Some Quality Methods for Affymetrix Microarrays

R. Gentleman

January 10, 2007

1 Introduction

For those of you using some variant of Unix, this lab will be very easy. You will simple need to load the affyQCReport package, locate the CEL files and then run the `affyQAReport` function. This will produce a nice report that you can read, and try to understand.

For those of you using Windows, things will be a bit harder. We can spend some time, if you want trying to set up your Windows computer to have a working Latex etc, which is what is needed for the automatic report generation. Or, you can spend some time reading in the data and doing various things to it.

We have two data sets that you can use. One is small, and is contained in the affydata package. It is called `Dilution` and contains four microarrays. It has the advantage of being small, and hence reasonable on computers with limited processing capabilities.

The second data set is much larger and comes from a study on CLL, you will find the files in a folder that has been supplied to you, and you will need to run all of the code (hence it is a bit more realistic).

We demonstrate the different functionality, here, only on the Dilution data.

2 The **simpleaffy** package

The simpleaffy package contains code that implements and uses some of the QC/QA procedures that are recommended by Affymetrix.

Its use is fairly straightforward, and it has a very good vignette, and other documentation that you can read to become familiar with the different functions.

Simple use is as follows:

```
> library("simpleaffy")
> library("affydata")
> data("Dilution")
> Dilution

AffyBatch object
size of arrays=640x640 features (48010 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=4
number of genes=409600
```

```
annotation=hgu95av2
notes=
```

We see that the Dilution data is an instance of the **AffyBatch** class. And now we use the qc function.

```
> qcStats = qc(Dilution)
> qcStats

An object of class "QCStats"
Slot "scale.factors":
[1]  2.636213e+04 3.650470e-207 3.650470e-207 6.213191e-174


Slot "target":
[1] 100


Slot "percent.present":
20A.present 20B.present 10A.present 10B.present
   48.79208    49.82178    49.37822    49.75842


Slot "average.background":
     20A      20B      10A      10B
94.25323 63.63855 80.09436 54.25830


Slot "minimum.background":
     20A      20B      10A      10B
89.52555 60.01397 77.32196 49.22574


Slot "maximum.background":
     20A      20B      10A      10B
97.66280 68.18998 83.24646 57.62283


Slot "spikes":
    AFFX-BioB-3_at AFFX-BioC-3_at AFFX-BioDn-3_at AFFX-CreX-3_at
20A       9.450289       9.466491        9.442115       9.434839
20B       5.463099       5.519242        4.826695       4.434308
10A       5.463099       5.519242        4.826695       4.434308
10B       5.285038       5.367343        4.817450       4.544132


Slot "qc.probes":
    AFFX-HSAC07/X00351_3_at AFFX-HSAC07/X00351_M_at AFFX-HSAC07/X00351_5_at
20A                7.409595                6.755193                7.250279
20B                5.712357                5.397154                6.064566
10A                5.712357                5.397154                6.064566
10B                5.922709                4.844184                5.492612
    AFFX-HUMGAPDH/M33197_3_at AFFX-HUMGAPDH/M33197_M_at
```

2

```
20A                     8.226581                    5.762284
20B                     5.372438                    6.560110
10A                     5.372438                    6.560110
10B                     5.158105                    5.933775
      AFFX-HUMGAPDH/M33197_5_at
20A                     6.380310
20B                     6.594109
10A                     6.594109
10B                     6.153654


Slot "bioBCalls":
20A.present 20B.present 10A.present 10B.present
        "P"         "P"         "P"         "P"
```

You can also plot the output, and it is shown in Figure 1. This is a nice visual summary of the data, and it can be used to help interpret the output, or to show it to others involved in the analysis.

3   Other plots

The affy package provides an *RNA degradation plot*, this plot looks at each array and tries to assess the amount of RNA degredation, as manifested in changes in hybridization over the probes (laid out 3' to 5'). In this plot you are looking to see if any arrays are really different from the others as this will indicate potential problems. These plots cannot be interpreted between experiments and the slopes etc are meaningless.

```
> rnaDeg = AffyRNAdeg(Dilution)
```

The plot is shown in Figure 2

4   Metrics from **affyPLM**

The affyPLM provides yet another set of per array diagnostics that can be used to help assess array quality. There is extensive documentation in that package to help interpret and understand the different approaches that are taken there.

First we need to fit the basic model.

```
> library("affyPLM")
> dataPLM = fitPLM(Dilution)
```

There are two basic plots that should be examined, the Normalized Unscaled Standard Error (NUSE) plot and the Relative Log Expression (RLE) plot. In the NUSE plot low quality arrays are those that are significantly elevated or more spread out, relative to the other arrays. NUSE values are not comparable across data sets. In the RLE plot an array that has problems will either have larger spread, or will not be centered at M = 0, or both.
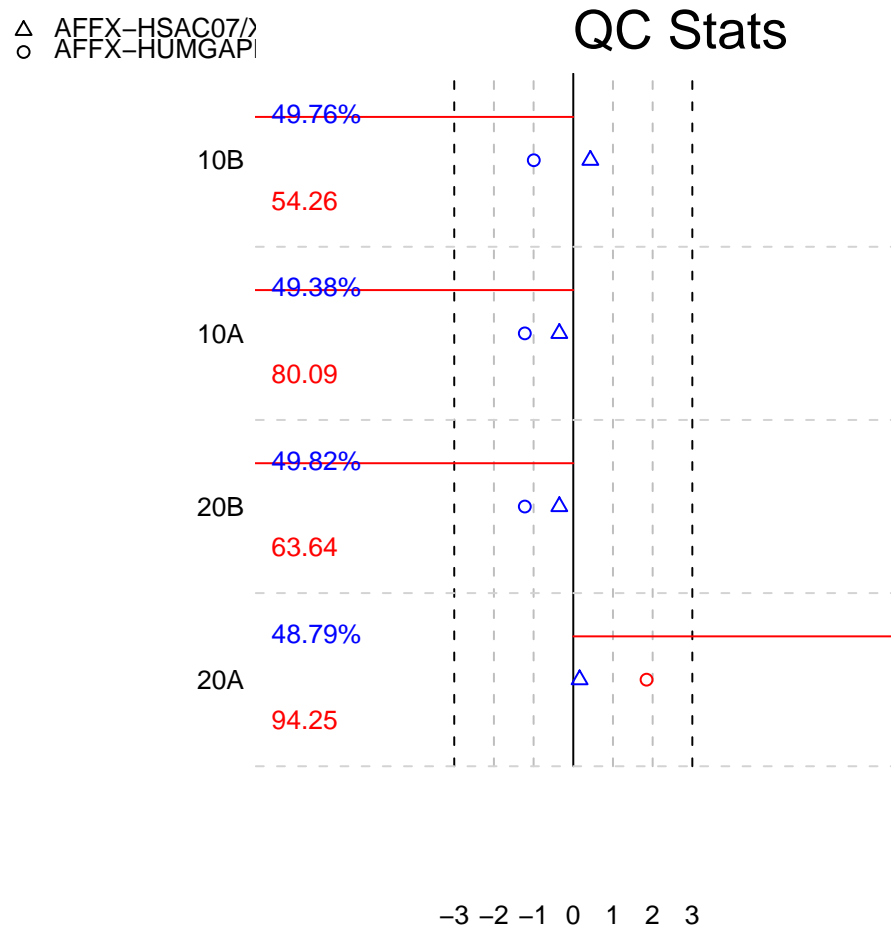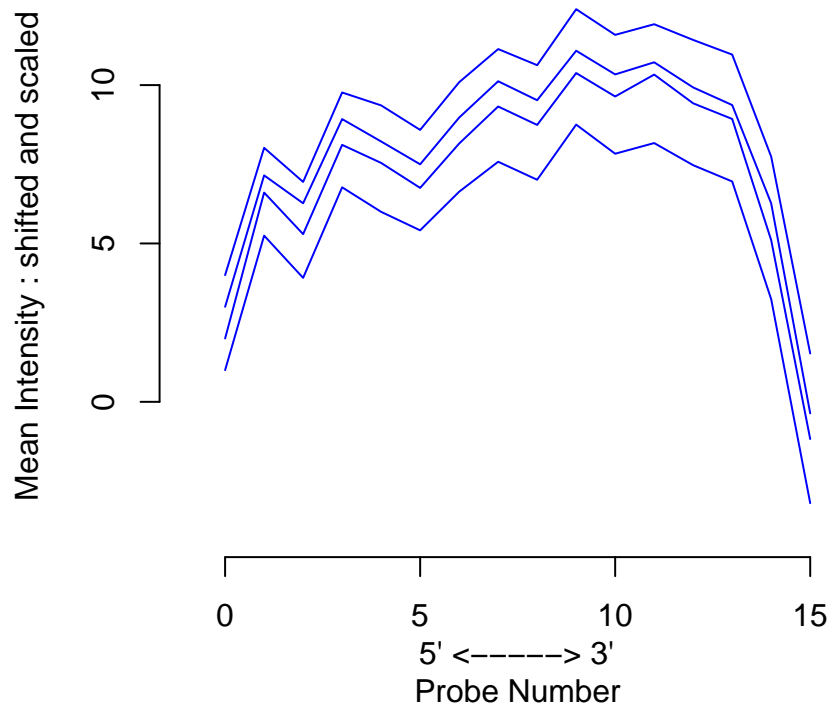
Figure 1: The simpleaffy QC plot.
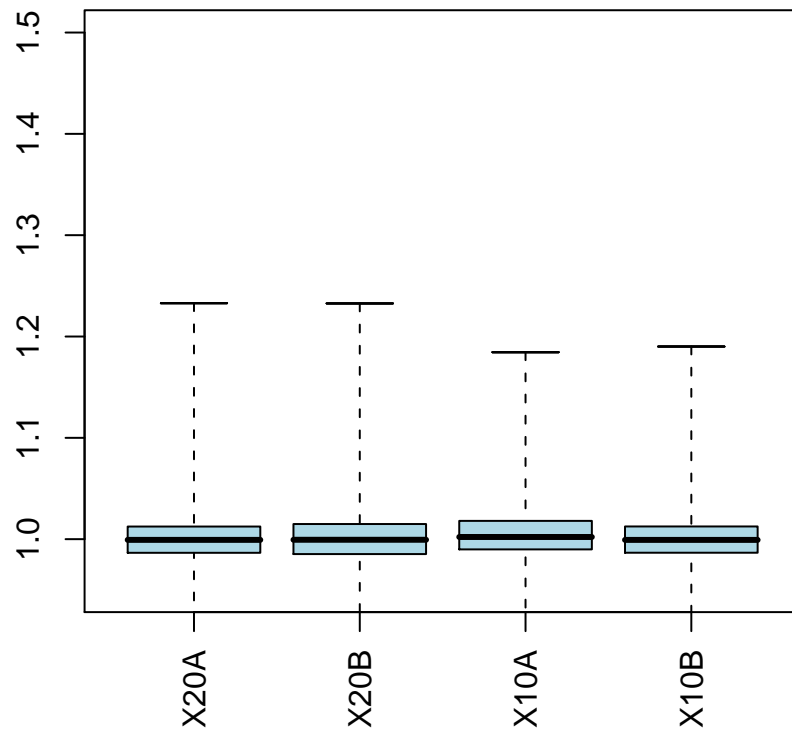
Figure 2: The RNA degradation plot.

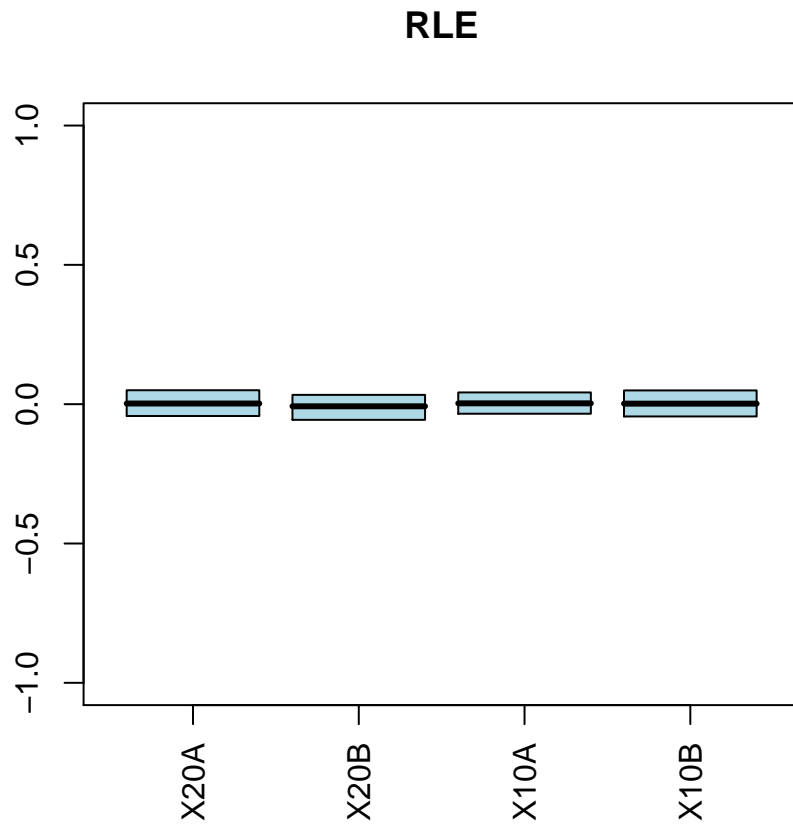Figure 3: The NUSE plot.

RLE

Figure 4: The RLE plot.

5 The simple way

So, if you have a working Latex, then you can do the following:

```
> library("affyQCReport")
> library("affydata")
> data(Dilution)
> rep1 = affyQAReport(Dilution)
> openQAReport(rep1)
```

This provides the data in different formats, and a textual description of the output. In some cases we can automate recommendations et. There are also a few more plots that are produced to help better interpret the data.

6 Using your own data

You can use your own data (or the supplied CLL data) by reading in the data in the CEL files and creating an `affyBatch` object. The function `read.affybatch` can be used to do this.

7 Session Information

The version number of R and the packages and their versions that were used to generate this document are listed below.

```
R version 2.5.0 Under development (unstable) (--)
powerpc-apple-darwin8.8.0

locale:
C

attached base packages:
[1] "splines"  "tools"     "stats"      "graphics"  "grDevices" "utils"
[7] "datasets" "methods"    "base"

other attached packages:
    affyPLM        gcrma matchprobes hgu95av2cdf    affydata  simpleaffy
  "1.11.13"      "2.7.1"     "1.7.4"    "1.15.0"    "1.11.1"     "2.9.1"
 genefilter     survival        affy      affyio     Biobase
   "1.13.7"       "2.30"   "1.13.12"     "1.3.1"   "1.13.29"
```