Lab: Preprocessing and quality control

Wolfgang Huber, Robert Gentleman

June 6, 2004

For this lab, you need the packages *estrogen*, *arrayMagic*, *vsn*. There are two alternative, partially overlapping paths through this lab, one for people with a preference for Affymetrix platform, and one for people using other array platforms (e.g. spotted and/or two-color arrays).

Exercise 1. For Affymetrix people

```
> library(estrogen)
> openVignette("estrogen")
```

This contains exercises for Affymetrix preprocessing, rudimentary quality control, up to the production of a gene list. Working through these exercises should take you 20-60 min, depending on your experience with R and the affy package. If you have brought your own data, process it in a similar way.

Exercise 2. For spotted-array people

```
> library(arrayMagic)
> openVignette("arrayMagic")
```

This contains exercises for spotted chip preprocessing and quality control. Working through these exercises should take you around 30 min, depending on your experience with R. If you have brought your own data, process it in a similar way.

Exercise 3. Go through the exercises in the vignette for the package *vsn*. Section 5 is on the comparison of different normalization methods. Using either the lymphoma data that comes with the package, or your own data, produce plots like Fig. 9 that also take into account

- loess normalization
- print-tip wise loess normalization
- print-tip wise vsn normalization

- with and without subtraction of the 'local background'
- subtraction of a smoothed version (2D-loess) of the 'local background'

Exercise 4. Instead of the plots of the previous exercise, use *receiver operating characteristic curves* to compare the performance of different preprocessing methods for finding differentially expressed genes. A receiver operating characteristic curve is a plot of false positive rate (on the x-axis) versus true positive rate (on the y axis).
In many cases, the "truth" is not actually known for every gene on the chip. We can however use the following estimates

$$\begin{aligned} \mathrm{TP} &= \mathrm{P}(1 - \mathrm{FDR}) && (1) \\ \mathrm{FP} &= \mathrm{P} \cdot \mathrm{FDR} && (2) \end{aligned}$$

where TP and FP are the number of true and false positives, P is the length of the genelist select with a certain method at a certain threshold, and FDR is the false discovery rate estimated through a permutation procedure, e.g. from the function *fdc* in the package *arrayMagic*.