

# Microarray statistical analysis using R and limma

Luc Janss  
ETH Zurich  
or [www.lucjanss.com](http://www.lucjanss.com)  
May 2006.

This is a summary of the exercises from the ARK-Genomics course “Design and Analysis of Microarrays”, Edinburgh 2-4 May 2006. It can also serve as a general short guide for analysis of a 2x2 design in particular, and for analysis of any designs using R and limma in general.

The design of the arrays supplied for the exercise was given as:

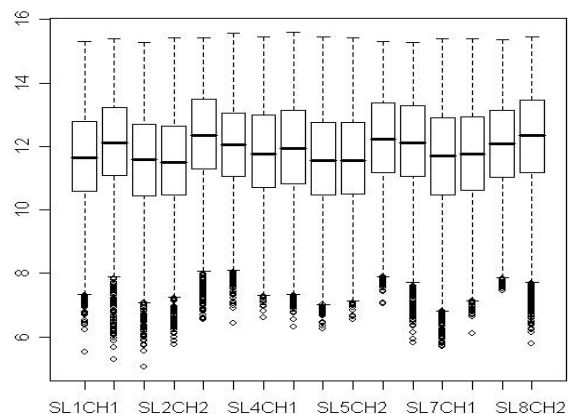
Slide nr	On Ch1 (cy3)	On Ch2 (cy5)
1	Cont-day5	REF
2	REF	Cont-day5
3	REF	Cont-day7
4	Cont-day7	REF
5	REF	Atten-day5
6	Atten-day5	REF
7	REF	Atten-day7
8	Atten-day7	REF

Where REF=biological reference sample; Cont=control sample; Atten=sample after challenge with attenuated APV strain; day5, day7 are two sample days.

## Data exploration

Most of the data exploration is not repeated here. One important exploration is at the right, showing boxplots of the raw intensity data by channel. Every pair comes from one slide, and as can be seen in the design, every two consecutive pairs are dye swaps. This information is important to judge whether a lowess normalisation on all genes, equalizing for each slide the red/green levels and centralizing the M-values looks sensible:

- when different levels in the two channels are consistent between dye swaps, we can argue that these differences may have biological meaning (expression is affected predominantly in one direction) and that it may not be sensible to equalize the red/green levels using a (simple) lowess normalisation (a more complicated normalisation using a slide and its dye swap jointly would be required, but was not explained in the course);



- when different levels in the two channels are not consistent between dye swaps, we can argue that these differences are probably artefacts (some slides are simply too green or too red), and that the levels may be equalized using a simple lowess normalisation.

Here the second appears to be the case; the level in the dye swaps is not consistent, e.g. the first slide (boxplots 1 and 2) shows more red, but its dye swap (boxplots 3 and 4) does not show more green. This is also seen at other slides and their dye swaps. In conclusions, for this data we can be pretty confident that any (small) differences in red and green levels are artefacts and may be equalized using a lowess normalisation, forcing centralized M values and forcing (roughly) as much up- and down regulation.

## **Lowess normalisation**

Also the results from lowess normalisation are not extensively repeated here. The exercise to study selected top genes intended to show how these selections change due to lowess normalisation, e.g.:

- A slide which has a bias in one dye (such as slide 1 which is too red), will show predominantly up-regulated genes when genes are selected on raw data; after lowess correction the selected genes are more evenly distributed over up- and down regulated ones.
- Slides which shows some curved M-A plot (such as slide 8) may for instance show down-regulated genes only at low A values, and up-regulated genes only at high A values; after lowess correction also this gets more evenly distributed over low- and high A values.

## **Analysis using t-test, and computation of FDR**

### **Reference design**

Data here comes from a design in which all samples are (initially) matched with a biological reference sample. This biological reference sample may be from completely different animals with different treatments and conditions, and comparison of our actual samples with this biological reference is not of interest. What can be of interest, though, is the comparison between particular slides, for instance one with the combination REF:Cont-Day5 (slide 2) and one with the combination REF:Cont-Day7 (slide 3). Suppose that for a particular gene:

- REF:Cont-Day5 gives  $M=+1$  (2x up regulated compared to biological reference)
- REF:Cont-Day7 gives  $M=+2$  (4x up regulated compared to biological reference)

As said, the  $M=+1$  and  $M=+2$  per se are not of interest, because they indicate up regulation of this gene compared to the biological reference (whatever it may be), but what is of interest is that the second is more up regulated compared to the first: we can say that *relatively* day7 gives an extra 2x up regulation (or 1 extra M-unit).

The bottom line is that most of the time we can ignore the biological reference, and that we can directly compare between slides and interpret differences as for instance a difference between day5 and day7.

As a primer, the data can be analysed using t-tests. Bearing in mind that we can directly compare between slides and forget about the biological references, the following four comparisons can be of interest:

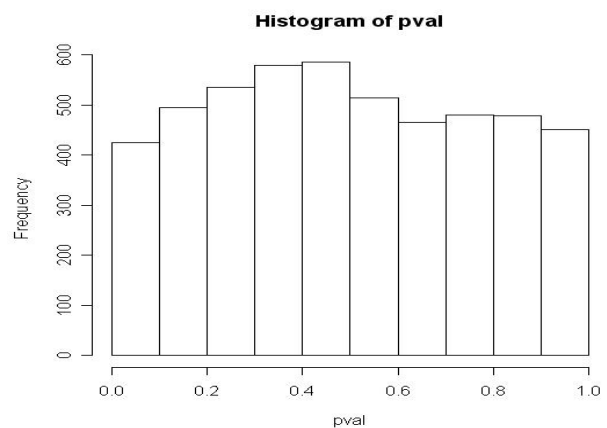
- Comparison 1: day 5 vs day 7 within control (slide 1 and 2 vs 3 and 4)
- Comparison 2: day 5 vs day 7 within attenuated (slide 5 and 6 vs 7 and 8)
- Comparison 3: control vs attenuated within day 5 (slide 1 and 2 vs 5 and 6)
- Comparison 4: control vs attenuated within day 7 (slide 3 and 4 vs 7 and 8)

To take account of dye swaps, signs of M values are reversed when originating from a swap.

Note: more comparison could in principle be made using t-tests, for instance to compare all day 5 (slide 1, 2, 5, 6) with all day 7 (slide 3, 4, 7, 8). But such a crude comparison would have undesirable effects when the number of slides gets unequal; for instance if the group “day 5” has more slides from control than from attenuated, the mean of “day 5” slides becomes affected by the control-attenuated difference. This kind of comparisons are therefore more properly made when also correcting for the treatment effects, and this is the subject of linear models as used further below.

### Comparison 1

For this comparison the commands were supplied, saving all the p-values. A rough first impression whether the entire pool of genes contain some up regulated ones can be obtained by making the histogram plot of the p-values as shown at the right and for which commands were supplied.



This particular histogram is an important one to remember, *because it shows a comparison where not much is happening*. If we make a rough FDR computation, for instance for selecting p-values  $< 0.10$ , we compute that the expected number of false positives is around 500 ( $0.10 \times 5008$ ), and the histogram shows that we do not even reach this level of 500; in other words all p-values  $< 0.10$  are likely false positives, or the whole result from this comparison appears to be just random noise.

### Use of q-values for estimate of FDR (comparison 1)

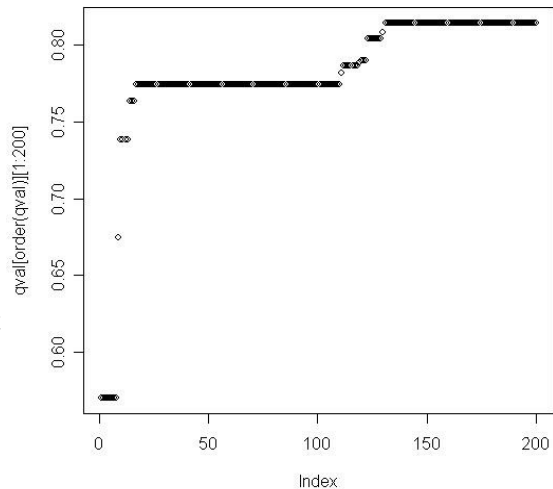
Q-values can be used to say roughly the same as the histogram of p-values but in a more direct and simple way, and for every desired number of genes selected (once you understand the concept). These are the first 10 lowest q-values for the comparison 1:

```
> qval[order(qval)][1:10]
[1] 0.5699485 0.5699485 0.5699485 0.5699485 0.5699485 0.5699485 0.5699485
[8] 0.5699485 0.6747791 0.7387332
```

These q-value say that:

- If you select the first 2 top genes, FDR is about 50% (look at the second q-value), or from the top 2 we expect 1 to be false and 1 to be true.
- If you select the top 5, FDR is also about 50% (5<sup>th</sup> q-value), or from the top 5 roughly 2-3 are false and the other 2-3 are true.
- If you select the top 10, FDR is about 70% (10<sup>th</sup> q-value), or from the top 10 about 7 are false and 3 are true.
- etc.

It can be seen that by selecting a larger group both the number of false positives and the number of true positives increases; it depends on the data how these rates are. This comparison 1 does not show a lot to be going on: after the top 5 the number of false positives already increases faster than the number of true positives and we can at best reckon with a few true positives (among many false ones). A plot as shown on the right is also convenient to study q-values (showing the first 200 q-values using `plot(qval[order(qval)][1:200])`). This plot says that initially FDR is around 50%, or from every 2 genes you take 1 is false, and FDR then quickly jumps to 75-80%, or from every next 10 genes you take about 8 are false. This also says that for this particular comparison we can not make a very sharp selection of differentially expressed genes.



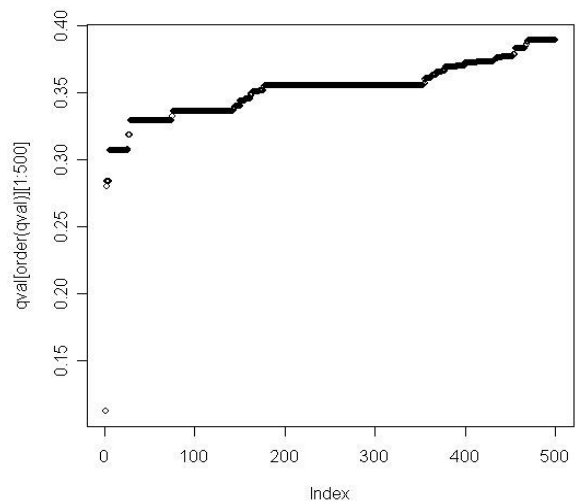
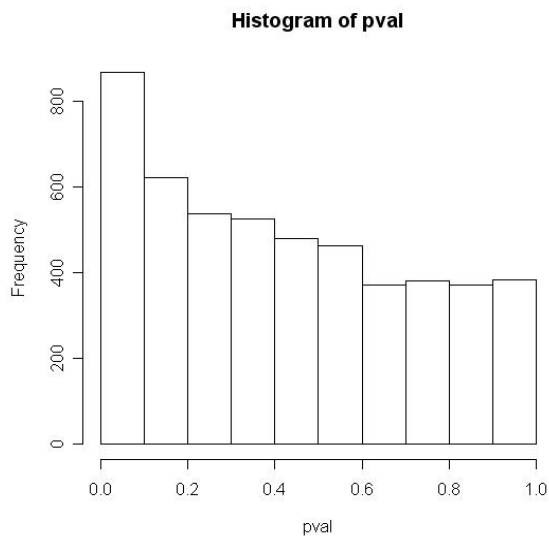
Note (1): this conclusion based on q-values is somewhat more optimistic than the conclusion based on the histogram of p-values; the conclusion based on q-values is the more reliable one. Roughly, however, the conclusion from each approach is the same, saying that this comparison 1 does not show a lot of differential expression and saying that the false detection rate is high.

Note (2): there are some different ways of computing q-values. In bad cases like this comparison 1, it can also happen that the computed q-values are directly near 1, also indicating that (virtually) all findings are false positives.

### Comparison 3

Below are shown the histogram of p values (left) and the first 500 q values (right) for the comparison 3. The histogram of p values shows some differential expression is going on here: in the first group with p-value  $< 0.10$  we expect around 500 false positives, but find over 800 (869 to be precise) positives, and this excess of positives must be real positives. The FDR roughly estimated when selecting all genes with p-value  $< 0.10$  (giving 869 genes) is  $500/869 = 58\%$ . Other rough estimates of FDR based on the p-value histogram are for instance:  $100/193 = 52\%$  when selecting all genes with p-value  $< 0.02$  (giving 193 genes), or  $25/54 = 46\%$  when selecting all genes with p-value  $< 0.005$  (giving 54 genes). The plot of q-values shows a FDR level around 35% for the first 300-400 genes. This means that up to that point roughly for every 3 genes we select more we find one false positive and 2 true positives: so this is relatively favourable because the number of true positives then increases faster than the number of false positives. After the first 1500 genes FDR has increased above 50% and the situation becomes unfavourable as we will then add more false positives than true positives to our selection.

Note: In cases with real good separation of differentially expressed genes, q-values may drop  $< 0.10$ , indicating that the false detection rate is low when selecting the first genes. This is not the case in these comparisons where minimum FDR is around 30%. This may be due to the limited power of this comparison (using only 4 slides), or some particular feature of the data.



## Analysis using Limma

The usual and default linear model analysis produces results from overall-comparisons between days (the pooled results from comparisons 1 and 2 listed above for the t-test analysis), and between treatments (the pooled results from comparisons 3 and 4 listed above). If you do not like this pooling, an interaction term can be added to make a specific day effect within a particular treatment, or a specific treatment effect within a particular day; this is explained below.

The default model (without interaction) is build up as follows:

- The model has a “mean effect” (usual symbol  $\mu$ ) which refers to our statistical reference group, and which we take here to be Cont-Day5. This means Cont-Day5 samples have the mean effect and nothing else.
- Then there is a “Day7 effect” which is the extra effect (of course, it can be negative) when the sample is from Day7 instead of Day5. Samples from Cont-Day7 then have the “mean effect” + “Day7 effect”.
- Then there is an “Attenuated effect” which is the extra effect (of course, it can be negative) when the sample was treated with Attenuated virus instead of control. Samples from Atten-Day5 then have the “mean effect” + “Attenuated effect”.
- Samples from Atten-Day7 will have the mean effect + day7 effect + Attenuated effect. That is, compared to our statistical reference (Cont-Day5, which has “mean effect”), these samples deviate (mean effect + ...) by having the Day7 effect as well as the Attenuated effect.

The building of the design matrix is now straightforward because this actually is only specifying for each sample which effects are present (1) or not (0). For instance samples from Cont-Day5 have 1 for the mean, 0 for Day7 effect and 0 for Attenuated effect, etc. However, we also have to take account of dye swaps as explained below.

## Reference design (again) and accounting for dye swaps

Suppose that we assign our statistical mean to the group Cont-Day5 and that a particular gene in this group has M value of +1; this M value in itself is again meaningless because it expressed the up/down regulation compared to our biological reference. What is clear however is that a dye swap should then have value  $M=-1$ . Hence, if we assign +1 x mean effect to samples from REF:Cont-Day5 (call this the non-swapped sampled), then swapped samples from Cont-Day5:REF will have a -1 x mean effect.

Also for other effects, signs are swapped when the slide is a dye swap. For instance consider that in Cont-Day7 this particular gene has M value +2, so the dye swap should show -2. For the non-swap this is a mean effect (+1) and an extra +1 day7 effect; for the swap all signs reverse and it gets a -1 x mean effect and -1 x day7 effect, which totals to an effect of -2.

General rules for assigning -1's and +1's:

- make a decision what to call swaps and what original slides. For instance call those having REF on cy3 as original, while those with REF on cy5 are called swaps.
- “original” slides get 1's for the mean, and 1's for the effects where appropriate
- “swaps” get -1's for the mean, and -1's for the effects where appropriate
- In the end result, 1's and -1's should be in the same positions.

## Design matrix

The following design matrix is the correct one for analysing day and treatment effects in this design:

slide	mean	day7	Atten
1	-1	0	0
2	1	0	0
3	1	1	0
4	-1	-1	0
5	1	0	1
6	-1	0	-1
7	1	1	1
8	-1	-1	-1

For an interaction we can add an extra effect, for instance to the day7+Atten combination (slides 7 and 8), adding a fourth column which has a 1 and -1 only for slide 7 and 8. This last effect will pull out genes which are particularly differentially expressed in the attenuated samples on day7, instead of being simply “overall differentially expressed” in the attenuated samples. Note: the interaction term can also be made by multiplying vectors for the main terms, after which the correct signs (-1's) need to be added; this can conveniently be done by also multiplying with the mean vector, which has all the -1's in the correct positions.

With the exercises given we can now obtain p-values and top lists again, the latter using topTable. Note: the P.Value from topTable is actually a q-value, when topTable is supplied with the appropriate adjust.method option.