Supplementary methods

Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis

Andrew J. Holloway[†], Alicia Oshlack[†], Dileepa S. Diyagama, David D. L. Bowtell, Gordon K. Smyth

1 Introduction

Here we show a single gene example of the non-linear model fit. A gene is taken from the set of arrays in the cDNA experiment and in the affy experiment using the dilution series described in the paper. Briefly, there are six mixtures of RNA from the Jurkat and MCF7 cell lines starting with 100% MCF7 and decreasing in the amounts shown in table 1. For the cDNA arrays each mixture is hybridized against a pure Jurkat sample and a dye swap replicate is done for each hybridization giving 12 arrays in total. For the Affy arrays each mixture is hybridized to an array twice with two extra arrays that are the pure Jurkat reference giving 14 arrays in total.

Data is extracted, background corrected and normalized in the bioconductor package limma as outlined in the paper. This produces an MAList object `MA` for the two-colour arrays and an expression matrix `e` for the Affy arrays produced using the affy package in biocomductor. For each gene we have 12 log ratios (M) for the cDNA arrays and 14 expression level estimates (E) for the Affy arrays. To show the computations we give a complete session, including commands and output in the R programming environment for one gene on the array. The gene we are looking at it the LCK gene (Figure 3a in the main paper) which appears on both platforms as gene number 152 on the cDNA arrays and gene 4418 on the Affy arrays.

```
> M <- MA$M[152,]
> M
      A1_r1        A1_r2        A2_r1        A2_r2        A3_r1     A3_r2
 5.749777042 -6.483113303  4.379823378 -4.049547767  3.445271245 -3.406163837
      A4_r1        A4_r2        A5_r1        A5_r2        A6_r1     A6_r2
 2.304004820 -2.182671899  1.077180085 -1.115646322  0.069668611 -0.007702478
> E <- e[4418,]
> E
    A1_r1     A1_r2     A2_r1     A2_r2     A3_r1     A3_r2     A4_r1     A4_r2
 5.550927  4.731717  7.912309  7.632732  8.311479  8.308532  8.703293  8.702519
    A5_r1     A5_r2     A6_r1     A6_r2      B_r1      B_r2
 9.916827  9.796382 11.188361 11.163835 11.124529 11.083071
```

It can be seen that the M values for the dye swaps of the cDNA arrays change the sign of the differntial expression. Also, as the mixtures progress from A1 to A6 and more Jurkat in is the mixture, the magnitude of the differential expression decreases as expected. For the Affy arrays the expression level increases as more Jurkat is present in the mixture indicating the gene is more highly expressed in Jurkat.

We define the concentrations of MCF7 in the sample as `C`. We also include the dye swap term d which unswaps the dyes *insilico*. For the cDNA arrays we also estimate the gene specific dye effect by including the term δ=`del` in the nonlinear model. To fit the model we use the nls function in R and estimate R (the expression ratio of MCF7 to Jurkat) and δ.

```
> C<-c( 1.00, 1.00, 0.94, 0.94, 0.88, 0.88, 0.76, 0.76, 0.50, 0.50,
0.00, 0.00)
> d<-rep(c(1,-1),6)
> fit <-nls(M~del+d*log2(C*R+(1-C)),start=list(R=0.01,del=0))
> summary(fit)

Formula: M ~ del + d * log2(C * R + (1 - C))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
R    0.013186   0.002352   5.607 0.000225 ***
del -0.018260   0.107062  -0.171 0.867976
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3709 on 10 degrees of freedom

> log2(0.01318600)
[1] -6.244849
```

For this gene on the cDNA array the estimate of the log ratio of MCF7 to Jurkat is -6.24. The estimate of the gene specific dye effect is -0.018 and the residual standard error σ is 0.3709.

For the Affy array analysis we estimate the expression levels of MCF7 and Jurkat.

```
> C<-c( 1.00, 1.00, 0.94, 0.94, 0.88, 0.88, 0.76, 0.76, 0.50, 0.50,
0.00, 0.00,0.00,0.00)
> fit<-  nls(E~log2(C*M+(1-C)*J),start=list(M=10,J=10))
> summary(fit)

Formula: E ~ log2(C * M + (1 - C) * J)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
M   37.313      5.685   6.564 2.68e-05 ***
J 2143.576    146.613  14.621 5.21e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 0.3171 on 12 degrees of freedom
>  log2(37.31260/2143.57623)
[1] -5.844213
```

For this gene the estimate of the log ratio of MCF7 to Jurkat is -5.84 and the residual standard error ϕ is 0.3171.

It can be seen that the fold change for the Affy arrays is nearly the same as for the cDNA arrays. In this example σ is slightly larger than ϕ for the direct design and this would increase if a reference design was used.

The pure error sum-of-squares is defined as the sum of the difference of each measurement from the mean for that sample squared. ie

$$SS_{PE} = \sum_{i=1}^{n} (y_{ki} - \bar{y}_{k\cdot})^2 \tag{1}$$

where y are the M values or E values, i is the array and k is the mixture A1 to A6. For the cDNA array the pure error is calculated as follows.

```
> M <- d*(M-dye)
> PE <- anova(lm(M~factor(C[1:12])))
> PE
Analysis of Variance Table

Response: M
                 Df Sum Sq Mean Sq F value    Pr(>F)
factor(C[1:12])  5 48.420   9.684  175.96 2.024e-06 ***
Residuals         6  0.330   0.055
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> SS.PE<-PE$Sum[2]
> SS.PE
[1] 0.3302161
> SS.LOF <-  1.375668 - SS.PE
> SS.LOF
[1] 1.045261
```

For this example the pure error sum-of-squares is on 6 degrees of freedom while the lack of fit estimate is on 4 degrees of freedom. The F statistics for lack-of-fit is calculated as

```
> F.LOF <- (SS.LOF/4)/(SS.PE/6)
> F.LOF
[1] 4.748077
> pf(F.LOF,4,6,lower=FALSE)
[1] 0.04538153
```

This gene has one of the highest fold changes and the lack-of-fit is only just significant.