Integrin beta7 Data:
Direct comparison with dye-swaps

Gordon Smyth
16 August 2005

1. Aims

This case study illustrates the use linear modeling to allow and correct for gene-wise dye-effects in the analysis. It also compares background correction methods and the illustrates the use of a spot types file to highlight or to remove control spots.

2. Required data

The integrin beta7 data set is required for this lab and can be obtained from Data/integrinbeta7.zip. You should create a clean directory, unpack this file into that directory, then set that directory as your working directory for your R session using setwd() or otherwise.

3. The integrin beta7 experiment

This experiment was conducted by the Erle Lab in UC San Francisco. This experiment aims to study the cell adhesion molecule integrin alpha4/beta7 which assists in directing the migration of blood lymphocytes to the intestine and associated lymphoid tissues. The goal of the study is to identify differentially expressed genes between the alpha4/beta7+ and alpha4/beta7- memory T helper cells. The study hypothesizes that differentially expressed genes may play a role in the adhesion or migration of T cells. Further details and results of the experiments can be found in Rodriguez (2004}.

The data set given here is a subset from the original dataset consisting of 6 replicated slides from different subjects. Complete information about the array platform and data from each of the individual arrays is available from GEO (accession number GSE 1039). Each hybridization involved beta 7+ cell RNA from a single subject (labeled with one dye) and beta7- cell RNA from the same subject (labeled with the other dye). Target RNA was hybridized to microarrays containing 23,184 probes including the Operon Human version 2 set of 70-mer oligonucleotide probes and 1760 controls spots (e.g., negative, positive and normalization controls spots). Microarrays were printed using 12x4 print-tips and are thus partitioned into a 12x4 grid matrix. Each grid consists of a 21x23 spot matrix that was printed with a single print-tip.

Each of the arrays were scanned using an Axon GenePix 4000B scanner and images were processed using GenePix 5.0 image processing software. The data comprises 6 GenePix gpr output files. Each gpr file contains 23,184 rows and 56 columns; rows correspond to probes (spots) while columns correspond to different statistics from the image analysis output. The gpr files also contain probe names and IDs.

4. Read the targets file

```
library(limma)
targets <- readTargets("TargetBeta7.txt")
targets
```

Exercise. Read the Limma User's section on targets files.

5. Read the data

A filter 'f' is defined so that any spot which was flagged as ``bad'' by the GenePix operator is given zero weight.

```
f <- function(x) as.numeric(x$Flags > -75)
RG <- read.maimages(targets$FileName, source="genepix", wt.fun=f)
RG$printer <- getLayout(RG$genes)
RG$printer
```

The data read by read.maimages() differs slightly from read.GenePix() in the marray package because read.maimages() reads mean foreground intensities for each spot while read.GenePix() reads median foreground intensities, although this difference should not be important here.

6. Locate the control probes

```
spottypes <- readSpotTypes()
spottypes
RG$genes$Status <- controlStatus(spottypes, RG)
plotMA(RG)
plotMA(RG,array=2)
```

Exercise. See the Limma User's Guide section on spot-types files. These provide a method for locating control spots or any other spots of interest by matching patterns using the probe IDs or names.

7. Try different background correction methods

```
RGsu <- backgroundCorrect(RG, method="subtract")  # the default
RGno <- backgroundCorrect(RG, method="none")
RGne <- backgroundCorrect(RG, method="normexp", offset=25)
```

Background correction is more important than often appreciated because it impacts markedly on the variability of the log-ratios for low intensity spots. MA-plots using RGsu will show fanning out of log-ratios at low intensities. When using background subtraction, many spots are not even shown on the MA-plots because the background corrected intensities are negative leading to NA log-ratios. Fanning out of the log-ratios is undesirable for two reasons. Firstly it is undesirable than any log-ratios should be very variable, because this might lead those genes being falsely judged to be differentially expressed. Secondly, the empirical Bayes analysis implemented in

eBayes() delivers most benefit when the variability of the log-ratios is as homogeneous as possible across genes.

Exercise. Examine closely the MA-plots from the three background correction methods. Notice that subtracting produces a decreasing fan effect with intensity while not background correcting produces an increasing fan effect. The 'normexp' produces a more balanced stabilisation of the variances. It also preserves all the data. To examine all the MA-plots efficiently, you may find it helpful to use the following commands, which write all the MA-plots to png disk files in compact format:

```
plotMA3by2(RGsu, prefix="MAsu")
plotMA3by2(RGno, prefix="MAno")
plotMA3by2(RGne, prefix="MAne")
```

8. Normalize

Print-tip loess normalization.

```
MA <- normalizeWithinArrays(RGne)
```

9. Form the design matrix

The basic design matrix simply records which arrays are dye-swapped:

```
design <- modelMatrix(targets, ref="b7 -")
design
```

We will find however that it is much better to include a "dye effect" term in the model. For direct two-color designs, this always takes the form of an intercept term:

```
design2 <- cbind(Dye=1, Beta7=design)
```

10. Find differentially expressed genes

```
fit <- lmFit(MA, design2)
fit <- eBayes(fit)
topTable(fit, coef="Beta7", adjust="fdr")
```

The gene names are very long, so it is convenient to truncate them:

```
tab <- topTable(fit, coef="Beta7", adjust="fdr")
tab$Name <- substring(tab$Name,1,20)
tab
```

Exercise. Check to see if the dye-effect is significant for many probes. Repeat the differential expression analysis without the dye-effect intercept term to see how much difference it makes.

Exercise. Repeat the normalization and differential expression analysis with ordinary background subtraction.

11. Removing control spots

It is usually wise to remove uninteresting control probes from the data before fitting the linear model. This reduces the amount of multiple testing and usually makes the variances more stable so that empirical Bayes is more effective.

```
isGene <- MA$genes$Status=="Gene"
MA[isGene,]
```

Exercise. Repeat the differential expression analysis without the control probes.

Acknowledgements

Thanks to Yee Hwa Yang for making the beta7 data available.

References

1. Rodriguez, M., Paquet, A. C., Yang, Y., and Erle, D. J. (2004).Differential gene expression by memory/effector T helper cells bearing the guthoming receptor integrin alpha4/beta7. BMC Immunology 4, 5-13.