Microarray experiments: considerations for experimental design

Sanushka Naidoo[a*], Katherine J. Denby[b] and Dave K. Berger[a]

Microarrays are useful tools to investigate the expression of thousands of genes rapidly. Some researchers remain reluctant to use the technology, however, largely because of its expense. Careful design of a microarray experiment is key to generating cost-effective results. This article explores issues that researchers face when embarking on a microarray experiment for the first time. These include decisions about which microarray platform is available for the organism of interest, the degree of replication (biological and technical) needed and which design (direct or indirect, loop or balanced block) is suitable.

Introduction

Initially conceived and launched ten years ago,[1] microarray technology has become an attractive choice for researchers to screen the expression of thousands of genes simultaneously. During its short history, the technology has made invaluable contributions to various scientific fields. One example is evident in human cancer research and the development of a prognostic tool based on gene expression profiles in early breast tumours.[2] This assists doctors in predicting whether severe cancer is likely to develop that warrants aggressive response such as chemotherapy and hormone treatments, and prevents low-risk patients from receiving harsh treatments unnecessarily, because surgery and radiotherapy are sufficient in these cases. Concurrent with the sequencing of whole genomes, microarray technology has become progressively more sophisticated, allowing high-density arrays and consequently high throughput of data. Despite recent advancements of the technology, several questions remain, especially to those researchers embarking on microarray experiments for the first time. The design of the experiment depends, first, on the biological question being asked, as well as the organism being studied. Different microarray platforms exist and selection of the correct design influences the analysis of the data to obtain biologically significant results. This review aims to assist those researchers wishing to employ microarrays for their biological organism of interest by outlining the principles of experimental design.

Microarray platforms

Microarrays are conceptually quite simple and may be regarded as a large-scale reverse Northern blot. There are several types of microarray platforms: spotted microarrays, such as cDNA microarrays and oligonucleotide arrays, and the Affymetrix GeneChip® system, which involves synthesis of oligonucleotides directly onto the microarray support. In South Africa, two microarray facilities are available: the African Centre for Gene Technologies (ACGT) facility at the University of Pretoria (http://microarray.up.ac.za), and capar in the University of Cape Town's Molecular and Cell Biology Department (http://www.capar.uct.ac.za). Both centres are capable of producing cDNA and oligonucleotide microarrays at high densities.

Figure 1 illustrates a typical microarray experiment that uses a cDNA microarray platform. cDNA fragments, representing different genes, are amplified and spotted at high density onto microscope glass slides with special surface chemistry that allows binding of the spotted DNA. Two different cDNA populations derived from independent RNA samples are labelled with red (Cy5) and green (Cy3) fluorescent dyes, respectively, and hybridized to the slide. The array is subsequently washed and scanned by lasers that excite the different dyes. A fluorescent signal is computed for each spot on the array and the ratio of Cy3:Cy5-induced fluorescence for each spot corresponds to the relative amount of transcript in the samples. In microarray experiments, the selection of candidate genes depends on the criteria set by the researcher to describe differential expression. Previously, those genes that satisfied the criteria of having a fold change greater than two were considered differentially expressed. However, the role of statistics in determining the significance of results has become increasingly important and only those genes that are shown to be differentially expressed with statistical support across replicates are selected.[3] It is for the latter reason that any microarray experiment would benefit from the expertise of a statistician able to advise on the experimental design and subsequent analysis for a particular biological question.

For those organisms with little or no genome sequence available, arrays can be constructed by picking clones from a cDNA library and amplifying the insert cDNAs prior to spotting. The identity of selected clones can be determined after microarray analysis.[4] It is important to normalize the cDNA library before preparing the microarray in order to reduce the redundancy of clones. Redundant clones contribute only to increased expense during amplification of the library. The preparation of microarrays from a normalized cDNA library is a viable strategy especially for uniquely South African organisms demonstrated in the case of the desiccation-tolerant plant *Xerophyta humilis*.[5] Another way to generate a normalized cDNA library is by using a technique such as suppressive subtractive hybridization (SSH). An SSH library is created by subtracting the transcripts common to both samples so that the resulting cDNA clones are derived from transcripts present in one sample (the tester), such as disease tissue, but not in the other (the driver), for instance healthy tissue. The SSH technique allows the detection of low-abundance, differentially expressed transcripts and may identify essential regulatory components in a number of biological processes.[6] Yang *et al.*[7] successfully combined SSH and microarrays to identify genes differentially expressed in breast cancer cell lines. Microarrays have also been used to screen clones derived from SSH libraries to identify up-regulated genes in banana and pearl-millet during defence responses.[8]

As an alternative to preparing your own cDNA libraries, arrays can be created from amplification of sequenced cDNA clones called expressed sequence tags (ESTs). Currently, there are

[a]Botany Department, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0002, South Africa.
[b]Department of Molecular and Cell Biology, University of Cape Town, Private Bag, Rondebosch 7701, South Africa.
*Author for correspondence. E-mail: sanushka.naidoo@fabi.up.ac.za

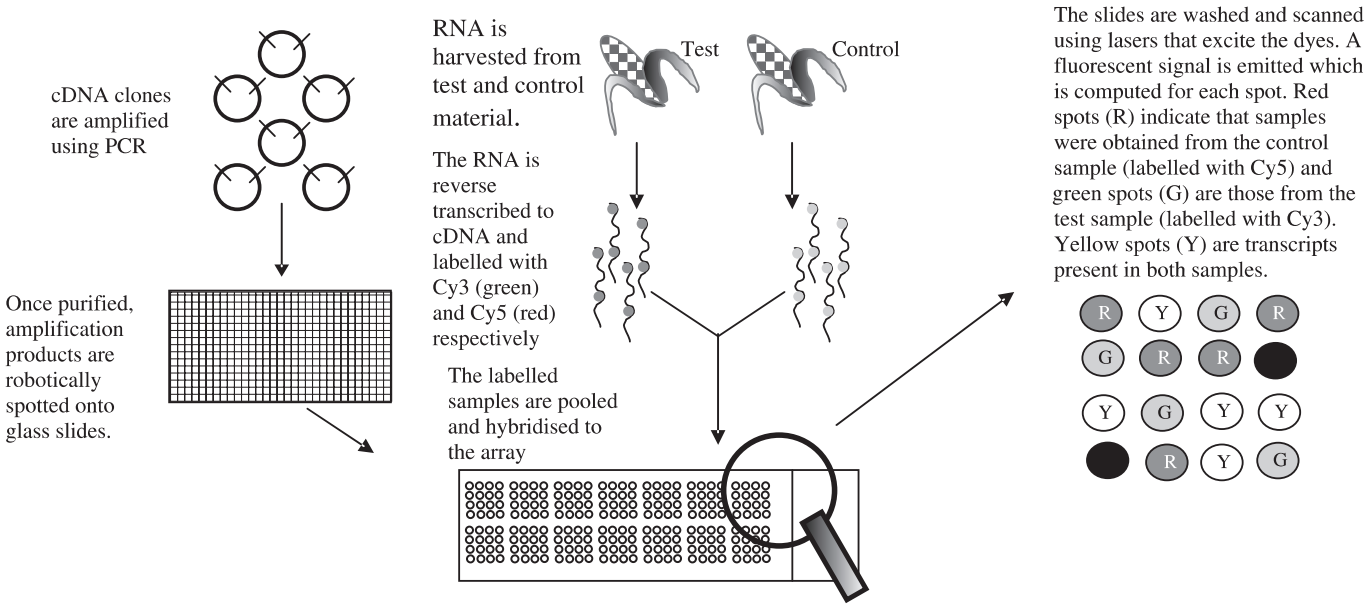

Fig. 1. An example of a typical microarray experiment using dual-colour-labelled cDNA transcripts hybridized to glass slides containing amplified cDNA fragments.

several million ESTs from various organisms in the NCBI public collection (http://www.ncbi.nlm.nih.gov/dbEST/). Ideally, each EST should represent a unique gene, referred to as a unigene set. Unigene sets for most genomes were initially assembled using software that identifies unique clones in EST databases. With the availability of whole-genome sequences, new unigene sets are becoming available. Some clones are genomic clones representing predicted genes for which no EST has been identified.[4] The advent of whole-genome sequences also allows one to custom-design arrays with genes predicted or known to be involved in a particular biological process. Kidson *et al.* (pers. comm.[c]) customized an array consisting of 384 amplified ESTs involved in eye-development. Other EST collections, like that consisting of 6000 *Arabidopsis* ESTs, have a wider application (S. Naidoo, unpubl.[d]). However, an expensive step in cDNA microarray analysis is the amplification of the EST set. This promotes the case for preparing smaller custom arrays rather than using large collections.

Affymetrix GeneChip® technology uses a series of 25mer oligonucleotides,[9] which are designed using a computer algorithm to represent known or predicted open reading frames. This technology is limited to organisms with a significant amount of genome information. There are between 10 and 20 different oligonucleotides representing each gene to control for variation in hybridization efficiency due to factors such as GC content. A control for cross-hybridization with similar short sequences in transcripts other than the one being probed for is a mismatch oligonucleotide next to each oligonucleotide with a single base pair change at its centre. Under stringent hybridization conditions, this control should not hybridize to the exact match cDNA. The level of expression of each gene is calculated using a procedure provided by the Affymetrix software, which computes the weighted average of the difference between the perfect match and mismatch. The high-density arrays are constructed on silicon wafers using a technique called photolithography and combinatorial chemistry. The process used to prepare the arrays is expensive and processing requires a proprietary hybridization station, scanner and software, which places a constraint on the number of slides that can be purchased for replication and availability to South African researchers. The target cDNA is labelled using amplified RNA and only a single sample is hybridized to each chip. Although Affymetrix GeneChip® arrays can accommodate a higher density of genes and are probably considered the 'gold standard' of microarray technology, they are limited to model organisms, whereas cDNA methods can be applied to any organism, are cheaper and more repetitions can be achieved. This enhances statistical analysis and can be more flexible in design.[4]

Spotted oligonucleotides have grown in popularity and are considered a hybrid technology, combining the uniformity of Affymetrix GeneChips® and the versatility of cDNA microarrays. Their use also removes the variability inherent in amplification of cDNA clones. This technology involves spotting 50–70mer oligonucleotides onto glass slides. Subsequent probe preparation and hybridization is similar to that of cDNA microarrays. Hughes *et al.*[10] found 60mer oligonucleotides were able reliably to detect transcript ratios at one copy per cell in complex biological samples. These results are in accordance with data obtained with robotically printed cDNA arrays.

Recently, Yauk *et al.*[11] compared six microarray platforms, two cDNAs and four oligonucleotides (including 25mer Affymetrix microarrays, 30mer spotted microarrays, and 60mer oligonucleotides synthesized *in situ*). The objective of this exercise was to determine whether gene expression profiles are influenced more by biology than by artefacts of the technology. There was a significant difference in the ability of the different platform types to detect differential expression in the two very different cell types that were used for the study. More differentially expressed genes were identified using the oligonucleotide- than the cDNA-based platforms. The validation exercises using Northern hybridizations and RT-PCR (reverse transcriptase polymerase chain reaction) supported the suggestion that cDNAs are less sensitive than the oligonucleotide platforms. These authors concluded that with high-quality microarrays and the appropriate normalization methods, the primary factor determining variance is biological rather than technical. This provides reassurance that if one cannot afford the Affymetrix platform, biologically meaningful data can still be obtained using cDNA microarrays or spotted oligonucleotide arrays.

[c]Kidson, S., Department of Human Biology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town.

[d]Naidoo, S., Botany Department, Forestry and Agricultural Biotechnology Institute, University of Pretoria.

The Affymetrix system may not be the first choice for South African researchers, primarily because of the cost and the restriction to model organisms. Spotted oligonucleotides provide a good alternative and are commercially available for organisms with a large amount of sequence information. The cDNA microarray platform remains the only possibility for organisms with limited sequence information. Given the cost of microarray experiments, it is important that attention be given to the design of the experiment. Typically, one wishes to obtain the best possible results with the resources available to answer the question of interest. The aspects of design discussed below pertain to two-colour dye experiments such as those using the cDNA and spotted oligonucleotide platforms. The Affymetrix system, which uses a single labelled sample during hybridization, is not discussed further.

Questions on design

Several questions should be addressed when one embarks on a microarray experiment. Logically, the first is, what exactly is the researcher investigating, that is, what is the biological question or hypothesis being posed or tested? Will the microarray experiments be able to answer the question and how will the experimental results contribute to the research as a whole? Would an alternative method be better, such as quantitative RT-PCR, SAGE (serial analysis of gene expression), cDNA-AFLP (amplified fragment length polymorphism) or ddRTPCR (differential display reverse transcriptase PCR)? In order to determine precisely what comparisons are being made, Yang and Speed[12] advise that the priority of the different scientific questions being asked should be identified along with the types and number of samples available.

On the technical side, another important consideration is whether the RNA sample is limiting and whether the process prior to hybridization — that is, RNA isolation, RNA extraction and labelling — are optimized for the organism of interest. If one wishes to identify a few genes to work on, one should determine, in advance, which method will be appropriate to verify the data obtained from the experiments. This is because a considerable amount of RNA is required for Northern hybridization whereas quantitative RT-PCR remains the method of choice for other researchers owing to the small amount of starting material required. Other experiments, such as those which compare expression profiles, rely on the strength of the statistical analysis to make conclusions and do not require verification.[13]

The data from spotted microarray experiments may have to be normalized prior to analysis owing to variability in labelling efficiency of the two different dyes. For this purpose, control spots are often necessary. The researcher has to determine what types of controls would be most appropriate for the tissue type being used. External or spike-in controls aid the researcher in determining whether labelling and hybridization has worked well.[12] This method uses genes from an organism different from the one being studied or else synthetic genes with no significant regions of homology to genes on the microarray, to prevent cross-hybridization. These genes are spotted onto the microarray and their corresponding RNA transcripts are included in both the target samples, which are subsequently labelled with the red and green dyes and hybridized to the slide. Spike-in artificial RNA controls and corresponding DNA targets to be spotted are commercially available, for example, Lucidea™ Universal ScoreCard™ (Amersham Biosciences). Hybridization results in predictable red and green fluorescence intensities at the target spots relative to the different concentrations of spiked RNA added to the samples. This controls for labelling efficiency but does not control for the difference in the amount of RNA in the two samples being hybridized. Negative controls (no DNA or DNA that is unlikely to cross-hybridize, e.g. from an unrelated organism) are sometimes included on spotted microarrays to determine the background fluorescence and whether the hybridization conditions are stringent enough.

Internal controls may be housekeeping genes or genes known to be constitutively expressed between the test and control sample. Housekeeping genes are required for fundamental cellular processes in different cell types and tissues. The expression of housekeeping genes does not depend on the physiological, developmental or pathological state of the tissue, for instance, actin and GAPDH in some cell types.[12] One problem with housekeeping genes is that they tend to be highly expressed and may not represent the genes of interest, which tend to be expressed less and are more likely to be subjected to intensity-dependent bias. The selection of an appropriate housekeeping gene for a particular condition is a contentious issue, since so-called housekeeping genes do not remain constitutively expressed under some conditions.[14] One way to identify internal control (housekeeping) genes is to data-mine previous microarray experiments for genes whose expression levels do not vary under various treatments. This approach is useful for model organisms for which there is a large amount of publicly available microarray data (such as *Arabidopsis* and human).

It is advisable first to test candidate internal control housekeeping genes for stable expression in the tissues of interest using sensitive methods such as quantitative RT-PCR. Vandesompele *et al.*[15] developed a procedure in Microsoft Excel® to analyse real-time quantitative RT-PCR data of putative housekeeping genes. They tested 10 commonly used housekeeping genes and confirmed that normalization using a single housekeeping gene was unreliable. Their procedure, which is also applicable to microarray data, uses the geometric mean of relative expression levels from carefully selected housekeeping genes to calculate a normalization factor.[15]

Normalization removes unwanted systematic bias from microarray data. This includes within-slide normalization to remove effects of dye bias and spatial bias (e.g. variation due to different print tips in a slide spotting run). Arrays with a large number of spots representing a substantial portion of an organism's genome can be normalized based on the assumption that most of the genes on the array should not be differentially expressed and should thus remain yellow.[3] If this assumption holds, then a linear or non-linear regression can be applied. The linear regression method, referred to as total intensity normalization, assumes that the relationship between the Cy3 and Cy5 channels is linear. However, this is not true for most microarray experiments. ANOVA models have also been applied for normalization.[16,17] Normalization of spatial biases can also be incorporated into the latter ANOVA models. LOWESS (LOcally WEighted Scatterplot Smoothing; also known as loess) is a commonly used non-linear regression method applied to microarray data, and performs a series of local regressions in overlapping windows through the range of the data.[18] The regression is then joined to form a smooth curve. Spatial biases can also be corrected separately using the LOWESS regression. In customized arrays containing a small number of genes biased towards a certain condition, for example, disease or salt-stress, control spots are required for normalization.[12] These could be a set of validated housekeeping genes. However, several that are expressed at a range of intensity levels should be used in order to perform a non-linear normalization. Additionally, prior to analysis, slides are subject to between-slide normalization,

which allows comparison of multiple arrays on an equal footing. Only basic normalization issues in the context of experimental design have been covered in this review; the reader is referred to Stekel,[3] Yang *et al.*,[19] Quackenbush,[20] and Futschik and Crompton[21] for normalization in the context of microarray data analysis.

Replication

Replication is necessary in order to apply a statistical test and reduce the variability inherent in microarray experiments. There are different levels of replication: technical and biological.[22] One type of technical replication is spot duplication. If space permits, cDNAs can be spotted in duplicate on every slide and the degree of conformity between duplicate spot intensities is a good indicator of the quality of the slide and hybridization. It is advisable, however, that duplicate spots be well spaced rather than spotted adjacently as this facilitates inspection of the degree of variability across the slide. Replicate slides hybridized with target RNA from the same preparation are also considered technical replication. Statisticians prefer the latter type of technical replication, as replicate spots on the same slides are not independent of each other.[22] Biological replicates could be hybridizations performed using RNA from independent preparations from the same source, or preparations from different sources, such as different organisms or different versions of a cell line. The latter type of biological replication encompasses greater variation in measurements. For instance, an experiment investigating drug treatment in mice is subject to the variation within the mice population such as differences in immune system, sex, and age. The greater variability inherent in this form of replication contributes to a broader generalization of the experimental results.[15] Typically, a researcher should use biological replicates to validate generalizations of conclusions and technical replicates to reduce the variability of these conclusions.[22]

Pooling RNA from a number of similar sources is often unavoidable in order to have sufficient material for a single hybridization. One way of overcoming the problem of limited starting material is RNA amplification.[23] Pooling may also be used to reduce the number of arrays in order to save on cost.[24] However, a single pool of many samples does not allow for the estimation of technical and biological variability. Shih *et al.*[25] show statistically that there is a loss of degrees of freedom and a decrease in power when pooling and suggest that, if pooling is used, the number of different pools should not be too small and the number of individuals should be increased to compensate for this.[25] The decision to pool is at the discretion of the researcher as it is sometimes not appropriate to combine samples. For example, when studying the effect of a drug on cancer patients, the gene expression in different patients is of interest. In this case, hybridizations with individual samples should be carried out. On the other hand, in an investigation of two inbred homozygous ecotypes of *Arabidopsis,* differences between the individual plants are not of interest, so pooling may be justified.

Another form of replication, dye-swap replications, are hybridizations that are repeated with the dye assignments reversed in the second hybridization. This method is useful to reduce the systematic differences in the red and green intensities, which have to be corrected during normalization.[26] A dye-swap replication can be performed for both a technical and biological replicate. Dye-swap pairs are not routinely warranted and may be excluded when indirect comparisons, such as those involving a common reference sample, are performed, because this design is based on differences between slides and the repeatable colour bias is removed during the analysis.[26] Similarly, a balanced block design negates the use of dye-swap replication as the design inherently compensates for the dye effect.[24] Balancing the dyes using the latter method is favoured over repeating each comparison with a dye swap, as this requires the use of more slides and thus increases the expense of the experiment.[26]

Design types

In cancer studies, Golub *et al.*[27] identified three categories to which a microarray experiment can be assigned, depending on the objective. These categories (class comparison, class discovery and class prediction) are applicable to most microarray experiments regardless of the organism being studied. In class comparisons, researchers are interested in comparing samples with each other.[19] An example of this is comparing gene expression profiles in wild-type mice with a mutant strain. Class predictions involve using the expression profiles generated by class comparisons and applying a multigene statistical model to determine to which class a new sample belongs. One strategy is, first, to make a class comparison to identify genes differentially expressed between cancer patients who respond to a particular treatment and those who don't. Subsequently, a commonly used class prediction approach would involve developing a univariate statistical model to identify a subset of genes that would help predict whether a new patient will respond to that therapy on the basis of their tumour expression profile.[24] Class discovery involves studies in which the samples are not predefined according to different classes before the microarray experiment. The objective is to discover clusters of the samples based on gene expression profiles. Once the classification is made, the next step is to characterize the cluster. An example of this would be a set of tumour samples that one wishes to divide into sub-classes based on gene expression profiles.[28] Other studies that investigate which classes of genes are co-regulated, for example in a time-course experiment, are also considered as class discovery. When samples have to be co-hybridized as in the case of spotted microarrays, careful design for pairing and labelling samples is required. Designs may involve direct or indirect comparisons and more than one option may answer the same question.

Direct comparison

Yang and Speed[12] stress the importance of deciding whether to use direct (within slides) or indirect (between slides) comparisons. In our laboratory, investigation of differentially expressed genes in a mutant *Arabidopsis* plant involved a direct comparison design. Figure 2A illustrates the comparisons made. The mutant RNA sample was co-hybridized with the wild-type RNA sample on the same slide. For the repeat slide, the same comparison was made with the dye assignments reversed. The platform used for this experiment was a cDNA microarray containing 500 *Arabidopsis* ESTs and several controls including a commercially available spike-in control (Lucidea™ Universal ScoreCard™, Amersham Biosciences) and several housekeeping genes, e.g. actin and β-tubulin. Negative controls of mouse genes, with no known homology to *Arabidopsis,* were also included. Spots were duplicated on the slide. Each sample contained leaf material pooled from 6–8 plants and two types of biological replication were performed: one using independent RNA preparations of the leaf material from trial 1 and the other using leaf material harvested from a completely different trial (trial 2). Two technical replications were performed per biological replicate. In total, twelve slides were used for this study. The correlation between all the mutant:wild-type gene expression ratios in each of the

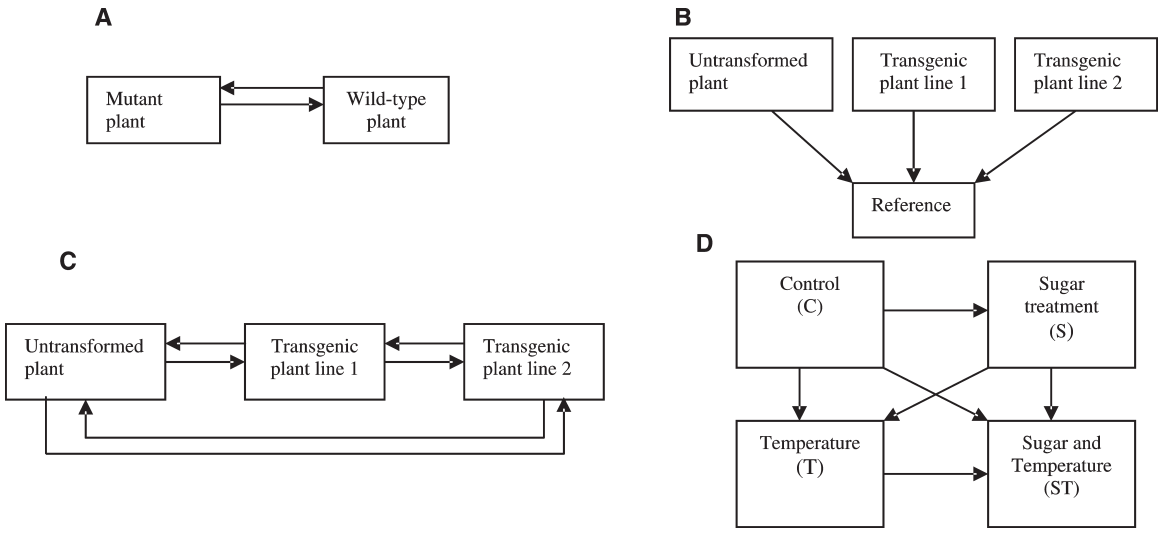

Fig. 2. Diagrammatic representations of the designs of four microarray experiments. Each microarray slide is represented by an arrow. The head of the arrow indicates that the sample was labelled with Cy5, while the tail represents a sample that was labelled with Cy3. **A**, Direct comparison between a mutant and wild-type *Arabidopsis* plant. **B**, An indirect comparison using a reference design. **C**, A loop design that investigated differentially expressed genes in transgenic plant lines. **D**, A factorial experiment that investigated the interaction between two factors: temperature and sugar.

replicates was calculated (Table 1). It is evident that the correlation between biological replicates derived from independent trials was less than that for biological replicates derived within the same trial. Thus, it is advisable that, when making generalizations, a biological replicate be included which is independent of the first. The data were analysed using a mixed model ANOVA[17] and approximately 2% of the genes arrayed were regarded as differentially expressed at a significance threshold of $-\log_{10}(P)$ equal to 5 (that is, $P < 0.00001$, Bonferroni adjusted to correct for multiple testing) (data not shown).

Table 1. Correlation between technical and biological replicates in a direct class comparison between a mutant and a wild-type *Arabidopsis* plant.

Comparison	Correlation
Duplicate spots on 1 slide	0.93
Between two technical replicates (slides)	0.92
Between biological replicates from within a trial	0.84
Between biological replicates from independent trials	0.72

Reference design

A commonly used means of indirect comparison for microarray experiments is a reference design.[29] This design uses an aliquot of a common reference RNA; the intensity of hybridization of a test RNA sample is compared to hybridization of the reference RNA to the same spot. A reference sample should be in large supply and is sometimes prepared by constructing complex mixtures of RNA in order to achieve maximum hybridization to the array. Such reference samples are commercially available, e.g. the Stratagene® Universal Mouse reference RNA set. Another method of preparing reference samples involves pooling aliquots of test samples that are to be investigated. Thus, every sample present in the test sample is present in the reference sample and so the relative amounts of each RNA species will be the same. This implies that in any comparison of test versus reference, the RNA concentrations will not be greatly different as each test sample is represented in the reference sample, a strategy which facilitates normalization.[29]

Figure 2B illustrates a reference design. For example, suppose we wish to identify genes that are differentially expressed in two transgenic plant lines. Samples from the untransformed line, transgenic plant line 1 and transgenic plant line 2 can be individually compared to a reference sample in this case made up of a pool of equal amounts of RNA from each sample. In a reference design, the reference sample is labelled with the same dye each time. It is generally assumed that any remaining dye bias not removed by normalization affects all the arrays similarly and does not bias comparison between the samples.[30] However, Dombowski *et al.*[31] have recently suggested that gene-specific dye bias exists in microarray reference designs. If this is a significant variable, then microarray data have to be validated before conclusions can be made or a reverse-dye comparison could be incorporated in a biological replicate to account for the dye effect on specific genes.[32] There are two steps connecting two samples in a reference design, so each comparison can be made equally efficiently. An advantage of this method is that as long as the amount of reference sample is not limiting, the design can be extended to handle large numbers of samples and in class discovery experiments samples from a new class can be added at a later stage.[29]

Balanced block design

A drawback of the reference design is that half of the hybridizations are used for the less interesting sample, the reference.[24] An alternative is a balanced block design. In a simple situation, suppose we wish to compare 4 mutant mice with 4 wild-type mice. One could hybridize on each array one mutant sample with a wild-type sample. Half the arrays should have the mutant samples labelled with the red dye and the wild-type samples with the green dye. In the other half of the slides, the samples should be hybridized with the labelling reversed. One disadvantage of the balanced block design is that cluster analysis of the expression profiles cannot be performed effectively. The common reference design is more amenable for the latter purpose as the relative expression measurements are consistent with regard to the same reference. Without a common reference, as in the balanced block design and direct comparison, comparison of samples on different arrays can be skewed by variation in size and shape of corresponding spots on different arrays.[24] The balanced block design is most effective when comparing two classes and can accommodate n samples of each type using n arrays. The advantage of the balanced block design is that half the number of slides can be used compared to a reference design or direct comparison. However, the balanced block design loses

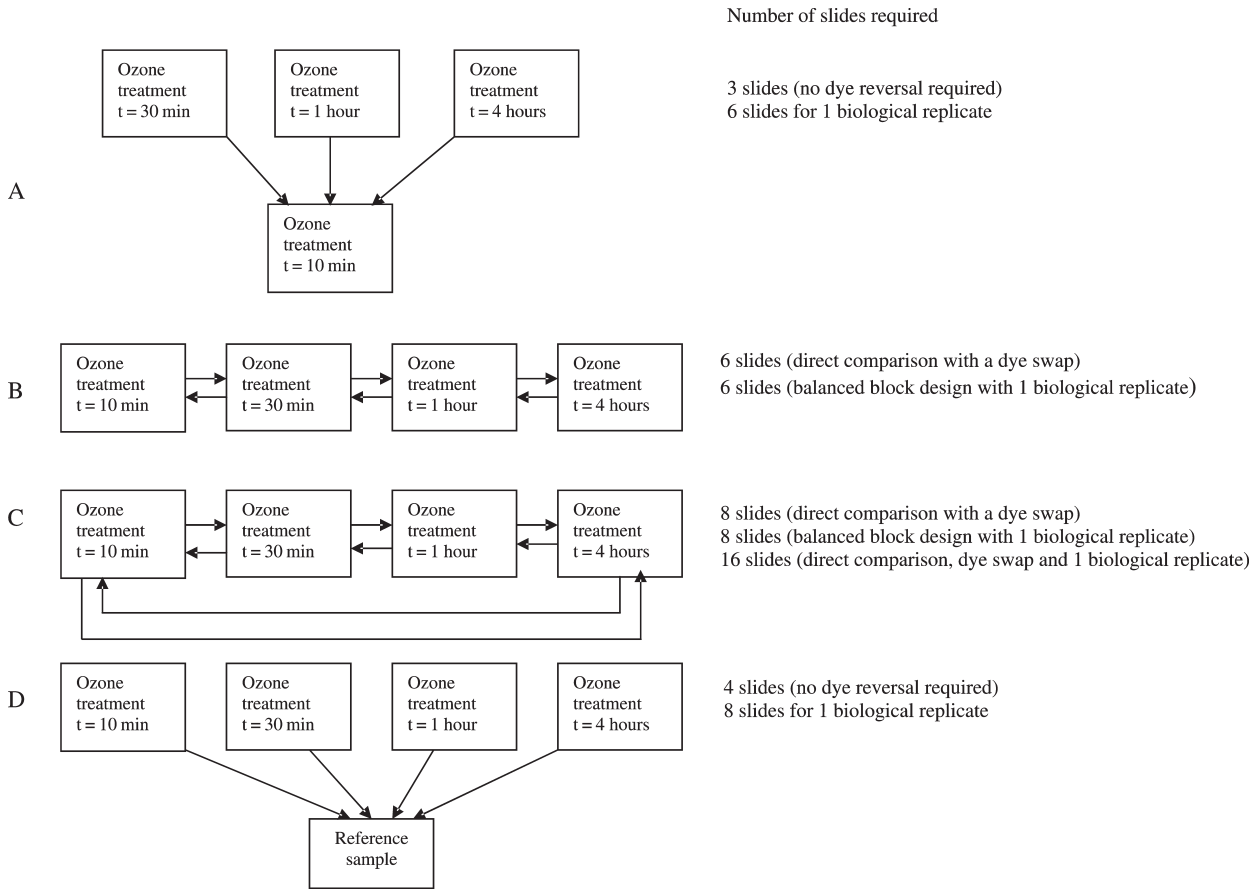

Fig. 3. Possible designs and the minimum number of slides required for a time-course experiment. Design **A** uses the first time-point as a reference, whereas design **B** compares consecutive time-points. A loop design is indicated in **C** and a reference design in **D**. Each box represents a sample whereas the arrow represents a slide. The head and tail of the arrows correspond to samples labelled with Cy5 and Cy3 dyes, respectively.

to the reference design when there is large variability between samples and when the number of samples and not the number of arrays is in limited supply.[24]

Loop design

A loop design involves array hybridizations that link the samples together in a loop. The comparisons being made control for variation in spot size and sample distribution patterns using a statistical model.[33] The example illustrated in Fig. 2B could be designed in a loop-wise fashion. This is illustrated in Fig. 2C. This design uses two aliquots of each sample and n arrays are used to study n samples. It is advisable to repeat the loop with the dye assignments reversed using the same sample (technical replicates) or employing a balanced block design by performing the loop with the biological replicates labelled with the reverse dyes to account for any dye bias. Comparison of two samples far apart in the loop is inherently more variable in a loop design and is more susceptible to failure if there are two or more bad quality arrays. This can result in the collapse of the loop, which would have to be solved by repeating the bad quality arrays.[30]

Factorial experiments

The previous types of design have been single factor experiments (typical factors being time, genotype, tissue type or treatment); experiments investigating two or more factors require a more complex design. Factorial experiments can be used to study the expression profiles resulting from single factors or those arising from the combined effect of two or more factors.[12] For example, suppose we wish to investigate the growth of bacterial cells under two conditions: high sugar content and high temperature. Figure 2D illustrates the comparisons that can be made. Let C denote expression of the untreated control sample; S the expression of samples grown in media containing high sugar content; T the expression of those samples grown at high temperature; and ST the expression of bacteria exposed to both conditions simultaneously. Then, the impact on gene expression of sugar treatment (S) in the absence of high temperature (T) can be assessed by log (S/C), and similarly the effect of high temperature can be estimated from log (T/C) in the absence of the effect of sugar treatment. The effect of the factor S in the presence of T is measured by log (ST/T) and a similar calculation can be made for factor T. The interaction of the two treatments, which is in effect measuring the extent to which the differential expression of a gene induced by sugar is dependent on whether the high temperature (T) is present, is indicated by: log (ST/T) – log (S/C) = log (ST × C/T × S). The same experiment can be repeated with the samples labelled with dyes reversed, after which the data for the two experiments can be combined to normalize the dye bias. Subsequently, the same calculation can be performed to determine the effect of treatment.[12]

Time-course experiments

Several designs are possible for time-course experiments, which depend on the comparisons of interest and the number of time points.[22] Most studies are aimed at identifying co-regulated genes, which falls under gene class discovery.[24] For example, in an experiment to investigate the effect of ozone treatment on cells over time, the schemes represented in Fig. 3 could be used. The design in Fig. 3A would be suitable if one were interested in the relative changes between time-points 2, 3 and 4 and the start.

Table 2. Some published examples of microarray designs.

Design	Type of study	Question	Replication	Number of slides	Reference
Direct comparison	Single-factor experiment class comparison	Which genes are differentially expressed in *Phythophthora infestans*-infected leaves of *Arabidopsis*?	2 spots/gene/slide 2 technical replicates Individual plants pooled (no biological replication) Dye swap?: yes	4	35
Loop	Two-factor experiment class comparison	How many genes are differentially expressed within and between natural populations of teleost fish?	2 spots/gene/slide 1 technical replicate Individual fish compared (15 biological replicates) Dye swap?: yes	60	36
Balanced block in a loop design	Single-factor experiment class comparison	Which genes are differentially expressed in parasitic and infectious larval stages of the common canine parasite *Ancylostoma caninum*?	2 spots/gene/slide 1 technical replicate 2 strains assessed (1 biological replicate) Dye swap?: no	24	37
Reference	Single-factor experiment class comparison	Which genes are preferentially expressed in the retina?	2 technical replicates (minimum of 1 biological replicate) Dye swap?: no	18	38

However, if comparisons between consecutive time-points are of interest, then a sequential comparison (Fig. 3B) or a loop design (Fig. 3C) may be more appropriate. A reference design could also be used (Fig. 3D) but, like the loop design, would require four slides whereas designs A and B use three. The dye bias would have to be removed in a loop and sequential design, necessitating the use of more slides, with dye assignments reversed. Deciding between a reference or loop design is influenced by several factors; Kerr and Churchill[33] provide ANOVA models to evaluate the microarray design and assist in selecting a loop or reference design for particular experimental objectives. Vinciotti *et al.*[34] evaluated a loop versus a reference design in two sets of microarray experiments and concluded that the loop design attained a higher precision than the reference design. These authors advise how simple loop designs can be extended to more complex experimental designs.

Table 2 lists examples of microarray designs employed by different researchers, according to the degree of replication, number of slides and aim of the experiment.

Sample size

Researchers have to consider the number of slides to use for a particular investigation. In microarray experiments, the relative expression levels across hybridizations vary greatly across genes, so sample size is an important question to address.[12] Power analysis can be used to determine the number of replicates required in an experiment given that an estimate of the technical variability is known.[3]

A common approach is to consider a null hypothesis for every gene in a microarray experiment. For example, in the experiment illustrated in Fig. 2A, the null hypothesis could be that a given gene is not differentially expressed between the mutant and wild-type plants. In this type of class comparison experiment, we would be interested in identifying those genes that do not support the null hypothesis. False positives would be genes identified as being differentially expressed when they are not, whereas false negatives would be genes that are identified as not differentially expressed when they actually are. False-positive results, where the null hypothesis is rejected when it is true, may be referred to as type I errors. The confidence of a statistical test is the probability of not getting a false-positive result (i.e. the probability of accepting the null hypothesis when it is true). False-negative results, where the null hypothesis is accepted when it is false, are called type II errors. The power of a statistical test is the probability of not obtaining a false-negative result (i.e. the probability of not accepting the null hypothesis when it is false). While type I errors can be controlled explicitly when we select a significance level for the statistical test (e.g. a 1% significance threshold), type II errors are controlled implicitly via the experimental design. The power of an experiment relies on the number of replicates used. Thus, the number of replicates chosen is determined by the power one wishes to attain in the analysis.[3] When a more stringent significance threshold is set, greater confidence but less power is achieved and, conversely, a lower significance threshold means less confidence and greater power. Depending on the experiment in question, one can judge as to whether a type I or a type II error is more acceptable. For example, if the purpose of the experiment is to identify genes involved in disease resistance to a certain plant pathogen and much time and money will be spent researching each chosen candidate gene, then it is important that type I errors (false positives) are avoided. If the microarray is being used as a diagnostic tool for cancer, however, then type II errors (false negatives) are less desirable as a patient falsely diagnosed as being cancer-free could develop a fatal tumour, which would otherwise have been treated.[3] The reader is directed to Stekel,[3] Tempelman[32] and Zien *et al.*[39] as useful guides to determining the number of microarrays needed to ascertain differential gene expression.

Analysis

The particulars of analysis have not been discussed in this review; however, more software is emerging which is open-source, user-friendly and can be applied to various types of microarray design. These include the TM4 microarray software suite (http://www.tigr.org/software/tm4), the Gene Expression Pattern Analysis Suite (http://gepas.bioinfo.cnio.es), Bioconductor (www.bioconductor.org) and R (www.r-project.org).

While this review aims to give non-statisticians an overview of how to approach microarray experimental design and select design parameters for particular types of experiment, it is advisable to enlist the assistance of a statistician at the very beginning of a microarray experiment. Expertise in this area is growing as microarray technology generates more interest among statisticians. It is encouraging to note that the capacity for successful microarray experiments exists in South Africa. The quality of publications generated from the two local microarray facilities attests to this.

We acknowledge the support of the Mellon Foundation and the National Research Foundation.

1. Schena M., Shalon D., Davis R.W. and Brown P.O. (1995). Quantitative monitoring of gene expression patterns with a cDNA microarray. *Science* **270**, 467–470.
2. van't Veer L.J., Dai H., van de Vijver M.J. He, Y.D., Hart A.A.M., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen, A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530 – 536.
3. Stekel D. (2003). In *Microarray Bioinformatics,* pp. 73–130. Cambridge University Press, Cambridge.
4. Gibson and Muse (2002). In *A Primer of Genome Science,* pp. 122–181. Sinauer and North Carolina State University.
5. Collett H., Shen A., Gardner M., Farrant J.M., Denby K. J. and Illing N. (2004). Towards transcript profiling of desiccation tolerance in *Xerophyta humilis:* construction of a normalized 11 k *X. humilis* cDNA set and microarray expression analysis of 424 cDNAs in response to dehydration. *Physiologia Pl.* **122**(1), 39–53.
6. Diatchenko L., Lukyanov S., Lau Y.F. and Siebert P.D. (1999). Suppression subtractive hybridisation: a versatile method for identifying differentially expressed genes. *Methods Enzymol.* **303**, 349–380.
7. Yang G.P., Ross D.T., Kuang, W.W., Brown, P.O. and Weigel, R.J. (1999). Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acids Res.* **27**(6), 1517–1523.
8. van den Berg N., Crampton B.G., Ingo H., Birch P.R.J. and Berger D.K. (2004). High throughput screening of SSH cDNA libraries using DNA microarray analysis. *Biotechniques* **37**(5), 818–824.
9. Lipshutz R.J.S., Fodor S., Gingeras T. and Lockhart D. (1999). High-density oligonucleotide arrays. *Nature Gen.* **21**, 20–24.
10. Hughes T.R., Mao M., Jones A.R., Burchard J., Marton M.J., Shannon K.W., Lefkowitz S.M., Ziman M., Schelter J.M., Meyer M.R., Kobayashi S., Davis C., Dai H., He Y.D., Stephaniants S.B., Cavet G., Walker W.L., West A., Coffey E., Shoemaker D.D., Stoughton R., Blanchard A.P., Friend S.H. and Linsley P.S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotech.* **19**(4), 342–347.
11. Yauk C.L., Berndt M.L., Williams A. and Douglas G.R. (2004). Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32**(15), e124. doi: 10.1093/nar/gnh123.
12. Yang Y.H. and Speed T. (2003). An introduction to microarray bioinformatics, Part 1: design of microarray expression experiments. In *DNA Microarrays, A Molecular Cloning Manual,* pp. 523–525. Cold Spring Harbor Laboratory, New York.
13. Rockett, J.C. and Hellmann G.M. (2004). Confirming microarray data – is it really necessary? *Genomics* **83**(4), 541–549.
14. Wu S-H., Ramonell K., Gollub J. and Somerville S. (2001). Plant gene expression profiling with DNA microarrays. *Plant Physiol. & Biochem.* **39**, 917–926.
15. Vandesompele J., De Preter K., Pattyn F., Poppe B., Van Roy N., De Paepe A. and Speleman F. (2002). Accurate normalisation of real-time RT-PCR by geometric averaging of multiple internal control genes. *Genome Biol.* **3**(7), 00341.
16. Kerr M.K., Martin M. and Churchill G. (2000). Analysis of variance for gene expression data. *J. Comp. Biol.* **7**(6), 819–837.
17. Wolfinger R., Gibson G., Wolfinger E.D., Bennett L., Hamadeh H., Bushel P., Afshari C. and Paules R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comp. Biol.* **8**(6), 625–637.
18. Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836.
19. Yang Y.H., Dutoit S., Luu P., Lin D.M., Peng V., Ngai J. and Speed T. (2002). Normalisation for cDNA microarray data: a robust and composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**(4), e15.
20. Quackenbush J. (2002). Microarray data normalisation and transformation. *Nature Gen.* **32**(2), 496–501.
21. Futschik M. and Crompton T. (2004). Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol.* **5**(8), R60.
22. Yang Y.H. and Speed T. (2002). Design issues for cDNA microarray experiments. *Nature* **3**, 579–588.
23. Livesey F.J. (2003). Strategies for microarray analysis of limiting amounts of RNA. *Brief Funct. Genomic Proteomic.* **2**(1), 31–36.
24. Dobbin K. and Simon R.M. (2003). Experimental design of DNA microarray experiments. *Biotechniques* **34**, 16–21.
25. Shih J.H., Michalowska A.M., Dobbin K., Ye Y., Qiu T.H. and Green J.E. (2004). Effects of pooling mRNA in microarray class comparisons. *Bioinformatics* **20**(18), 3318–3325.
26. Dobbin K., Shih J.H. and Simon R. (2003). Statistical design of reverse dye microarrays. *Bioinformatics* **19**(7), 803–810.
27. Golub T.R. and Slonim D.K (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
28. Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J. and Gillanders, E. (2000). Molecular classifications of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540.
29. Churchill G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Gen.* **32**, 490–495.
30. Dobbin K. and Simon R. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* **18**, 1438–1445.
31. Dombowski A.A., Thibodeau B.J., Starcevic S.L., Novak R.F. (2004). Gene specific dye bias in microarray reference designs. *FEBS Lett.* **560**, 120–124.
32. Tempelman R.J. (2005). Assessing mixed statistical precision, power, and robustness of alternative experimental designs for two colour microarray platforms based on mixed effects models. *Vet. Immunol. Immunopathol.* **105**, 175–186.
33. Kerr M.K. and Churchill G.A. (2001). Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**, 123–128.
34. Vinciotti V., Khanin R., D'Alimonte D., Liu X., Cattini N., Hotchkiss G., Bucca G., De Jesus O., Rasaiyaah J., Smith C.P., Kellam P. and Wit E. (2004). An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* **21**(4), 492–501.
35. Huitema E., Vleeshouwers G.A.A., Francis D.M. and Kamoun S. (2003). Active defence responses associated with non-host resistance of *Arabidopsis thaliana* to the oomycete pathogen *Phythophthora infestans. Mol. Plant Path.* **4**(6), 487–500.
36. Oleksiak M.F., Churchill G.A. and Crawford D.L. (2002). Variation in gene expression within and among natural populations. *Nature Gen.* **32**, 261–266.
37. Moser J.M., Freitas T., Arasu P. and Gibson G. (2005). Gene expression profiles associated with the transition to parasitism in *Ancylostoma caninum* larvae. *Molecular Biochem. Parasitol.* **143**, 39–48.
38. Chowers I., Gunatilaka T.L., Farkas R.H., Qian J, Hackam A.S., Duh E., Kageyama M., Wang C., Vora A., Campochiaro P.A. and Zack D.J. (2003). Identification of novel genes preferentially expressed in the retina using a custom human retina cDNA microarray. *Invest. Ophthalmol. Vis. Sci.* **44**(9), 3732–3741.
39. Zien A., Fluck J., Zimmer R. and Lengauer T. (2002). Microarrays: how many do you need? In *Proc. RECOMB 2002.* Association for Computing Machinery, New York.