Analysis of microarray data

Rebecca E. Walls*[1], Stuart Barber[1], Mark S. Gilthorpe[2] and John T. Kent[1]

[1] Department of Statistics, University of Leeds
[2] Biostatistics Unit, University of Leeds

1 Introduction

In the past, scientists have been restricted by a lack of suitable technology to conducting genetic analyses on only a very few genes at any one time. However, recent years have seen the advancement of DNA microarray technology, which allows the examination of thousands of genes simultaneously, with the aim being to identify genes which are expressed differently from one sample to another.

Gene expression is the term used ideally to describe the imperative process in which a gene converts the coded information stored in its *deoxyribonucleic acid* (DNA) sequence into essential proteins, which are needed to perform and regulate most basic functions. However, it is often the intermediate molecule *messenger ribonucleic acid* (mRNA) through which the protein-coding instructions from the gene are transmitted, that is referred to as 'expression', despite being only a precursor to the protein. In order to identify differences between samples, one would wish to be able to quantify the levels of expression in a sample, and microarrays aim to achieve this through measuring the abundance of the mRNA in that sample.

Briefly, a typical DNA microarray experiment would involve the arrangement of minuscule amounts of the genes of interest, usually printed in a rectangular array of spots, on a single microscope slide. The printed spots are called *probes* and each usually represents one gene sequence. This poster concerns the use of microarrays in a comparative two-channel experiment, testing a treated sample against a control on the same slide. Each of the two different mRNA samples, whose gene sequences are unknown, are labelled with a fluorescent dye (red for treatment and green for control) and mixed together, before being washed over the slide. Matched mRNA molecules will hybridize to any complementary DNA sequence on the slide and the solution will stick to the slide. Any unmatched mRNA will not hybridize and is washed away. A laser scanner is then used to measure both the red and green fluorescent signal emitted at each point on the chip. If a particular gene is highly expressed, it produces many molecules of mRNA, which hybridize to the DNA on the microarry and generate a very bright fluorescent area. Genes that are somewhat less expressed produce fewer mRNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of the messenger molecules have hybridized to the DNA, indicating that the gene is inactive. By comparing the intensity levels of the emitted fluorescent lights between the samples it is hoped that one might be able to identify any differences in gene structures across the various samples. Any spot whose intensity is different between the two channels corresponds, by inference, to a gene that is differentially expressed in the treated versus control, possibly due to a treatment effect.

2 Analysis of a microarray experiment

Suppose we let X_{gR} and X_{gG} denote the red (treatment) and green (control) intensities measured for the gth gene. The *fold change* is defined as $R_g = \frac{X_{gR}}{X_{gG}}$. If we were to plot a scatter plot

of X_{gG} (along the x-axis) and X_{gR} (along the y-axis) for $g = (1, \ldots, G)$, we would expect most genes to fall along the identity line $X_R = X_G$, which suggests that they are expressed to the same degree in both samples. The differentially expressed genes are outliers that lie far from the identity line. If the point lies substantially above the line, it suggests that the gene is expressed to a greater extent in the treated sample than the control. Alternatively, if the point lies substantially below the line, this implies the opposite. Similarly, if we were to calculate R_g for each g, most genes would have values close to one, indicating no difference in expression. Those whose R_g is large (e.g., $R_g > 2$) suggest that these genes are *overexpressed* or *upregulated* in the treated sample, and those whose R_g is small (e.g., $R_g < 0.5$) suggests that these genes are *underexpressed* or *downregulated* in the treated sample.

Early researchers in the microarray field noticed substantial differences in intensity measures between the two channels even amongst microarrays that were treated exactly alike. The differences can usually be linked to systematic effects inherent in the process such as mRNA preparation, the concentration and amount of DNA put on the microarrays, labelling efficiency, hybridization efficiency, lack of spatial homogeneity of the hybridization on the slide, scanner settings, saturation effects and background fluorescence, all of which increase the challenge of analysing microarray data.

Furthermore, there is a dye bias present in almost all 2-channel experiments. Since the red dye molecule is larger in size that that of the green, fewer red molecules effectively bind to the sample. Thus, the intensities from the green channel tend to be generally higher than those from the red channel. The magnitude of the difference tends to depend on the overall intensity, along with contributing factors such as the scanning properties of the dyes, fluctuations in processing procedures or the settings of the scanner. Clearly, the effects of these systematic sources of variation need to be removed to improve the comparability between channels and this is known as *normalization*.

Figure 1 shows data from a typical 2-channel microarray experiment, with the raw data in the left panel. The small clusters of points suggest that the intensities for the replicated spots for each gene have similar measures. This is to be expected and any departure from this may imply some location bias, which could have arisen through experimental procedures. Whilst it seems that the outliers are quite apparent, the departure of the fitted linear model from the identity line is what we aim to improve upon through normalization.

An MVA plot, as seen in the centre panel of Figure 1, is another useful tool in assessing the need for normalization of the two channels. We construct a scatterplot of $M_g = X_{gR} - X_{gG}$ against $A_g = (X_{gR} + X_{gG})/2$, in other words, the difference between the red and green intensities for each spot against the arithmetic mean for each spot. If there is no systematic dye bias, the points on the MVA plot should be randomly scattered around the $M = 0$ line. Here, the funnel shaped distributions of the points around the identity line suggest that the variance is dependent on the mean, specifically, as the mean intensity increases so too does the variance, and an appropriate transformation is often necessary to stabilize the variance. A further concern is that the raw data is often heavily skewed to the right and clearly non-normal, as illustrated in the histogram in Figure 1. It is hoped that a transformation will help to substantially reduce the skewness, prior to normalization.

A logarithmic transformation, $X \to \log(X)$, (often to base 2), tends to be the preferred convention within the scientific field, yet we question whether this choice is perhaps somewhat arbitrary and not always the most appropriate. Here, whilst a $\log_2$ transform substantially reduces the skewness, some skewness does remain, whereas a transformation of the form $X \to X^{-\frac{1}{2}}$ results in a distribution of the intensities that possesses a more desirable, symmetrical form. Moreover, working with transformed intensities does has other advantages: the variation of transformed intensities tends to be less dependent on the magnitude of the values and it im-

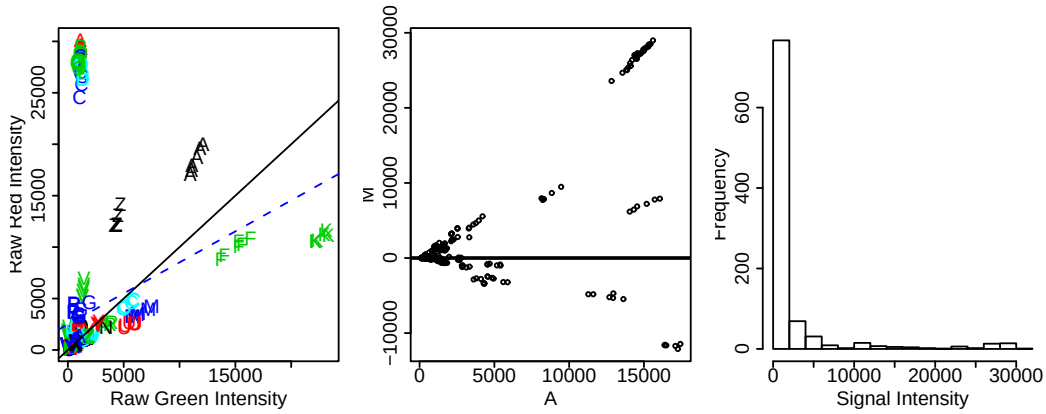

Figure 1: Left: Scatterplot of the raw data. Points denoted by the same character in the same gray-scale correspond to spot replicates. The solid line is the identity line and the dashed line is a linear model fitted to the data. Centre: An MVA plot for the raw data with the line $M = 0$. Right: Histogram of the raw intensities.

proves variance estimation. If we consider the MVA plots of Figure 2, after the log transform, there still appears to be some evident pattern, whilst after the negative square root transform, the points appear to be more randomly dispersed.

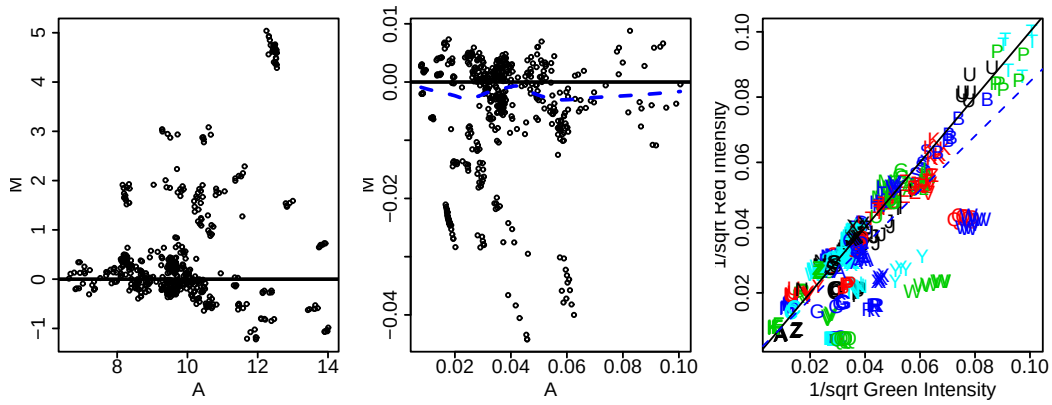

Figure 2: Left: MVA plot after $X \rightarrow \log(X)$ transformation. Centre: MVA plot after taking $X \rightarrow X^{-\frac{1}{2}}$ transform with a loess smoothed curve (dashed line). Right: Scatterplot of the data after transformation with fitted linear model (dashed line).

The loess curve on the centre plot of Figure 2 and the fitted linear model on the scatterplot of the transformed data suggest that there is still some evidence of bias in the data and some form of calibration is required. Many normalization methods exist, the simplest being *global* schemes, which are based around the assumption that spot intensities for a pair of channels are linearly related with no intercept. A result of this assumption is that it should be possible to correct any lack of comparability by simply adjusting every single spot intensity by the same amount, called the *normalizing factor*, regardless of its intensity level. However, both the loess curve and the linear model imply that the normalizing factor needed to adjust the low-intensity measurements is not the same as the factor needed to adjust high-intensity measurements. Hence, *intensity-*

dependent normalization schemes encompass those more sophisticated methods in which the normalizing factor is a nonlinear *normalization function* of the intensity level: $X \to f(X)$.

Many of these methods can be implemented simply in widely-used computer packages and incorporate a logarithmic transform as part of the function. We have been investigating whether using an alternative transformations provides a superior final outcome, based on the discussion above.

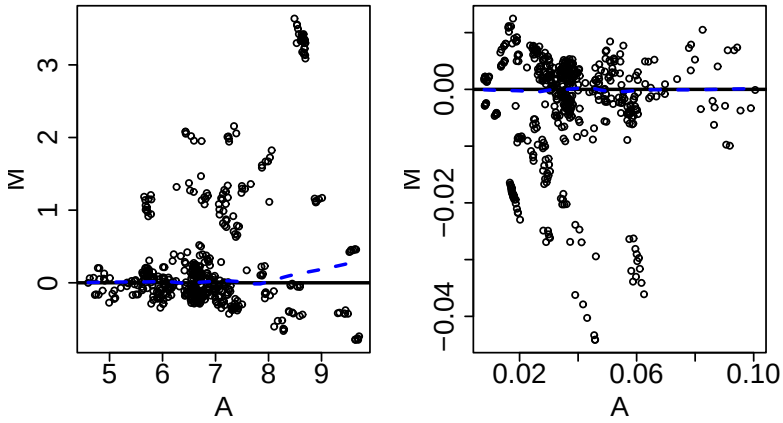

Figure 3: Left: MVA plot after loess normalization on $\log_2$ transformed data. Right: MVA plot after loess normalization on negative square root transformed data. In both plots the dashed lines are loess curves fitted to the data

Whilst the loess curve fitted to the data in the left plot follows the desired $M = 0$ line for the lower intensities, it does deviate for the higher intensities, suggesting some non-constant variance still inherent in the data, whilst in the right plot, the curve follows the $M = 0$ line, hopefully indicative that the variance is now independent of the mean.

3 Future work

We are experimenting with methods to distinguish between within-gene variability and between-gene variability, in order to investigate whether the rather simplistic linear model that is usually assumed is appropriate, or whether a more complicated model is indeed necessary to describe all the data. We are also hoping to explore the possibility of using the empirical Bayes approach of Johnstone and Silverman (2004), both as a way to denoise large datasets before normalization, and as a thresholding procedure on the log ratios after normalization, to identify the truly differentially expressed genes, as an alternative to the somewhat arbitrary boundaries 0.5 and 2 used as convention.

References

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley.

Johnstone, I.M. and Silverman, B.W. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist*, **32**, 1594-1649.