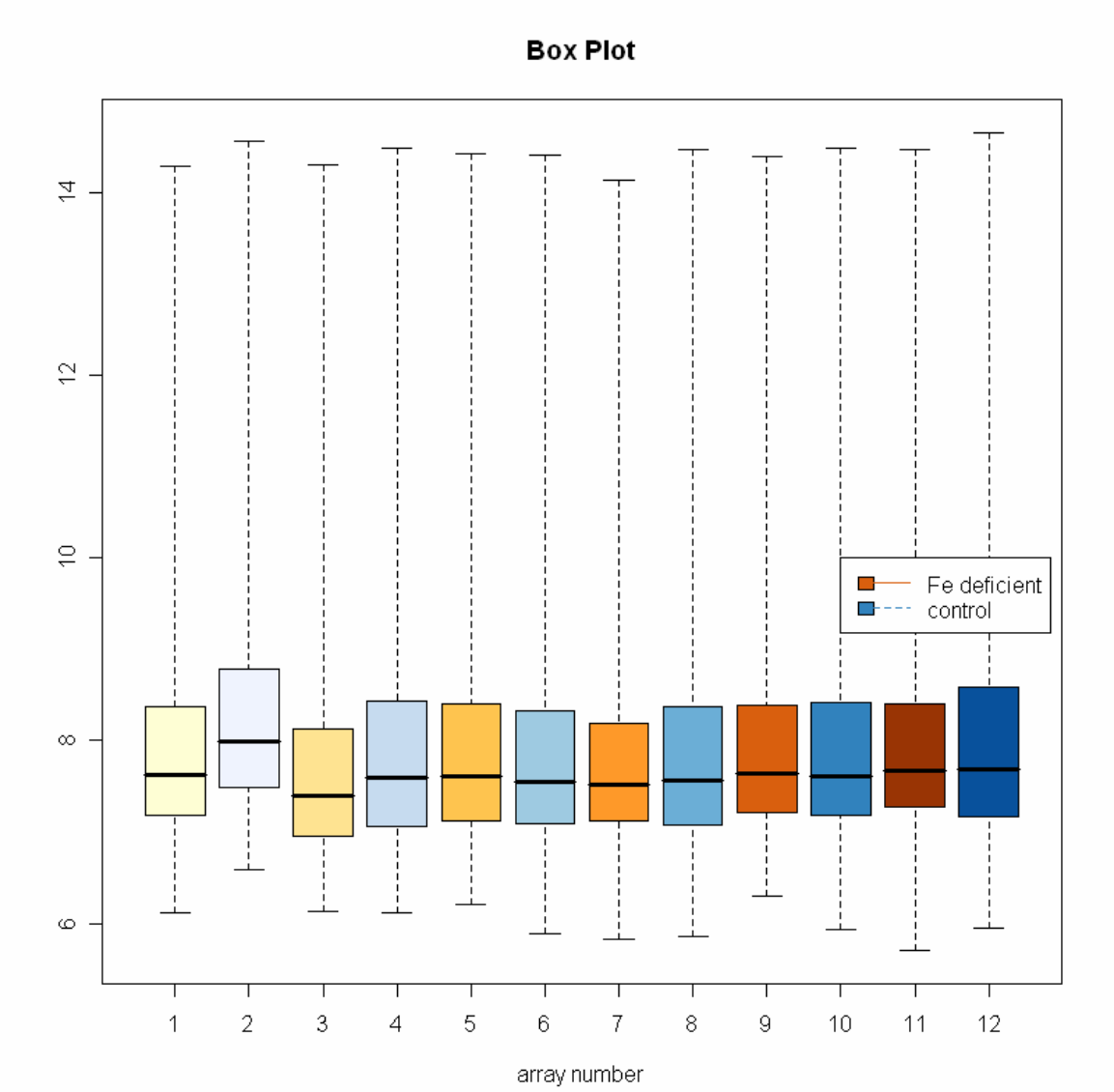
# Quality Assessment

*[ Referentie: Michiel E. Adriaens, Processing Affymetrix GeneChip Data ]*

## Raw data plots

To check if the quality of the data is acceptable a number of statistical test are done.

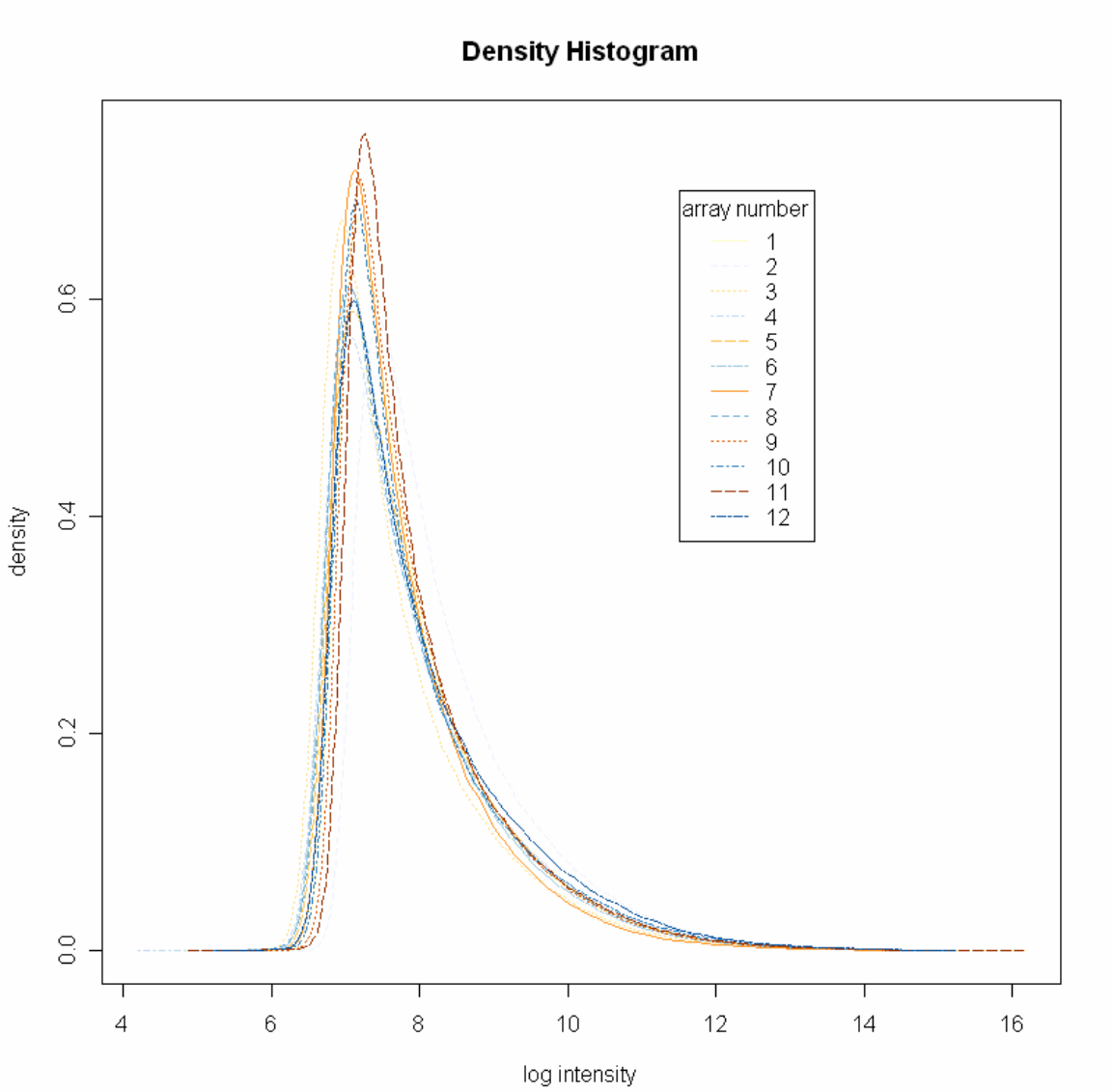
*Box plot*
First, a box plot of the raw data is plotted:




Compared tot the other arrays in the Fe deficient group, the interquartile range and median intensity of array 3 are slightly shifted downwards. Array number 2 is slightly shifted upwards when compared to the other arrays in the control group.

The box plot also shows that arrays number 7 and 11 have a smaller interquartile range while array number 12 has a bigger interquartile range when compared to the other arrays.

*Density Histogram*
Next, the density histogram is plotted:

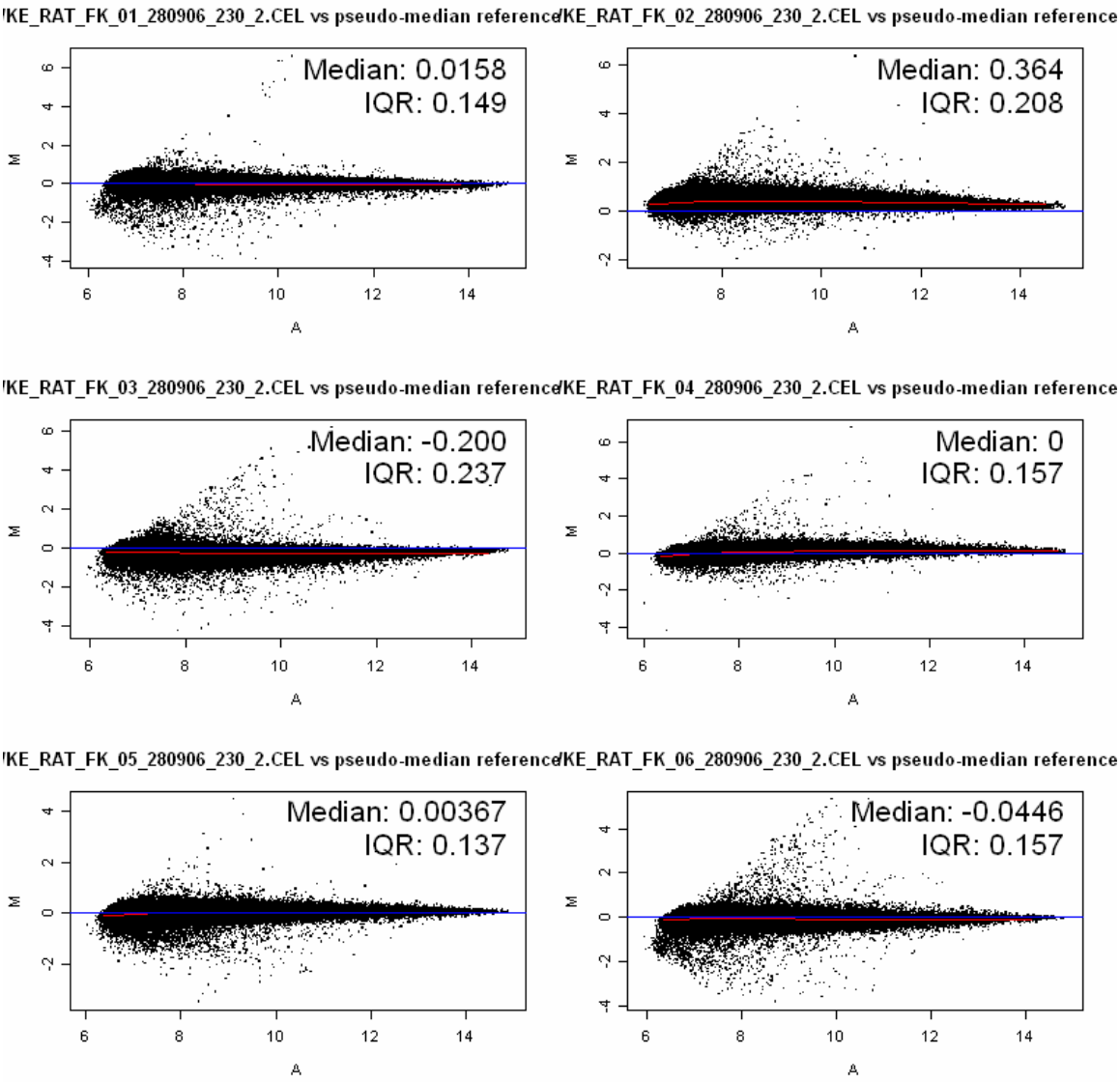



The curves in the density histogram are very comparable. All the curves have a maximum density at a log intensity value of about 7.5, except for the curve of array number two which is shifted slightly to the right and has a maximum density at a log intensity value of about 8.

*M-A plots*

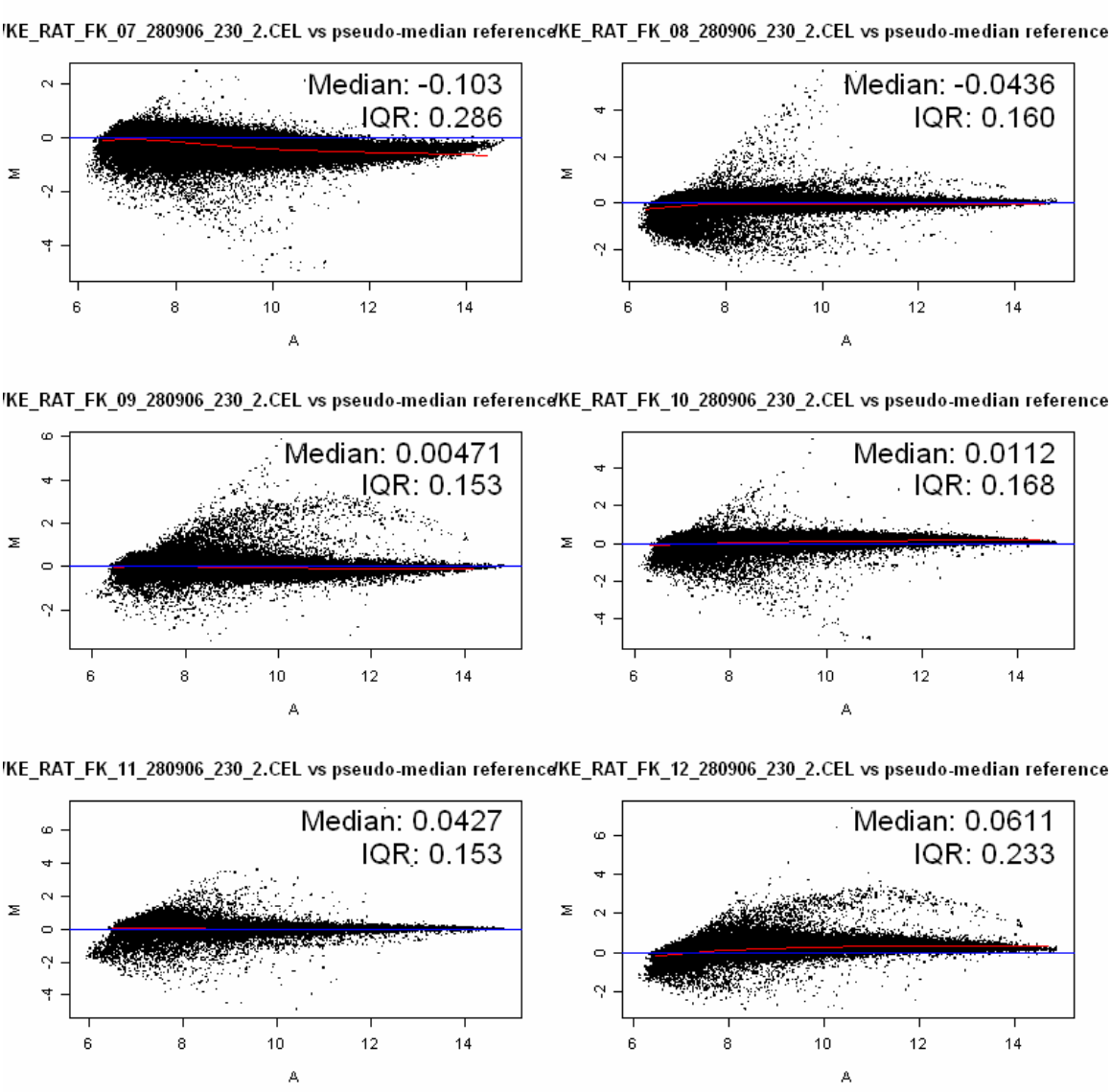
M-A plots are used to assess the difference of the $\log_2$ intensity of two arrays for each probe (M) versus the mean of the $\log_2$ intensities of two arrays for each probe (A). Ideally these plots are shaped like a comet and symmetric about the line M = 0.

Array 1-6:




For arrays 1 to 6 the largest deviations can be seen in array 2 and there is a small deviation in array 3. The other arrays all are symmetric about the line M=0.

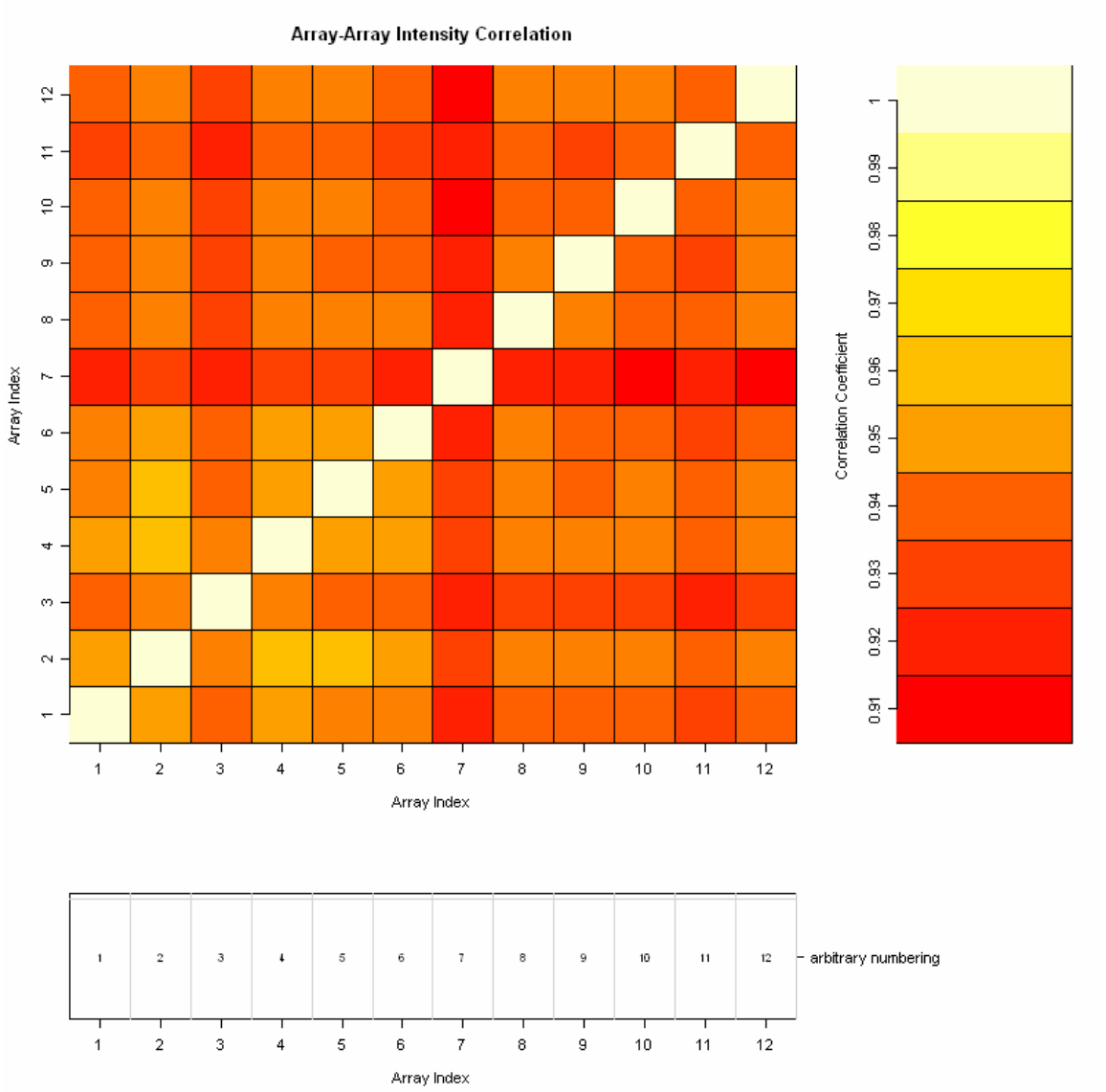
Array 7-12:




For arrays 7 to 12 the largest deviations can be seen in array 7 and 12.

*Correlation plot*

In the correlation plot the correlation coefficient for each array with each of the other array is visualized. A high correlation is to be expected, especially within the groups.




The correlation plot shows that array 7 has the lowest correlation with the other arrays. Arrays 3 and 11 also show slightly lower correlation.

Based on these plots only array number 7 is of questionable quality. Arrays number 3, 11 and 12 show a slightly lower quality in some of the plots, so an eye will be kept on these arrays.

*Affimatrix style quality assessment*

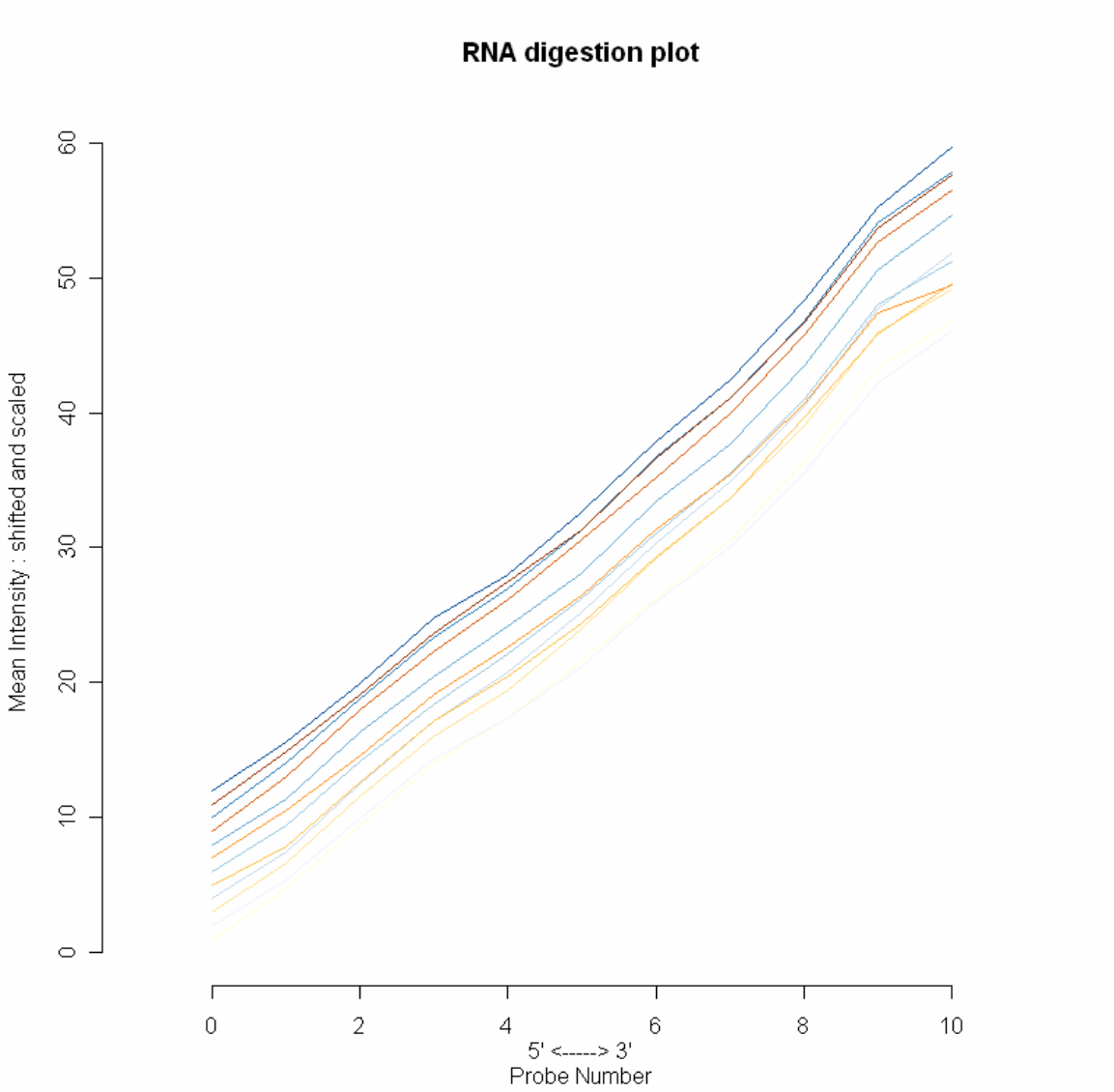
To further assess the quality of the data the AffyMetrix algorithms where used to calculate the average background, the scale factors, the percentage present call and 3'/5' ratios. These values can be seen in the following table:

| **Array Number** | **Average Background** | **Scale Factors** | **Percentage Present calls** | **3'/5' ratios of control probes (beta actine, GADPH)** | |
|---|---|---|---|---|---|
| Criterion | Comparable for all chips > ok | Within 3 fold> ok | Comparable for groups > ok | Below 3 > ok | |
| 1 | 98.83704 | 0.5231616 | 62.17885 | 1.593941 | -0.032297092 |
| 2 | 119.41148 | 0.3999762 | 62.99238 | 1.404486 | -0.044394124 |
| 3 | 84.43426 | 0.6383073 | 60.81868 | 1.847646 | -0.011128770 |
| 4 | 88.37134 | 0.4699430 | 63.84450 | 1.881658 | 0.083003442 |
| 5 | 91.76367 | 0.5021018 | 63.07920 | 1.542759 | -0.039093610 |
| 6 | 88.71948 | 0.5448165 | 62.58079 | 1.520448 | -0.002624696 |
| 7 | 95.83127 | 0.7470292 | 57.28158 | 1.925460 | 0.018510004 |
| 8 | 89.32572 | 0.5099548 | 62.60651 | 1.841503 | 0.010006354 |
| 9 | 103.57337 | 0.5389596 | 61.21419 | 1.784199 | 0.059047204 |
| 10 | 97.89478 | 0.4469118 | 63.23354 | 1.709405 | 0.084026246 |
| 11 | 107.02650 | 0.5024593 | 61.18846 | 1.623564 | 0.048440280 |
| 12 | 95.40536 | 0.3895214 | 65.93460 | 1.550901 | 0.085502890 |

The table shows the calculated values with the corresponding criteria for acceptable quality of the data. It can be seen that all arrays qualify as acceptable according to these AffyMetrix algorithms.

*RNA digestion plot*

The main purpose of the RNA digestion plot is to highlight differences in the laboratory treatments of the arrays. The lines in the plot should be roughly parallel but not horizontal.




The curves are almost parallel and straight, indicating that the data is of acceptable quality. **(????)**

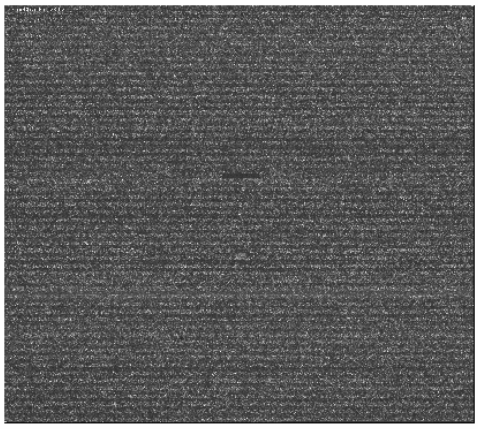
*Probe Level Mode*

By fitting an RMA model to the arrays useful data is created. By taking the difference of the observed signal and the fitted signal the residual for each probe is calculated.

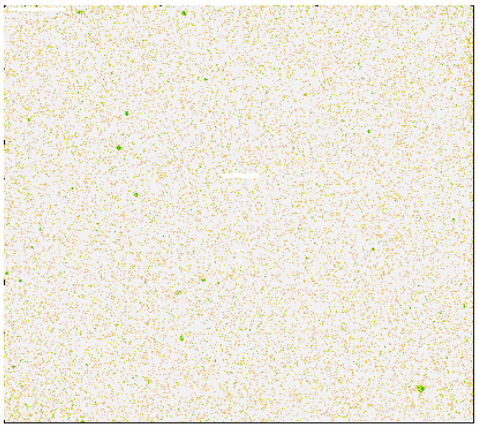
The weight and residual of each probe can be plotted to create pseudo-images. These are very useful for detecting artifacts on arrays that could pose potential quality problems.

First, the pseudo-images of array number 4, which is assumed to be of good quality, are plotted:

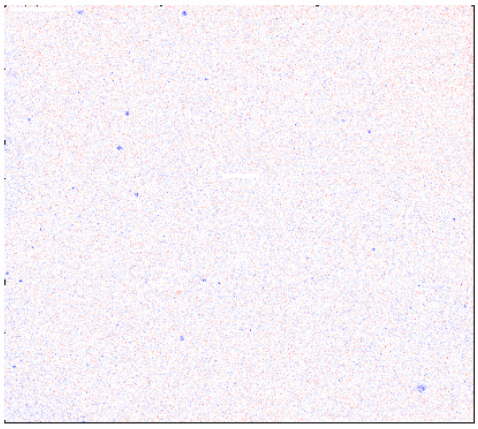
WKE_RAT_FK_04_280906_230_2.CEL



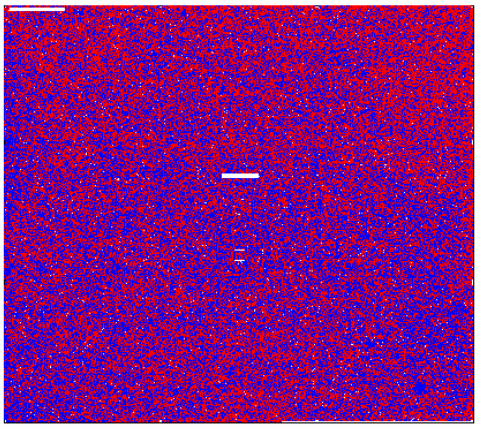
WKE_RAT_FK_04_280906_230_2.CEL



WKE_RAT_FK_04_280906_230_2.CEL



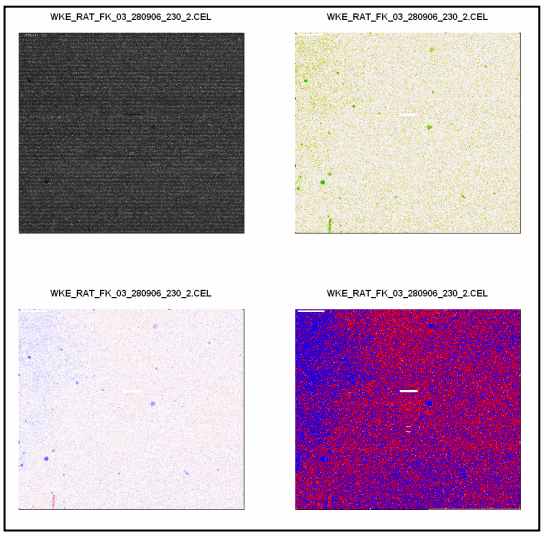
WKE_RAT_FK_04_280906_230_2.CEL



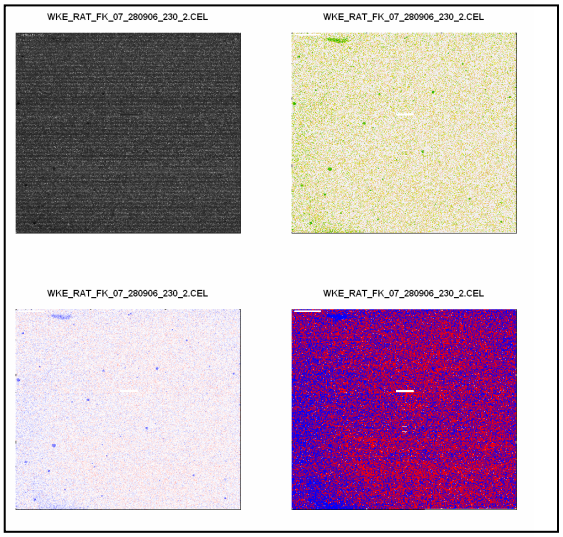
No deviations like stains can be seen in these pseudo images and the images all have light colors. This means the residuals and weights are low and the quality can be assumed to be acceptable.

Next, the pseudo images of the questionable arrays (3, 7, 11 and 12) are plotted.

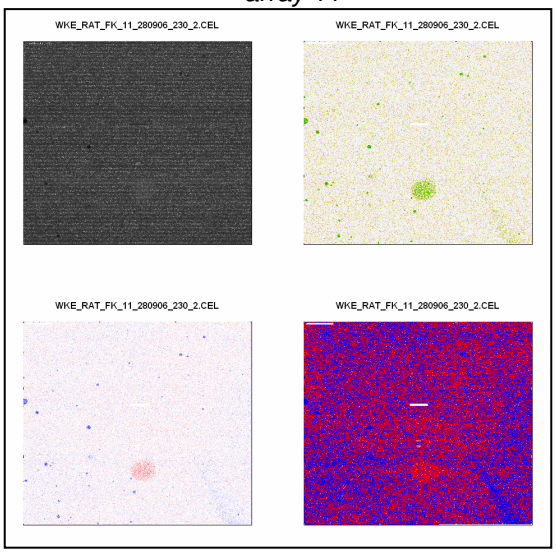
*array 3*



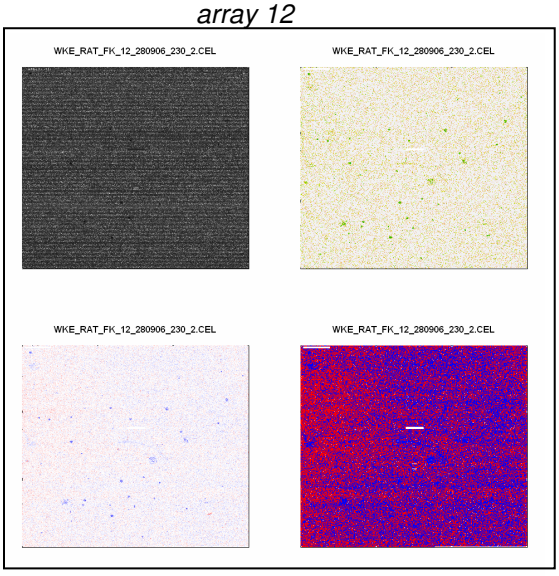

*array 7*




*array 11*

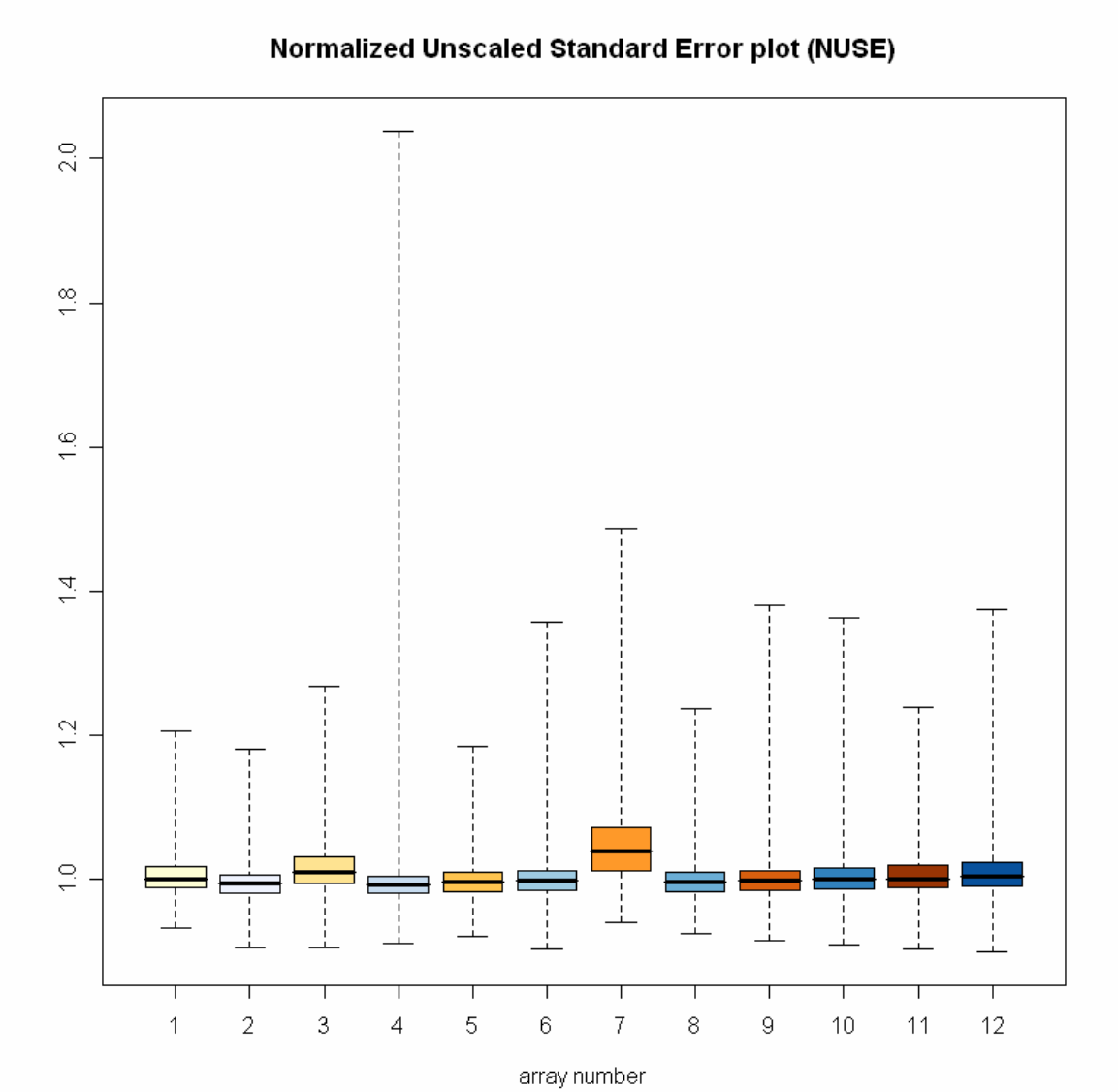



*array 12*




These pseudo-images all show small deviations in the form of stains (array 11), a darker color on some parts of the array (array 3 and 12) or completely darker colored (7).
However, these deviations are small and do not show that major problems occurred during the experiment. After background correction, normalization and summarizing over all of the arrays these deviations should have very little effect on the quality of the data.

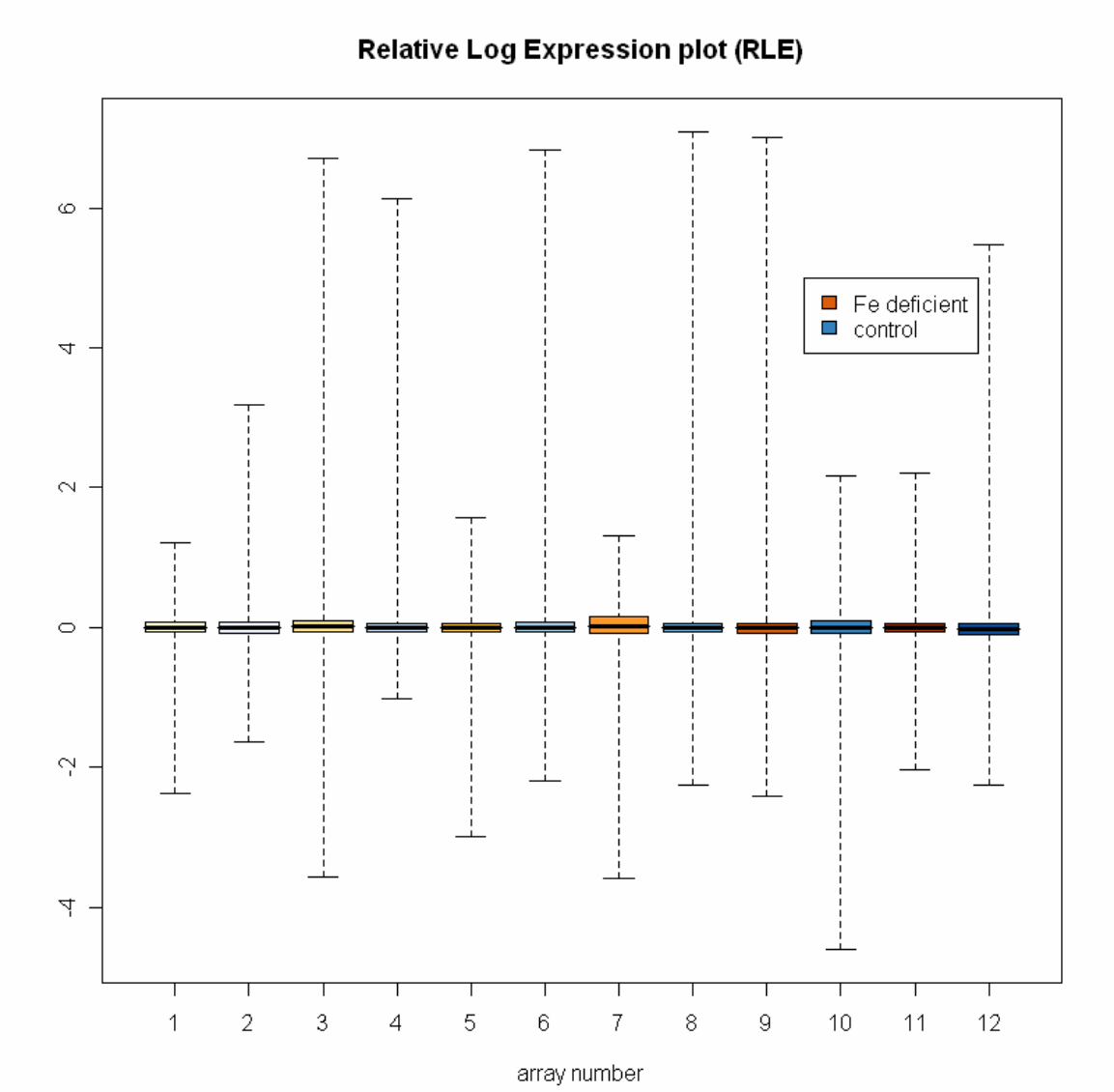
*Normalized unscaled standard error (NUSE)*
The NUSE plot is made using the standard errors of the residuals of each probe. The median of the NUSE should be around 1 for acceptable chips. A NUSE higher than 1.05 is an indication for a bad chip.




As can be seen, most chips have an NUSE of about 1. The NUSE of chip 3 is slightly higher, about 1.01. Again, chip 7 shows the lowest quality with a NUSE of about 1.04. This is still below the threshold of 1.05, so the quality is acceptable.

*Relative expression (RLE)*

The RLE plot is made by comparing the expression value on each array against the median expression value for that probeset across all arrays. Since the expression of each gene should be about the same the RLE is zero for good chips.

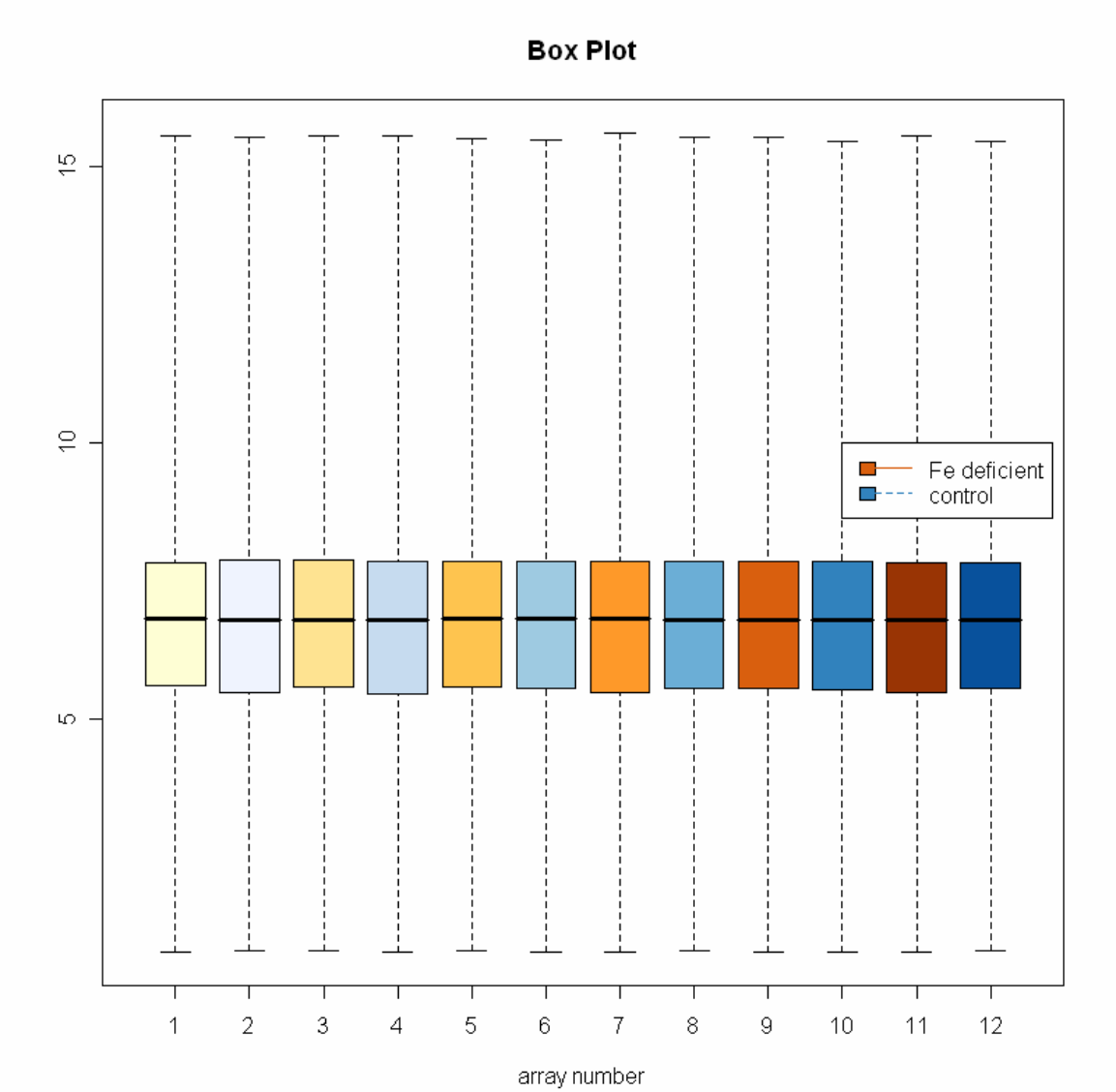


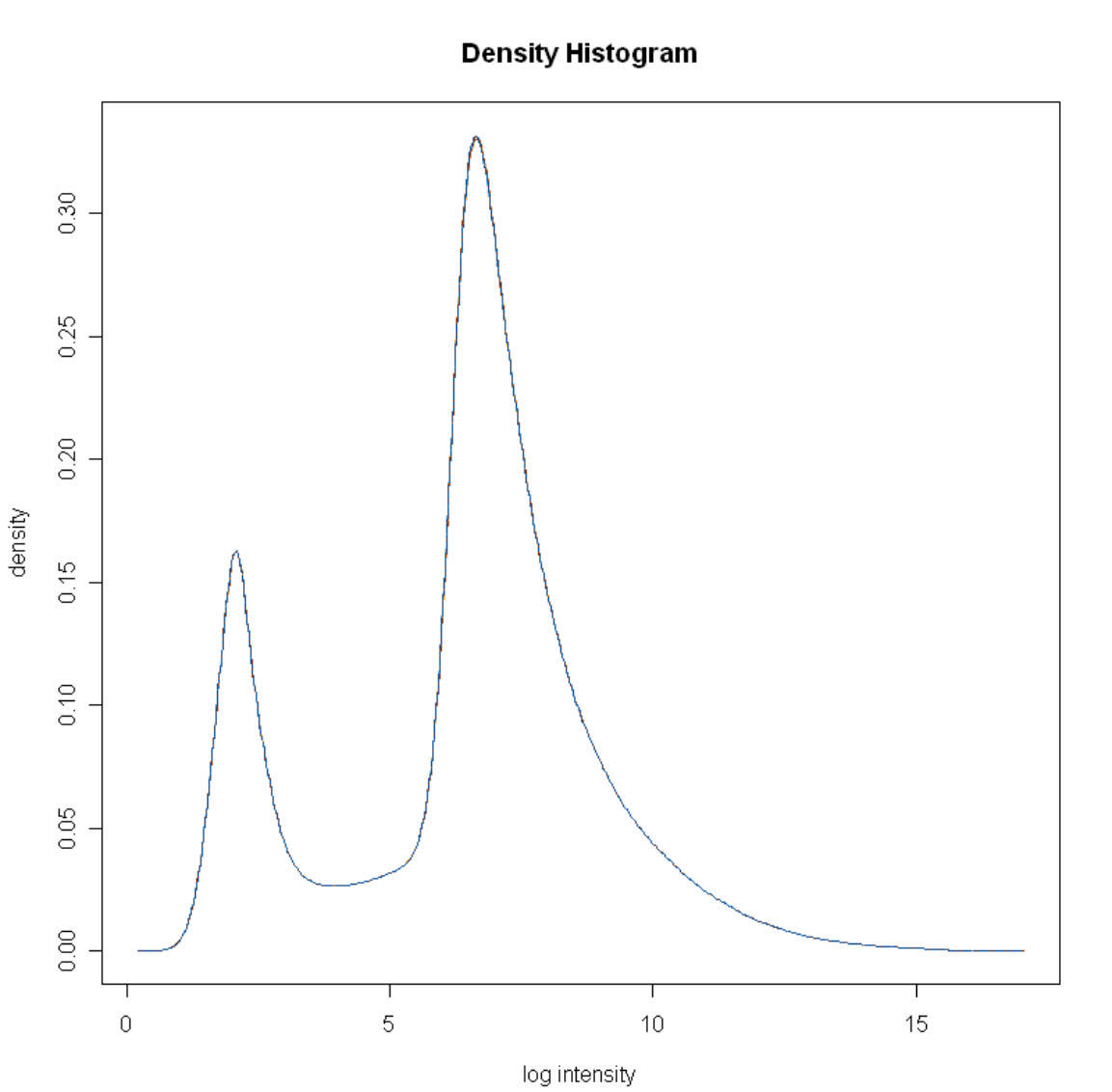

All the chips have an RLE very close to 0. Again, chip 7 shows a lower quality because the interquartile range is slightly larger, but the quality is still acceptable.

Combining all of the statistical tests it can be concluded that there is no array of unacceptable quality. Array number 7 is questionable in some of the test, but since it is always below the threshold of unacceptable data there is no reason to exclude it from further research.

## Background corrected and normalized data plots

After GCRMA background correcting and quantile normalization the following box plot en density histogram are plotted:



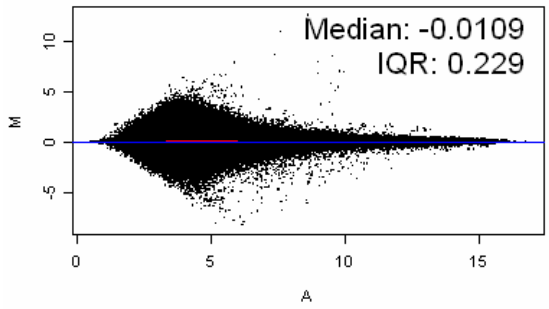



The box plots look almost the same and the shape of the density curves is the same for all arrays.
The two peak signal with a valley in between is something often seen in RMA background correction. GCRMA was developed to improve this bad response to low intensity signals, but the there still is a rather large valley. **(????)**

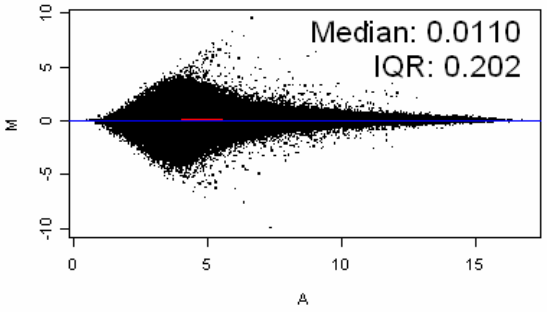
Using the background corrected and normalized data the following M-A plots are plotted:

Array 1 - 6:

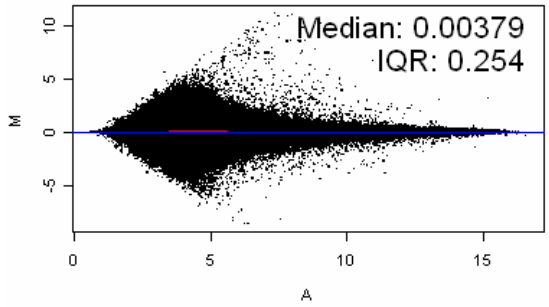
'KE_RAT_FK_01_280906_230_2.CEL vs pseudo-median reference



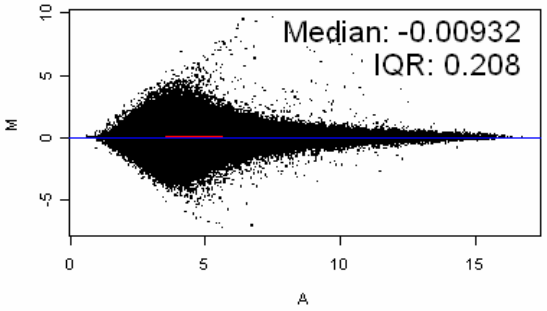

'KE_RAT_FK_02_280906_230_2.CEL vs pseudo-median reference



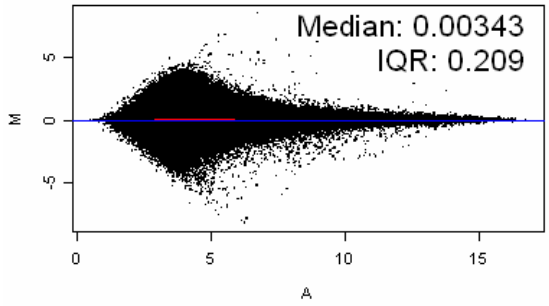

'KE_RAT_FK_03_280906_230_2.CEL vs pseudo-median reference



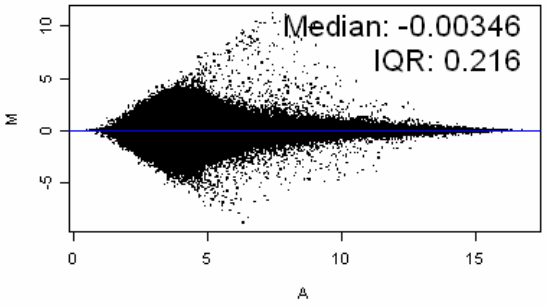

'KE_RAT_FK_04_280906_230_2.CEL vs pseudo-median reference



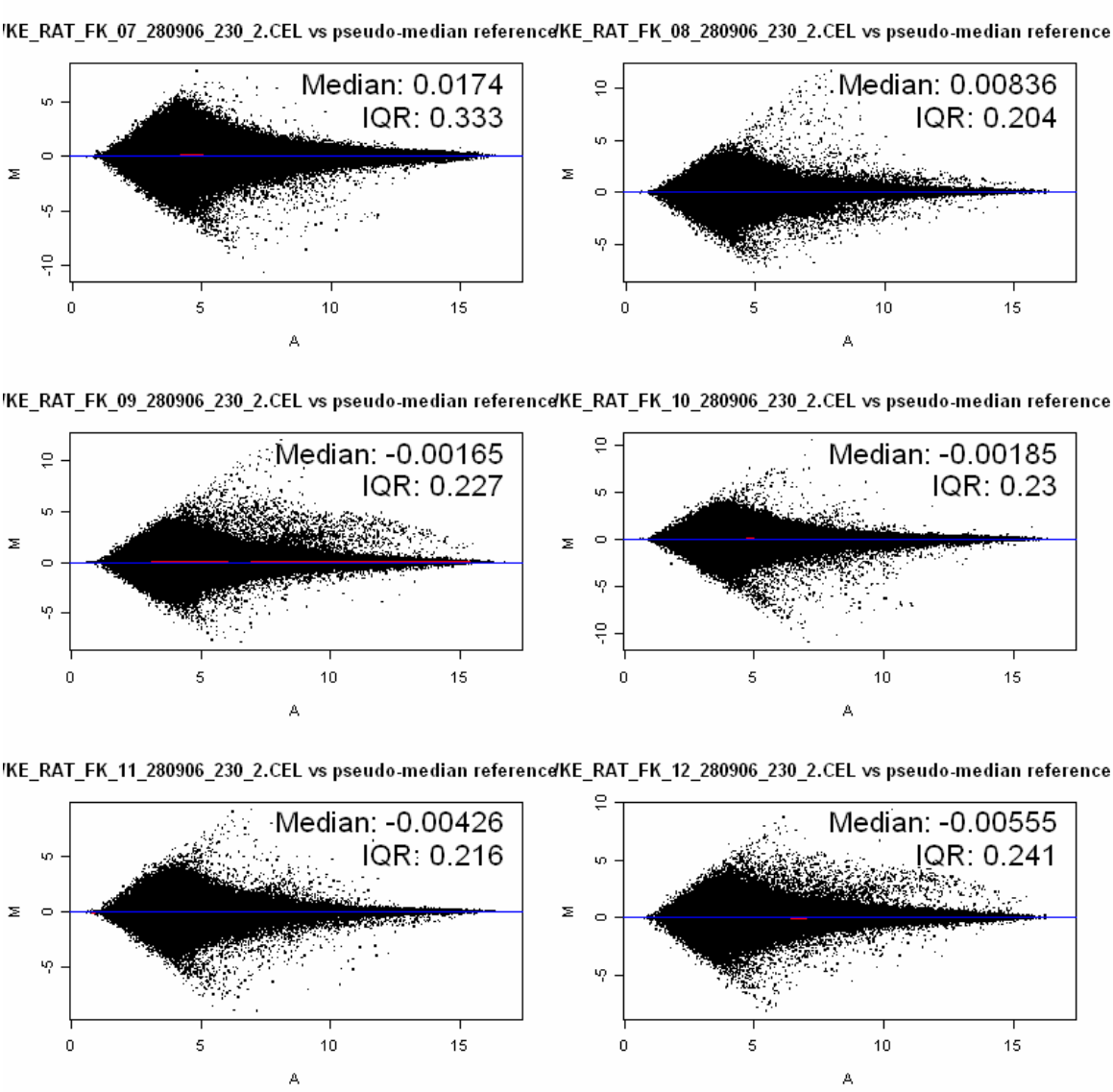

'KE_RAT_FK_05_280906_230_2.CEL vs pseudo-median reference




'KE_RAT_FK_06_280906_230_2.CEL vs pseudo-median reference



Array 7-12:




After normalization and background correcting the M-A plots are all almost perfectly symmetric about the M = 0 line.

It can be concluded that the data is of acceptable quality and useful for further research, also after background correction and normalization.