOX 4 provides two illustrations in which the aim of the experiment sometimes leads directly to the best design.

In most experiments, however, several designs that seem equally suitable can be devised, and we need some principles for choosing one from several possibilities. Such experiments are discussed below, where we focus on the question of identifying differentially expressed genes. We begin by explaining some design principles, which we subsequently illustrate with examples.

## Direct versus indirect comparisons

The key issue in designing a cDNA microarray for which more than one design is appropriate is to decide whether to use direct or indirect comparisons; that is, whether to make the comparison within or between slides (FIG. 1). We begin by discussing this comparison in the simplest case — treatment T versus control C. For the purposes of subsequent discussion, the terms ‘treatment’ and ‘control’ are used broadly, to include comparisons between treated (for example, with a drug) and untreated cells, between wild-type and mutant samples (including samples from knockout or transgenic animals) or between two tissues (for example, tumour versus normal).

## Box 4 | Design choices from scientific aims

The two designs represented below are best answered by a common reference design.

**Case 1:** Use of meaningful biological control (Ctl).

**Samples:** Liver tissue from mice treated with cholesterol-modifying drugs and from untreated (Ctl) mice.

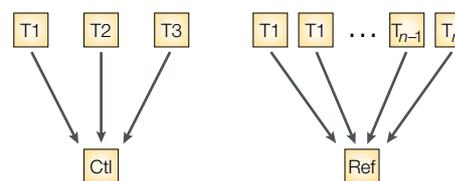
**Question 1:** The expression of which genes differs between the treated and untreated (Ctl) mice?

**Question 2:** Which genes respond similarly to two or more treatments, when compared to wild-type?

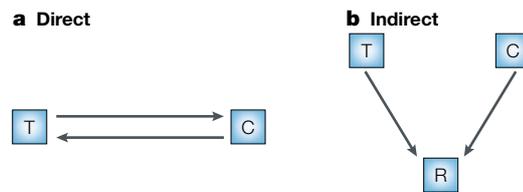
**Case 2:** Use of universal reference (Ref).

**Samples:** Tissue from different tumours.

**Question:** What are the tumour subtypes?



CHANNEL  
cDNA microarrays have paired hybridization intensity measurements that are taken from two wavelength bands after laser excitation at two wavelengths. These two sources of data are known as channels. By contrast, measurements of radiolabelled hybridization products are single channel, as are the Affymetrix microarrays.



**Figure 1 | Direct versus indirect designs.** Two possible designs that compare gene expression in two cell-population samples T and C. **a** | In a direct comparison, the differential expression of the genes in samples T and C is measured directly on the same slide (in a single experiment). **b** | In an indirect comparison, expression levels of samples T and C are measured separately on two different slides. The log ratio  $\log_2(T/C)$  is estimated by the difference  $\log_2(T/R) - \log_2(C/R)$ . R, reference.

Suppose that we plan to do two hybridizations, and that the quantity of mRNA that is available for the experiment is not a limiting factor. To carry out a direct comparison, we might label T with Cy5 and C with Cy3 and hybridize them together (T–C) on both slides, as shown in FIG. 1a. For any particular gene, we would then obtain two independent estimates of the log ratio  $\log(T/C)$  for that gene. If the VARIANCE for one such measurement is  $\sigma^2$ , then the variance of the average of two independent measurements is  $\sigma^2/2$ . Conversely, if we make use of a common reference, for example R, then our two hybridizations would be T–R and C–R, as shown in FIG. 1b. In this case, the log ratio,  $\log(T/C)$ , for any gene is the difference of two independent log ratios from the equation  $\log(T/C) = \log(T/R) - \log(C/R)$ . As before, if the variance for a single log ratio is  $\sigma^2$ , it follows that the variance of the difference of two independent log ratios is  $2\sigma^2$ . To summarize, with two hybridizations, we obtain a measure of the log-ratio of a gene with variance  $\sigma^2/2$  by doing two direct comparisons, and the log ratio of a gene with variance  $2\sigma^2$  by doing two indirect comparisons. The fact that these variances differ by a factor of 4 is the key difference between direct and indirect comparisons. In practice, we do not always observe a factor of 4 difference because the independence assumption is often not satisfied if target mRNAs are from the same biological extraction, a point that will be discussed in more detail later.



**Figure 2 | Dye-swap replications.** Dye-swap experiments involve two hybridizations for two mRNA samples, in which dye assignment is reversed in the second hybridization. **a, b** | Two sets of two replicates; dye-swap replication is shown in **b**.

**VARIANCE**  
The most common statistical measure of variability of a random quantity or random sample about its mean. Its scale is the square of the scale of the random quantity or sample. The square root of the variance is known as the standard deviation.

Most of the remainder of our discussion focuses on which RNAs to assign to fluorors and which to hybridize together, and follows from the simple observation above on the relative merits of direct and indirect comparisons. For simplicity, we assume that all target mRNAs are independent biological replicates.

**Dye-swap experiments.** Dye-swap replications (FIG. 2), in which each hybridization is done twice, with the dye assignments reversed in the second hybridization, are useful for reducing systematic bias<sup>12–14</sup>. Most cDNA microarray experiments show systematic differences in the red and green intensities, which require correction at the normalization step. It is very unlikely that this normalization can be done equally well for every spot on every slide, leaving no residual colour bias. Whether averaging over dye-swap pairs will leave an experiment more or less prone to these biases will depend on the extent of this colour bias, and its repeatability across slides. For this reason, we recommend the use of dye-swap pairs wherever possible. Alternatively, random dye assignments to samples can be used, effectively to include the bias in random error. Importantly, direct comparisons of replicates of slides with the same labelling should be avoided because unadjusted colour bias might persist and accumulate. With indirect comparisons, repeatable residual colour bias should be removed when between-slide differences are taken; as shown in the following equation [ $\log(A/R) + \text{residual colour bias}$ ] – [ $\log(B/R) + \text{residual colour bias}$ ] =  $\log(A/R) - \log(B/R)$ .

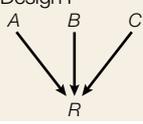
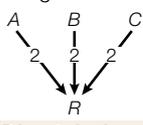
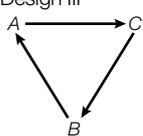
**Constraints.** For a proposed design to be acceptable, it has to satisfy two types of constraint: physical and scientific (BOX 3). First, it must be clear that the experiment is feasible, given the number of slides and the amount of mRNA expected to be available for use. There might be limitations of other kinds, for example on the number of RNA labellings that can be made. In practice, the number of slides and the amount of mRNA are the main physical constraints.

Second, it will have to be clear that the scientific questions that motivate the experiment can be answered if a given design is used — the more important questions will be answered more precisely than the less important ones. The illustrations that follow show that a combination of direct and indirect comparisons is often the best practical solution to a design problem.

**Single-factor designs**

We now present a series of examples that illustrate how the principles of design fare in different contexts. We follow the layout of design parameters set out in BOX 3 to show that the precise balance of direct and indirect comparisons in a given context should be determined by making the estimated comparisons more precise, subject to the scientific and physical constraints of the experiment. Although we hope that these principles are clear, and that the examples provide guidance, we cannot present all technical details, such as calculations of the variances of comparisons (see [supplementary Box online](#) for

Table 1 | **Single-factor experiments**

Design choices	Number of slides	Units of material (number of samples)	Average variance
<b>Indirect designs</b>			
Design I 	3	$A = B = C = 1$	2.00
Design II 	6	$A = B = C = 2$	1.00
<b>Direct design</b>			
Design III 	3	$A = B = C = 2$	0.67

Variance of estimated effects for three different designs of single-factor experiments.  $\sigma^2$  was set to 1 throughout.

an example of variance calculation). Although not difficult, these are beyond the scope of this review. Again, our examples focus on the question of identifying differentially expressed genes in various experiments.

**Comparisons among sources of mRNA.** We begin by considering an experiment in which three mRNAs from three sources are compared, and we suppose that all pairwise comparisons are of equal interest. This type of experiment could arise, for example, when investigating the differences in expression between three different regions (A, B and C) of the brain (FIG. 2). The scientific aim of this experiment is, therefore, to identify genes that are differentially expressed in these brain regions. The main interest is in identifying genes with differential expression in (A–B), (B–C) or (A–C) comparisons. TABLE 1 shows a few design choices, in which R is a common reference source of mRNA. In this table, we assume that the variance for log ratios within a slide for a given gene is  $\sigma^2$ . Each table entry is the average variance that is associated with the three pairwise comparisons of interest:  $\log(A/B)$ ,  $\log(B/C)$  and  $\log(A/C)$ . Note that, because all pairwise comparisons are of equal interest, the main scientific constraint in the experimental design is that they can all be estimated. Depending on the physical constraints, different design choices will be made. For example, if an investigator has unlimited amounts of reference material, but only one sample of RNA from each of A, B and C, then design I is the only possible choice out of the three presented in TABLE 1. However, if the investigator has two samples of RNA from the A, B and C regions, then both designs II and III are feasible (but design II will use twice as many slides). However, direct comparison (design III) will provide more precise comparisons between the regions and will reduce the number of slides that is required.

**LOOP DESIGN**  
 A design that involves mRNA samples labelled 1, 2, 3, ..., n, hybridized together in pairs (1,2), (2,3), ..., (n–1,n), (n,1).

With more than three sources of mRNA, the situation becomes more complex. The so-called reference designs are analogous to designs I and II, and compare each of the three or more sources of mRNA to a fixed reference source. The analogue of design III — which we call the ‘all-pairs design’ — is unlikely to be feasible or desirable for a large number of comparisons because of the amount of mRNA that would be required. For example, with six sources of mRNA, there are 15 pairwise comparisons that require five units of each target mRNA, for seven there are 21 that require six units, and so on. Alternative classes of designs that involve far fewer slides include the LOOP DESIGNS of Kerr and Churchill<sup>15</sup>, in which the graph is a single loop that connects successive pairs of vertices. However, the larger loop designs necessarily have long paths between some pairs of vertices, and consequently, some comparisons are much less precise than others. Therefore, instead of regarding the problem of choosing a design as a competition between classes of designs (such as reference, loop and all-pairs), a more productive approach is to ask which comparisons are of greatest interest and which are of lesser interest, and to seek a design that gives higher precision to the former and lower precision to the latter. We illustrate these issues with a discussion of short time-course experiments below.

**Time-course experiments.** In time-course experiments, the design choices depend on the comparisons of interest. Scientific constraints definitely matter, along with physical ones (for example, if the number of hybridizations is restricted), and the best design can crucially depend on the number of time points. TABLE 2 shows a range of design choices. Design II in TABLE 2 involves hybridizations between consecutive time points, whereas design I uses T1 (where T is treatment) as a common reference. When the main focus of the experiment is on the relative changes between T2, T3, T4 and the initial time point T1, design I is the better choice. However, if more subtle variations from one time point to another are of greater interest, then design II will be preferable. This is an illustration of how the comparisons of greatest interest determine the best design.

The choice becomes less obvious when four hybridizations can be done. Design III illustrates a common reference approach, whereas design IV is similar to using T1 as a common reference, with one extra direct hybridization between T2 and T3. Design V is an example of a loop design and design VI offers a mixture of direct and indirect comparisons that lead to some comparisons being more precise than others. TABLE 2 also shows the precision that is associated with each pairwise comparison. The choice between designs V and VI clearly depends on the comparisons of interest, as the average variance of comparisons is the same. For example, design V is preferable if comparisons between consecutive times are of more interest than those that are two time units apart.

At present, most microarray experiments<sup>3,4</sup> use reference designs, as they have the advantage of easy analysis

Table 2 | Time-course experiments

Design choices	t versus t + 1			Comparisons t versus t + 2		t versus t + 3	Average variance
	t <sub>1</sub> /t <sub>2</sub>	t <sub>2</sub> /t <sub>3</sub>	t <sub>3</sub> /t <sub>4</sub>	t <sub>1</sub> /t <sub>3</sub>	t <sub>2</sub> /t <sub>4</sub>	t <sub>1</sub> /t <sub>4</sub>	
Design I — T1 as common reference 	1.00	2.00	2.00	1.00	2.00	1.00	1.5
Design II — direct: sequential 	1.00	1.00	1.00	2.00	2.00	3.00	1.67
Design III — common reference 	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Design IV — T1 as common reference 	0.67	0.67	1.67	0.67	1.67	1.00	1.06
Design V — direct: loop 	0.75	0.75	0.75	1.00	1.00	0.75	0.83
Design VI — direct: mixed 	1.00	0.75	1.00	0.75	0.75	0.75	0.83

Variance of estimated effects for six different designs of time-course experiments. Designs I and II involve only three slides and the remaining designs involve four.  $\sigma^2$  was set to 1 throughout.

and interpretation without the need for statistical tools. However, because of frequent cross-disciplinary collaborations in microarray data analysis, it is not unreasonable to expect that statistical tools will become available for combining slide data across many experiments<sup>14</sup>. The two examples above illustrate some of the possible design choices that need to be made, and the general considerations that lead to a decision. We turn now to a more complex kind of experiment, with its own design questions.

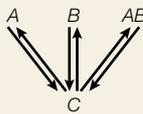
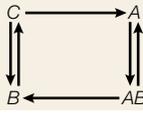
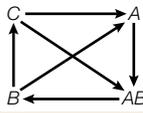
**Multifactorial designs**

The previous two examples are known in the statistics literature as single-factor or one-way designs, in which the factor (brain region in the first example and time in the second example) has three or more levels or values (regions A, B and C in the first example; times T1, T2, and so on, in the second). A more complex class of designs arises when two or more factors are considered jointly, and each factor has two or more levels. In a sense, these could be regarded as one-way designs with four or more levels, but the special nature of the levels singles out certain comparisons for attention above others. These are called factorial experiments, and are used to study differences that not only are caused by single factors, but also result from the joint effect of two or more factors. Any discussion of factorial experiments necessarily includes a study of the idea of interaction. (Loosely speaking, interaction refers to the way in which the joint effect of two factors differs from what might be predicted on the basis of their effects alone. Interaction is therefore an idea that is, to a great extent, model dependent and scale dependent).

Factorial experiments were introduced by R. A. Fisher in 1926 (REF. 8), and studied extensively by him and his collaborator<sup>16</sup>. They arise frequently enough to warrant separate discussion. Here, we focus on understanding the interactions and the relationship between separately and jointly administered treatments (cf. Glonek and Solomon<sup>17</sup>).

**2 × 2 factorial experiments.** Suppose that we have two ways of treating a cell line — for example, by adding different growth factors. If we let C denote mRNA that is derived from the untreated (control) cells, and A and B denote mRNA that is derived from the cells that were treated by the first and second method separately, we can then use AB to denote cells that were treated with both factors simultaneously. TABLE 3 shows a few examples of factorial experiments. The impact of the first treatment on gene expression can be assessed by studying the relative expression of a given gene in samples A and C in the absence of the second treatment, and also by comparing the relative expression of that gene in samples AB and B, that is, in the presence of the second treatment. To measure the extent to which these relative expression levels differ, we look at the difference of the log ratios log(A/C) and log(AB/B). The difference between log(AB/B) and log(A/C): log(AB/B) – log(A/C) = log(AB × C/A × B) is called the interaction of the two treatments on this log scale. We can think of the interaction as a measurement of the extent to which the presence or absence of factor B affects the differential expression of a gene in response to the presence or absence of factor A. The terms log(A/C) and log(B/C) are defined as the main effects of factors A and B, respectively. Note that sometimes different

Table 3 | **2 × 2 factorial experiments**

Design choices	Main effect A	Main effect B	Interaction A.B
<b>Indirect design</b>			
Design I 	0.50	0.50	1.50
<b>A balance of direct and indirect design</b>			
Design II 	0.67	0.43	0.67
Design III 	0.50	0.50	1.00
Design IV 	N/A	0.30	0.67

Variance of estimated effects for four different designs of 2 × 2 factorial experiments.  $\sigma^2$  was set to 1 throughout.

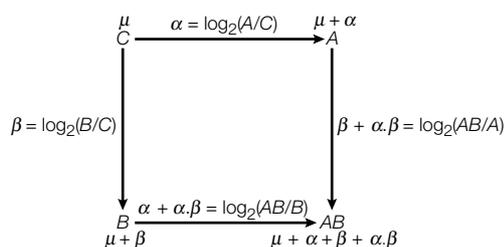
**SUMMARY STATISTIC**  
A numerical summary of some aspect of an experiment, typically an estimate of a parameter.

definitions (parameterizations) are used to describe these concepts and, for ease of explanation and interpretation, we have chosen the simplest definition (BOX 5). Ultimately, it does not matter which parameters are used, as long as care is taken when the estimates are interpreted.

In the treated-cell-line experiments of the kind we have discussed, we have four sources of mRNA: C, A, B and AB. Let us suppose that our main scientific interest is to identify genes for which the difference of the log ratios  $\log_2(AB/B)$  and  $\log_2(A/C)$  is not small. We can do

**Box 5 | Parameters for a 2 × 2 factorial experiment**

As well as looking at the effects that are caused by single factors, factorial experiments also provide information about the joint effects of two or more factors. To define parameters for a 2 × 2 factorial experiment, let us imagine that C denotes control, A and B denote singly treated cell samples, and AB denotes the doubly treated cell samples, as described in the main text. We denote the expected log intensities for a generic gene in these four samples by  $\mu, \mu + \alpha, \mu + \beta$  and  $\mu + \alpha + \beta + \alpha\beta$ , respectively. It follows that the expected values of the log ratios  $\log_2(A/C), \log_2(B/C), \log_2(AB/B)$  and  $\log_2(AB/A)$  are  $\alpha, \beta, \alpha + \alpha\beta$  and  $\beta + \alpha\beta$ , respectively, and that the interaction  $\log_2(C \times AB/A \times B)$  is  $\alpha\beta$ . (Note that this is not the standard parameterization for 2 × 2 factorial experiments, but here it is the most suitable one.) Table 3 refers to the estimation of the parameters  $\alpha, \beta$  and  $\alpha\beta$ , which we describe as the main effects for the factors A and B, and the interaction A.B.



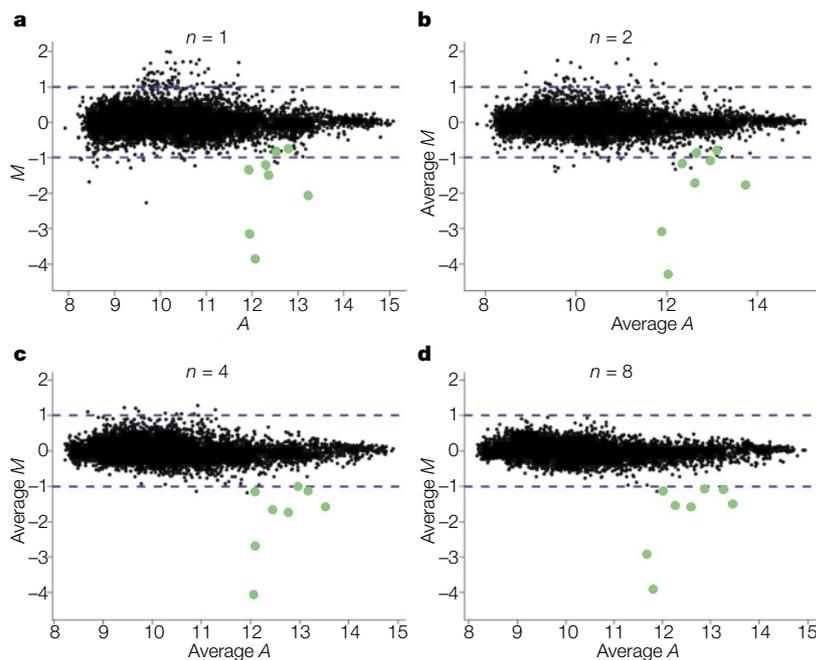
this in many ways, some of which are shown in TABLE 3. Note that all designs in TABLE 3 involve six slides. Designs II and IV give the smallest variance for the interaction term  $\log_2(AB \times C/A \times B)$ , whereas differential expression owing to A is not even estimable in design IV. Also, notice that design I is by far the worst for estimating the interaction, but that it uses less mRNA (two units from each source, compared with three in all of the others). The design of choice here depends on the level of interest in the main effects of A and B rather than in their interactions, assuming that any constraints on the number of slides or amount of mRNA available are satisfied. In general, design II (or its analogue with dye swaps between C and A, and B and AB, rather than between C and B, and A and AB) will probably be the design of choice that offers good precision for all comparisons. However, differential expression owing to one of A or B will be estimated more precisely than that resulting from C, as the designs are not symmetrical.

**Variability and replication**

Why should replicate slides be used in microarray experiments? And how many replicates should be used? Replicates reduce variability in SUMMARY STATISTICS and, importantly, the data obtained from replicate slides can be analysed using formal statistical methods. It is more difficult to say how many replicates should be done, although Lee *et al.*<sup>18</sup> indicate that three replicates are sufficient.

The cDNA microarray system is rather variable at the individual gene level. Expression of a gene might vary fourfold in one hybridization, but only 1.3-fold in a second independent hybridization, and twofold in a third. If we wish to determine which genes are differentially expressed between two samples of mRNA, for example, in the same tissue type from a knockout and from a wild-type animal, and have some assurance that our determinations are not false positives, then replication is essential. In essence, replication allows averaging, and averages are less variable than their component terms.

One common form of replication in cDNA microarray experiments involves putting replicates of the same spot (cDNA probe) on each slide<sup>19</sup>. Data from replicate spots can be extremely valuable for monitoring and improving the overall quality of the experimental data, but adjacent spots can never be full replicates for the following reasons. Nearly all aspects of the experiment (printing, general hybridization and scanning conditions) will be shared by spot replicates, and as a result, any systematic effects of these conditions on the measurements will also be shared. The consequent lack of independence of the measurements greatly reduces their value for broader statistical inference. Of course, sharing of conditions also applies to a lesser extent to experiments on different slides, but different hybridizations, even of identically prepared material, usually lead to rather different data, and it is these replicate hybridizations that are of real interest here. Note that, if duplicate spots are to be used, it is advisable to have them well spaced and not adjacent, as this would give a better reflection of the variability across the slide.



**Figure 3 | Averaging replicates reduces variability.** Plots of log ratios  $M = \log_2(\text{KO}/\text{WT})$ , averaged across replicate slides, against overall intensity  $A = \log_2\sqrt{(\text{KO} \times \text{WT})}$ , which is similarly averaged, are shown. The green spots correspond to eight genes that were known to be differentially expressed between the two mRNA sources (knockout (KO) and wild-type (WT) liver). The example shown here is based on data from the Apo AI experiment in REF. 5. The numbers of replicate slides ( $n$ ) shown are **a** | 1, **b** | 2, **c** | 4 and **d** | 8.

Lack of replication greatly restricts our ability to use formal statistical tests to decide whether a given intensity log ratio is significantly different to zero. In particular, replication is essential to estimate the variance of the log ratios across slides. Attempts to assess the significance of log ratios using data from only a single slide depend on unverifiable modelling assumptions (see REF. 20 for further details) and, in general, fail to take into account the most important source of variation — between-slide variability. By contrast, suitably defined, standard statistical methods, such as *t*-tests, are applicable to analysing data from replicate slides, although some changes to these methods need to be made<sup>20,21</sup>. When we replicate, we usually have a random sample of different mRNA from cell samples, and under these circumstances, we can extrapolate from our sample to the population of all such cell samples. In this sense, replication is intimately connected with the statistical extrapolation from sample to population.

**Technical replicates.** As explained above, and consistent with statistical tradition<sup>8</sup>, replication is a highly desirable feature of comparative microarray experiments. There are several forms of replication, and we briefly review them here. The differences lie in the degree to which the data might be regarded as independent, and in the populations that are represented by the experimental samples. Given that replicate hybridizations are almost invariably carried out by the same person, using the

same equipment and protocols, and frequently at about the same time, it is inevitable that replicate data will share many features. Most of the differences listed below concern the target mRNA samples.

Technical replicates between slides refers to replication in which the target mRNA is from the same pool, that is, from the same extraction. We have observed that there are characteristic, repeatable features of extractions and, therefore, conclude that technical replicates generally involve a smaller degree of variation in measurements than the biological replicates described below. Consequently, they do not provide the independence of data that gives the fullest benefits of averaging, and shared systematic features of technical replicate samples will remain even after averaging.

**Biological replicates.** The term ‘biological replicates’ usually refers to hybridizations that involve mRNA from different extractions — for example, from different samples of cells from a particular cell line or tissue. In many cases, this is the most convenient form of genuine replication. Provided the sample labelling is carried out separately for mRNA from different extractions, this approach will lead us as close to independent experimental results as is feasible in this context. Therefore, we strongly recommend biological replication as the principal source of replicate slides.

The term can also mean that the target mRNA comes from different individuals or different versions of a cell line. This form of biological replication is different in nature from the biological replication described above, and typically involves a much greater degree of variation in measurements. For example, experiments with mice have to deal with the inevitability of the hormonal and immune systems of individual mice being in different states or their tissues being in different states of inflammation. Most of the variation might seem unnecessary, as it can make real expression differences harder to discern, but from the perspective of the generalizability of conclusions, for example, to an entire inbred strain of mice, this might be the appropriate form of replication for some experiments.

We should note that, even in this case, if a common reference design is being used, logs of ratios in which the numerators come from independent, biological replicates, will still show some correlation because they share some unique features (they share a batch of reference mRNA as their denominator). This can be seen in FIG. 3, which shows plots of log ratios  $\log_2(\text{KO}/\text{WT})$  averaged across replicate slides, against overall intensity  $\log_2\sqrt{(\text{KO} \times \text{WT})}$ , similarly averaged. Each of the replicate slides involves mRNA from a different experimental animal, hybridized with the same reference mRNA<sup>5</sup>. The green spots correspond to eight genes that are known to be differentially expressed between the two mRNA sources (knockout and wild-type liver tissue). As the sample size (here, the number of mice) increases, the cloud of points around the horizontal axis shrinks. This makes it easier to distinguish real change and random variation about zero. Note that, with  $n = 1$  replicate, the cloud extends beyond  $\pm 1$  on the log base 2 scale, that is, twofold in either direc-

tion. By contrast, with  $n = 8$  replicates, the cloud is largely contained in the range  $\pm 0.7$  on the log scale, that is,  $\sim 1.6$ -fold in either direction. If the eight replicate data sets had been genuinely independent, which they are not in this case (the reference RNA is shared across all eight mice), then we would have expected a much greater reduction in the size of the cloud. It would be reduced to  $\pm 0.35$  on the log scale or a 1.3-fold change in either direction.

The type of replication to be used in a given experiment affects the precision and the generalizability of the experimental results. In general, an experimenter will want to use biological replicates to obtain averages of independent data and to validate generalizations of conclusions, and perhaps technical replicates to assist in reducing the variability. Given that there are usually several possible forms of technical and biological replication, we need to be careful when deciding how much replication of a given kind is desirable, subject to experimental and cost constraints. For example, if a conclusion that is applicable to all mice of a certain inbred strain is sought, experiments that involve many mice, preferably a random sample of this mouse population, must be carried out. Extrapolating to all mice of that strain from results on a single mouse, even when several mRNA extractions are used, has well-known dangers associated with it.

#### Power and sample-size determination

Having chosen a form of replication that is suited for the experiment under consideration, the experimenter needs to determine the sample size, that is, the number of slides to use. In general, a POWER CALCULATION requires that the experimenter states: the variance of individual measurements; the magnitude of the effect to be detected; the acceptable false-positive rate; and the desired power of the calculation, that is, the probability of detecting an effect of the specified (or greater) magnitude. The question of sample size is difficult to answer in the context of microarray experiments, as the variance of the relative expression levels across hybridizations varies greatly from gene to gene. Even if the experimenter knew these gene-specific variances in advance (which they could not in any detail), they would usually be unable to nominate in advance the gene expression changes that are of interest. So, two vital components of the standard power calculations are absent: the variance  $\sigma^2$  of the individual log ratio measurements and the magnitude of the effects of interest for individual genes.

Experimenters might wonder how many hybridizations they need to do to have a 90% chance of detecting a twofold differential expression. This question could be answered provided we knew the variance  $\sigma^2$  for the differentially expressed genes. One way to get over this impasse is to specify a value of  $\sigma^2$  (which is necessary for power calculations) that is not too small, on the basis of past experience with that experimental system, for example, the MEDIAN or upper quartile of the variances across genes. Doing a power calculation with the upper quartile variance, experimenters would be able to assert that their number of replicates gives them a certain power for detecting differential expression of greater than a stated value, for all but the 25% most variable genes.

This discussion of power and sample size raises a question that can be better addressed with an analysis that examines the trade off of power against false-positive rate. This is a standard statistical question, but in the context of microarray experiments, in which validation of results is routine, there is a special twist. In situations where mRNA samples for the experiments are scarce and the verification method is straightforward and relatively cheap, the experimenter might be willing to accept a higher false-positive rate on the grounds that sorting out true from false positives is not so difficult. In such cases, the number of replications needed can be reduced. A non-standard approach to determining the number of microarrays needed to ascertain differential expression is presented in Zien *et al.*<sup>22</sup> These authors base their analysis on a model of the variability in (high-density short oligonucleotide) array data; in particular, they do not include gene-specific variances in expression.

#### Statistical design in microarray practice

To what extent have the principles we have described been used in microarray practice? It has to be said that unreplicated microarray experiments are still the most common. In part, this is because researchers seem reluctant to 'waste' a hybridization by replicating one that has already been done, when they could do a new and different one. But perhaps the main reason for the lack of replication is the wide use of clustering methods to analyse the data<sup>3,4,23</sup> that do not seem to fall into the standard framework of statistical inference. Many of these experiments are unashamedly exploratory, and therefore do not seek statistical support for their conclusions. However, it should be pointed out that there is an element of 'effective replication' in time-course or other sets of similar experiments.

Several authors have replicated their comparisons: Callow *et al.*<sup>5</sup>, Conklin *et al.*<sup>24</sup> and Pritchard *et al.*<sup>25</sup>, to mention just a few. Factorial experiments were used by Jin *et al.*<sup>14</sup> and Boldrick *et al.*<sup>26</sup> The experiments of Jin *et al.*<sup>14</sup> include replication but have no common reference mRNA sample, whereas Boldrick *et al.*<sup>26</sup> do not replicate their hybridizations but do include a common reference sample. In another study, Hughes *et al.*<sup>1</sup> duplicated every hybridization they carried out and, in addition, made use of an error model (not mentioned in this review) (see supplementary material in REF. 1 for further information on this model). An optimistic assessment might be that statistical design principles are trickling into the microarray world.

Our distinction between expression comparisons carried out within- and between-slide needs to be supplemented by one of even greater magnitude, and hence importance. In their study, Jin *et al.*<sup>14</sup> took the bold step of treating the signals from the two channels of their cDNA experiment as two separate sources of data; they did not convert to ratios or log ratios, but kept both of the signals for their analysis. As a result they detected two types of effect: those that are estimated within hybridizations and that are therefore based on ratios within hybridizations (of age, in this case), and others

#### POWER CALCULATION

A calculation that leads to the probability that a null hypothesis that is being tested will be rejected in favour of the alternative, under specified assumptions that imply that the alternative hypothesis is true.

#### MEDIAN

The middle value in a set of numbers ordered in value from smallest to largest. If there are an even number of numbers, the median is the average of the middle two after ordering.

that are not based on ratios within hybridizations (effects of sex and strain, in this case). In our view, this approach can only be adopted after a very thorough multi-slide normalization of all the single channels, because there are many systematic non-additive spatial and intensity-dependent hybridization biases that will only disappear when ratios of single-channel readouts are considered.

Setting aside the normalization issue, designing for single-channel cDNA microarray experiments is analogous to designing for high-density short oligonucleotide microarray or nylon membrane experiments at the whole slide level. In this review, we have concentrated on the within-slides design. Combining the two design types is beyond the scope of this review, although we note that there is much experience and some theory on design for two-stratum experiments in the statistics literature<sup>27,28</sup>.

**Conclusion**

In our view, the main design issue with cDNA microarray experiments is the determination of which mRNA samples should be hybridized together on the same slide, bearing in mind the objectives of the experiment and the constraints on reagents and materials. We have discussed this question, here, in the context of several types of

experiment. The next most important question concerns replication. Different types of replicates contribute to the analysis in different ways, and we have attempted to explain these differences and to make recommendations for different types of experiments, emphasizing biological replication. We do not believe that a straightforward analogue of traditional statistical power analysis can be used in the microarray context. We have explained why and have outlined a weaker calculation. Finally, we briefly addressed a design issue that is unique to cDNA microarray experiments — the use of dye-swap experiments — and briefly reviewed the extent to which the ideas presented here are evident in the microarray literature.

How can we expect this topic to evolve? It is undoubtedly true that increasingly complex microarray experiments are being carried out — for example, the ones with many factors or those that combine the factorial and the time-course experiments (as discussed above). The analysis of such experiments is already challenging, but it will be even more difficult to provide well-founded design recommendations for them. In a sense, the tools and techniques of statistical design must continue to develop as more and more imaginative experiments are devised by biologists, but they must always be based on the sound principles of analysis of these experiments.

1. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
2. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
3. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
4. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
5. Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.* **10**, 2022–2029 (2000).
6. Redfern, C. H. *et al.* Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proc. Natl Acad. Sci. USA* **97**, 4826–4831 (2000).
7. Kerr, M. K. & Churchill, G. A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201 (2001). **The first paper to present the statistical principles of experimental design in the context of microarray experiments. Analysis involves a linear model for log intensities. Loop designs are introduced and compared with common reference designs.**
8. Fisher, R. A. The arrangement of field experiments. *J. Min. Agric. Gr. Br.* **33**, 503–513 (1926).
9. Cox, D. R. *Planning of Experiments* (Wiley, New York, 1958). **A classic book about the statistical design of experiments.**
10. Box, G. E. P., Hunter, W. G. & Hunter, J. S. *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building* (Wiley, New York, 1978). **A modern classic on the statistical design and analysis of experiments.**
11. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
12. Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
13. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, E15 (2002).
14. Jin, W. *et al.* The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.* **29**, 389–395 (2001). **The authors carry out a mixed model analysis of variance on single channel log intensities, including age, sex and strain. Of these, age effects were estimated within arrays, whereas sex and strain effects were estimated between arrays. No single-channel between-slide normalization was carried out. The authors found strong evidence for differential dye effects.**
15. Kerr, M. K. & Churchill, G. A. Statistical design and the analysis of gene expression microarrays. *Genet. Res.* **77**, 123–128 (2001). **The authors apply classical statistical experimental design to cDNA microarray experiments and keep Cy3 and Cy5 spot intensities separate in the analysis. The study assumes global normalization is adequate.**
16. Yates, F. *The Design and Analysis of Factorial Experiments* Technical Communication 35 (Commonwealth Bureau of Soils, Harpenden, Herts, 1937). **A classic book on factorial experiments.**
17. Glonek, G. F. V. & Solomon, P. J. *Factorial Designs for Microarray Experiments* Technical Report (Department of Applied Mathematics, University of Adelaide, South Australia, 2002). **The first careful treatment of optimal design for factorials.**
18. Lee, M. L., Kuo, F. C., Whitmore, G. A. & Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA* **97**, 9834–9839 (2000).
19. Black, M. A. & Doerge, R. W. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold changes for cDNA microarrays. *Bioinformatics* (in the press). **The authors limit their discussions to replicate spots within a slide.**
20. Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statist. Sinica* **12**, 111–139 (2001).
21. Wolfinger, R. D. *et al.* Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comp. Biol.* **8**, 625–638 (2001).
22. Zien, A., Fluck, J., Zimmer, R. & Lengauer, T. *Microarrays: How Many do you Need?* Proceedings of RECOMB 2002 (Association for Computing Machinery, New York, 2002). **Using non-standard power analysis, this paper answers the question posed in its title.**
23. Friddle, C. J., Koga, T., Rubin, E. M. & Bristow, J. Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Natl Acad. Sci. USA* **97**, 6745–6750 (2000).
24. Redfern, C. H. *et al.* Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proc. Natl Acad. Sci. USA* **97**, 4826–4831 (2000).
25. Pritchard, C. C., Hsu, L., Delrow, J. & Nelson, P. S. Project normal: defining normal variance in mouse gene expression. *Proc. Natl Acad. Sci. USA* **98**, 13266–13271 (2001).
26. Boldrick, J. C. *et al.* Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl Acad. Sci. USA* **99**, 972–977 (2002).
27. Hinkelmann, K. & Kempthorne, O. *Design and Analysis of Experiments* Vol. 1 *Introduction to Experimental Design* (Wiley, New York, 1994).
28. Bingham, D. & Sitter, R. R. Design issues for fractional factorial split-plot experiments. *J. Quality Technol.* **33**, 2–15 (2001).
29. The chipping forecast. *Nature Genet.* **21** (Suppl.) (1999).
30. Youden, W. J. in *Precision Measurement and Calibration: Statistical Concepts and Procedures* Vol. 1 of Special Publication 300 (ed. Ku, H. H.) 146–151 (National Bureau of Standards, United States Department of Commerce, Washington, DC, 1969).

**Acknowledgements**  
We thank S. Dudoit and N. Thorne for discussions and assistance during the course of this review. We also thank M. J. Callow from the Lawrence Berkeley National Laboratory and members of J. Ngai's lab — D. Lin, E. Diaz and J. Scolnick — for providing the data used in the figures. In addition, we are grateful to D. Bowtell for feedback on many design issues. This work was supported in part by the National Institutes of Health.

 **Online links**  
**FURTHER INFORMATION**  
**Terry Speed's lab:**  
<http://www.stat.berkeley.edu/users/terry/zarray/html/index.html>  
**Access to this interactive links box is free online.**