Fundamentals of experimental design for cDNA microarrays

Gary A. Churchill

doi:10.1038/ng1031

Microarray technology is now widely available and is being applied to address increasingly complex scientific questions. Consequently, there is a greater demand for statistical assessment of the conclusions drawn from microarray experiments. This review discusses fundamental issues of how to design an experiment to ensure that the resulting data are amenable to statistical analysis. The discussion focuses on two-color spotted cDNA microarrays, but many of the same issues apply to single-color gene-expression assays as well.

Sources of variation in microarray experiments

The design of a two-color microarray experiment can be considered as having three layers. Figure 1 shows an example of an experiment that compares the effects of two treatments—A and B—on gene-expression profiles in a mouse tissue. At the top layer of the experiment are the experimental units, the two mice to whom each treatment is applied. The term 'treatment' pertains to any attribute, such as the sex or strain of the organism, of primary interest in the experiment. The mice were selected to be representative of a population of mice and, if possible, the treatment should be assigned using a randomizing device such as a coin toss. Assigning at least two mice to each treatment group ensures that there is biological replication in the experiment. In the middle layer, two RNA samples are obtained from each mouse. These technical replicates may be two independent RNA extractions or two aliquots of the same extraction. The RNA samples are assigned to two different dye labels, indicated by the red and green test tubes. They are then paired (one red and one green) and mixed for co-hybridization on microarray slides. The bottom layer of the experiment involves the arrangement of array elements on the slides. In this example, duplicate spots of each cDNA clone have been printed side by side.

The many sources of variation in a microarray experiment can be partitioned along these three layers. Biological variation (top layer) is intrinsic to all organisms; it may be influenced by genetic or environmental factors, as well as by whether the samples are pooled or individual. Technical variation (middle layer) is introduced during the extraction, labeling and hybridization of samples. Measurement error (bottom layer) is associated with reading the fluorescent signals, which may be affected by factors such as dust on the array. Valid statistical tests for differential expression of a gene across the samples can be constructed on the basis of any of these variance components, but there are important distinctions in how the different types of tests should be interpreted. If we are interested in determining how the treatments affect different biological populations represented in our samples, statistical tests should be based on the biological variance. If our interest is to detect variations within treatment groups, the tests should be based on technical variation. For example, Olesiak *et al.*[1] employed both types of tests to look at variation between and within natural populations. Tests based on measurement error variance can also be constructed but are of limited utility[2]. For most questions of interest, the higher two levels of variation are appropriate for constructing tests, and hence good designs should incorporate replication at the higher layers.

Experimental units and treatments

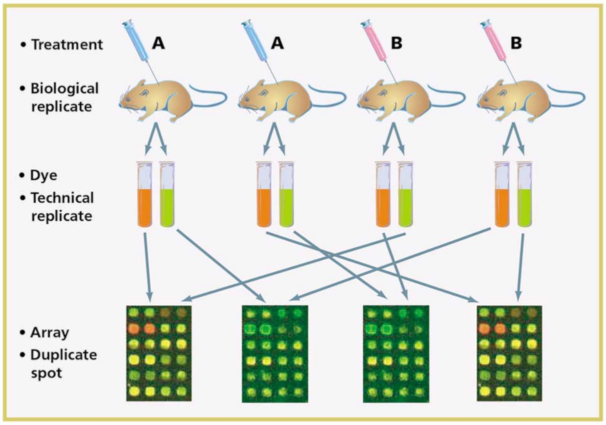
The correlation observed between ratios of fluorescent intensity from duplicate spots on a single microarray slide will typically exceed 95%. This is often interpreted as a demonstration that microarray assays are reproducible. However, if the same target sample is divided and hybridized to two different microarray slides, the correlation across hybridizations is likely to fall to the 60 to 80% range, somewhat lower if the dye labeling is reversed. Correlations between samples obtained from individual inbred mice may be as low as 30%. If the experiments are carried out in different laboratories, the correlations may be lower still.

These decreasing correlations reflect the cumulative contributions of multiple sources of variation. It is tempting to avoid biological replication in an experiment because results will appear to be more reproducible. The apparent increase in statistical power is illusory, however, and significant findings may simply reflect chance fluctuations in the particular animals chosen for the experiment. In general, it is appropriate to take steps to vary the conditions of the experiment—for example, by assaying multiple animals—to ensure that the effects that do achieve statistical significance are real and will be reproducible in different settings[3].

Identifying the independent units in an experiment is a prerequisite for a proper statistical analysis, as any hidden correlations in the data can lead to bias and inflated levels of statistical significance. Statistical independence is a relative concept. For example, hybridizations of the same target sample to multiple slides may be viewed as independent replicates if the intent is to characterize that sample accurately. However, in an experiment where the question of interest concerns a biological comparison at the whole-organism level (for example, a comparison of gene-expression profiles between genetically altered and control animals), the technical replicates from any one sample may no longer be regarded as independent.

Details of how individual animals and samples were handled throughout the course of an experiment can be important to

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA (e-mail: garyc@jax.org).



Bob Crimi

Fig. 1 A schematic representation of the three layers of design in a simple microarray experiment. In the top layer, biological units (mice) are assigned to treatment groups (A and B). In the middle layer, two RNA samples are obtained from each mouse. These technical replicates are differentially labeled and hybridized in pairs to microarrays. Each pairing involves a direct comparison of an A mouse to a B mouse, and the dye labels are reversed in two of the four comparisons. The bottom layer of the experiment design is represented by the array images, in which the duplicate spotting of clones is apparent.

identify which biological samples and technical replicates are independent. In general, two measurements may be regarded as independent only if the experimental materials on which the measurements were obtained could have received different treatments, and if the materials were handled separately at all stages of the experiment where variation might have been introduced[4].

As an example, consider a cell line that is divided into eight equal samples. Four are assigned to one treatment, and the remaining four receive a second treatment. The eight aliquots are handled separately throughout the entire experimental procedure, and each is measured in triplicate. This results in 24 total observations, but there are eight experimental units. Now consider a cell line that is divided into two aliquots, each one receiving a different treatment. The material is further subdivided into four aliquots per treatment group, each of which is processed and then measured in triplicate. Again we have 24 observations, but now there are only two independent experimental units.

A simple way to assess the adequacy of a design is to determine the degrees of freedom (df). This is done by counting the number of independent units and subtracting from it the number of distinct treatments (count all combinations that occur if there are multiple treatment factors). If there are no degrees of freedom left, there may be no information available to estimate the biological variance, the statistical tests will rely on technical variance alone, and the scope of the conclusions will be limited to the samples in hand. If there are 5 df or more, you are in good shape (see Box).

In some circumstances, a large number of experimental units may be available, perhaps more than can be measured individually, in which case we have the option to form pools of individual samples. In other cases, pooling may be a necessity owing to the limited availability of RNA. Pooling the original experimental units creates new units, the pools. Pooling can reduce the biological component of variation, but it cannot reduce the variability due to sample handling or measurement error.

In a two-sample comparison, we could consider making two large pools of all available units and measuring each pool multiple times. This is a poor design, as it does not allow estimation of the between-pool variance. By pooling all the available samples together we have minimized the biological variance, but we have also eliminated all independent replication. It is better to use several pools and fewer technical replicates.

Pairing samples for hybridizations

The ability to make direct comparisons between two samples on the same microarray slide is a unique and powerful feature of the two-color microarray system. By pairing samples, we can account for variation in spot size that would otherwise contribute to the error. However, it is often impractical to make all possible pairwise comparisons among the samples, because of cost or limitations in the amount of sample. Thus, an important

Allocating resources in a microarray experiment

The precision of estimated quantities depends on the variability of the experimental material, the number of experimental units, the number of repeated observations per unit and the accuracy of the primary measurements[4]. The basis for drawing inferential conclusion is the residual error (or mean squared error, MSE), which quantifies the precision of estimates and thus allows one to determine whether estimated quantities are significantly different in the statistical sense. In a microarray experiment, the residual error can be decomposed into three components of variance corresponding to the three layers of the design (Fig. 1). The first component is the intrinsic variation of the biological units within a treatment group, which we will denote by

$$\sigma_B^2.$$

The second component, denoted by

$$\sigma_A^2,$$

represents the variation between technical replicates and includes effects due to the extraction and labeling of RNA as well as array to array variation. The third component, denoted by

$$\sigma_e^2,$$

represents the measurement error within a single array. We note that the last two components of variance could be combined or decomposed in different ways, but this particular breakdown will serve our immediate purpose of deciding how to allocate replication and repeated measurements in a microarray experiment. In general, the MSE is the square root of the sum of these components. In multifactor experiments, however, the MSE may depend on which factor or combination of factors is being tested.

The magnitude of the residual error is an unknown quantity. The most straightforward and reliable method of estimating it is through the observed variation between independent experimental units. An important quantity for establishing the precision of the estimated MSE is the residual degrees of freedom (df). For a single-factor experiment with N animals divided into p treatment groups, the residual df are $N - p$. The effect of having to estimate the residual error, and hence of being uncertain of the true variation, is to multiply the true variation by

$$1 + \frac{1}{df}.$$

(This is effectively the difference between a z-test and a t-test.)

Although it is generally recommended to have no fewer than 5 residual df, it is quite common to see fewer in microarray experiments, even to the extreme of having no residual df at all. In the latter case, some strong (that is, questionable) assumptions about the variability in the experiment must be made in order to draw conclusions that can be generalized.

Replication and/or repetition of measurements at various levels in the experiment can increase precision. The most direct

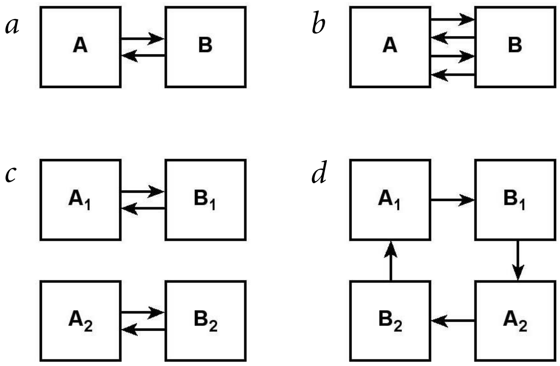


Fig. 2 Experimental designs for the direct comparison of two samples. Boxes, representing mRNA samples, are labeled as varieties A or B. Subscripts indicate the number of independent biological replicates of the same treatment. Arrows represent hybridizations between the mRNA samples and the microarray. The sample at the tail of the arrow is labeled with red (Cy5) dye, and the sample at the head of the arrow is labeled with green (Cy3) dye. This figure shows a dye swap (***a***), a repeated dye swap (***b***), a replicated dye swap (***c***) and a simple loop design (***d***). For example, in *a*, sample A, labeled red, and sample B, labeled green, are hybridized to one array, and then sample A, labeled green, and sample B, labeled red, are hybridized to another.

step in designing an experiment is to decide how many technical replicates will be measured and how these will be paired together on arrays.

Although the problem of arranging the direct comparisons may seem bewildering, following a few simple guidelines will ensure that a design is effective[5]. The efficiency of comparisons between two samples is determined by the length and the number of paths connecting them[5,6]. It is most efficient to make the comparisons of greatest interest directly on the same array. Contrasts between samples that are never directly compared in an experiment are possible, provided that there is a path of comparisons linking them. Potential biases can be minimized by balancing dyes and samples[7]. To achieve balance, create an even number of technical replicates from each biological sample and assign equal numbers of these to each dye label.

Dye swaps. A simple and effective design for the direct comparison of two samples is the dye-swap experiment[7] (Fig. 2*a*). This design uses two arrays to compare two samples. On array 1, the control sample is assigned to the red dye, and the treatment sample is assigned to the green dye. On array 2, the dye assignments are reversed. This arrangement can be repeated by using four (or six or more) arrays to compare the same two biological samples (Fig. 2*b*). This repeated dye-swap experiment is useful for reducing technical variation but should not be confused with the replicated dye-swap experiment in which independent biological samples are compared (Fig. 2*c*). The latter experiment accounts for both technical and biological variation in the assay. It may be more difficult to achieve statistical significance using the replicated dye-swap experiment, especially if the biological variation is substantial. But the advantage is that while conclusions from the repeated dye-swap experiment are limited to the samples that were assayed, those from the replicated experiment apply to the biological population from which the samples were obtained.

Reference sample designs. In the most widely used experimental design for microarrays, all the direct comparisons are made to a reference sample using the same orientation of dye labeling (Fig. 3*a*). In this design, dye effects are confounded with treatment effects[5,6]. Using two arrays in a dye-swap configuration to compare each sample (Fig. 3*b*) provides technical replication and avoids confounding of effects. Brem *et al.*[8] used this design in a study of genetic effects on transcription levels in yeast.

We and others[5,7,9,10] have argued that the reference samples are not necessary and that the practice of making all comparisons to a reference sample can lead to inefficient experiments. Fully half of the measurements in a reference experiment are made on the reference sample, which is presumably of little or no interest. As a consequence, technical variation is inflated four times relative to the level that can be achieved with direct comparisons. Despite this inefficiency, reference designs have a number of advantages. The path connecting any two samples is never longer (or shorter) than two steps; thus all comparisons are made with equal efficiency. Reference designs can be extended (as long as the refer-

method is to increase the number of experimental units. The MSE decreases in proportion to the square root of the sample size (if experimental units are costly, this can be a problem). It is also possible to increase precision by taking measurements on multiple technical replicates obtained from the experimental units. However, this approach cannot reduce the biological variance component, and the gain achieved by taking repeated measurements of single RNA samples will be limited.

Lastly, the precision of measurements obtained on a single array may be improved by printing multiple spots of the same clone. This will not impact on the other variance components. Pooling of samples is another strategy that increases precision by reducing the variability of the experimental material itself. In the absence of empirical data on the effects of pooling, we can only speculate, but it seems reasonable to suppose that the between-pool variance for a pool size of *k* experimental units will be approximately

$$\sigma^2_{pool} = \frac{1}{k^a}\,\sigma^2_B$$

for some constant $0 < a < 1$. In the case $a = 0$ pooling will have no effect, and in the case $a = 1$ the variance is reduced in direct proportion to the pool size.

To determine the optimum arrangement of an experiment, we may wish to consider the costs of various components. Let C_I represent the cost of an experimental unit and C_M be the cost of measurement for a single technical replicate. Now suppose that we have created *n* pools of *k* individuals each, and each pool will be measured using *m* technical replicates on microarray slides with *r* repetitions of each clone. The MSE of this experiment will be

$$MSE = \sqrt{\left(\frac{\sigma^2_B}{k^a} + (\sigma^2_A + \sigma^2_e/r)/m\right)\Big/n} \quad ,$$

and the cost of the experiment is cost $= n \bullet kC_1 + n \bullet mC_M$.

These formulae may be useful for planning experiments, but they depend on detailed knowledge of the variance components. Nonetheless, this exercise suggests some general guidelines for allocating the resources in a microarray experiment.

- When measurement is expensive and/or the individual measurements are very precise, it is preferable to add experimental units rather than technical replicates.
- When the variablity of measurements is greater than the variability between experimental units, technical replication and repeated measurements will effectively increase precision.
- When variability between individual samples is large and the units are not too costly, it may be worthwhile to pool samples. The effectiveness of pooling is offset if precious degrees of freedom are lost from the experiment. Pooling does not impact technical variation, and so it is recommended that several technical replicates from each pool be obtained when possible.

Lastly, it is hard to overstate the importance of independent biological replication, in the form of multiple individuals or multiple pools within each treatment group, as a means to achieve adequate power and validity for statistical tests.

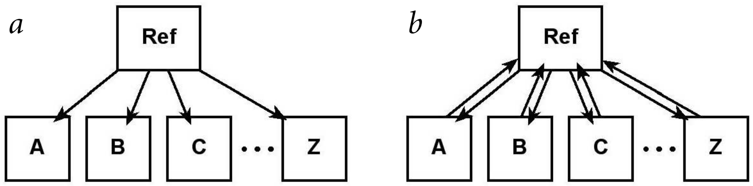


Fig. 3 Experimental designs using a reference RNA sample. Boxes represent RNA samples, and arrows represent microarrays, as in Fig. 2. ***a***, The standard reference design uses a single array to compare each test sample (A, B, C, and so on) to the reference RNA. ***b***, A variation uses a dye swap for each comparison.

ence sample is available) to assay large numbers of samples that are collected over a period of time. From a practical perspective, every new sample in a reference experiment is handled in the same way. This reduces the possibility of laboratory error and increases the efficiency of sample handling in large projects.

The most important considerations in choosing an appropriate reference sample are that it should be plentiful, homogeneous, and stable over time. Reference samples have been constructed using complex mixtures of RNA obtained from tissues or cell lines in an attempt to ensure that they can 'light up' every spot on an array. Another strategy is to form a pooled reference from the samples that will be assayed in the experiment. This ensures that every sample present in the test samples will be represented in the reference sample and that the relative amounts of each RNA species will be similar. The advantage of this approach is that we are not comparing samples that have widely different RNA concentrations, which obviates some of the difficult issues of normalization (see review by J. Quackenbush, pages 496–501, this issue)[11]. Use of a biologically relevant reference sample can also motivate the choice of this experimental design.

Loops. The simple loop design (Fig. 2*d*), in which samples are compared to one to another in a daisy-chain fashion, can be an efficient alternative to the reference design[6]. In general, small loops provide good average precision. However, depending on the goals of the experiment, large loops may be inefficient. For example, if an investigator wants to compare every pair of samples, loops become inefficient when there are more than 10 samples. In addition, the estimation efficiency of a simple loop is greatly reduced by loss of just a single array. Designs that interweave two or more loops together or combine loops with reference designs improve efficiency and robustness by creating multiple links among the samples. The difficulty presented by loop designs is that the deconvolution of relative expression values is not always intuitive. However, the availability of software tools that can analyze general designs reduces this concern.

Printing the slides

Although a thorough discussion of which probes to spot on a microarray slide is outside the scope of this review, it is a critical aspect of design. The arrangement of spots on a slide (Fig. 4) raises design issues that can impact on normalization and analysis of microarray data[12]. Groups of spots printed by the same pins and/or subarrays that may have been printed at different times lead to correlations among the spots that should be taken into account. Repeated spotting of the same clone on an array increases precision[2] of the measurements if the spot intensities are averaged. It can also minimize problems caused by scratches, dust, and other mishaps that can contaminate the surface of microarray slides. Repeated spots should be dispersed over the microarray surface to minimize correlations; however, repeated spots should always be considered as correlated observations and treated accordingly in analysis. The use of internal control spots may also help ensure the quality of the data and can provide information for calibrating the results of an analysis. Data should not, however, be standardized using internal controls. The control spots themselves are subject to random variation, and the process of standardization can induce spurious correlations in data that are otherwise uncorrelated. (This point seems to bear repeating in the literature every 50 years or so[13,14].)

Randomization

Randomization of treatment assignments and random sampling of populations form the physical basis for the validity of statistical tests[15]. It is most crucial to apply randomization or random sampling at the stage of assigning treatments to the experimental units. If the treatment is something that can be applied to the units (for example, injection of a drug), then a carefully randomized experiment can lead to causal inferences; that is, we can conclude that the drug causes the observed effects. If the treatment is already attached to the units (for example, the sex and strain of a mouse), then the conclusions of the study are limited to associations; that is, sex is associated with differential expression of, for example, androgen receptors. The valid scope of the conclusions in such studies is contingent on how well the population of interest (both in its mean behavior and its diversity) is represented by the sample of animals in the experiment. True random sampling of populations is an ideal that is difficult to achieve, but often a good representative sample can be obtained.

Randomization can be used at other stages in the microarray experiment to help avoid or minimize hidden biases. When multiple technical replicates are used, the dye assignments can be randomized. Assigning the first sample obtained to Cy5 and the second one to Cy3 has an obvious potential to introduce biases. Slides are often printed in batches that can vary in their overall quality and even within a batch, the order and position on the printing device can affect results. In the study of Oleksiak *et al.*[1], slides were numbered (1 through 48 in print order) and for each hybridization, a slide was chosen by drawing a numbered slip of paper from a hat.

Lastly, we could consider randomizing the arrangement of spots on an array. Fisher[15] warns of the potential biases that can arise due to regular arrangements. Ideally, each slide in an experiment might have clones printed in a different arrangement. But, because of the nature of the printing devices and logistics of tracking spot identities, randomization would be impractical. The possibility of position effects within the array is not far-fetched, but it may be a reality that we simply have to accept with awareness.

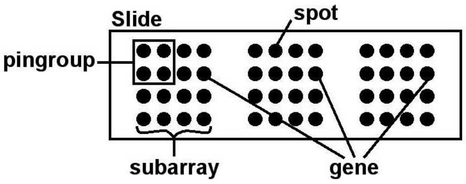


Fig. 4 Common features in the layout of a microarray slide. The fundamental units on which measurements are obtained are spots containing cDNA clones fixed to a glass substrate. Spots may vary in size, shape and concentration of DNA. Robotic printing devices used to generate the spots often work with multiple printing tips or pins. Printing may also be done in blocks (subarrays), which were perhaps printed on different days. The same clone (gene) may be printed multiple times on a single slide.

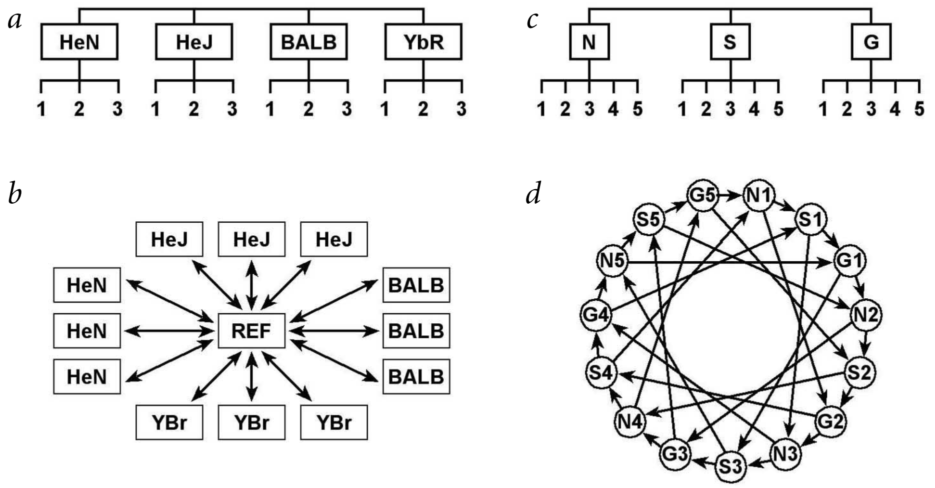


Fig. 5 Two examples of one-way classifications with replication. ***a***, In the tumor survey, three tumor samples (1–3) were obtained from mice belonging to four strains (HeN, HeJ, BALB, YbR). ***b***, In this survey, each sample was compared using a dye swap to a common reference sample. ***c***, In the fish population study, samples were obtained from fish in each of three populations (N, S and G). ***d***, Direct comparisons among the samples were arranged as a series of loops.

Analysis

This review has not touched on issues of analysis, which are, however, discussed elsewhere in this supplement (see review by D.K. Slonim, pages 502–508, this issue)[16]. A well-designed experiment will often suggest a suitable method of analysis. Analysis-of-variance models for microarray data were introduced by Kerr *et al.*[17], and this framework has been extended to account for correlations and multiple sources of variance[18,19]. Software tools are available for general-purpose analysis of experiments with multiple sources of variance[14], but proper application of these methods is often not trivial. Until standards of microarray design and analysis evolve further, we recommend that analysis should be carried out in collaboration with a statistician.

Examples

The following examples of strategies for microarray experiment design illustrate some of the points discussed above. Each of these experiments has its strengths and weaknesses from a design perspective, and the commentary provided here is intended to highlight these points. More examples of designs that use both direct and reference sample comparisons can be found at http://www.tigr.org/pga.

Pairwise comparisons. Callow *et al.*[20] carried out experiments to identify differentially expressed genes between genetically altered strains of mice and wildtype controls. RNA samples were obtained from eight transgenic and eight control mice, providing 14 residual df (16 mice minus 2 treatments) for testing the treatment effect. Each of the 16 samples was labeled with Cy5 and compared to a Cy3-labeled pool of RNAs from control mice. An alternative design for this situation would have been to arrange eight direct comparisons as dye swaps between transgenic and wildtype mice. This alternative design provides technical replication, avoids dye bias and does not require a reference sample.

Kerr *et al.*[21] describe an experiment comparing gene expression in two cell lines, one of which has been treated with a toxin. Three aliquots were obtained from each cell line and directly compared on pairs of microarrays using dye swaps. The experiment lacks biological replication because the aliquots are not independent. With the extensive technical replication in this experiment, relative differences as small as 12% are detectable. A large number of differences were detected in this experiment. While some are certainly biologically relevant, others may reflect chance fluctuations between the two cell lines and may not be reproducible in repetitions of the experiment. The design could be improved by using six independent cell lines with random assignment of three to the toxin treatment.

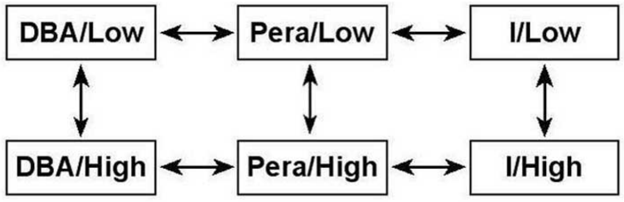


Fig. 6 Experimental design for a 3 × 2 factorial experiment. Direct dye-swap comparisons are made within strains between low- and high-fat diet conditions. Comparisons between strains are restricted to strains susceptible to gallstones (Pera) versus resistant (DBA and I) strains.

One-way classifications. Figure 5 illustrates designs of two different experiments involving a single multi-level treatment factor. The first design is from a survey of mouse mammary tumor samples (G.A. Churchill, unpublished work). The treatment factor is strain, with four levels and each strain represented by three independent tumor samples, providing 8 df for testing differences among strains. RNA from each tumor was compared directly to a reference sample using two arrays in a dye-swap arrangement.

The second design is from a study of variation in natural populations of teleost fish[1]. There are three populations, and five fish were sampled from each, providing 12 df to test for differences between the populations. In this experiment, the direct comparisons were arranged as loops. Each sample was measured using four technical replicates, and dye assignments were balanced. In both examples, within-treatment group differences can be tested relative to technical variation. These examples illustrate how experiments with similar structure at the biological level can be arranged using very different pairing strategies among the technical replicates.

Experiments with multiple factors. A study was carried out to compare gene expression in liver tissues of mice from a gallstone-susceptible strain (Pera) and gallstone-resistant strains (DBA and I) on low-fat and high-fat diets (Fig. 6*b*; B.J. Paigen, pers. comm.). Each of the six combinations in this 2 × 3 factorial experiment was represented by two independent pools of three mice providing 6 df (12 pools minus 6 groups) for testing biological effects. Direct comparisons among the strains in this study are restricted to resistant versus susceptible. The two resistant strains can still be contrasted, but with less efficiency than the comparisons of interest.

Jin *et al.*[9] studied expression patterns in two strains of *Drosophila*, using both sexes and two ages. Several hundred flies representing each of the eight combinations of these factors were used to create pooled RNA samples. Twenty-four microarrays were used to compare 48 independent labeling reactions, six per pool, obtained from these RNA samples. All direct comparisons were made between the two age groups using six arrays, two with dye labeling reversed. There are no direct comparisons between flies that differ in sex or strain. This design reflects the primary interest of the researchers in the effects of aging on gene expression and a secondary interest in other factors. In the analysis of this interesting design, different error terms are used to test the age, sex, strain, and sex-by-strain interaction effects[9]. An alternative design that includes replicate pools and direct comparisons across all of the treatment factors to achieve equity in precision of comparisons was proposed by Churchill and Oliver[22]. These examples show that the arrangement of pairings in a multifactor experiment can differentially impact on the precision of different comparisons.

Conclusions

In summary, the problem of designing a microarray experiment can be decomposed into three distinct layers. First, replication of biological samples is essential in order to draw conclusions that are valid beyond the scope of the particular samples that were assayed. Second, technical replicates increase precision and provide a basis for testing differences within treatment groups. Third, duplication of spotted clones on the microarray slides increases precision and provides quality control and robustness to the experiment. Full disclosure of the details of sample preparation and handling is important to help identify the independent units in an experiment and to avoid inflated estimates of significance or artifactual conclusions.

Statistical quantification of evidence is widely accepted as a standard requirement in scientific investigation and is preferred over the qualitative description of observations. A carefully designed experiment provides a sound basis for statistical analysis and lends itself to simple and powerful interpretation. Putting experimental design principles into practice is not difficult, and there are often several design alternatives that will work well for any given situation. The following are some important points to keep in mind.

Use adequate biological replication. A common mistake is to generate an excess of technical replication with little or no independent replication of the biological samples. This is akin to studying the difference in heights of the two sexes by repeatedly measuring one man and one woman.

Make direct comparisons between samples whose contrasts are of most interest and use short paths to connect any samples that might be contrasted.

Use dye swapping or looping to balance dyes and samples.

Always keep the goals of the experiment in mind. Experiments that are constructed to address a particular question are more likely to be simple and interpretable compared to experiments compiled from a haphazard set of conditions.

Acknowledgments

Support for this work was provided by the US National Institutes of Health. The analogy of measuring one man and one woman is attributed to Peter Petraitis.

1. Oleksiak, M.F., Churchill, G.A. & Crawford, D.L. Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261–266 (2002).
2. Lee, M.T., 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).
3. Rosenbaum, P.R. Replicating effects and biases. *Am. Stat.* **55**, 223–227 (2001).
4. Cox, D.R. *The Design of Experiments* (Wiley, NY, 1958).
5. Kerr, M.K. & Churchill, G.A. *Biostatistics* **2**, 183–201 (2001).
6. Yang, Y.H. & Speed, T.P. *Nature Rev. Genet.* **3**, 579–588 (2002).
7. Kerr, M.K. & Churchill, G.A. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**, 123–128 (2001).
8. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
9. Jin, W. *et al*. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.* **29**, 389–395 (2001).
10. Simon, R., Radmacher, M.D. & Dobbin, K. Design of studies using DNA microarrays. *Genet. Epidemiol.* **23**, 21–36 (2002).
11. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32**, 496–501 (2002).
12. 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).
13. Tanner, J. Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *Appl. Physiol.* **2**, 1–15 (1949).
14. Pearson, K. Spurious correlation between indices. *Proc. R. Soc. Lond.* **60**, 489 (1897).
15. Fisher, R.A. *The Design of Experiments* 6th edn (Oliver and Boyd, London, 1951).
16. Slonim, D.K. From patterns to pathways: gene expression data analysis comes of age. *Nature Genet.* **32**, 502–508 (2002).
17. Kerr, M.K., Martin, M. & Churchill, G.A. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**, 819–837 (2000).
18. Wolfinger, R.D. *et al.* Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625–637 (2001).
19. 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).
20. Callow, M.J. *et al*. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.* **10**, 2022–2029 (2000).
21. Kerr, M.K. *et al*. Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sinica* **12**, 203–218 (2002).
22. Churchill, G.A. & Oliver, B. Sex, flies and microarrays. *Nature Genet.* **29**, 355–356 (2001).