affyQCReport: A Package to Generate QC Reports for Affymetrix Array Data

Craig Parman and Conrad Halling

April 30, 2008

Contents

1 Introduction

This document describes an R package for generating QC reports. The goal of this project is to create a tool to allow users of the popular Affymetrix GeneChip[1] arrays to quickly access the data quality of a batch of processed arrays. The package makes use of the *affy* Gautier et al. (2004) package for reading cell files and generating several of the plots. The QC plot from the package *simpleaffy* Miller (2005) is also used. Several new plots are generated and a printable pdf file is created.

The functions in the package will work on normalized or non-normalized data. The results of these QC procedures will change slightly depending on whether normalization has been done. When in doubt it is recommended that the procedures be run before and after normalization.

The example data included in the *affydata* package can be used to generate a example report.

---

[1]`www.affymetrix.com/`

2 Getting Started

After starting R, the package should be loaded using the following.

```
> library(affyQCReport)
```

This will load *affyQCReport* as well as the *affy* and *simpleaffy* packages and their dependencies. The example data named `Dilution` which is an object of class `AffyBatch` his loaded with the following data command.

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

To generate an example report simply use the method `QCReport`

```
R> QCReport(Dilution,file="ExampleQC.pdf")
```

Any valid `AffyBatch` object can be used as long as the corresponding CDF environment is also available. Phenotypic data contained in the `AffyBatch` object will be used to group arrays, but it is not required. If the `AffyBatch` object needs to be created from the cel files, a call directly to the various forms of the `ReadAffy` method can be used. For example the graphical user interface widget can be used for input as shown below.

```
R> QCReport(ReadAffy(widget=TRUE))
```

The methods for creating `AffyBatch` objects are described further in the *affy* package documentation.

The report consists of 6 pages. The first page consists of a list of the sample names and an index number that is used to identify each array in later plots. The second page consists of two plots made using the *affy* package. The first is a box plot of the **pm** intensities and the second plot consists of density plot of the log these intensities. The third page is the QC plot generated with the *simpleaffy* package. This plot shows the $3' : 5'$ ratios for spiked-in and control genes specific to the array type. Additionally the percentages of present gene calls and background levels are given.

The next two pages are generated by analyzing the intensities of the positive and negative control elements on the outer edges of the Affymetrix arrays. The fourth page contains box plots of the intensities of these positive and negative elements. The fifth page is a plot of the "center of intensity" (**COI**) for the positive and negative border elements. The sixth page is a heat map of the array-array Spearman rank correlation coefficients of the array intensities. The arrays are ordered using the phenotypic data (if available) in order to place arrays with similar samples adjacent to each other. Arrays of similar expression patterns will have a higher correlation coefficient.

If desired each page can be separately generated by a single function call with the `AffyBatch` object as the argument. For example the following command will generate the titlepage.

```
R> titlePage(Dilution)
```

3 Figure Details

This section will describe the details of each page of the report and the function call to generate the individual pages. An expample of each page is shown in a figure.

3.1 Report page 1

The first page simple list the names of the arrays and assigns an index number to be used in future plotting. The names taken from the data set by use of the *sampleNames* method of the *affy* package. These sample names and indexes are also listed on several other plots. An example is shown in Fig. 1. The plot is generated with the following command.

```
R> titlePage(Dilution)
```

3.2 Report page 2

The second page consists of two plots. The first is a boxplot plot of the all **pm** intensities and the second plot consists of kernel density estimates of these intensities. Both of these methods are defined in the *affy* package. These plots are useful assessing the overall signal quality for the arrays. Any array with a low average intensity or a significantly different shaped density would be suspect. An example is shown in Fig. 2. The plot is generated with the following command.

```
R> signalDist(Dilution)
```

3.3 Report page 3

The third page is the QC plot from the *simpleaffy* package. This plot shows the $3' : 5'$ ratio for spiked-in and control genes specific to the array type. Additionally the percentage of present gene calls and background levels are given. An example is shown in Fig. 3.The plot is described in detail in the document *QC and Affymetrix data* included in the *simpleaffy* documentation. The following is an excerpt from that document describing the plot.

> The figure is plotted from the bottom up with the first chip being at the base of the diagram and the last chip in the QCStats object at the top. If the standard steps for generating a QCStats object are followed, then this corresponds to the order of your samples in the AffyBatch object. Dotted horizontal lines separate the plot into rows, one for each chip. Dotted vertical lines provide a scale from -3 to 3. Each row shows the %present, average background, scale factors and GAPDH / β-actin ratios for an individual chip.

3

AffyBatch QC Report

Array Index	Array Name
1	20A
2	20B
3	10A
4	10B

Wed Apr 30 02:21:21 2008
Produced by AffyQCReport R Package

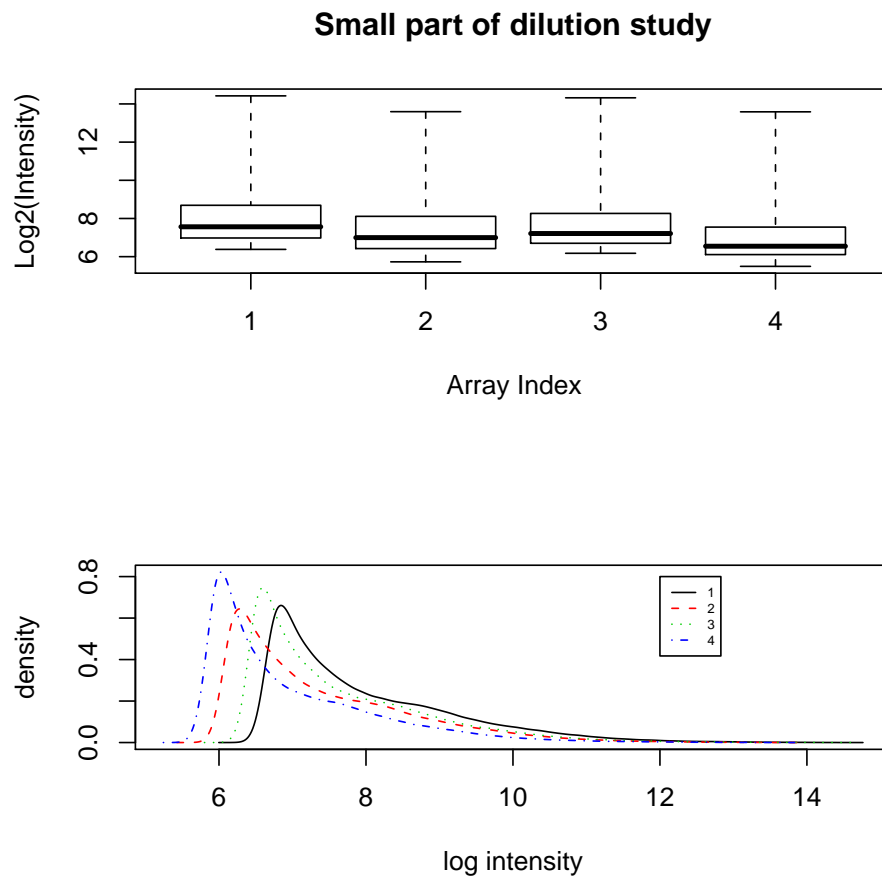Figure 1: First page: Table of arrays in the data set.

Figure 2: Second page: Boxplot and histograms of **pm** intensities.

- GAPDH $3' : 5'$ values are plotted as circles. According to Affymetrix they should be about 1. GAPDH values that are considered potential outlier (ratio > 1.25) are coloured red, otherwise they are blue.

- β-actin, $3' : 5'$ ratios are plotted as triangles. Because this is a longer gene, the recommendation is for the $3' : 5'$ ratios to be below 3; values below 3 are coloured blue, those above, red.

- The blue stripe in the image represents the range where scale factors are within 3-fold of the mean for all chips. Scale factors are plotted as a line from the centre line of the image. A line to the left corresponds to a down-scaling, to the right, to an up-scaling. If any scale factors fall outside this Ś3-fold regionŠ, they are all coloured red, otherwise they are blue.

- % present and average background, are listed to left of the figure.

The plot is generated with the following command.

```
R> plot(qc(Dilution))
```

3.4   Report page 4

The next two pages are generated by analyzing the positive and negative control elements on the outer edges of the Affymetrix arrays. For each array the intensities for all border elements are collected. Elements with an intensity greater the 1.2 times the mean for that group are assumed to be positive controls. Elements with a signal less that 0.8 of the mean are assumed to be negative controls. This method of separation into positive and negative controls is used so that exact details of the arrangement of these elements is not required. Elements falling in between these cut offs are not used in further calculations.

The fourth page consists of box plots of the positive and negative elements. The means and standard deviations of the intensities for each array should be comparable. Large variations in the positive control can indicate non-uniform hybridization or gridding problems. Variations in the negative controls indicate background fluctuations. The plot (shown in Fig. 4) is generated with the following command.

```
R> borderQC1(Dilution)
```

3.5   Report page 5

As a further test, the elements are separated based on which edge of the array they are located. The mean values for the left, right, top, and bottom elements are calculated for positive and negative controls. Once the elements are separated into positive and
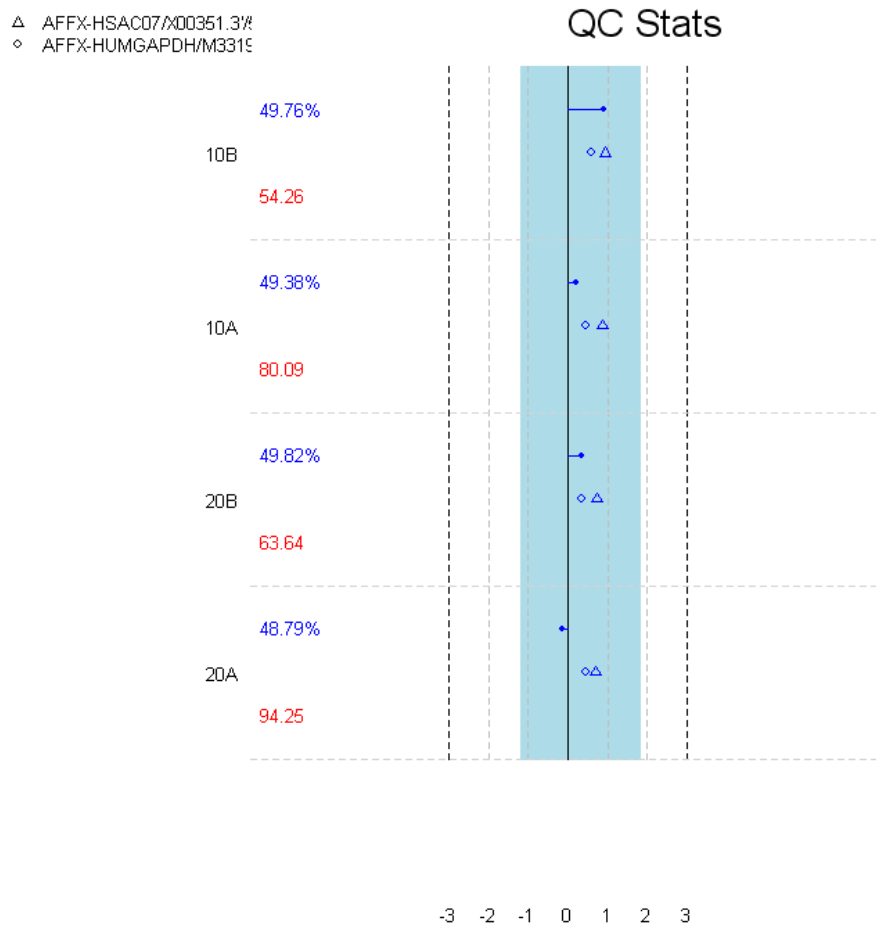
Figure 3: Third page: *simpleaffy* QC plot of $3' : 5'$ ratios and percent present calls.
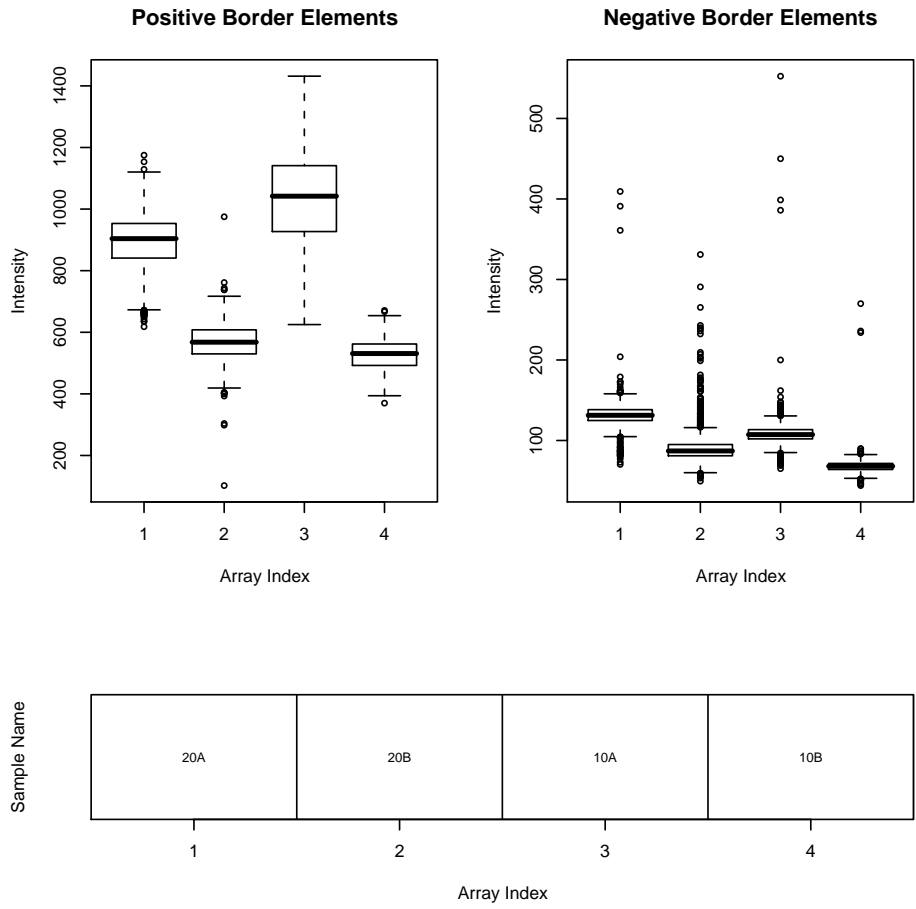
Figure 4: Fourth page: Boxplot of positive and negative feature intensities.

negative controls, and further divided by the four locations, the "center of intensity" (**COI**) for the controls is calculated. If the hybridization is uniform across the array, the location the **COI** for the positive elements will be located at the physical center of the array. Any spatial variations in the hybridization, such as those caused by a bubble being present during hybridization, will cause the **COI** to move from center. Another cause to the **COI** being off center is a slight misalignment of the grid used to determine the cell intensities.

The **COI** is plotted on a relative scale where the point (0,0) is the center and 1 and -1 represent the edges of the array. Some variation to the **COI** is expected but an array with visible intensity variations stands out in these plots as an outlier. Any array that where the **COI** has coordinate with and magnitude greater that 0.5 is flagged by labeling the data point with the array index.

A similar plot is made for the negative controls. This plot is a measure of the uniformity of the background across the array. Again arrays where the **COI** has coordinate with and magnitude greater that 0.5 is flagged. An example is shown in Fig. 5. The plot is generated with the following command.

```
R> borderQC2(Dilution)
```

3.6   Report page 6

The sixth page is a heat map of the array-array Spearman rank correlation coefficients. The arrays are ordered using the phenotypic data (if available) in order to place arrays with similar samples adjacent to each other. Self-self correlations are on the diagonal and by definition have a correlation coefficient of 1.0. Data from similar tissues or treatments will tend to have higher coefficients. This plot is useful for detecting outliers, failed hybridizations, or mistracked samples. See in Fig. 6 for an exaple. Of course caution must be used in deciding if an array should be discarded, because the differences in the expression patterns might be due to interesting biology, not a processing error.

The plot is generated with the following command.

```
R> correlationPlot(Dilution)
```

Positive Elements

Y Center of Intensity position

X Center of Intensity position

Negative Elements

Y Center of Intensity position

X Center of Intensity position

Sample Name

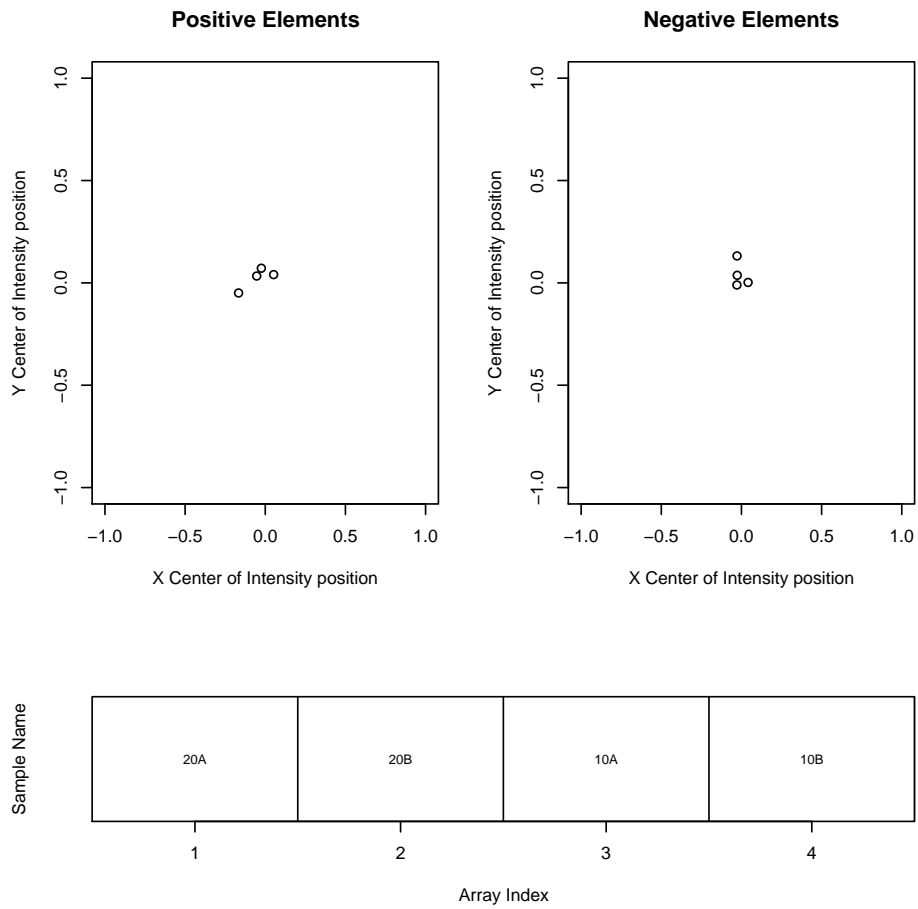| 20A | 20B | 10A | 10B |

1 2 3 4

Array Index

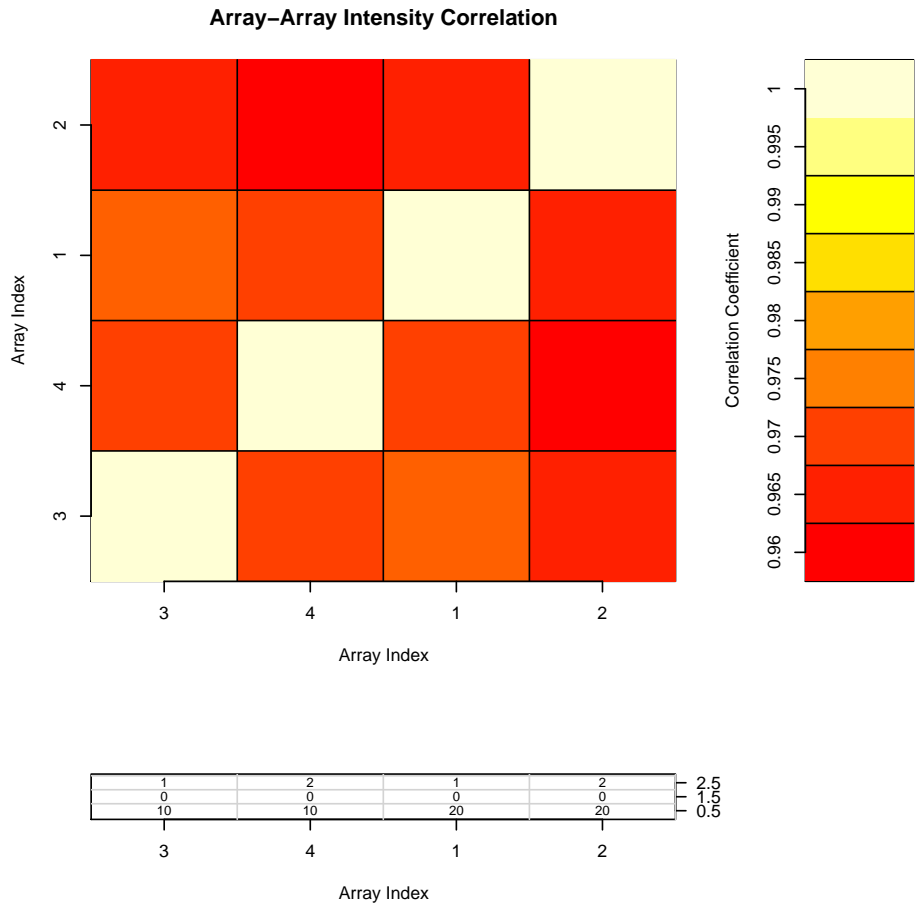Figure 5: Fifth page: "Center of intensity" for positive and negative feature intensities.

Figure 6: Sixth page: Array-array Spearman rank correlation coefficients

References

Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy–analysis
of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.
http://bioinformatics.oupjournals.org/cgi/content/abstract/20/3/307.

Crispin J Miller. simpleaffy, 2005. http://bioinformatics.picr.man.ac.uk/simpleaffy/index.jsp.