Review: Practical Design and Analysis of 2-Colour cDNA Microarray Experiments

Routley, Ben and Muldoon, Mark R.

2007

MIMS EPrint: **2007.105**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Review: Practical Design and Analysis of 2-Colour cDNA Microarray Experiments

Ben Routley and Mark R. Muldoon

Abstract

This review paper, is aimed at biological researchers who are interested in or have begun to use cDNA microarrays for their investigations. Large microarray studies typically involve a multi-disciplinary team with various groups performing different aspects of the same experiment. This approach means that microarrays are less accessible to new researchers than more traditional biological techniques. This review aims to make current techniques of statistical design, normalisation and linear analysis of cDNA microarray experiments accessible to a wider community. These methods will be illustrated with examples that use freely-available packages implemented in Bioconductor and R.

Background

Identifying gene transcript (mRNA) regulation has been largely possible due to the increasing use of DNA microarrays or gene chips; first developed at Stanford University (Schena *et al.* 1995). They provide a vehicle for exploring genomic transcript levels and are one of only a few methods that are able to identify novel changes in gene transcripts. The method is based on the phenomenon of preferential complementary base pairing, known as hybridisation, and produces its signal by parallel hybridisation of labelled targets to specific probes that have been immobilised on a solid surface in an ordered array. Each *probe* corresponds to either a complete transcript or to part of a transcribed sequence which is tethered onto the array and the *target* is a labelled pool of DNA that is complementary to mRNA. The most versatile micoarray platform is cDNA microarrays in which cDNA sequences are printed on a glass slide using a robotic "arrayer".

A typical dual-hybridisation study compares two samples (or targets) on the same microarray slide (or probe) enabling the biologist to make direct, quantitative comparisons between two expression patterns. The samples are said to be derived from different *treatments* where the term treatment encompasses, for example, conditions or time-points. Each transcript from the two treatment groups is tagged with a distinct fluorescent dye and hybridised to the probes on the array. Fluorescence measurements are then recorded using a scanner to yield a quantitative estimate of expression for each treatment and for every probe. These are often combined to determine ratios or relative abundances of transcripts. The most common dyes for microarray studies, Cy3 and Cy5, have non-linear sample labelling and hybridisation kinetics, which means that they do not provide equal sensitivity across the whole range of transcripts in a sample. More specifically, they have differential labelling and scanning efficiencies and also exhibit gene-specific bias (Tseng *et al.* 2001). To combat this, the roles of the dyes are often exchanged and the procedures of hybridisation and scanning repeated, known as a *dye-swap*. Taking a suitable average of both dye-swap pair ratios removes dye-bias, giving more reliable results. If a dye-swap has not been performed, gene-specific dye-bias cannot easily be removed. The contribution and cause of gene-specific dye-bias to the underlying variation has not been properly characterised however there has been recent research in this area aimed at modelling this effect (Dobbin *et al.* 2005; Martin-Magniette *et al.* 2005).

An alternative microarray technology uses oligonucleotide chips. These were pioneered by Affymetrix, who developed a method for synthesising high density arrays using photolithography, directly on the chip (Lipshutz *et al.* 1999). Design and analysis of cDNA microarrays differ significantly from oligonucleotide chips so only the former will be considered here. Please refer to the Nature supplements The Chipping Forecast I and II for a more detailed introduction.

Logarithmic transformation of fluorescent intensities is widely accepted to be the first step in normalisation, making subsequent interpretation easier. The raw intensity ratios may not reflect variations in expression levels accurately because, for example, a gene up-regulated by a factor of 2 will show a ratio of 2 but a down-regulation by the same factor will appear as 0.5. By using the logged intensity ratios, up-regulation and down-regulation are treated symmetrically.

up-regulated by a factor of 2	e.g. 4:2 = 2 (ratio of 2)	log2(2) = 0.69
down-regulated by a factor of 2	e.g. 2:4 = 0.5 (ratio of 0.5)	log2(0.5) = -0.69

Table 1. Taking the log of expression ratios, produces a symmetric scale.

Logarithmic transformation instantly makes the data easier to handle and interpret. There has been much research to accurately predict the distribution of spot-intensities using variance-stabilising transformations (Durbin *et al.* 2002) however a log-normal distribution is a good approximation of the bulk of microarray data (Hoyle *et al.* 2002). It has been shown that log-transformations stabilise the variance at high levels, but highly variable results can be expected at low expression levels (Kerr and Churchill 2001). If the purpose of an investigation is to determine those genes that are highly expressed, log transformation is thus a suitable approach. If however one wants a more sensitive estimate of weakly expressed genes, a *variance stabilising transformation* may be necessary (Durbin et al. 2002). The choice of base value for the log is not that important, as long as it is consistent throughout an experiment. Many groups use log 10 or natural log (ln), however log to the base 2 is a appealing choice because genes up- (or down-) regulated by a factor of two will appear as a log change of 1 (or –1, respectively).

Sources of Variation

Before an experiment can be designed, the sources of variation to which it is prey must be identified. Experimental variation can be divided into *technical* and *biological* sorts and, in addition, one should allow for a separate Measurement Error. To quantify these sources of variation, it is necessary to perform independent replication.

Technical Variation

Technical variation is introduced during the extraction, labelling and hybridisation steps (Churchill 2002). Technical variation is also known as systematic variation because it arises within the specified methods. Due to the large number of steps in a microarray experiment there are many areas where variation can be introduced. The technical or systematic errors can be increased in two ways (Cochran and Cox 1992):

1. Additional fluctuations of a random nature. Such fluctuations should reveal themselves in the estimate of error.
2. Consistently biased technique. Estimates of error cannot detect biases because they are measures of precision rather than accuracy. Bias can be reduced using an accurate protocol and through randomisation.

In order to prevent any treatment from systematically biasing the results, treatments must be randomly assigned to arrays. This can be done simply with a chance device such as a coin or die, though the researcher may use their judgement in omitting randomisation where there is real knowledge that the results will not be affected (Cochran and Cox 1992).

Biological variation

Biological variation is intrinsic to all organisms and in many cases represents the focus of the experiment. In the case of transcript regulation, variation at this level is the information the experimentalist wishes to determine. In response to treatment or environmental factors, all organisms exhibit biological variation, characterised by differential gene expression.

Measurement error

Measurement error (ME) is an estimate of the repeatability or precision of the measurements. Independent sampling or replication of the experiment steadily decreases the error associated with the difference between the average results of the treatments, provided there is no bias in the technique (Cochran and Cox 1992).

Replication

Independent sampling or replication is a relative concept that depends on the objectives of the experiment. For example, hybridisations of the same target sample to multiple slides may be seen as independent replication if the intent is to characterise that particular sample accurately (Churchill 2002). One might refer to such a procedure as *technical replication* as it will provide information about the variation that arises when the same sample is analysed many times. But if one wishes to characterise the underlying biological variation in addition to technical replicates, one will also need *biological replicates:* multiple samples drawn from biologically distinct sources (*e.g.* from distinct individuals distinct culture dishes).

Due to the length and complexity of a microarray experiment, it is crucial to check that the results were not obtained by mere chance fluctuations, but rather arise from genuine underlying biological variation. Technical replication can be used to obtain an average measurement from each probe or to quantify systemic variation. It has been estimated that at least three technical replicates are required precisely to estimate differential gene expression in a single sample (Lee *et al.* 2000). With limited resources, it may not be possible to perform technical replication on each sample, in which case it may be preferable to carry-out additional biological replication. The costs can also be significantly reduced by not performing a repeated dye-swap (technical replication) on each sample: this will be outlined in more detail below.

Biological replication is necessary to quantify independent biological samples to provide statistical confidence for the estimated gene expression profile. Averages over multiple independent biological replicates converge to a more precise estimate of the underlying level of expression and the measurement error gives confidence to these values by describing the genuine biological variation. To obtain an expression profile that is representative of the larger population, it is necessary to perform biological replication at least twice. Such independent biological replication can increase the cost of an experiment, but is essential if one wishes to characterise a population from multiple samples rather than a single specific sample. Researchers must always consider the cost of repeating the experiment using biological replicates if this is the goal. The purpose or objectives of an experiment are crucial to determine the choice of replicates and experimental design. For example, experiments to verify the precision of a laboratory technique (technical variation) will be constructed differently from those that aim to characterise gene expression over several treatments (biological variation).

The choice of whether to perform technical replication at the expense of biological replication may also depend on the variance of the measurement error obtained from technical replication. If pilot studies show the variation of measurements from technical replicates is substantial, it can be assumed the experimental technique has not been optimised and is an important source of variation. In this case, resources should be put into obtaining more precise measurements by using technical replicates at the expense of analysing fewer independent biological samples. If however the variance of measurements from technical replicates is small, the technique appears to be consistent, so more resources could be put into performing biological replicates. The variation of biological replicates can also influence the choice of replicates in an

experiment. If the samples are taken from a biologically homogenous (say, monoclonal) population, biological replication will not give any further information so it may be preferable to concentrate on obtaining precise estimates through technical replication.

Pooling

Due to the instability of mRNA, it can be difficult to extract sufficient material for hybridisation, especially if the sample is to be spread over several replicates. The mRNA required for even a single array may prove unachievable for small organisms. In such a circumstance, the mRNA from several samples could be pooled to make up the volume needed, but this practical constraint may alter the objectives of the investigation. After pooling, the experimentalist would no longer be able to make inferences about the individual samples, but only about the population from which they were drawn. This restriction may not be too important because often the purpose of analysing individual samples is to make inference on the population. On the other hand, when one wishes to chracterize a population, pooling can reduce the overall costs of an experiment because arrays are usually more expensive than the generation of the bio-samples (Peng *et al.* 2003). The cost of an experiment can be substantially reduced by measuring the same number of pooled samples on a smaller number of arrays.

Pooling multiple replicates will also have the effect of decreasing the population variance and diminishing random fluctuations. The researcher must decide at the outset whether the potential loss of information about individuals is out-weighed by the increase in statistical power and cost efficiency of design (Peng et al. 2003).

Experimental Design

Thorough experimental design is essential to facilitate statistical inference during the latter stages of the process. This field has been the subject of rigorous research over the past fifty years therefore there are some common techniques that are now being applied to microarrays. An efficient design is one in which; 1) assignment of dyes and samples (targets) is arranged so the effects of the dyes are eliminated and 2) comparisons of interest can be made using the smallest possible number of arrays , thus minimising the error variance of the comparisons (Bretz *et al.* 2003).

Comparing Two Treatments

Due the parallel nature of dual-hybridisation microarrays, the most efficient design to compare two samples, A and B, is to directly compare them on the same array (Yang and Speed 2002; Dobbin *et al.* 2003; Smyth *et al.* 2003). This approach allows the researchers to examine the relative abundance of the two samples on the same array. The log expression ratio fore the *i*-th gene, $\log (1{:}2)_i$ is calculated directly from the ratio of the logged intensities for each probe $_i$.

$$\log_2 (1{:}2)_i = \log_2 (\text{sample1}_i^{\text{red}} / \text{sample2}_i^{\text{green}})$$

In the analysis that follows , log ratios for direct comparisons of measurements from the same array will be taken to have a nominal variance of 1.

$$\sigma^2 (1{:}2)_i = \sigma^2 (\text{sample1}_i^{\text{red}} / \text{sample2}_i^{\text{green}}) = 1$$

As mentioned above, dye effects can influence measured expression levels so, where possible, a dye-swap should be performed on the same two samples. Taking the mean log expression ratio on each probe for both dye-swaps will remove the effects of the dye.

$$\log_2 (1{:}2)_i = 1/2 (\log_2 (\text{sample1}_i^{\text{red}} / \text{sample2}_i^{\text{green}}) + \log_2 (\text{sample1}_i^{\text{green}} / \text{sample2}_i^{\text{red}}))$$

The log ratio is estimated using two independent measurements, which has the effect of halving the variance:

$$\sigma^2 (1{:}2)_i = \sigma^2 (\text{sample1}_i^{\text{red}} / \text{sample2}_i^{\text{green}}) + \sigma^2 (\text{sample1}_i^{\text{green}} / \text{sample2}_{i\,\text{red}}) = 1/2$$

In the vocabulary of classical statistics, a dye-swap design is a *complete block design*, taking the form of a *2x2 Latin Square*. The arrays and dyes in a dual-hybridisation microarray experiment are homogenous so the experiment can be partitioned into these groups. This simple design plan eliminates dye and array effects from the measurements.

	Red dye (cy5)	Green dye (cy3)
Array 1	Sample1	Sample2
Array 2	Sample2	Sample1

Table 2. A Latin Square design to compare two samples (S1 and S2) directly.

Comparing More Than Two Treatments

Reference Designs

When there are more than two treatments to be compared, not every treatment can appear on every array (Kerr and Churchill 2001) and so comparisons must be indirect: this is known as an *incomplete block design*. In this circumstance, each sample could be compared to a reference sample. Of great importance is the choice of the reference sample, which should be plentiful, homogeneous and stable over time (Churchill 2002). The reference must contain transcripts for all of the probes on the array, otherwise non-represented probes will have a value of zero for the reference channel, resulting in a loss of information for that probe (Sterrenburg *et al.* 2002). As an alternative to RNA, genomic DNA (gDNA) is being used more frequently as a reference standard, particularly for organisms with a small genome (Williams *et al.* 2004).

The plan for a reference design can depend on whether the experimentalist is interested in the reference measurements themselves or whether they are only used as a vehicle to analyse other samples. If the reference is being used as a mere vehicle to allow comparisons between the other samples, it may not be necessary to perform a dye swap for each hybridisation (Tseng *et al.* 2001; Sterrenburg *et al*. 2002). With unlimited resources, a dye-swap should be performed because it corrects for dye-bias (see Pritchard *et al,* 2001 and Zhou *et al,* 2003) and reduces the variation of the log ratios (Yang and Speed 2002) however, an efficient design can be found using only single orientation arrays. Examples of this strategy use one dye to label the reference and the other dye to label the samples of interest. Critics of this design note that it is impossible to separate the systemic bias in each dye from the variation in the sample: dye effects are confounded with sample effects (Dombkowski *et al.* 2004). But with a limited number of arrays and resources, single orientation strategies can produce valuable designs because the dye bias will affect each sample equally so will cancel-out when comparisons are made between samples. A design plan of this type also makes analysis easier and more robust (Dobbin *et al.* 2003).

If the experimentalist *is* interested in the measurements collected on the reference, each hybridisation should be performed as a dye-swap in order to eliminate dye-bias.

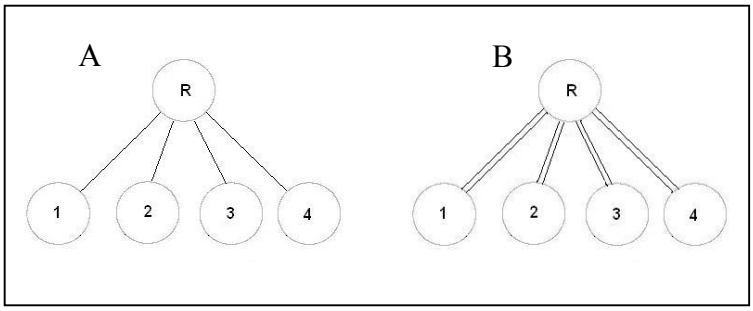


Figure 1 shows two reference designs. The plots represent the treatments as circular nodes, which are connected to other nodes via lines, with each line representing an array. In design A, we are not interested in information collected on the reference R and so need not perform dye-swaps. In design B, we *are* interested in the reference

measurements and so each comparison is performed as a dye-swap.

A major advantage of reference designs is that they are easily extended (Kerr *et al.* 2000): additional treatments can be added to the experiment by comparison to the reference sample. A major drawback of reference designs is their extreme asymmetry: half of the measurements in the experiment are devoted to the reference, a sample which may be of little or no intrinsic interest (Kerr and Churchill 2001; Vinciotti *et al.* 2004).

Loop Designs

A *loop design* involves comparing treatments to one another in a daisy-chain fashion (Churchill 2002; Yang and Speed 2002). Loop designs collect twice as much data on the treatments of interest (when compared to the reference design), resulting in more precise estimates (Kerr and Churchill 2001; Vinciotti et al. 2004). A practical problem with a loop design is that if a slide fails or is damaged, the precision of many other estimates can be strongly affected. This is a major consideration because of the many stages of a microarray experiment, at each of which problems could occur.

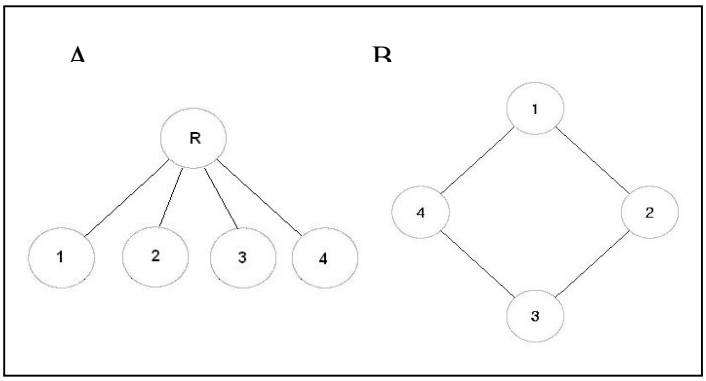


Figure 2 contrasts a reference design (A) with a loop design (B). Half of all the measurements in A are collected on the reference sample, R, while design B compares each sample directly to another. In B each sample is measured twice, so the resulting estimates of log expression ratio will have greater precision.

It is not necessary to perform a dye-swap on each comparison in a loop design however this will increase the overall robustness and precision of the estimates. An efficient design can also be achieved by *balancing* the dyes with respect to samples in the experimental plan. A balanced dye design is one in which half of each sample is labelled with each of the two fluorescent dyes and their input is balanced in the design so that the numbers and orientation of the dyes are equal. This can be more efficient than a plan which uses a dyes-swap on each comparison because a balanced dye approach still removes gene-specific-dye bias and, for a fixed number of arrays, allows more independent biological samples to be measured (Dobbin et al. 2003).

For a small number of treatments, the *standard loop design* is most efficient in that it results in the least amount of variance of per-spot differential expression estimates when compared to reference designs (Churchill 2002; Yang and Speed 2002; Wit and McClure 2004). For example, if an experimentalist wishes to obtain the log ratio between sample 1 and 2 (log(1:2)), using a reference design, it would have to be calculated through the difference of the two log ratios of sample 1 with sample R (log(1:R)) and sample 2 with sample R (log(2:R)), both assumed to be independent random variables with normal distribution. See table 3. The variance of the difference for two independent comparisons is $2\sigma^2$, because variances from independent distributions are additive. Loop designs do not use a reference to make comparisons, so all ratios are estimated directly or through a combination of direct comparisons. Using a loop design, the same comparison log(1:2) has a measurement error variance of σ^2: half that of a reference design. The number of paths separating two treatments in a loop design can inflate the variance considerably. For example log(1:3) may be estimated as a linear combination of log(1:2) and log(2:3), increasing the total variance to $2\sigma^2$.

Design	Log Ratio	Variance

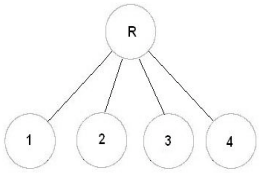
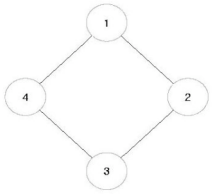
	$\log(1:2) = \log(1:R) - \log(2:R)$ $\log(1:3) = \log(1:R) - \log(3:R)$	$2\sigma^2 = \sigma^2(1:R) + \sigma^2(2:R)$ $2\sigma^2 = \sigma^2(1:R) + \sigma^2(3:R)$
	$\log(1:2) = \log(1:2)$ $\log(1:3) = \log(1:2) + \log(2:3)$	$\sigma^2 = \sigma^2(1:2)$ $2\sigma^2 = \sigma^2(1:2) + \sigma^2(2:3)$

Table 3 compares the efficiency of two different microarray designs.

To increase the efficiency of loop designs, the log-ratios can be computed by combining more than one estimate, each of which arises from an independent path (two paths are said to be *independent* if they do not have any lines of nodes in common) through the design's diagram (Khanin and Wit 2004). So, for example, log(1:2) is best estimated via a weighted combination of the directly measured log(1:2) and also via log(1:4), log(4:3) and log(3:2) (see table 3):

$$\text{log ratio}(1:2) = 3/4(1:2) + 1/4((1:4) + (4:3) + (3:2))$$

This *optimal variance estimator* yields an estimate with variance:

$$\text{var}(1:2) = (3/4)^2 + (1/4)^2 = 0.75$$

Standard loop designs become less efficient (when compared to reference designs) as the number of treatments increase because the variance of an indirect comparison depends on the number of steps in the path to which it corresponds. For comparisons between samples at opposite sides of the loop there must be many intermediate comparisons, each contributing to an inflated variance. This problem can be overcome by choosing designs that interweave two or more loops together or combine loops with reference designs (Churchill 2002,Vinciotti, 2004 #28, Khanin, 2004 #39). These are known as *interwoven loop designs* and can greatly improve the efficiency and robustness of standard loops by creating multiple links among the samples. Choosing an optimal design for such interwoven loop experiments is not a trivial task. but an be accomplished with the help of (freely-available) software or with assistance from statisticians.

For example, members of BBSRC/EPSRC UK funded MARIE consortium have developed an online-tool for choosing an optimal design, available from http://exgen.ma.umist.ac.uk/. The tool is designed to find optimal designs for two-channel microarrays, given a number of arrays and treatments (Wit and McClure 2004). It is based upon functions from the smida package (Wit and McClure 2004), designed for use with the free statistical package R (R Development Core Team 2005) and available at http://www.stats.gla.ac.uk/~microarray/book/smida.html.

The analysis of results from loop designs can also be difficult as the de-convolution of relative expression values is not always intuitive (Churchill 2002). Members of the MARIE consortium have also developed an on-line companion to the design tool mentioned above that recreates the design plan and uses it to provide optimal estimates of the ratios of expression level between different treatments. This will be discussed in greater detail below.

Transformation and Normalisation

Normalisation is an essential step for microarray analysis because the objects of study are usually relative levels of expression. Also, the process is necessary to adjust the data to remove any known bias that has been introduced during the experimental procedure.

The two main effects that need to be normalised are dye and spatial effects. It is widely accepted that normalisation should first be applied to each array (within-slide normalisation) and then to all the arrays in a complete experiment (between-slide normalisation).

Before the intensities are normalised, background correction is often employed in much of the literature to give a "true" indication of the signal intensity. It may seem natural to subtract the background intensities from the foreground however, as each signal—the background as well as the foreground—has an associated noise, background subtraction may increase the variance of the estimated ratio. Recent research in this area suggests that the background signal should be treated as an independent variable (Konishi 2004) and should not be corrected (Qin and Kerr 2004). A second problem with background correction is that the estimated background signal may exceed some spot intensities; producing negative signal intensities. Depending on the choice of background estimate, negative expression levels may be observed for as many as half the unexpressed genes, which may constitute a large part of the data (Durbin et al. 2002; Qin and Kerr 2004). Finally, the elimination of this step simplifies the experimental process allowing greater time for data analysis.

Within-Slide Normalisation

Spatial normalisation. Spot intensities can have significant variation due to the probe's location on the array. Particular areas of an array can be affected by "cover slip effects" and other position-dependent sources of variation so that there is variability on the array surface or local differences in hybridisation efficiency across the array (Quackenbush 2002). To give more accurate estimates, these spatial effects can be removed by smoothing the whole array so that the variation in spot intensities across the array is uniform. This type of normalisation removes the spatial trend over the whole array and corrects for more variable regions caused by systemic effects. For a more detailed overview, see Wit and McClure 2004. (See supplementary material figure 2.12.)

Intensity normalisation. Dye-bias can significantly bias the observed expression levels so it is necessary to normalise-away these systemic effects. Spot intensity also plays a role in introducing bias into the data: it manifests itself through a curved *M-A plot* as shown in figure 3. . The MA-plot is a 45° rotation of a scatter plot, increasing the space available to represent the range of values, see figure 5 (Yang *et al.* 2002).

log ratio (M)	$M_i = \log_2 (sampleA_i^{red} / sampleB_i^{green})$	$M = \log_2 (sampleA_i^{red}) - \log_2 (sampleB_i^{green})$
Mean log intensity (A)	$A = 1/2(\log_2 (sampleA_i^{red} * sampleB_i^{green}))$	$A = 1/2(\log_2 (sampleA_i^{red}) + \log_2 (sampleB_i^{green}))$

Table 4. Construction of M and A values from each spot for MA-plot.
(See supplementary material figure 2.3.).

A typical M-A plot of the non-normalised intensity ratios (3b) shows both the asymmetric dye-effects and that these effects are dependent on intensity. To remove intensity-dependent bias, Loess normalisation can be used to give consistently normalised data (Yang et al. 2002; Smyth et al. 2003). Loess corrects the curvature of the M-A plot by estimating, for each mean log intensity (A) separately, the dependence of the log ratio (M) on A: one then subtracts this function, A(M), to correct the log ratio (M) values for each spot (Quackenbush 2002). This has the effect of smoothing the M-A plot so that the log-ratio is linearly dependent on intensity.

$$M_{i\,norm} = M_{i\,raw} - A(M_i)$$

where $M_{i\,raw}$ is the raw log ratio, $M_{i\,norm}$ the normalised log ratio for each probe, *i.*

Intensity normalisation does not attempt to correct for gene-specific or spatial bias. (See supplementary material figure 2.6 and 2.11.)

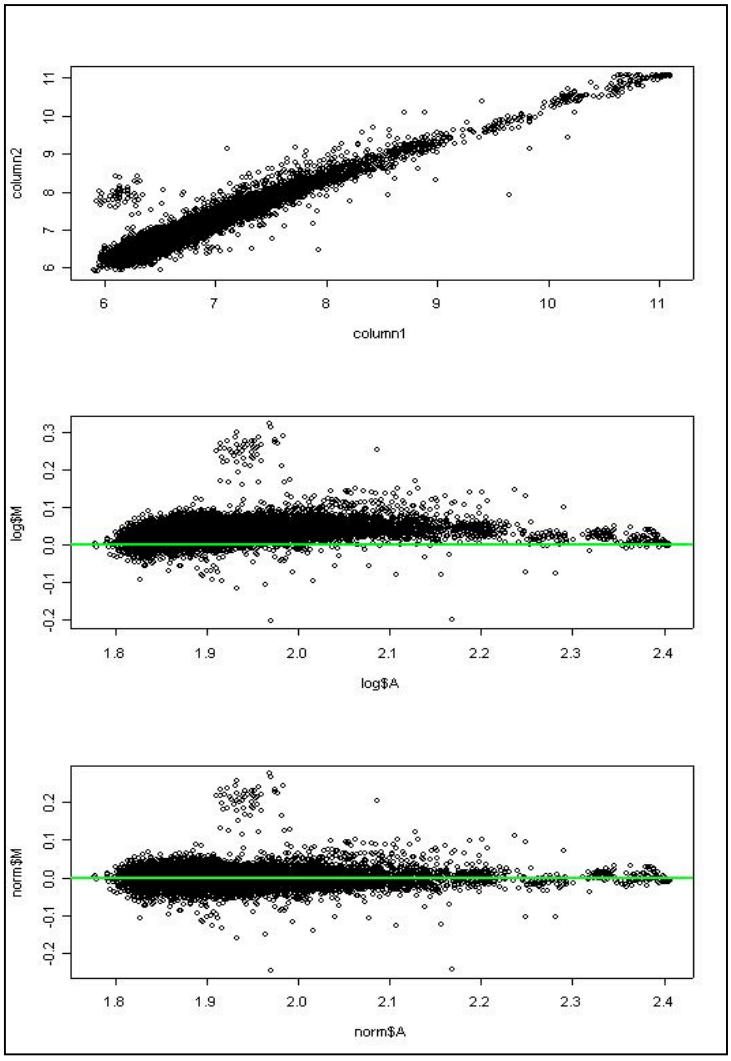


Figure 3, shows three illustrations of two channels from a single array. 3a is a scatter – plot of the raw intensities however most of the data is condensed into a small area on the plot. The MA-plot of raw intensities, 3b, spreads the data over a wider area on the graph so it is possible to identify the slight curvature due to intensity effects. A plot of the normalised intensities, 3c, shows the data to be intensity-normalised.

The relationships between treatment groups are very important for choosing the approach for intensity normalisation. If the bulk of the data is assumed to be non-differentially expressed, between two closely related treatment groups, then the Lowess algorithm can be applied to the dataset as a whole. If, however, one is comparing two very different treatment groups and so expects to observe a great many differentially expressed genes, an invariant set of probes—that is, a set of probes whose expression level is not expected to vary despite the large differences between the two treatments—should be used to estimate the normalisation function. Using an invariant set of genes may have the result of increasing the noise in the data (Yang et al. 2002) so it is necessary to choose a large set of invariant probes to counterbalance any natural variation not due to dye effects (Wit and McClure 2004) .

Print-tip-group normalisation. cDNA's are deposited onto the slide in blocks known as print-tip-groups. The groups consistently print particular regions or grids on the array and it is to be expected that some systematic bias will be introduced at this level. Variation at the print-tip level may be caused by inconsistencies in the length and opening of the tips or deformation after hours of printing (Yang et al. 2002). Normalisation by print-tip group is a useful approach because it not only normalises for each print-tip group but also serves as a proxy for spatial normalisation and is often used to correct for both types of bias together. An invariant set of genes should be used to obtain the normalisation function if it is expected that the bulk of the data will be differentially expressed.

The process estimates $A_i(M_i)$, from a Lowess fit to the M-A plot for each print-tip-group.

$$M_{i\ norm} = M_{i\ raw} - A_j(M_i)$$

where A*j* is the *j*th print-tip group and M*i* the *i*th probe. (See supplementary material figure 2.4, 2.6b and 2.8.).

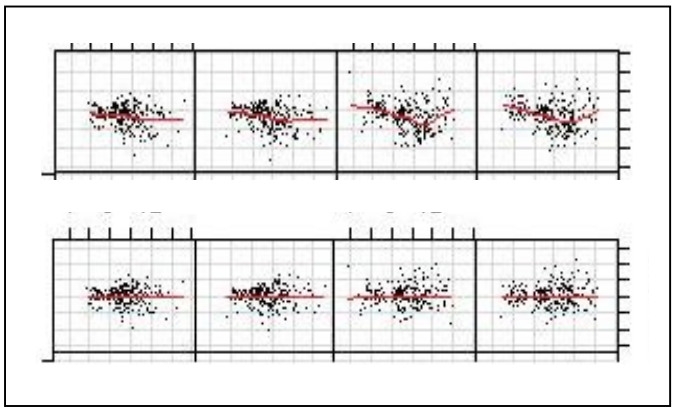


Figure 4 shows the effects of print-tip group normalisation for the grids on the first row of a particular array. Figure 5a shows that there is significant variation in print-tip-groups. Figure 4B is the same plots after normalised by print-tip groups.

The result of print-tip-group normalisation is to centre the log-ratios from each group on zero but it is likely that the spread or variance of each distribution will be different. To correct this it may be necessary to rescale each grid so that a particular grid is not given undue weight. Suitable scaling factors include the ratio of the variance for a particular grid to the geometric mean of the variances for all grids (Quackenbush 2002) or the Mean Absolute Deviation (MAD) (Yang et al. 2002).

Dye-swap normalisation adjusts for some dye-effects including gene-specific dye-bias, unequal labelling, and scanning properties, but does not correct for spatial bias on the array and assumes the variances of both ratio distributions are the same.

$$M_{i\ dye\text{-}swap\text{-}norm} = \tfrac{1}{2}(M_{i\ norm}^{forward} + M_{i\ norm}^{reverse})$$

As mentioned above, dye-swap experiments may not be possible due to the limited amount of samples or cost. If however, a dye-swap has been performed this type of normalisation should be used. This strategy does not assume the bulk of data is non-differentially expressed, but relies on the assumption that the number of up and down-regulated genes will be the same (**$c \approx c^1$).**

Between-slide Normalisation

Once the within-slide bias has been removed, it is necessary to re-scale the intensities so that they are comparable across arrays to biological replicates and other treatment groups. This type of normalisation affects the spread or variance of the distribution for each slide, so is considered a scaling normalisation. A common scale normalisation is to divide the intensity value by the total intensities on the slide.

$$M_{i\ norm} = M_{i\ norm} / M_{total}$$

More robust methods are discussed below.

Mean Absolute Deviation. Re-scaling can be achieved between arrays with the same method used to re-scale grids on the same array (Yang et al. 2002). Methods that re-scale the data may be undesirable in some cases because they can have the effect of increasing the variance. Before between-slide normalisation is carried out, it may be preferable to assess the spread of the

distributions to check if re-scaling is necessary. (See supplementary material figure 2.10 and 2.13)

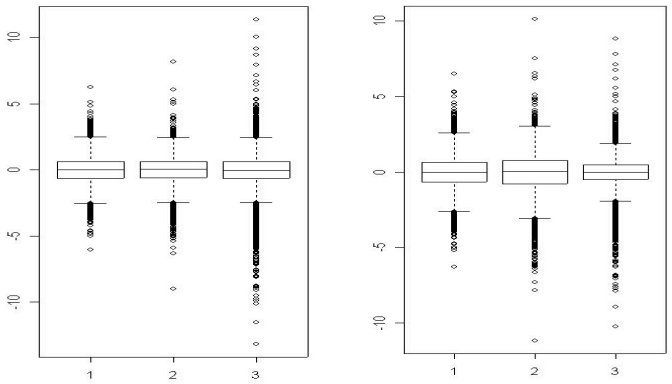


Figure 5 shows box-and-whisker plots of three array distributions that have been scale normalised. In figure 5b, the distributions are unequal but after scaling is applied, they have a similar distribution (5a).

Quantile normalisation. A distribution can be ordered into equal sized intervals called *quantiles.* Quantile normalisation transforms two distributions to the same scale in a rather strong sense: it rescales the data, perhaps nonlinearly, so that certain quantiles agree. However, as samples may have a different expression scales, this may be too violent a procedure. As an alternative, it is recommended that one calibrate quantile normalisation over a set of genes whose expression level is thought to be constant across all slides and then rescale all of the data using this calibrated normalization (Wit and McClure 2004). (See supplementary material figure 2.13.)

Gene Inference

The first step in most microarray analysis is to combine technical or biological replicates in order to obtain a more precise estimate of variability.

For technical replicates, the final estimate of log-expression-level for each gene can be derived from independent replicates as well as duplicated spots within the array. Depending on the purpose of the experiment, it may be necessary to evaluate the variation of each gene across technical replicates. This variation may be used to estimate the precision of the experimental technique. For most experiments, the mean expression level for each probe, averaged over all technical replicates is used to obtain a more precise estimate of gene expression. The standard deviation of each replicate from the mean is not generally used for technical replicates, as the replicates are intended to remove variation rather than characterise it.

Taking the mean of biological replicates is not widely used because this removes information on the variation of each spot value. The variation of biological replicates for each spot is crucial to give statistical confidence to the estimated differential gene expression profile. There are many statistical tests designed to assess differential gene expression and decide whether it exceeds expected levels of variability.

Mean and Variance

The simplest method for identifying differentially expressed genes is to rank the genes on the basis of the normalised log-ratio, M.

$$M_i = \log_2 (\text{sampleA}_i^{\text{red}} / \text{sampleB}_i^{\text{green}})$$

To detect differentially expressed genes, this method uses a *threshold fold-change*, *T*, (usually greater than a two-fold) value (DeRisi *et al.* 1997). Such thresholds are often chosen in an arbitrary way and have little biological foundation.

Methods that focus on the mean log-ratios as the sole indicator of differential expression often assume that the measurement error is normally distributed with a constant coefficient of variation. This assumption need not be true, so statistical variability should be taken into account.

The variability of log values is not generally constant over replicates and genes with larger variances have a good chance of giving a large log-ratio even if they are not differentially expressed (Smyth et al. 2003).

The mean taken from replicates can be combined with other measures such as standard deviation to give more clarity to the data: a low standard deviation provides an indication of high confidence in a spot measure. Ranking genes by the mean within a threshold of standard deviation is a widely used criterion. Examples of this strategy often use a *z*-transformation of each probe value (subtract the mean and divide by the standard deviation) and declare significant variation in those genes which have a *z*-score whose absolute value is at least 2-3, indicating the number standard deviations away from the mean (Draghici 2002; Yang *et al.* 2002).

$$z_i = (M_i - \mu_i)/\sigma_i$$

where μ_i is the mean log-ratio and σ is the standard deviation for probe each probe, *i.*

Researchers are encouraged to apply strict thresholds at first and then sequentially lower the cut-offs to gradually expand the results to reduce the false predictions.

The *t*-test

A better approach than using fold change or *z*-score is to calculate a statistic based on replicate data and rank genes according to a critical value or confidence level (Leung and Cavalieri 2003). Of these, the *t*-test is the most widely used. The *t*-test is a simple statistically-based method for detecting differentially expressed genes. The *t*-statistic is estimated for each probe on the array by combining information across biological replicates. The per-probe *t*-test provides a robust framework as it is not affected by the heterogeneity in variance across probes. The key concept for the *t*-test is the standard error of the mean (SEM), which is related to the observed variation in log expression ratio across all replicates for the *i*-th probe:

$$SEM_i = \sigma_i/\sqrt{n_i}$$

Here σ_i is the standard deviation of the n_i different replicated estimates of the log expression ratio. The SEM will tell us how far the observed mean may reasonably be expected to differ from its true value. The *t*-statistic is thus:

$$t_i = \mu_i / SEM_i$$

where μ_i is the mean log-ratio for probe *i*. Standard computer packages that do *t*-tests usually provide both the test statistic *t* and a probability value (*p-value*). The *p*-value is the probability of getting a value of the test statistic as extreme or more extreme (that is, further from zero) than the observed value by chance, given the null hypothesis (H_0) of no true differential expression.

A drawback to the standard *t*-test is that probes for which the variance of the log expression ratio is very small may give rise to large values of the *t*-statistic even when the genes are not differentially expressed (Lonnstedt and Speed 2002; Leung and Cavalieri 2003). One of the most widely used variations on the standard *t*-test is Welch's *t*-test, which is designed to handle unequal variances, but still assumes the expression ratios are log-normally distributed (Pan 2002). Alternative *t*-like tests can be used such as the Mann-Whitney and Wilcoxon tests, both of which work under much weaker assumptions about the distribution of log-expression levels. These methods are very attractive in that they do not depend on the assumption of log-normally distributed expression levels, but, when the data *are* log-normally distributed, are somewhat less powerful than the *t*-test: that is, they are less likely to detect small changes in expression.

There are many modifications to the *t*-test in the literature to more accurately estimate the measurement error using information from the data set. Significance Analysis of Microarrays (SAM) estimates gene-specific fluctuations by defining a statistic based on the ratio of change in gene expression to standard deviation for each probe plus a small constant (Tusher *et al.* 2001).

The positive constant, $\mathbf{s_0}$, is added to the denominator to increase the size of the standard deviation thus preventing over-estimation of *t*-scores for weakly-expressed genes:

$$d_i = \mu_i / (s_i + s_0)$$

where the statistic d_i is the product of the averaged log ratio μ_i and s_i the standard deviation over replicate data for each probe, i. The method is very similar to the *t*-test described above, but does not assume a normal distribution and uses permutation to achieve greater statistical significance.

Bayesian approaches to data analysis are becoming more popular whereby data from all the genes in a replicate set are combined into estimates of parameters of a posterior distribution. The parameter estimates are combined at the gene level to form a *B-statistic,* which is a Bayes log posterior odds (Lonnstedt and Speed 2002). The B-statistic favours genes with a high average and low variance and is more robust than the ordinary *t*-statistic.

A microarray experiment is a very good example of a multiple testing problem as the aim is to identify differential expression in any of several thousand genes. The problem is that in attempting to classify differential expression from such a large data set, it is impossible to avoid accidental false detections. For example, if one uses a critical *p*-value of 0.05, an array of 10,000 genes should be expected to yield 500 "significant" instances of differential expression even if none really are differentially expressed. To control the number of such false detections, a False Discovery Rate (FDR) (Benjamini and Hochberg 1995) algorithm can be applied to test statistics. There are various implementations of the FDR that re-estimate the *p*-values using the dependence structure between expression levels in order to control the false positive rate (also known as the Type 1 error rate) (Dudoit *et al.* 2002; Reiner *et al.* 2003). The use of an FDR algorithm is critical for microarrays, where the output from a study is usually a list of candidate genes whose differential expression must then, separately and expensively, be ratified in the laboratory. Application of a FDR algorithm will prevent time and resources being wasted on the study of false positives.

Linear models

As illustrated above, the log expression ratios can be estimated directly from a single array and also via a *linear* combination of ratios measured on other arrays. For example, for a reference design the log ratio of two non-reference samples A and B, log(A/B), is estimated through a linear combination of A with the reference, log(A/R) and B with the reference log(B/R).

$$\log(A/B) = \log(A/R) - \log(B/R)$$

In the context of simple loop designs, this concept allows us to estimate a particular log ratio both directly and also through a combination of all the other interactions in the loop.

For more complex designs, which may involve multiple inter-woven loops, this linear modelling approach allows comparisons between treatments to be estimated simultaneously for each gene via many different combinations of intermediate results. As the relationships between the ratios are assumed to be linear, it is possible to estimate unknown expression ratios indirectly, through a linear combination of more directly-measured pair-wise ratios. This type of linear model estimates gene expression with greater precision because the estimates combine several independent different routes through the design's diagram and converge to a local value for each gene.

The starting point for a linear model is to identify parameters on which to base the model. These parameters are known as *contrast-pairs* and they are used to define the treatment ratios to estimate differential gene expression. The number of contrast-pairs must: be one less than the number of treatments; represent all treatment groups and be independent of each other. These parameters are then used to construct a *design matrix,* which is used to define the relationships between the experimentally observed ratios and the contrast pairs. The details of these

relationships are crucial to formulate correct estimates, so it is customary to represent individual two-colour microarrays as arrows on the diagram that represents the design of the experiment: one dye is used for the condition at arrow's tip and the other for the condition at the tail. Thus, for example in figure 6c, the array A1 is used to compare treatments 1 and 2 directly: targets from treatment 1 will be labelled with the a red dye, while those from condition 2 will be labelled with a green dye. As shown in figure 6a, the expression ratios taken from a slide will always be taken to be ratios of the form (red/green), so, in the case of array A1, we will work with ratios of the form

(expression for treatment 1) / (expression for treatment 2).

When setting up the design matrix one converts paths through the design's diagram into rows of the matrix: if the path includes an arrow pointing in the forward direction the corresponding entry is a 1, while if the arrow points backward the entry is a -1. In the example below (figure 6), three arrays were used in the experiment (A1-3), each estimating a difference in expression directly on the array, as illustrated in 6A and 6C. For this example, the parameters were taken to be the log expression ratios 1to2 and 1to3. Next, for each array we calculate how these parameters can be estimated. Array1 measures log ratio(1:2) directly so can estimate parameter 1to2 (notation of 1) but cannot estimate parameter 1to3 (notation of 0). Array3 is a similar case to array1. Array2 measures log ratio(2:3) directly however it can be found also via a combination of 1to2 and 1to3. The indirect path is via 1to2 (notation of -1) and 1to3 (notation of 1).

The design matrix constructed from this information is then used to estimate the parameters via different routes and relationships. A least squares solution produces coefficient estimators that discriminate between the conditions/treatments of interest. This is essentially a regression procedure that finds the estimated log expression ratios by fitting a linear equation to the observed data. It finds the closest fit (in the sense of minimizing squared deviations) between the observed values (probe intensities) and the estimated values (linear model). *T-test*s can be applied to these estimators.

A.

$$A1 = \log (y_1^{\text{red (A1)}} / y_2^{\text{green (A1)}}) = d_{1to2}$$
$$A2 = \log (y_2^{\text{red (A2)}} / y_3^{\text{green (A2)}}) = d_{2to3}$$
$$A3 = \log (y_3^{\text{red (A3)}} / y_1^{\text{green (A3)}}) = d_{3to1}$$

Re-estimate path using independent parameters. In this case the parameters chosen are d_{1to2} and d_{1to3}.

B.

	d_{1to2}	d_{1to3}	
A1	1	0	A1 estimated directly via d_{1to2}
A2	-1	1	A2 estimated via d_{1to2} and d_{1to3}
A3	0	1	A3 estimated directly via d_{3to1}

C.

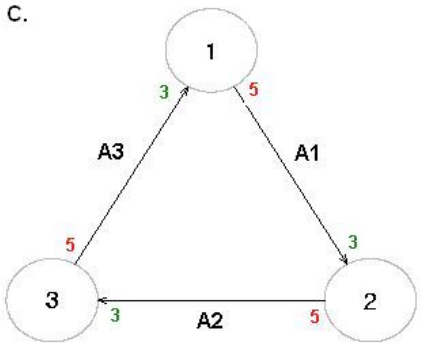


Figure 6 is an example of a simple linear model. 6A and 6C define the experimental design for three samples. The parameters d_{1to2} and d_{1to3} are used to define the paths to be estimated. The design matrix 6B, shows the relationships between these parameters and the design plan. The matrix uses the notation -1, 0 and 1 to specify the possible paths and direction.

There have been recent examples in the literature of linear approaches to microarray experiments (Smyth 2004; Vinciotti et al. 2004). One of the most straightforward ways to perform linear

modelling is through the Limma package (Smyth 2004) from Bioconductor (Gentleman *et al.* 2004) which allows the user to perform robust linear modelling to estimate gene expression differences. The methods allow a host of statistical tests to be applied to the estimates including an empirical Bayes method (E_Bayes) which moderates the standard errors of the estimated log-fold changes (Smyth et al. 2003). The method replaces the variance parameter with a posterior variance, estimated from the data. The resulting moderated *t*-statistic has the advantage over the standard *t*-statistic that large *t*-scores are less likely to arise merely from under-estimated sample variances. This is because the posterior variance offsets the small sample variances heavily in a relative sense while larger sample variances are moderated to a lesser degree (Smyth 2004).

As an alternative to the Limma methods, members of the MARIE consortium have developed an R routine to perform linear modelling of microarray data. Given an appropriate design matrix, the routine estimates the gene expression differences and also returns the variance of contrasts and a per-spot *p*-value (Vinciotti et al. 2004). The function is available at http://exgen.ma.umist.ac.uk/ however, we have also developed an online tool that walks the user through the construction of the design matrix and then performs linear modelling and returns the estimated log ratios by email.

Conclusion

Methods for the design and analysis of microarray experiments are constantly being updated and there have been numerous major advances in recent years. The savings available through the use of dye-balanced loop designs have ensured that this innovation in particular has been adopted rapidly. Advances made in normalisation schemes have also meant that researchers can obtain more accurate representations of the underlying biological variation. Linear models are an exciting new method to aid the detection of differentially expressed genes, especially over a range of realted treatments, as in time-course studies.

All of the calculations outlined above can be performed in the R environment for statistical computing (R Development Core Team 2005). In particular, the limma (Smyth 2004) and smida (Wit and McClure 2004) packages enable the user easily to perform all of the computations described above. An alternative linear analysis method is available as an R function from our website (http://exgen.ma.umist.ac.uk/). And, to complement this review, a walkthrough showing how to apply all of these methods is available as supplementary material from our website (http://exgen.ma.umist.ac.uk/). The MARIE website hosts a number of additional tools for design, normalisation and analysis, intended to make the methods above more accessible to working experimentalists.

References

1. Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." J R Statistical Society B **57**(1): 289-300.
2. Bretz, F., J. Landgrebe, et al. (2003). "Efficient Design and Analysis of Two Colour Factorial Microarray Experiments." Biostatistics **1**(1): 1-20.
3. Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." Nat Genet **32 Suppl**: 490-5.
4. Cochran, W. G. and G. M. Cox (1992). Experimental designs. New York ; London, Wiley : Chapman & Hall.
5. DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.
6. Dobbin, K., J. H. Shih, et al. (2003). "Statistical design of reverse dye microarrays." Bioinformatics **19**(7): 803-10.
7. Dobbin, K. K., E. S. Kawasaki, et al. (2005). "Characterizing dye bias in microarray experiments." Bioinformatics **21**(10): 2430-7.
8. Dombkowski, A. A., B. J. Thibodeau, et al. (2004). "Gene-specific dye bias in microarray reference designs." FEBS Lett **560**(1-3): 120-4.

9. Draghici, S. (2002). "Statistical intelligence: effective analysis of high-density microarray data." Drug Discov Today **7**(11 Suppl): S55-63.
10. Dudoit, S., Y. H. Yang, et al. (2002). "Statistical Methods For Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments." Statistica Sinica **12**: 111-39.
11. Durbin, B. P., J. S. Hardin, et al. (2002). "A variance-stabilizing transformation for gene-expression microarray data." Bioinformatics **18 Suppl 1**: S105-10.
12. Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.
13. Hoyle, D. C., M. Rattray, et al. (2002). "Making sense of microarray data distributions." Bioinformatics **18**(4): 576-84.
14. Kerr, M. K. and G. A. Churchill (2001). "Experimental design for gene expression microarrays." Biostatistics **2**(2): 183-201.
15. Kerr, M. K. and G. A. Churchill (2001). "Statistical design and the analysis of gene expression microarray data." Genet Res **77**(2): 123-8.
16. Kerr, M. K., M. Martin, et al. (2000). "Analysis of variance for gene expression microarray data." J Comput Biol **7**(6): 819-37.
17. Khanin, R. and E. Wit (2004). "Design of large time-course microarray experiments with two channels." not published.
18. Konishi, T. (2004). "Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment." BMC Bioinformatics **5**(1): 5.
19. Lee, M. L., F. C. Kuo, et al. (2000). "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations." Proc Natl Acad Sci U S A **97**(18): 9834-9.
20. Leung, Y. F. and D. Cavalieri (2003). "Fundamentals of cDNA microarray data analysis." Trends Genet **19**(11): 649-59.
21. Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-4.
22. Lonnstedt, I. and T. P. Speed (2002). "Replicated Microarray Data." Statistica Sinica **12**: 31-46.
23. Martin-Magniette, M. L., J. Aubert, et al. (2005). "Evaluation of the gene-specific dye bias in cDNA microarray experiments." Bioinformatics.
24. Pan, W. (2002). "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments." Bioinformatics **18**(4): 546-54.
25. Peng, X., C. L. Wood, et al. (2003). "Statistical implications of pooling RNA samples for microarray experiments." BMC Bioinformatics **4**(1): 26.
26. Pritchard, C. C., L. Hsu, et al. (2001). "Project normal: defining normal variance in mouse gene expression." Proc Natl Acad Sci U S A **98**(23): 13266-71.
27. Qin, L. X. and K. F. Kerr (2004). "Empirical evaluation of data transformations and ranking statistics for microarray analysis." Nucleic Acids Res **32**(18): 5471-9.
28. Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet **32 Suppl**: 496-501.
29. R Development Core Team, R. F. f. S. C. (2005). "R: A language and environment for statistical computing."
30. Reiner, A., D. Yekutieli, et al. (2003). "Identifying differentially expressed genes using false discovery rate controlling procedures." Bioinformatics **19**(3): 368-75.
31. Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.
32. Smyth, G. K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." Statistical Applications in Genetics and Molecular Biology **3**(1): Article 3.

33. Smyth, G. K., Y. H. Yang, et al. (2003). "Statistical issues in cDNA microarray data analysis." Methods Mol Biol **224**: 111-36.
34. Sterrenburg, E., R. Turk, et al. (2002). "A common reference for cDNA microarray hybridizations." Nucleic Acids Res **30**(21): e116.
35. Tseng, G. C., M. K. Oh, et al. (2001). "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects." Nucleic Acids Res **29**(12): 2549-57.
36. Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.
37. Vinciotti, V., R. Khanin, et al. (2004). "An experimental evaluation of a loop versus a reference design for two-channel microarrays." Bioinformatics.
38. Williams, B. A., R. M. Gwirtz, et al. (2004). "Genomic DNA as a cohybridization standard for mammalian microarray measurements." Nucleic Acids Res **32**(10): e81.
39. Wit, E. and J. D. McClure (2004). Statistics for microarrays : design, analysis, and inference. Chichester, John Wiley & Sons.
40. Yang, I. V., E. Chen, et al. (2002). "Within the fold: assessing differential expression measures and reproducibility in microarray assays." Genome Biol **3**(11): research0062.
41. Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Nucleic Acids Res **30**(4): e15.
42. Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nat Rev Genet **3**(8): 579-88.
43. Zhou, Y., F. G. Gwadry, et al. (2002). "Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: microarray analysis of dose- and time-dependent effects." Cancer Res **62**(6): 1688-95.