



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Experiments using microarray technology: limitations and standard operating procedures

Citation for published version:

Forster, T, Roy, D & Ghazal, P 2003, 'Experiments using microarray technology: limitations and standard operating procedures' *Journal of Endocrinology*, vol. 178, no. 2, pp. 195-204. DOI: 10.1677/joe.0.1780195

Digital Object Identifier (DOI):

[10.1677/joe.0.1780195](https://doi.org/10.1677/joe.0.1780195)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Journal of Endocrinology

Publisher Rights Statement:

Available under Open Access.
Copyright © 2013 by the Society for Endocrinology

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



REVIEW

Experiments using microarray technology: limitations and standard operating procedures

T Forster, D Roy and P Ghazal

Scottish Centre for Genomic Technology and Informatics (GTI), The University of Edinburgh, Medical School, Little France Crescent, Edinburgh EH16 4SB, UK
(Requests for offprints should be addressed to P Ghazal; Email: marilyn.horne@ed.ac.uk)

Abstract

Microarrays are a powerful method for the global analysis of gene or protein content and expression, opening up new horizons in molecular and physiological systems. This review focuses on the critical aspects of acquiring meaningful data for analysis following fluorescence-based target hybridisation to arrays. Although microarray technology is adaptable to the analysis of a range of biomolecules (DNA, RNA, protein, carbohydrates and lipids), the scheme presented here is applicable primarily to customised DNA arrays fabricated using long oligomer or cDNA probes. Rather than provide a comprehensive review of micro-

array technology and analysis techniques, both of which are large and complex areas, the aim of this paper is to provide a restricted overview, highlighting salient features to provide initial guidance in terms of pitfalls in planning and executing array projects. We outline standard operating procedures, which help streamline the analysis of microarray data resulting from a diversity of array formats and biological systems. We hope that this overview will provide practical initial guidance for those embarking on microarray studies.

Journal of Endocrinology (2003) **178**, 195–204

Microarray Standard Operating Procedures (M-SOPs)

The main purpose of this review is to outline technical procedures for the acquisition and analysis of array data. This is not an attempt to provide a comprehensive review of all aspects of microarray analysis. Several excellent and recent reviews of this nature are available (Duggan *et al.* 1999, Deyholos & Galbraith 2001, Quackenbush 2001, 2002). Microarray studies involve a complex multi-step process (Fig. 1). A successful microarray project is dependent on all steps of the process being accurately and consistently performed to maximise the reliability and significance of results. Consideration of steps upstream of data processing is necessary to ensure experimental consistency and these will be described briefly.

Experimental design must take into account the biological question under study, and should include statistical input to permit the required level of statistical significance to be obtained (Kerr & Churchill 2001, Churchill 2002, Glonek & Solomon 2002, Yee & Speed 2002). In this respect, experiments should be well controlled and replicated, and it is vital that consistency is applied to experimental logistics and timing, sampling and labelling to reduce sources of unwanted variation where possible. For example, if one proposes to perform time series

experiments with each chip hybridised with experimental and reference samples, thought must go into the correct selection of the reference material to ensure biological relevance to the study. Due consideration must be given to whether material is pooled or individually sampled.

The entire planning stage is as important as the subsequent implementation (see below) and omissions at this stage can easily lead to non-representative or false results. Planning of a study benefits from multiple inputs from biological researchers as well as statistician/bioinformaticians with experience in microarray technology.

Experimental sampling and extraction of RNA is a vitally important component of this process since successful microarray studies are dependent on the consistent extraction of high quality RNA. In broad terms, microarrays are performed on two basic biological systems: simple and complex. Simple biological systems are those where homogeneous cell populations are present, such as cell lines or purified cell populations. Sampling from simple systems is more likely to represent the expression level for the particular cell or tissue under study. Complex systems are typified by tissues and organs where there is a diversity of cellular substructures and mixed cellularity. Extraction of RNA from complex systems means that critical spatial and cellular information as to the origin of the signal is lost. This reduction of contextual information makes

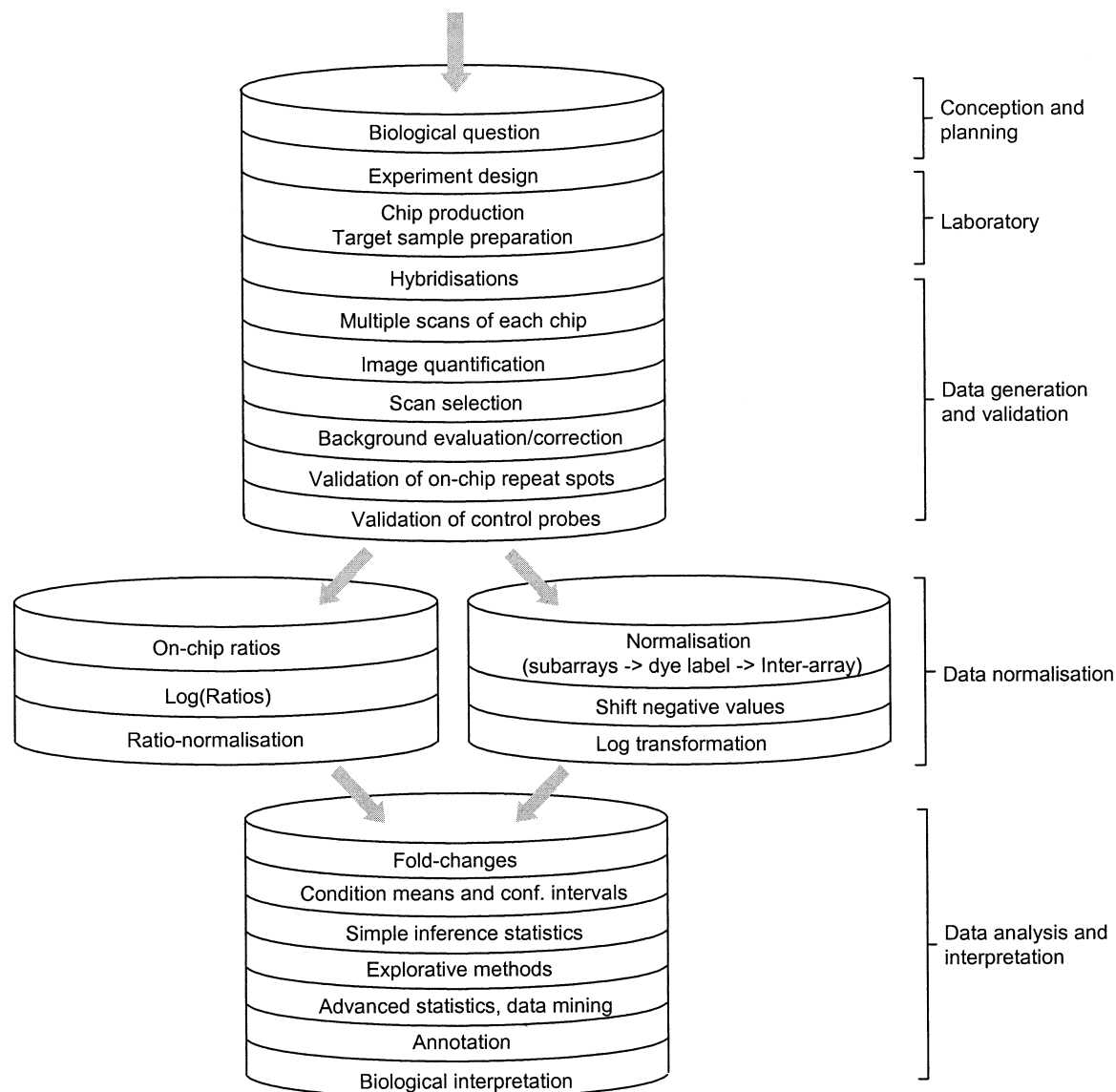


Figure 1 A complete microarray project workflow within our laboratory (GTI) is outlined. Shown here are the standard steps in top-to-bottom chronological order. The brackets to the right show a broader categorisation of these workflow processes. Depending on the purpose and question behind the project, individual steps like 'Advanced statistics' are skipped on occasion. This paper only describes the steps following hybridisation.

interpretation of data from complex systems problematic, and requires additional validation with cell-specific techniques to resolve questions of signal origin (Chuaqui *et al.* 2001). Sub-dissection techniques are being increasingly adopted to provide specificity to sampling from complex systems. For instance, laser-based methods for the precise microdissection of individual cells or tissue substructures are available, in addition to more traditional purification of cell populations using fluorescent markers. These approaches have enormous potential to improve resolution in microarray studies. Generally only limited (sub micro-

gram) quantities of RNA are gleaned from these sampling strategies – quantities that are usually too small for conventional labelling strategies. New amplification methods for the labelling of minute quantities of RNA are now being employed. However, it is becoming increasingly evident that even highly purified cell populations and apparently homogeneous cell lines may demonstrate complexity of phenotype and metabolism at the individual cell level. This variation is likely to encompass differences in RNA turnover, sublocalisation, splicing and translational activity. This only serves to highlight the importance of

standardising culture and purification methods as rigorously as possible to achieve consistency during sampling and extraction phases.

Regardless of the RNA sampling methods employed, it is important to apply rigorous quality control to purified RNA populations. For instance, the Bioanalyser system from Agilent Technologies (Cheadle Royal Business Park, Stockport, Cheshire, UK) is now commonly employed to check the quality and consistency of RNA samples. The resulting absorbance profile provides a useful means of assessing the suitability of RNA for labelling. At this stage, consistency during labelling and hybridisation steps is the starting point for the generation of consistent array data (Hegde *et al.* 2000).

The selection and production of the correct array format is important and a central feature of the process. The majority of custom arrays are produced by the direct deposition of nucleic acid probes as cDNA or long oligomeric sequences onto treated glass substrates. The production of reproducible arrays with current pin printing methods is challenging. In our own Centre we have introduced a number of quality control steps to ensure consistency of array production, but these are outside the scope of this review.

An essential theme is the requirement for microarray data to be MIAME (minimum information about a microarray experiment) (Brazma *et al.* 2001) compliant. In essence, this addition of standardised information about all stages of a microarray experiment allows for amalgamation of array data from different groups and sources in the public domain, ultimately permitting advanced and automatic data mining. Accordingly, there is an absolute necessity for the implementation of M-SOPs. The M-SOPs outlined here aid in the production of standardised project documentation, which ensures MIAME compliance for publication. In the following sections we outline in more detail the analytical steps of the workflow.

Data Generation and Validation

The chronological order of processes in a microarray project utilising customised arrays is given in Fig. 1. Approaches for individual and combined processing and analysis steps have recently been reviewed (Nature Genetics 2002, Speed 2002).

Array scanning and image quantification

The process of scanning an array is known as *image acquisition*, whereas the process of converting images to numerical data is referred to as *image quantification* or *processing*. The majority of microarray experiments involve the fluorescent detection of hybridised signal using confocal laser scanners. A wide variety of different scanning instruments are available, and a number of different

image acquisition and quantification packages are associated with them. In general, selection of image quantification parameters (e.g. 'adaptive', 'fixed circle', 'spot distance') should be carefully assessed and decided for each project as a whole, and will depend on array design, slide type and spot morphology. As an exception to this, a limited form of manual input is often required to fine-tune the layout of the template quantification grid for individual arrays and care should be taken to avoid user bias. Apart from this limited fine-tuning, it should be noted that the image *quantification* method should be identical for all slides constituting a project, whereas image *acquisition* parameters, for instance laser power and/or photo multiplier, can be optimised from slide to slide. For a comparative discussion of issues concerned with statistical image analysis we refer the reader to Glasbey and Ghazal (2003) and Yang *et al.* (2002b).

Scan selection

The procedure employed for image acquisition can have a significant effect on data analysis and interpretation. The aim here is to identify scanner settings that provide the best representation of signal distribution on the array. We find that either the frequently used trial-and-error approach or even single array scans can make potentially high quality experiments useless. Signal values across an array typically exhibit a density distribution approaching log-normal (a heavier right tail), and the best scan would ideally exhibit a high median value without distorting this shape, especially when nearing the scanner saturation point, which causes off-scale (censored) values among highly expressed genes. In practice, this reasoning can be used to define a maximum scan setting that lifts weakly expressed genes above the background noise level without causing saturation for highly expressed genes.

We use the following process which relates multiple scans of an individual array in a linear series. (1) Scan each array at five incremental laser powers or photo-multiplier-tube (gain, light amplification) settings S_1 to S_5 (a scanner mid range setting is usually a good starting point). (2) For each scan, perform identical image quantification (see previous section). (3) Create four scatter-plots: S_1 vs S_2 , S_2 vs S_3 , S_3 vs S_4 , S_4 vs S_5 . All axes need to be scaled identically. (4) Interpret numbers/graphs and select best scan.

A representative series of scatter-plots is shown in Fig. 2. If preferred, these graphs can be used for visual assessment of scan performance.

Numerical evaluation requires determination of the following simple statistics for each scan dataset (and combinations of scans for point 4). These are then used to form combined selection criteria to define optimal scanning conditions: (1) median value; (2) inter-quartile range (IQR); (3) count of saturated spots; and (4) linear regression

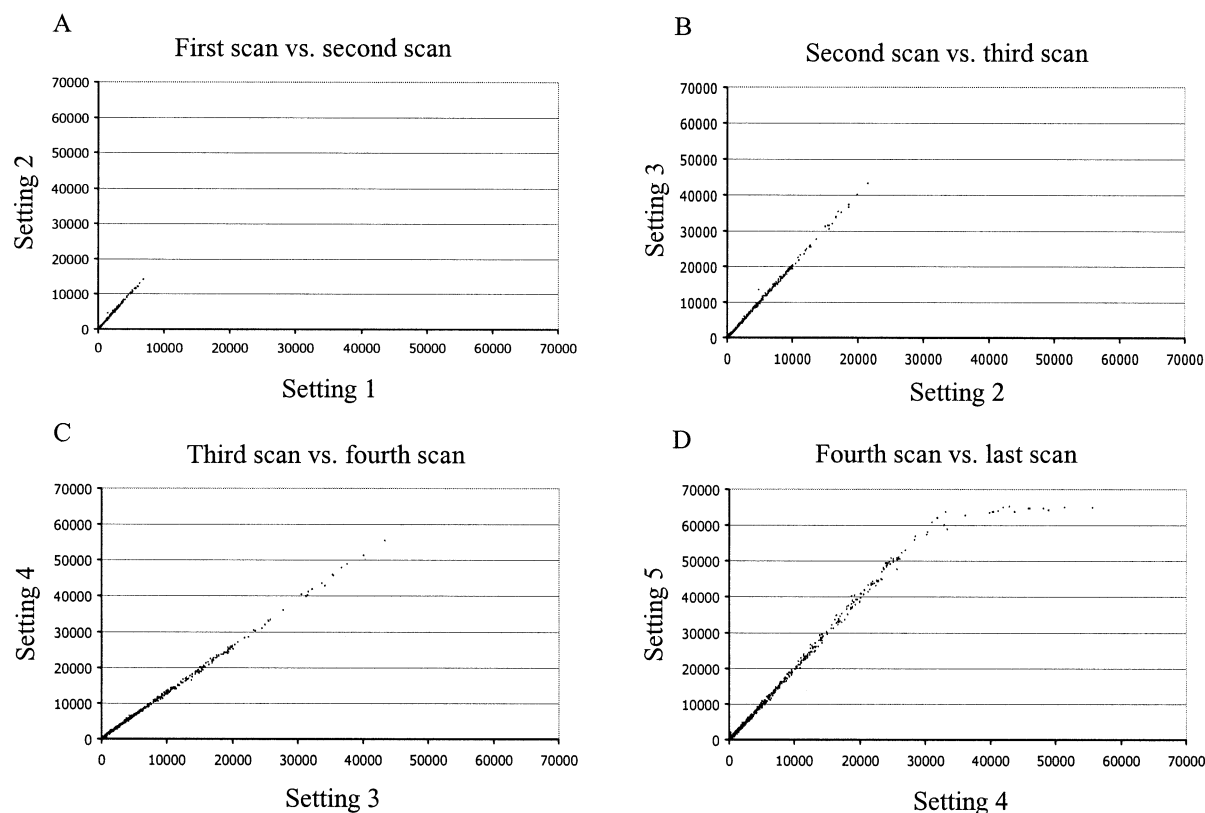


Figure 2 A series of four pair-wise graphs for five incremental (photo-multiplier-tube (PMT) settings for amplification of reflected light) scans of one chip are shown. For example, the first graph (A) plots data from the scan with the lowest setting against data from the scan with the next highest setting. We can see immediately that neither scan 1 nor scan 2 provide expression values across the possible range (0–65536). Nevertheless, it is also apparent that the increase in scan PMT has a strictly linear effect on the signal intensities. If we then examine the next two graphs (B and C) in this series, we can clearly see how the range of signal intensities increases to cover more of the available spectrum, lifting low to medium strength signals out of the possible background noise. The last graph (D) is not linear in the high signal range, meaning the highest scan setting in this series causes signal saturation in a number of spots. Based on this, we would here decide to carry forward data from scan setting 4 (y -axis in C and x -axis in D), the highest possible scan with the best representation of the data on this chip.

between pair-wise scans: r^2 and P (linear regression runs test).

Optimal scan parameters are defined from permutations of these values defining conditions which satisfy: the highest median value; the widest IQR; minimal or no saturation values; and, between adjacent scans, a high regression coefficient with a low ‘runs test’ P value (the ‘runs test’ determines whether the data differ significantly from the linear regression). The latter two parameters indicate at what setting the scanning breaks down, changes in signal becoming non-linear in relation to the next or previous scan.

In addition to this numerical evaluation, visual inspection of signal value distribution for each scan remains highly informative. Qualitative indicators for the interpretation of scan quality are: (1) there is no or only irrelevant saturation (scatter ‘cloud’ is linear) and (2) a dynamic, uninterrupted range of expression signals from

low to high is visible (this may not be true if the array contains only a subset of genes which are all expected to be expressed at a similar level).

Figure 2 provides examples of these visual properties. The highest scanner setting producing the optimal signal distribution is usually selected.

Background evaluation/correction

Fluctuations in background fluorescence can affect signal values. Background correction is therefore essential to obtain a noise-reduced signal value, but it is always best first to evaluate background noise after image quantification, since it will depend on the chosen software settings. A map of background values and a corresponding map for signal values (values as provided by quantification software) is our preferred option. These maps are based on plotting spot XY -coordinate vs value. Comparison of these

maps reveals if signal values are systematically influenced by noise, given the chosen software settings. If deemed necessary, our own approach requires subtracting individual background values from the corresponding signal value (both are usually provided separately by the image quantification software). Optionally, this subtraction can also be done with a spatial background average. One other common approach is to identify signal values that are smaller than the background mean plus two standard deviations. Negative values resulting from an adjustment could either be removed, flagged or shifted. Since we are frequently interested in information contained in borderline expression, we prefer not to select the removal option at this stage.

Replication of gene probes

As a definition in the context of microarrays, a 'probe' is the (partial) genomic sequence of a gene deposited and fixed on the glass slide, whereas the 'target' is the biological sample material.

Probe *variants* are defined as different nucleotide sequences representing the same gene. Signal strength is frequently sequence dependent. For this reason, averaging signal intensities is not appropriate and probe variants should be analysed separately until the final data comparison steps. Probe *replicates* have the same sequence, with multiple instances on the array. In theory, these should have identical expression, and their purpose is to increase the confidence in the reliability of the gene probe deposition process. Probe replicates are utilised in two major ways.

Our preferred option, given a sufficient number ($\sim 5+$) of biological samples per condition, is to calculate the median (in the case of three spot replicates this is the 'middle' value) value for each replicate set of gene probes on a chip and use this as the 'true value' for that gene probe. The median is a robust measure and will inherently ignore outlier values (generally probes for which the robotic deposition process failed) within a replicate set, in contrast to the arithmetic mean.

The second option is to include all replicate values in the full dataset (e.g. to apply ANOVA-type analysis methods later (Kerr *et al.* 2001)). Experimental error is then pooled with biological variation. It is important not to confuse this increase in the number of data values per gene with the true statistical sample size, which is merely the number of biologically independent target samples, irrespective of how many print replicates there are.

Control probes for array sensitivity/specificity and data normalisation

The inclusion of sufficient numbers of negative control probes (minimum of 10–30 independent gene probes) and a collection of positive control probes (ideally 30+ house-

keeping and/or spike controls) demonstrating a range of expression levels is required in many types of experiment and can aid analysis, as well as providing a measure for array quality and consistency. This is particularly appropriate where a large proportion of all gene probes on a chip are expected to be differentially expressed between conditions, e.g. time-series, developmental staging or viral genomes. In this case, control probes can serve as a subset of genes with theoretically constant expression for normalisation purposes.

We also utilise receiver-operator-characteristics (ROC) analysis, which is a quantitative assessment of expression detection in terms of sensitivity and specificity, wherever we have known sets of positive and negative controls to serve as a 'gold standard' for expression and non-expression (Wagner *et al.* 2002).

A simpler approach is to look at signal distribution on each array. Box-and-whisker plots for positive and negative control probes should show clear differentiation between them, and in a similar fashion for all arrays in the experiment.

Further flagging of probes

We routinely define detection threshold level as the 80th percentile of all negative control values (gene probes external to the biological system under investigation) on an array. Because we are flagging genes below this cut-off value rather than removing them, we consider this arbitrary percentile choice to be information gain rather than loss. With prior ROC-analysis, we may also set a custom detection threshold at a desired level of sensitivity and specificity.

Normalisation

In microarray analysis, normalisation (Schuchhardt *et al.* 2000, Quackenbush 2002) refers to a collection of processes that are used to adjust data means or variances for effects resulting from systematic non-biological differences between arrays, subarrays and dye-label channels. As a brief definition of these terms, an array is the entire set of gene probes on a chip. A subarray or print-tip group is a subset of these deposited by the same print-tip - they can easily be seen as distinct smaller arrays of probes within the full array. And lastly, dye-label channel refers to the fluorescence frequency of the target sample(s) hybridised to the chip. Experiments where two differently dye-labelled samples are mixed and hybridised to the same chip are commonly referred to as 'dual-dye experiments'. They result in a relative rather than absolute expression value for each gene on the array, often presented as the log of the ratio between 'red' channel and 'green' channel ($\log(R/G)$). Usually the second sample on an array is a common standard reference sample for all other samples, so that

expression for each gene on all arrays is in relation to the same baseline, and therefore directly comparable.

At this stage either ratiometric or absolute data normalisation can be performed. This decision is based on the original biological question and the experimental design constraints. Ratiometric analysis is most commonly performed, but we frequently use absolute values to assess magnitudes of expression.

Ratiometric methods

Ratiometric analysis is mainly employed in dual-dye experiments where one channel or array is considered in relation to a common reference. A ratio of expression for each gene is calculated between test and reference sample. This is followed by transformation of the ratio into $\log_2(\text{ratio})$ to symmetrically represent relative changes. The normalisation strategy adopted is influenced by the signal data.

For this, let sample 1 and 2 be R and G, and individual genes indexed i . NLR is the normalised log-ratio. Option 3 is our default approach.

(1) When the expected number of differentially expressed genes between samples is only a small fraction (< 10%) of the total number of genes present, and there are approximately balanced levels of up- and down-regulation between them the following applies:

$$\text{NLR}_i = \log_2(R_i/G_i) - \text{median}(\log_2(R/G))$$

In summary, the median of all $\log(\text{ratios})$ is subtracted from all individual $\log(\text{ratios})$, resulting in a new overall median $\log(\text{ratio})$ of zero.

(2) When a large fraction of all genes on the array is expected to be differentially expressed, or differential expression is one-way only, the above formula still applies, but R and G must be based on a subset of control probes with theoretically identical expression in both samples.

(3) Log-ratios can be non-linear with respect to signal magnitude. Use of the R-I plot, namely, $\log_{10}(R \times G)$ vs $\log_2(R/G)$ will show if this effect is present (Fig. 3). After applying a locally robust LOWESS (LOcally WEighted Scatterplot Smoother) function to this graph (Yang *et al.* 2002a), the following serves as normalisation procedure:

$$\text{NLR}_i = \log_2(R_i/G_i) - \log_2(2^{\gamma(x_i)})$$

where γ is the log ratio, x is the log signal level and $\gamma(x_i)$ is the result (value by value) of the LOWESS function estimating the dependence of the ratio on the signal level. If conditions as in option 2 apply and a large enough set of positive controls is available, $\gamma(x_i)$ can be derived from those. The LOWESS function is included in most statistical or microarray packages, and is based on the R-I plot. In other words, this function adjusts every value individually based on its level of intensity, straightening out any bends apparent in the plot, and centering the R-I plot around zero.

Provided a sufficient number of probes or control probes are contained within each subarray (print-tip group), the options above can be expanded to normalise these within an array, adjusting for possible print-tip or hybridisation gradient problems.

Absolute value methods

This approach represents single-dye experiments, or dual-dye experiments where there is no suitable reference for a channel or array. Sometimes a series of increasing/decreasing values over time or conditions is more relevant than a series of increasing/decreasing ratios. Therefore, we often normalise absolute signal values from a larger number of arrays/samples rather than ratios.

The normalisation assumptions as specified above are unchanged.

For instance in a series of arrays 'A' with individual arrays indexed by 'j', and individual genes within those arrays indexed by 'k', and S(NS) being the signal (normalised signal):

(4) the normalisation factor (NF) for each array/channel is

$$\text{NF}(A_j) = \frac{\text{Mean}(q_{0.75}(A))}{q_{0.75}(A_j)}$$

and then

$$\text{NS}_{jk} = S_{jk} \times \text{NF}(A_j).$$

In summary, we calculate the 75th percentile value for each chip and use the mean of all these as the reference value to which each chip is scaled. The required normalisation factor is the reference value divided by the 75th percentile value of the individual chip. This value is multiplied with all individual gene probe values on a chip.

(5) The same as (4), only A and A_j are subsets of selected control probes on the array with theoretically identical expression in all samples/arrays.

(6) LOWESS has mainly been used for correcting intensity-dependent dye-label effects between two samples on the same array. It is generally less useful in experiments with multiple conditions on multiple arrays, unless it is applied multiple times where the reference array has a different dye-label than other arrays, and all adjustments are made between the same reference sample and each test array.

If subarray (intra-array) normalisation is deemed necessary, our standard approach consists of identifying those control probes that exhibit a similar intensity profile across all the subarrays. This subset is then used for a simple scaling of the 75th percentile value of each subarray control subset to the median of the pooled 75th percentiles of all subarrays (see (4)).

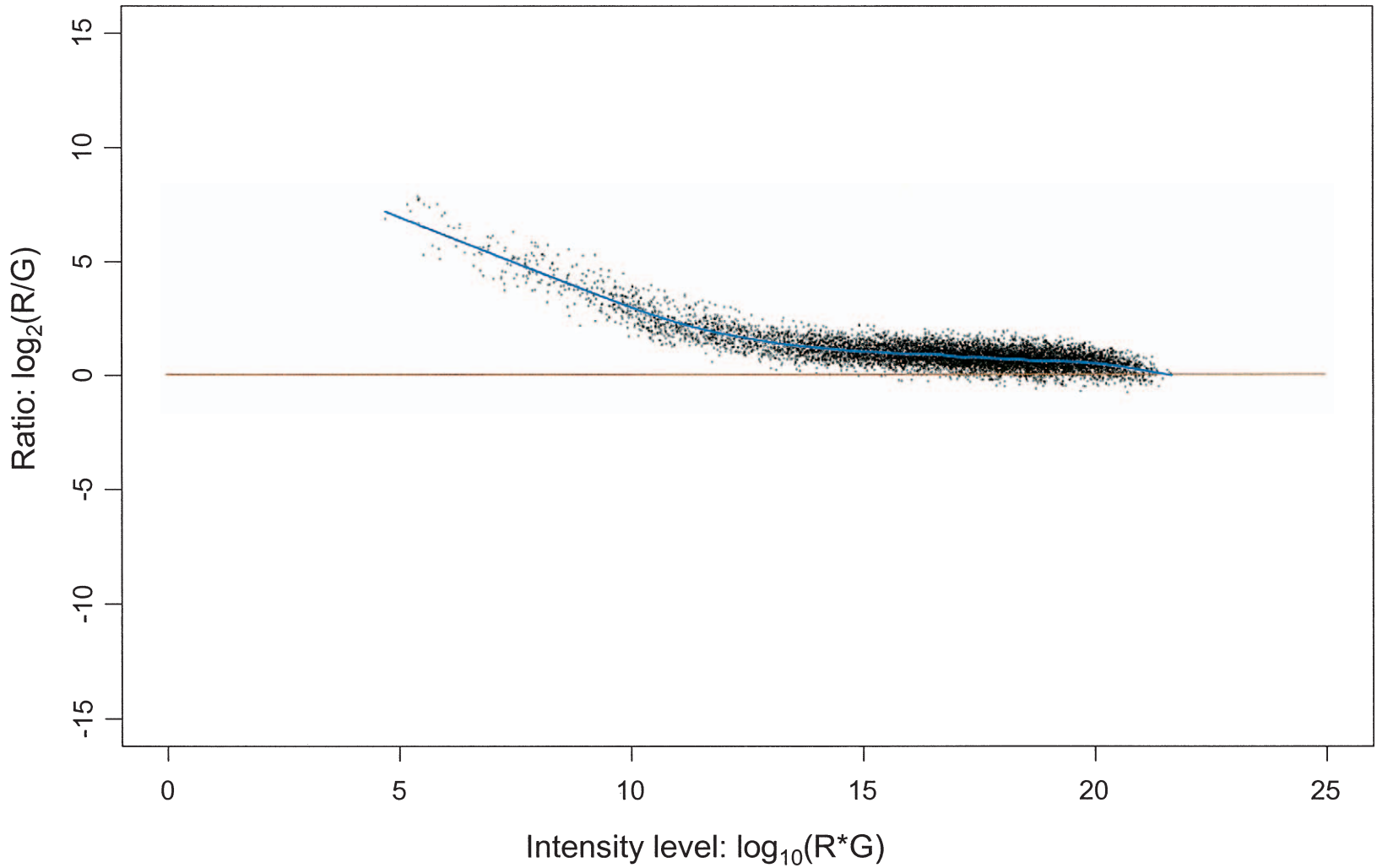


Figure 3 An example of an R-I plot to assess intensity-dependent dye-label incorporation. The R-I plot is commonly used to investigate the dependency of relative expression (between two differentially dye-labelled samples on one chip, or between any two chips) on the magnitude of expression. The magnitude of expression ('Intensity') is presented as the product of the values of both channels (x-axis), \log_{10} is merely used to simplify interpretation of scale. The relative expression ('Ratio') is expressed as usual: \log_2 of the ratio of the two channel values for each gene probe (y-axis). Any bend in the global distribution of the data points suggests that, depending on the magnitude of expression, systematic differences of dye-label incorporation or related effects are present. If so, a linear normalisation method is insufficient, and the LOWESS (LOcally WEighted Scatterplot Smoother) curve (blue line) is used as the basis for 'straightening out' this effect. In addition to this, a very irregular scattering of data points indicates that a large proportion of all gene probes is differentially expressed between the two. In this case, control probes (spikes, housekeeping, negative controls) are essential for normalisation purposes. The red line is the theoretical global log-ratio of zero around which all data points would be centred if average expression in both channels is identical; LOWESS aims to move the data to the red line.

Negative value correction

The initial background correction may lead to negative values on some or all arrays. Instead of removing them and thereby losing information that may still have potential use, all negative values of all arrays are plotted to ascertain a symmetrical distribution. We identify the quantile value at 0.5% of this distribution (percentile 0.005) as the lower cut-off point below which true outlier values are more likely. We then shift the data distributions of all arrays by the absolute of this value, into the positive range. This cut-off is arbitrary, therefore it is essential to flag all the probes that have gone from negative to positive as a result of this operation (on each array). This information is used for result interpretation at a later stage.

The remaining values below zero are considered to be definite outliers, and will become missing values in the log transformation.

Log transformation

This is usually applied to the ratio of two conditions on the same array. However, this can also be applied to absolute values from single-dye experiments. Log transformation changes positively skewed data distributions into symmetrical ones and it can reduce the increase of data variance over time in time-series experiments. The choice of \log_{10} , \log_e or \log_2 is user-dependent, although, due to convenience of scale interpretation, biologists often prefer the latter.

Analysis

Here we provide a brief outline of initial data analysis, which includes ratio-statistics (Chen *et al.* 2002, Dudoit *et al.* 2002). However, valuable information can also be contained in the absolute expression levels within an experiment. It is also important to note that for microarray projects designed to study defined gene pathways and interactions, a maximum of annotation and statistical reliability is required. We suggest the minimum result set for each gene should include: fold-changes, mean expression level per condition (with 95% confidence intervals, i.e. the 'true' mean will be within this interval with a probability of 95%), and *P* values from significance testing.

Data visualisation

Plots of log (condition 1) vs log (condition 2) with added fold-change thresholds and optionally fitted LOWESS regression curve work well to display the entirety of the array data (Fig. 4). In case of biological chip replicates, data points are usually the mean values of each gene in a given condition. Limited subsets of interesting genes can obviously be plotted by means of simple vertical bar charts

(log(ratio) or absolute values). This can be complemented by custom project-dependent graphs, often to integrate annotation data with gene expression results (e.g. GenMapp www.genmapp.org to integrate pathway diagrams and expression values). Expression profiles of genes within a condition or each gene across a number of conditions are covered by cluster analysis (see below).

Condition means and confidence intervals

These parameters are needed to present the level of expression of a gene, and enable better interpretation of fold-change. In most dual-dye experiments, the confidence interval can be calculated for the fold-change itself.

Calculation of 95% confidence intervals (CI95) for means can be done in different ways. If there are sufficient numbers (~ 10 or more) of observations for each condition, the common CI95 formula based on a *t*-distribution can be used. However, for smaller numbers of observations a bootstrap method is recommended (Davison & Hinkley 1997, Carpenter & Bithell 2000). Broadly speaking, the values making up both groups of observations are randomly reassigned to each group a large number of times, with the desired statistic (mean, CI95, *t* etc.) calculated for each such resampling run. The accumulated set of newly generated statistics is then used to 'estimate' the corresponding parameter for the original data.

Significance tests

A simple two-sample *t*-test or Welsh *t*-test is often the first tool of choice for statistical inference. An adjustment for multiple testing (Dudoit *et al.* 2003) will adjust obtained *P* values, but not the order of sorted significance values. If two conditions can be assumed to be dependent (e.g. cell lines) then paired *t*-tests can increase statistical power, even though measurements were not necessarily taken on the same subject.

A non-parametric test (e.g. Mann-Whitney-U) is an alternative with less power that nonetheless works better under the assumption that underlying distributions are non-symmetrical, but for the often very small numbers of observations in microarray studies (5 to 7) the resulting *P* values can be less useful as a filtering tool.

Statistical power can generally be improved by employing bootstrap versions of significance tests.

For most microarray studies, *P* values resulting from significance testing must be interpreted with care in cases with few biological replicates per group.

ANOVA-type methods (Kerr *et al.* 2001) are somewhat more involved, and appropriate where there is more than one experimental factor under investigation (e.g. treatment and dose, or biological replicates and hybridisation replicates). It is important to note that the expression of

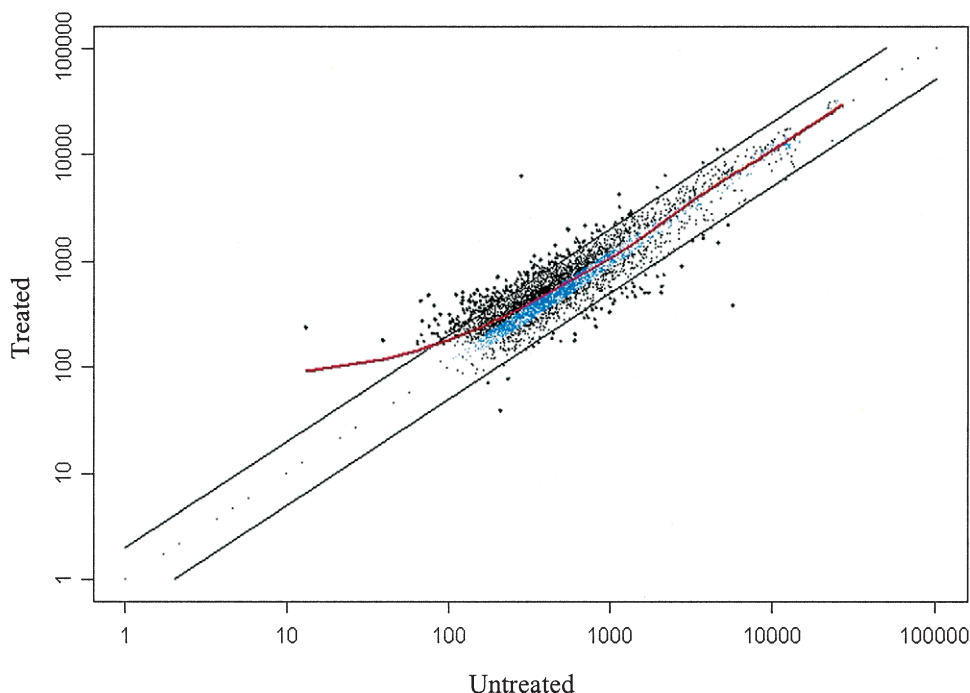


Figure 4 Log-log plot of untreated vs treated conditions. A simple log-log plot, which serves as an at-a-glance visualisation of the fold-differences in expression between any two conditions is shown. Conditions 1 and 2 are the average (arithmetic mean) for a population of untreated and treated mice respectively. Plotting in log-space avoids graphing problems usually associated with large ranges of data values. The central dotted line represents 'no differentiation', i.e. any gene on or near this line has no difference in expression between the two conditions. In order to enable us to distinguish individual dots more easily, we have also reduced the size of dots near this line, the most central ones are additionally coloured blue. The two solid lines represent the limits for twofold up-regulation and down-regulation respectively. Also shown in red is a fitted LOWESS (LOcally WEighted Scatterplot Smoother) line, which in this case only serves as a robust representation of central trends in the global dataset. It should be noted that after data normalisation, one usually expects most of the data points to be within the fold-limit lines, unless global expression changes between the two conditions in question are expected.

individual genes of interest is usually backed up by verification using other techniques such as RT-PCR, *in situ* hybridisation and Northern blotting.

Explorative methods

These largely consist of various clustering techniques (generally without statistical inference) performed on absolute or relative expression values (Quackenbush 2001), to identify genes or samples with similar expression profiles, indicating co-regulation or sample type respectively. If co-regulation or time-effects are of interest, a (graphical) principal-components-analysis can be used to gauge the number of clusters that may be contained within the data, and then this number serves as the input parameter for the number of expected clusters in a K-means or Self-Organising-Map (SOM) clustering approach. Due to the nature of explorative methods, we recommend using several combinations of algorithms and distance measures (SOM and hierarchical, both with

Pearson correlation and Euclidian distance as a minimum) in order to highlight different features in the data.

Conclusions

While not discussed in this review, a laboratory information management system with an associated relational database is key for recording workflow information consistently and retrievably. Gene expression/profiling databases should be publicly available and comply with international standards, which would make an important step in providing useful gene profiling data for the research community (see www.ncbi.nlm.nih.gov/geo, www.ebi.ac.uk/arrayexpress, www.gti.ed.ac.uk/cgi-bin/database/login.pl).

In microarray based experiments there are two major sources of variation, one involving the performance of the technology itself and the other the biological samples. The part of the M-SOPs outlined here aims to minimise

the former through appropriate data processing and analysis steps. Good laboratory protocols serve as a sound basis to minimise the amount of data 'manipulation' necessary and thereby increase efficacy. When used optimally, microarray technology allows for pathway determination, identification of specific gene alterations causative in disease, the ability to model molecular aspects of pathogenesis and the application of drug testing and ultimately it has the potential for being clinically informative for diagnosis and therapeutic treatment.

Acknowledgements

This work was supported in part by grants from SHEFC, the European Union and BBSRC to P G. We wish to thank all our colleagues and collaborators at GTI and BioSS for providing much of the foundation work.

References

- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al.* 2001 Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **29** 365–371.
- Carpenter J & Bithell J 2000 Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19** 1141–1164.
- Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS & Trent JM 2002 Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18** 1207–1215.
- Chuaqui RF, Bonner RF, Best CJM, Gillespie JW, Flaig MJ, Hewitt SM, Phillips JL, Krizman DB, Tangrea MA, Ahram M *et al.* 2001 Post-analysis follow-up and validation of microarray experiments. *Nature Genetics* **32** 509–514.
- Churchill G 2002 Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement* **32** 490–495.
- Davison AC & Hinkley DV 1997 *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- Deyholos MK & Galbraith DW 2001 High-density DNA microarrays for expression profiling. *Cytometry* **43** 229–248.
- Dudoit S, Shaffer JP & Boldrick JC 2003 Multiple hypothesis testing in microarray experiments. *Statistical Science* (In Press).
- Dudoit S, Yang YH, Callow MJ & Speed TP 2002 Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12** 111–139.
- Duggan DJ, Bittner M, Chen Y, Meltzer P & Trent JM 1999 Expression profiling using cDNA microarrays. *Science* **283** 83–87.
- Glasbey CA & Ghazal P 2003 Combinatorial image analysis of DNA microarray features. *Bioinformatics* **19** 194–203.
- Glonck GFV & Solomon PJ 2002 Factorial designs for microarray experiments. In preparation (http://mag.maths.adelaide.edu.au/fact_designs.pdf).
- Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N & Quackenbush J 2000 A concise guide to cDNA microarray analysis. *BioTechniques* **29** 548–562.
- Kerr K & Churchill G 2001 Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77** 123–128.
- Kerr K, Martin M & Churchill G 2001 Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7** 819–829.
- Nature Genetics 2002 The Chipping Forecast II. *Nature Genetics* **32** (Suppl) 461–552.
- Quackenbush J 2001 Computational analysis of microarray data. *Nature Review Genetics* **2** 418–427.
- Quackenbush J 2002 Microarray data normalization and transformation. *Nature Genetics* **32** 496–501.
- Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H & Herzel H 2000 Normalization strategies for cDNA microarrays. *Nucleic Acid Research* **28** E47.
- Speed TP 2002 *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton: CRC Press.
- Wagner E, Garcia Ramirez JJ, Stingley SW, Aguilar SA, Buehler L, Devi-Rao GB & Ghazal P 2002 Practical approaches to long oligonucleotide-based DNA microarray: lessons from herpesvirus. *Progress in Nucleic Acid Research* **71** 445–491.
- Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J *et al.* 2002a Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology* **3** 0062.1–0062.12.
- Yang YH, Buckley MJ, Dudoit S & Speed TP 2002b Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11** 108–136.
- Yee HY & Speed TP 2002 Design issues for cDNA microarray experiments. *Nature Review Genetics* **3** 579–588.

Received 30 January 2003

Accepted 8 May 2003