Linear Models in Microarrays: An Introduction

by James Wettenhall

15 October 2004
(with minor edits 23 November 2015)

1 Introduction

This document is intended to be only a very brief introduction to linear models in microarrays. For more detailed information see the limma User's Guide `https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf`.

2 M and A

We will first discuss M and A from the point of two-color cDNA microarrays, in which one can talk about comparisons within a slide (between the two colors) or comparisons between slides. Rather than representing microarray data for a cDNA slide with raw Red and Green intensities, it is better to use log intensities (base 2 is the standard base used for the logarithms). This gives a more symmetrical distribution about the mean values of $log_2 R$ and $log_2 G$ than you would get if you used R and G directly. Furthermore, we are most interested in differences between R and G and overall intensities of spots (the geometric mean of the Red and Green intensities), so we define M, the log differential- expression ratio as $M = \log_2(R/G)$ and A, the log intensity as $A = \frac{1}{2}\log_2(RG)$.

$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$
$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$

M can in fact be used to compare any pair of RNA types in the microarray experiments, whether they are on the same slide or on different slides. It is still called the log differential-expression ratio.

3 Linear Models

To illustrate the simplest case of a linear model, consider two slides for which the same hybridization has been performed for both slides with the same dye colors. In this case, the best estimate for the M value for each gene is simply the average of the two M values for that gene, one from each slide. If a dye-swap was performed, then, one of the M values would have to be multiplied by -1 before taking the average. The fitted M values in limmaGUI refer to the M values after the "averaging" has been done.

Things become more complicated when you want to estimate confidence in your average M values. The t statistic, B statistic and P value in the toptables in limmaGUI are used to provide an overall ranking of genes in order of evidence for differential expression. (By default, ranking is done by the B statistic.) In order to calculate these statistics, limmaGUI must consider all replicates of each gene (whether on the same slide or different slides) and consider the *variation* in M values as well as the magnitude of the M values to decide which genes are differentially expressed. If you just use M (the log differential-expression ratio) to rank genes from a microarray experiment, then you are ignoring all of the information about variability between replicates.

As well as using a linear model to estimate "average" M values, it is also possible to estimate M values for comparisons which were not directly performed in the experiment. If A is hybridized with B and B is hybridized with C, then you can estimate two M values with a linear model. One possible choice is to estimate M for the comparison (A,B) and estimate M for the comparison (B,C). But it would also be possible to estimate M for comparison (A,B) and M for comparison (A,C), even though there was no direct hybridization between A and C. In limmaGUI, this is known as choosing a parameterization. A parameterization can be specified in terms of simple comparisons between RNA types, e.g. (A,B) and (A,C), or for people with a bit of statistical experience, the parameterization can be specified in terms of a design matrix (by pressing an "Advanced" button.) The columns of the design matrix represent the parameters to be estimated by the linear model (e.g. M values for comparisons (A,B) and (A,C)) and the rows of the design matrix are the slides in the experiment. For the simple averaging scenario described at the beginning of this section, the design matrix would simply be a column of two 1's. If there was a dye-swap, then one of the 1's would become a -1. In limmaGUI you can try specifying a parameterization in the simple way (comparisons between pairs of RNA types) and then press the "Advanced" button to see what the corresponding design matrix looks like.