Biomathematics and Statistics Scotland

David A. Elston

Biomathematics and Statistics Scotland (BioSS; www.bioss.ac.uk) is a specialist organisation delivering consultancy, training and research in statistics, mathematical modelling and bioinformatics. BioSS forms a distinctive element of SCRI Group and plays a unique role in the Scottish research community, bridging the gap between research in the mathematically-based and traditionally more qualitative sciences such as biology.

BioSS manages its consultancy work under four broad scientific areas:

- plant science
- animal health and welfare
- ecology and environmental science
- human health and nutrition.

In each area, BioSS staff have a wide range of different types of interaction with scientists, ranging from the provision of short pieces of advice that allow BioSS expertise to guide a large number of scientific research projects, through to a smaller number of deep, collaborative relationships.

Our ability to support a large portfolio of projects in these four application areas is greatly enhanced by our training courses in quantitative methodologies. These courses increase the understanding and computational abilities of our collaborators, enabling them to perform many analyses with minimal guidance and to discuss their projects with BioSS consultants at a higher level.

BioSS manages its programme of applied strategic research in three broad themes:

- statistical bioinformatics
- systems and process modelling
- statistical methodology.

The research we carry out addresses generic issues encountered in our consultancy work that are not adequately addressed using standard methods. Each research theme is related to each of our four broad scientific application areas, demonstrating the wide applicability of BioSS research.

BioSS inputs to microarray experiments (Christine Hackett, Chris Glasbey, Graham Horgan, Mizanur Khondoker, Claus-Dieter Mayer & Jim McNicol)

Microarrays are one of the many remarkable tools that enable to us to probe ever deeper into the molecular activity taking place within the cells of all living organisms. Whilst microarray technology continues to advance, the quantitative issues raised are generic, and the progress we have made in addressing these issues demonstrates the interplay between BioSS research and consultancy. To simplify the exposition, we consider two colour microarrays, each of which contains spots of cDNA from 10,000s of genes. Extracts from two samples (for example, individual plants or bacterial cultures) are applied to every spot on a given array, and the intensity of colour of each of two dyes on these spots records the levels of gene expression associated with each sample. The samples can be chosen to allow us to find out about expression levels associated with particular types of individuals (different varieties, experimental treatment groups, or varying according to a quantitative trait such as weight). Irrespective of the many variations, microarrays enable scientists to measure the expression of thousands of genes simultaneously, leading to enormously high dimensional data, posing challenges for design and analysis that have been addressed in complementary ways by BioSS staff in Edinburgh, Aberdeen and Dundee.

Design Differences in expression levels between pairs of samples applied to the same array are estimated much more precisely than differences between arrays, hence a key decision in the design of two colour microarrays is the specification of such pairs. Early experiments favoured the so-called reference design in which all samples were compared to a single reference or control sample (for example, [A,R], [B,R], [C,R]). As this only allows indirect comparison of pairs not including the reference sample (for example, [A,B]), more general designs like loop designs (for example, [A.B], [B,C]… [Z,A]) have also been used which allows some comparisons to be estimated with increased precision.

Our work on the design of microarray studies has been to capitalise on information about the samples in two different settings. The first setting is where we have measured a single quantitative trait on each sample and wish to estimate a linear regression between trait values and expression level. We have demonstrated the benefits of ranking the samples in both directions (highest to lowest and lowest to highest), then constructing a loop design which uses pairs of approximately the same rank in the opposite direction. Furthermore, we have established the situations in which additional replication of samples with extreme trait values, at the expense of samples with trait values close to the mean, can be beneficial.

The second setting is where we have information about a large number of genetic markers for each sample. We have investigated the construction of distant pair designs in which the aggregate number of discrepancies between markers within pairs is as large as possible. We have written a computer programme to optimise this criterion for a barley population, using simulated annealing.

Preprocessing Microarray data in raw form almost inevitably contain artefacts due to imperfections in the technology which, if not addressed properly, can obscure biological signals. Preprocessing is therefore a key step in analysis, which BioSS has advanced through *combining multiple laser scans* and using nonparametric methods for *normalisations*.

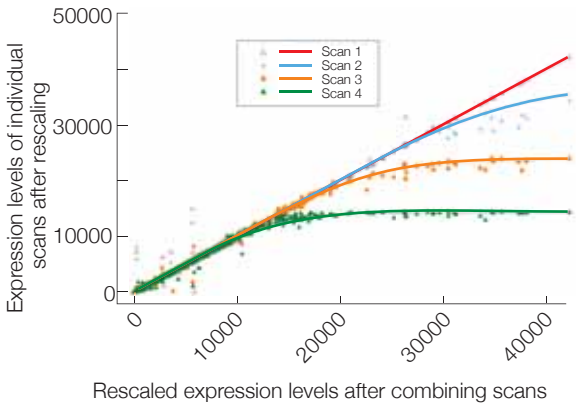

Figure 1 Example of rescaled multiple scan data with fitted nonlinear functional regression model shown by solid lines. Data from different scans with incremental PMT settings are related nonlinearly due to variant amounts of signal saturation.

Combining multiple scans Expression data for analysis are typically derived from a single laser scan of each hybridised cDNA microarray. However, because weakly expressed genes are better measured at high photo-multiplier tube (PMT) gain settings and highly expressed genes at low settings (Fig. 1), there are benefits in combining scans to obtain more sensitive data across the range of expression levels. We have developed a nonlinear functional regression model with errors based on the heavy-tailed Cauchy distribution for robustly estimating gene expression. Software is available via a web interface: www.bioss.ac.uk/ktshowcase/create.cgi

Normalisation is the process of removing systematic variation so that data from different microarrays are on a common scale. There are two main aspects: location normalisation to remove biases in mean expression levels, and scale normalisations to remove differences in the spread of observed expressions for given mean values. Typically, nonparametric methods such as *loess* are used for location normalisation and parametric transformations such as *arcsinh* for variance stabilisation. However as can be seen in Fig. 2a, patterns of variation are typically too complex to be adequately modelled parametrically.

We proposed a new normalisation method using a generalised additive model for location, scale and shape (GAMLSS; Fig. 2b). Simulation studies show that GAMLSS normalisation yields more powerful inference of differential expression than the standard parametric method.

Analysis The first analysis step following the normalisation is to find genes that exhibit evidence for differential expression. The standard statistical approach here is to apply a test for each gene, using data from that gene alone. As most microarray experiments tend to have small sample sizes, the estimate of the standard errors used in such tests will not be precise. For genes with small observed standard errors this value will tend to have been underestimated, leading to false positives, whereas for genes with large observed standard errors this value will tend to have been overestimated, leading to false negatives.

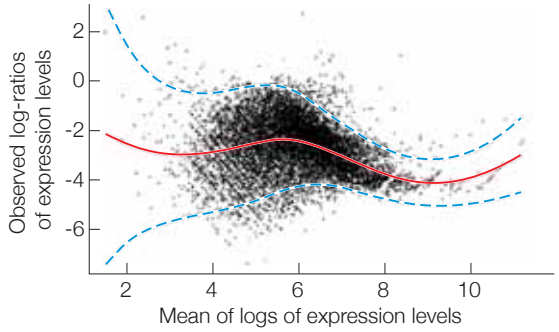

Figure 2a GAMLSS fit to data from a single microarray, showing the log-ratio of the observed colour intensities of the two samples for each gene plotted against the average of the two log intensities. The solid red line shows the location model and the blue dashed lines show the spread ascribed by the scale model.

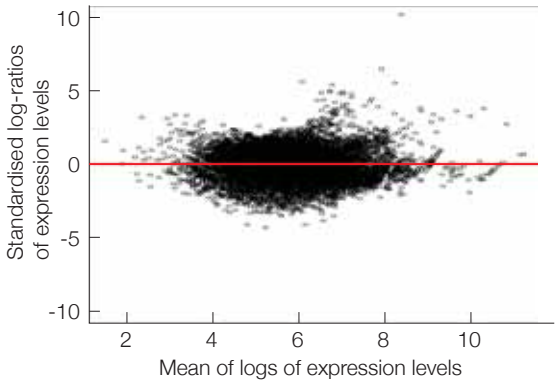

Figure 2b GAMLSS normalised log-ratio plotted against mean log-intensity (x). As intended, the majority of the normalised data (corresponding to genes not showing differential expression) are symmetrically distributed around the zero reference line and have homogeneous variability over the range of mean intensities.

For this reason we use *moderated t-tests* for comparisons of mean expressions in two treatment groups, and equivalent analyses for more complex designs. The process of moderation shrinks the standard error for each gene towards the average standard error observed across all genes to stabilise the analysis. For each gene we obtain a p-value which quantifies the statistical significance of the observed difference (technically, the probability of obtaining a test statistic as extreme as that observed if there were no true effect).

False discovery rate The next step is generally to select a cutoff, p_c, and declare that we have 'discovered' evidence of differential expression for all genes with p-values smaller than p_c. The value of p_c can be

David Elston

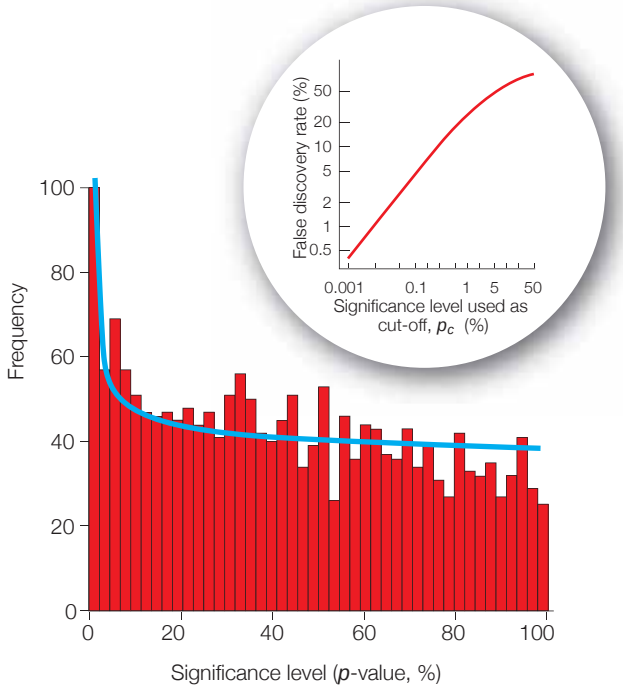

Figure 3 Observed (histogram) and modelled (solid blue line) distributions of significance levels for differential expressions of individual genes, together with (inset) the modelled relationship between gram false discovery rates (FDR) as a function of the significance level used as cut-off, p_c.

thought of as the false positive rate (the probability that a gene without differential expression gets incorrectly 'discovered'). Because the expression levels of so many genes are estimated in each microarray experiment, it is important also to know the proportion of discoveries which are likely to be false. This is the false discovery rate (FDR). The FDR associated with any value of p_c is estimated by modelling the distribution of p-values. Fig. 3 shows an example in which a significance cut-off of 5% corresponds to a FDR estimate close to 50%. The FDR estimate makes us wary of using 5% as a significance cut-off. With this data set, a FDR of 1% would need a p_c-value of around 0.005%.

Identification of regions controlling expression: eQTL analysis Microarrays provide an opportunity to merge the analysis of gene expression data with information on chromosome position provided by DNA markers. Typically such an analysis is carried out using data from a segregating population of offspring obtained from crossing two inbred parents. The outcome is an estimate of the regions of the genome controlling expression of each gene on the microarray.

Often, the locations of many of the genes on the microarray are known and the region associated with gene expression includes the known gene location (referred to as a *cis*-regulated gene). However, the region controlling gene expression is sometimes found to be separate from the gene location (referred to as *trans*-regulation). These sites may be major gene regulators, playing a central role in the molecular interactions taking place within cells and ultimately having effects on many traits exhibited by whole organisms. This information is currently being used by scientists working on completely sequenced species such as *Arabidopsis* to build gene regulatory networks, and should soon become more feasible in crop species such as barley.