



Sixth Framework Programme for Quality of Life and Management of Living Resources

Project no. LSHG-CT-2006-037686

EMERALD

Empowering the Microarray-Based European Research Area to Take a Lead in Development and Exploitation

EU Deliverable: D1.2, D1.3, D1.4, D1.5

Due Date:

D1.2 and D1.3 Jul 2007

D1.4 Nov 2007

D1.5 March 2008

Delivery Date:

June 2008

Version 1.0

Partner responsible: EBI

Author: Audrey Kauffmann, Wolfgang Huber

Work Package 1

Deliverable 1.2

Feature/Gene based quality metrics

1.2.1. Introduction

The aim of the Quality Metric and Ontologies work package (WP1) is to develop and disseminate quality metrics and tools for determining data quality and communicating data transformations. Specifically D1.2 aims at developing feature/gene based quantitative quality metrics. This report details the different metrics we are using to assess quality on this level.

1.2.2. Metrics

Quality problems usually affect many spots or even the whole array, very rarely single spots on one single array. We propose ways to assess the feature quality for any platforms and some metrics more specific to Affymetrix as this platform is the most widely used.

1.2.2.1. Genes mapping

The fact that a probe maps for a well known gene or for a putative gene whose annotation may be wrong could result in a different distribution of the expression values as the putative genes might not be expressed. A density plot of the probes mapped versus the unmapped probes can be performed to assess this probe effect.

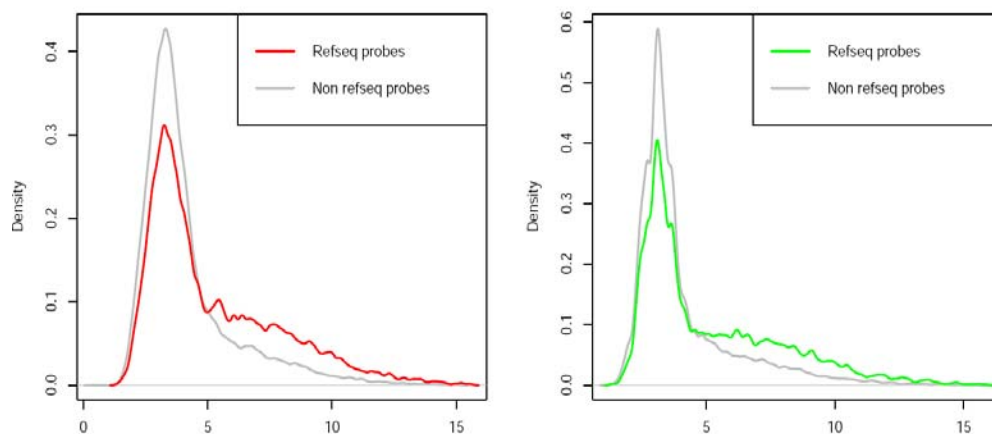
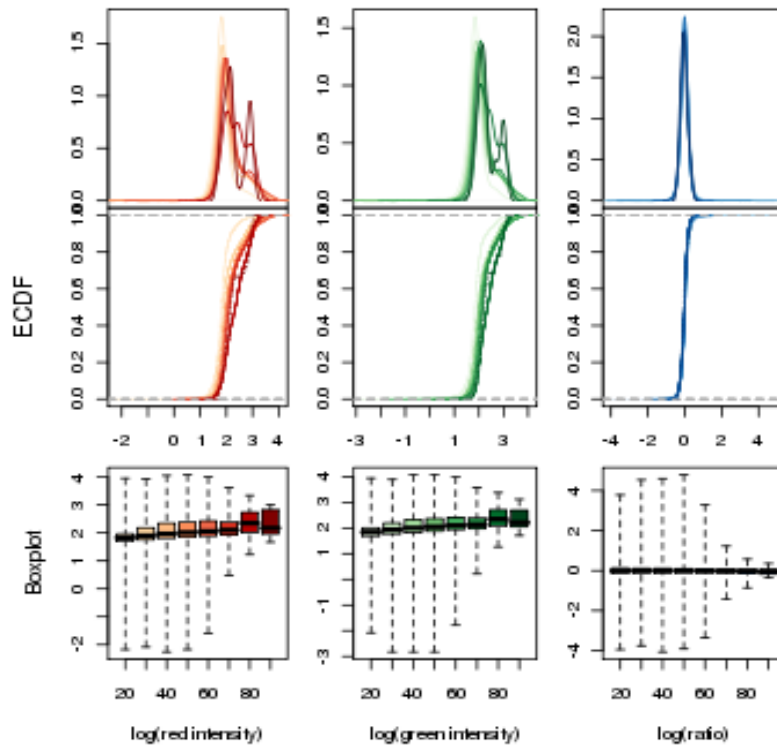


Figure 1: Example of density plots of the red intensity on the left panel and green intensity on the right panel, of all the probes sequences mapped to RefSeq sequences compared to all the probes for which the sequence could not be mapped to RefSeq sequences.

1.2.2.2. GC content effects

The content of the probes in guanine and cytosine can affect the hybridization and consequently the intensities [Pozhitkov]. However, after appropriate data preprocessing (background correction), the relative log expression values should not be affected. Boxplots and density plots function of the percentage of GC in the probe can be used to detect unusual effects of the content in GC of the probes.

Figure 2: Example of density plots, empirical cumulative distribution function



tion (ECDF) and boxplots of the intensities of probes by increasing percentage of GC in the sequence. From left to right, the red intensity, the green intensity and the $\log_2(\text{ratio})$ are represented.

1.2.2.3. Perfect Match - MisMatch

For each probe on the array that perfectly matches its target sequence, Affymetrix also builds a paired “mismatch” probe. The mismatch probes (MM) are not supposed to be hybridized as well as the perfect match ones (PM). The density plots of each group PM and MM allow to detect difference of overall expression of the different types of probes.

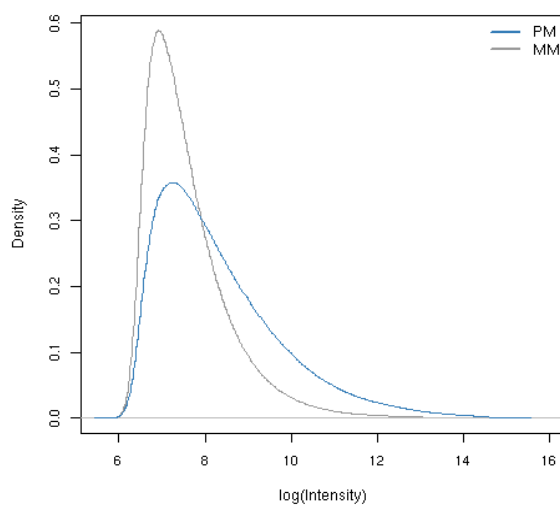


Figure 3: Example of density plots of the intensity of all the perfect match probes compared to all the mismatch probes of an Affymetrix microarray experiment.

1.2.2.4. RNA degradation

On their GeneChip, in addition to the conventional probe sets designed to be within the most 3' 600 bp of a transcript, additional probe sets in the 5' region and middle portion (M) of the transcript have also been selected by Affymetrix for certain housekeeping genes, including GAPDH and Actin. Since RNA degradation characteristically starts from the 5' end, one expects that 5' end probes show lower intensities than 3' end probes. However 3' to 5' probe intensity ratios can be quite variable, dependent on probe affinity differences within a probe-set. Therefore, a 3' to 5' trend can only be detecting after averaging over large number of genes. RNA degradation plots measure this trend [Gautier].

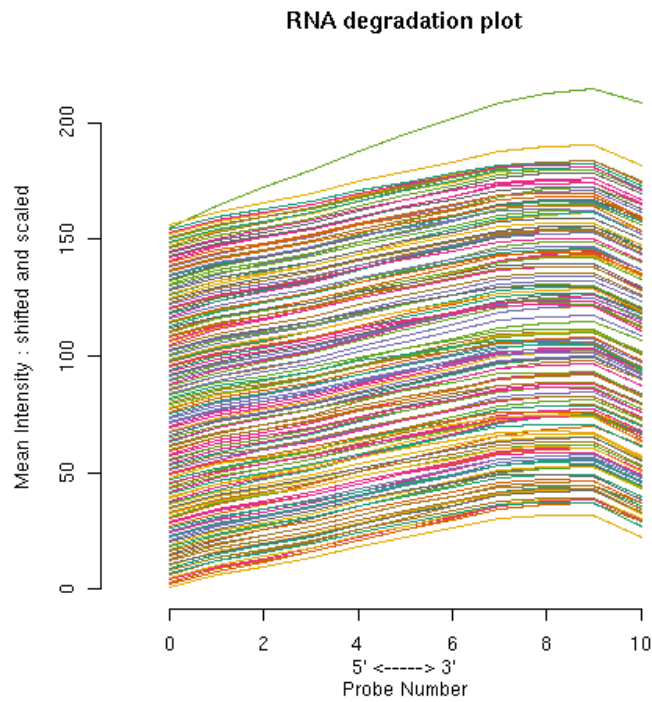


Figure 4: RNA degradation plot from the affy package. Representation of the average intensities of all the housekeeping genes from Affymetrix microarray for which probe sets in the 5' region, middle portion (M) and 3' region of the transcript are available. Each line represents one microarray.

Work Package 1
Deliverable 1.3
Whole array based quality metrics

1.3.1. Introduction

D1.3 aims at producing whole array based quality metrics. This report details different metrics we are using to assess quality on this level.

1.3.2. Metrics

1.3.2.1. Dependency between intensities and distribution of ratios

A M versus A plot for each array can be drawn with M and A defined as :

$$M = \log_2(I_1) - \log_2(I_2)$$

$$A = 1/2 (\log_2(I_1) + \log_2(I_2))$$

where, in the case of two colour arrays, I1 and I2 are the vectors of intensities of the two channels. In the case of one colour arrays, I1 is the intensity of the array studied and I2 is the intensity of a "pseudo"-array, which have the median values of all the arrays.

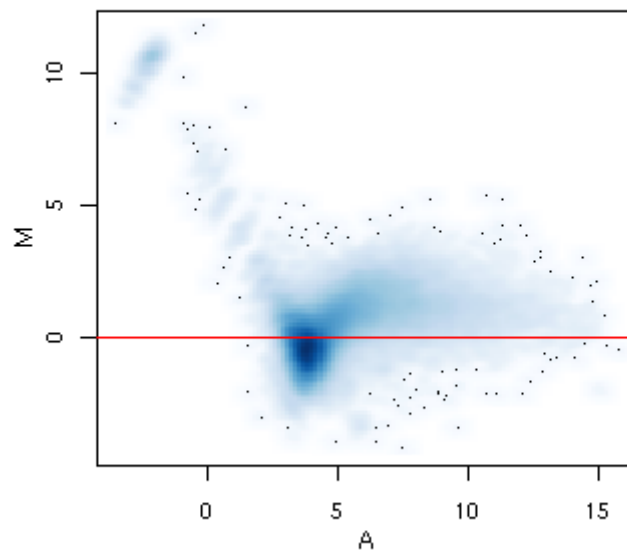


Figure 5: Example of MA-plot.

1.3.2.2. Spatial effects

To assess if there is any spatial effect on the chip, a false colour representation of the arrays' spatial distributions of feature intensities and boxplots of the intensities by row and columns of the array help in identifying patterns that may be caused by, for example, spatial gradients in the hybridization chamber, air bubbles, spotting or printing problems.

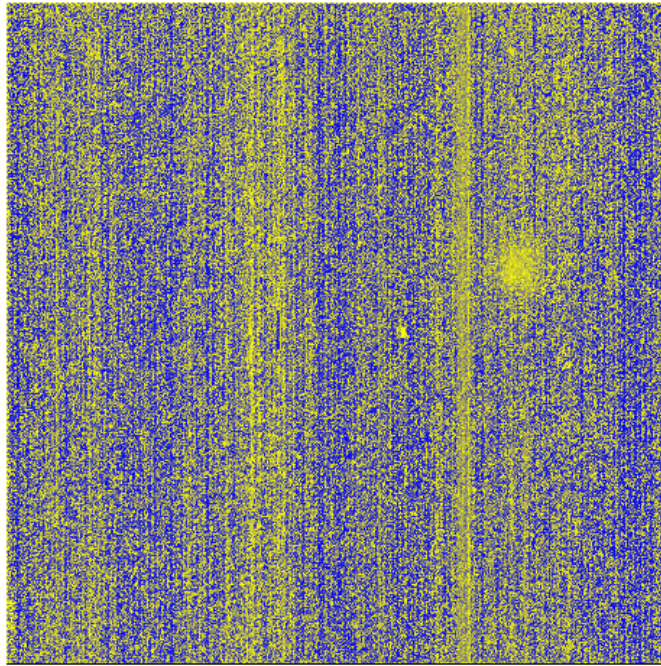


Figure 6: Example of spatial distribution of feature intensities of an Affymetrix chip.

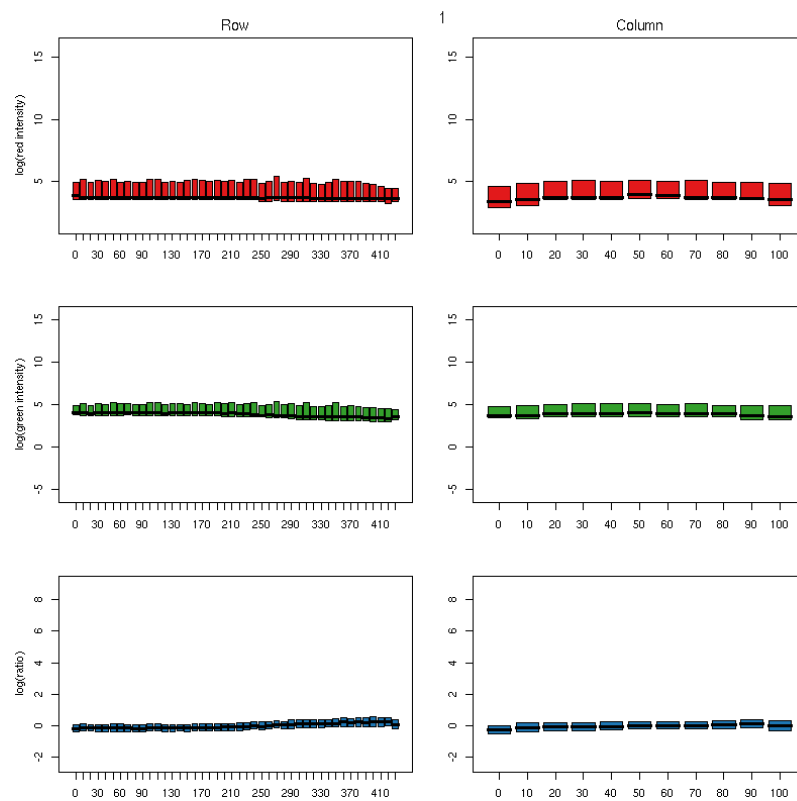


Figure 7: Example of print-tips boxplots of a two-colour Agilent chip. The first column contains the intensities by row and the second column contains the intensities by column. From top to bottom, the red intensity, the green intensity and the $\log_2(\text{ratio})$ are represented.

1.3.2.4. Homogeneity between arrays

To assess the homogeneity between the arrays, boxplots of the \log_2 intensities, the density estimate and the empirical cumulative distribution function (ECDF) plots are represented.

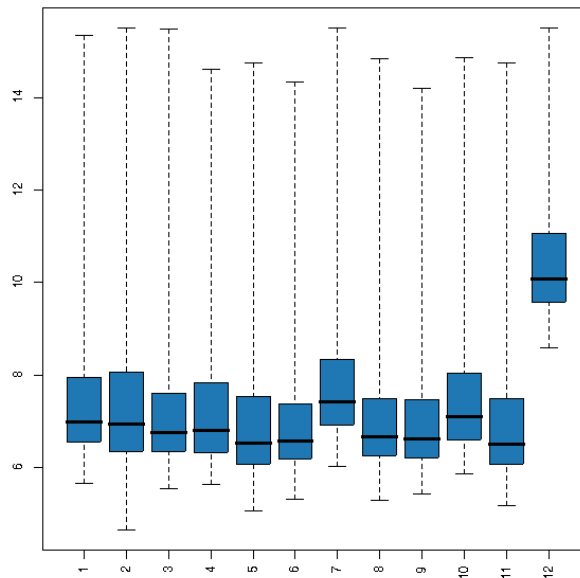


Figure 8: Example of boxplots of all intensities from a one-colour experiment.

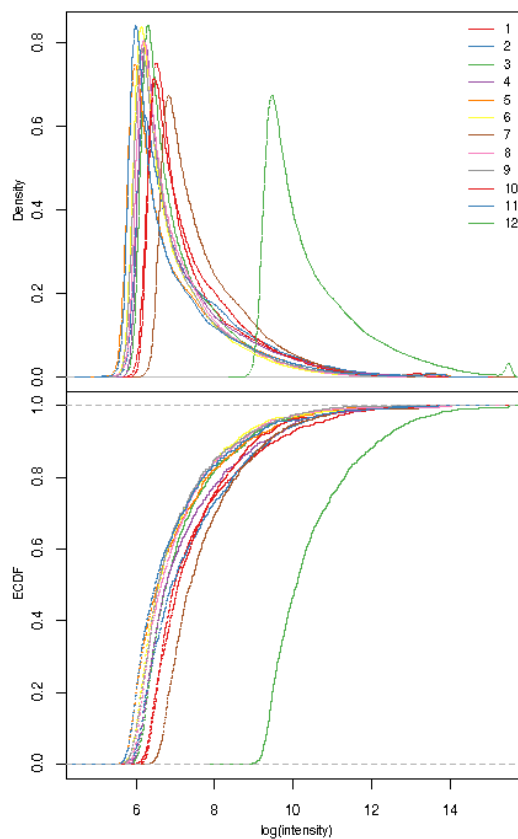


Figure 9: Example of density and ECDF representations of all intensities from a one-colour data set with 12 chips.

1.3.2.6. Between array comparison

A false color heatmap providing a comparison between arrays can be computed. Firstly, the distances between arrays are computed as the median absolute difference of the M -value for each pair of arrays.

$$d_{xy} = \text{mean}|M_{xi}-M_{yi}|$$

Here, M_{xi} is the M -value of the i -th probe on the x -th array. This plot can serve to detect outlier arrays.

Consider the following decomposition of M_{xi} : $M_{xi} = z_i + \beta_{xi} + \varepsilon_{xi}$, where z_i is the probe effect for probe i (the same across all arrays), ε_{xi} are i.i.d. random variables with mean zero and β_{xi} is such that for any array x , the majority of values β_{xi} are negligibly small (i. e. close to zero). β_{xi} represents differential expression effects. In this model, all values d_{xy} are (in expectation) the same, namely 2 times the standard deviation of ε_{xi} . Arrays whose distance matrix entries are way different give cause for suspicion. This dendrogram also can serve to check if, without any probe filtering, the arrays cluster accordingly to a biological meaning.

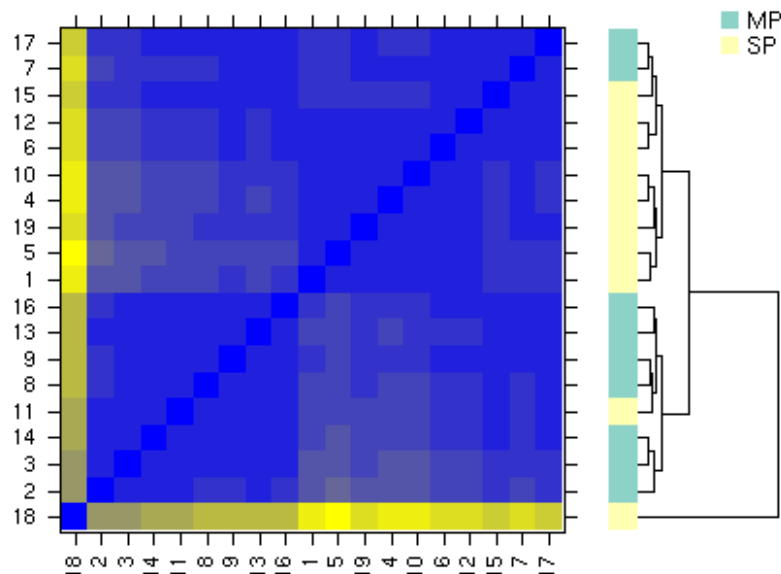


Figure 10: Heatmap of an Affymetrix data set where MP means Main Population and SP means Side Population from murine skeletal muscle and bone marrow cells.

1.3.2.7. Variance/Mean dependence

For each feature, the empirical standard deviation of the intensities of all the arrays on the y-axis versus the rank of the mean of intensities of all the arrays on the x-axis can be plotted.

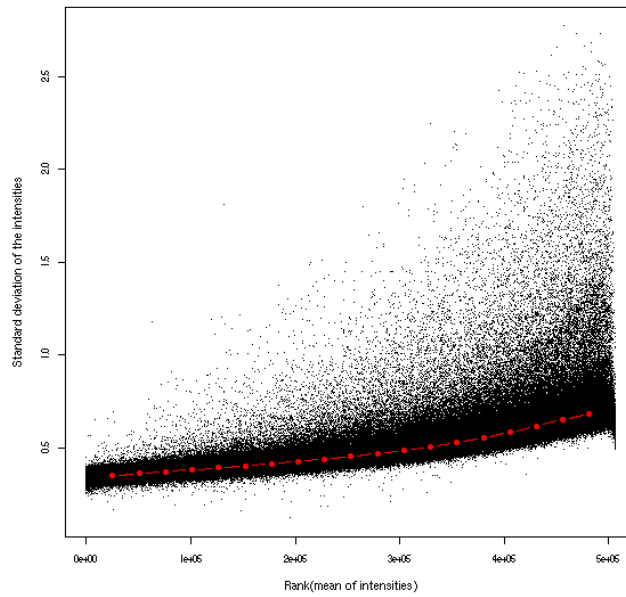


Figure 11: Example of the representation of the standard deviation function of the rank of the mean intensities. The red dots, connected by lines, show the running median of the standard deviation.

1.3.2.8. RLE

Another quality assessment tool are Relative Log Expression (RLE) values. Specifically, these RLE values are computed for each probeset by comparing the expression value on each array against the median expression value for that probeset across all arrays. Assuming that most genes are not changing in expression across arrays means ideally most of these RLE values will be near 0. Boxplots of these values, for each array, provides a quality assessment tool as shown in Figure 12. [Bolstad]

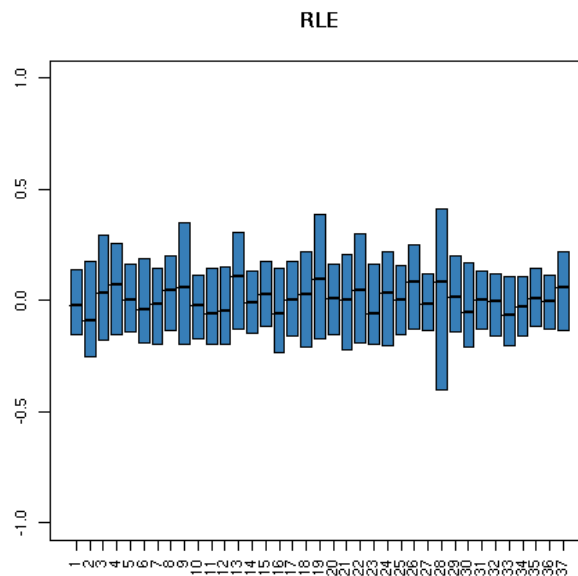


Figure 12: Example of a RLE boxplot from affyPLM package.

1.3.2.9. NUSE

Normalized Unscaled Standard Errors (NUSE) can also be used for assessing quality. In this case, the standard error estimates obtained for each gene on each array from fitPLM (Fitting Probe Level Models from the affyPLM package) are taken and standardized across arrays so that the median standard error for that genes is 1 across all arrays. This process accounts for differences in variability between genes. An array where there are elevated Standard Errors (SE) relative to the other arrays is typically of lower quality. Figure 13 shows boxplots of these values separated by array. This plot can be used to compare arrays. [Bolstad]

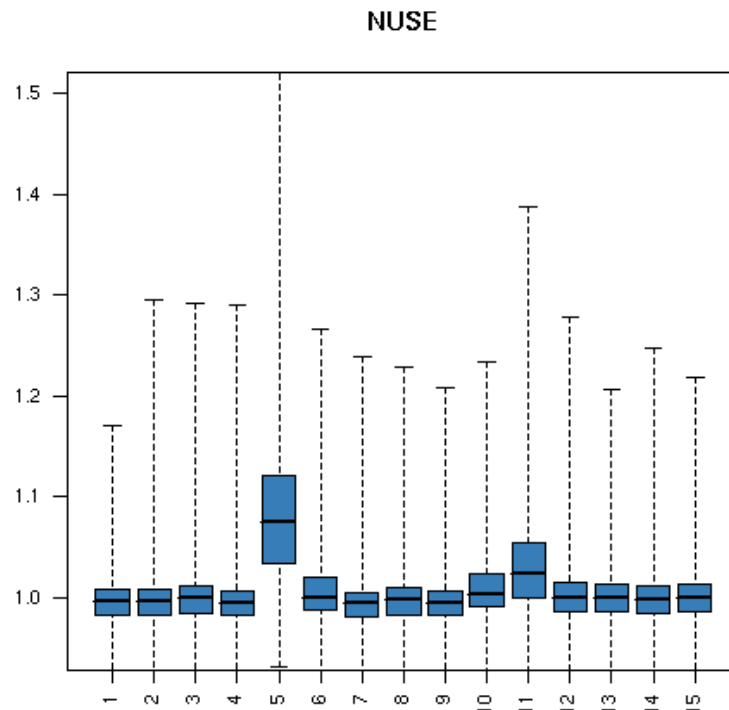


Figure 13: Example of a NUSE boxplot from affyPLM package.

1.3.2.10. QCStats

The QCStats [Wilson] object plotted on Figure 14 represents the following information:

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.

- GAPDH 3':5' values are plotted as circles. According to Affymetrix they should be about 1. GAPDH values that are considered potential outliers (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 ‘3-fold region’, they are all coloured red, otherwise they are blue.

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

Present/Marginal/Absent calls are generated by looking at the difference between perfect match (PM) and mismatch (MM) values for each probe pair in a probeset. Probesets are flagged Marginal or Absent when the PM values for that probeset are not considered to be significantly above the MM probes. As with scale factors, large differences between the numbers of genes called present on different arrays can occur when varying amounts of labelled RNA have been successfully hybridized to the chips. This can occur for similar reasons (differences in array processing pipelines, variations in the amount of starting material, etc.). The ‘% Present’ call simply represents the percentage of probesets called present on an array. As with Scale Factors, significant variations in % Present call across the arrays in a study should be treated with caution. Note that the absolute value is generally not a good metric – some cells naturally express more genes than others. [Wilson]

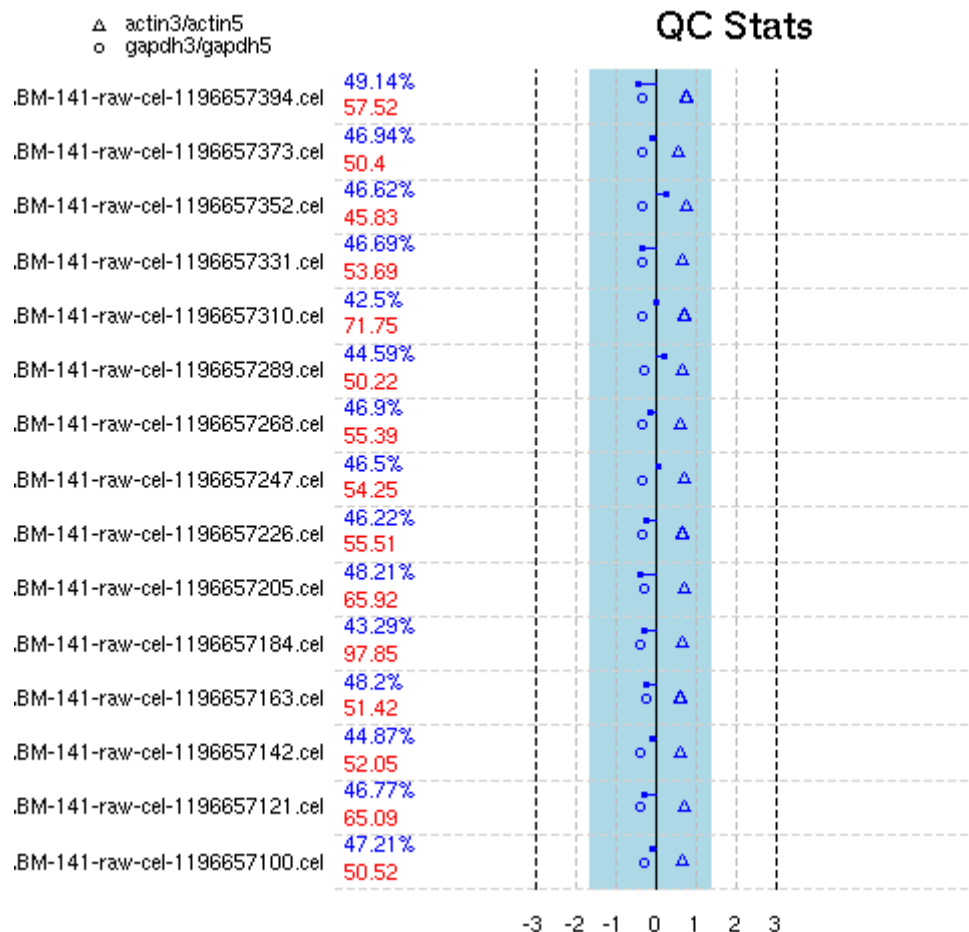


Figure 14: Example of a QCstats plot from the simpleaffy package.

Work Package 1

Deliverable 1.4

A description of “best practices” for microarray study

1.4.1. Introduction

The aim of D1.4 is to provide a description of the best practices for microarray study. The Bioconductor Case Studies book gives detailed examples on microarray analysis [Bioconductor book]. This book is a complete description of “best practices” for microarray study. D1.2 and D1.3 explained the type of quality metrics based on the feature or on the whole array. In this section we will explain how to interpret the plots described in sections D1.2 and D1.3.

1.4.2. Interpretation of the plots

1.4.2.1. Probes mapping

The density plot of the probes mapping for an existing gene is supposed to be shifted to the right compare to the density plot of the unmapped probes. This shift means that the overall intensities are higher for probes of known genes than for probes of putative genes or with unreliable gene annotation,, showing a better hybridization.

1.4.2.2. GC content effects

The boxplots or the density plots differents to each other according to the GC content mean that the intensities are dependent on the GC content. With proper data processing, this effect should not be seen on the log(ratios).

1.4.2.3. Perfect Match – MisMatch

The density plot of the PM probes is supposed to be shifted to the right compared to the density plot of the MM. As for probes mapping study, such a shift means that the overall intensities are higher for PM than for MM, showing a better hybridization of the PM probes.

1.4.2.4. RNA degradation

From the RNA degradation plot, it is important to identify any array that has a slope which is very different from the others. The indication is that the RNA used for that array has potentially been handled quite differently from the other arrays.

1.4.2.5. Dependency between intensities and distribution of ratios

We expect the mass of the distribution in an MA plot to be concentrated along the $M = 0$ axis, and there should be no trend in the mean of M as a function of A .

1.4.2.6. Spatial effects

Any gradient, smear, or border effect can be visually identified with the spatial representations. Furthermore, if there is no spatial effect, the boxes of the print-tips boxplots should have similar sizes and positions.

1.4.2.7. Homogeneity between arrays

If the arrays are homogeneous, the boxes from the boxplot of intensities (or $\log_2(\text{ratio})$ in case of two colour arrays) should have similar widths and y position. Arrays whose distributions on the density or ECDF plots are very different from the others should be considered for possible problems.

1.4.2.8. Between array comparison

From the dendrogram between arrays, if there is any clustering, it should be according to biological factors. An array with a higher distance to the other arrays is an outlier. It is also acceptable that there is no clusters at all.

1.4.2.9. Variance/Mean dependence

After vsn normalization, the red line of the variance versus mean plot should be approximately horizontal, that is, show no substantial trend. If the vsn normalization was used, then this plot is a useful QC step.

1.4.2.10. RLE

On the Relative Log Expression (RLE) plot, an array that has problems will either have larger spread, or will not be centered at $M = 0$, or both.

1.4.2.11. NUSE

On the Normalized Unscaled Standard Error (NUSE) plot, low quality arrays are those that are substantially elevated or more spread out, relative to the other arrays. NUSE values are not comparable across data sets.

1.4.2.12. QCStats

QCStats is fully described in the deliverable 1.3. Any metrics that is shown in red is out of the manufacturer's specific boundaries and suggests a potential problem, any metrics shown in blue is fine.

1.4.3. Scores

To help the reading and the understanding of the report, we have included the computation of scores associated with the plots. A microarray outlier detection can be computed for MA-plot, spatial distributions of the features intensities, boxplot, heatmap, RLE and NUSE.

For each of these plot, we can reduce each array to one value and then we can draw a boxplot of these values. The values which are more than 1.58 times the interquartile range divided by the square root of the number of arrays are considered as possible outliers arrays.

For the MA-plot, the values assessed are the mean of the absolute value of the M-values computed for each array. The mean and interquartile range (IQR) are used for the boxplots of intensities and NUSE. In the case of the heatmap, the values assessed are the sums of the rows of the distance matrix used to draw the dendrogram. We use periodogram to detect spatial effect from the spatial representations. Using periodogram from the spatial representations allows to detect any heterogeneity on the false color representations of the arrays' spatial distributions of feature intensities. Furthermore, if there is no spatial effect, the boxes of the print-tips boxplots should be homogeneous in location and scale.

In the case of the RLE plot, any array with a median RLE higher than 0.1 is considered as a possible outlier.

Work Package 1

Deliverable 1.5

Prototype software encoding these metrics

1.5.1. Introduction

D1.5 specifies the development of software libraries encoding the metrics described in D1.2 and D1.3. This report gives some details about the arrayQualityMetrics package that has been developed to achieve this aim.

1.5.2. arrayQualityMetrics

The arrayQualityMetrics package is available on Bioconductor (<http://www.bioconductor.org>). It provides a HTML report that contains all the plots and scores described previously.

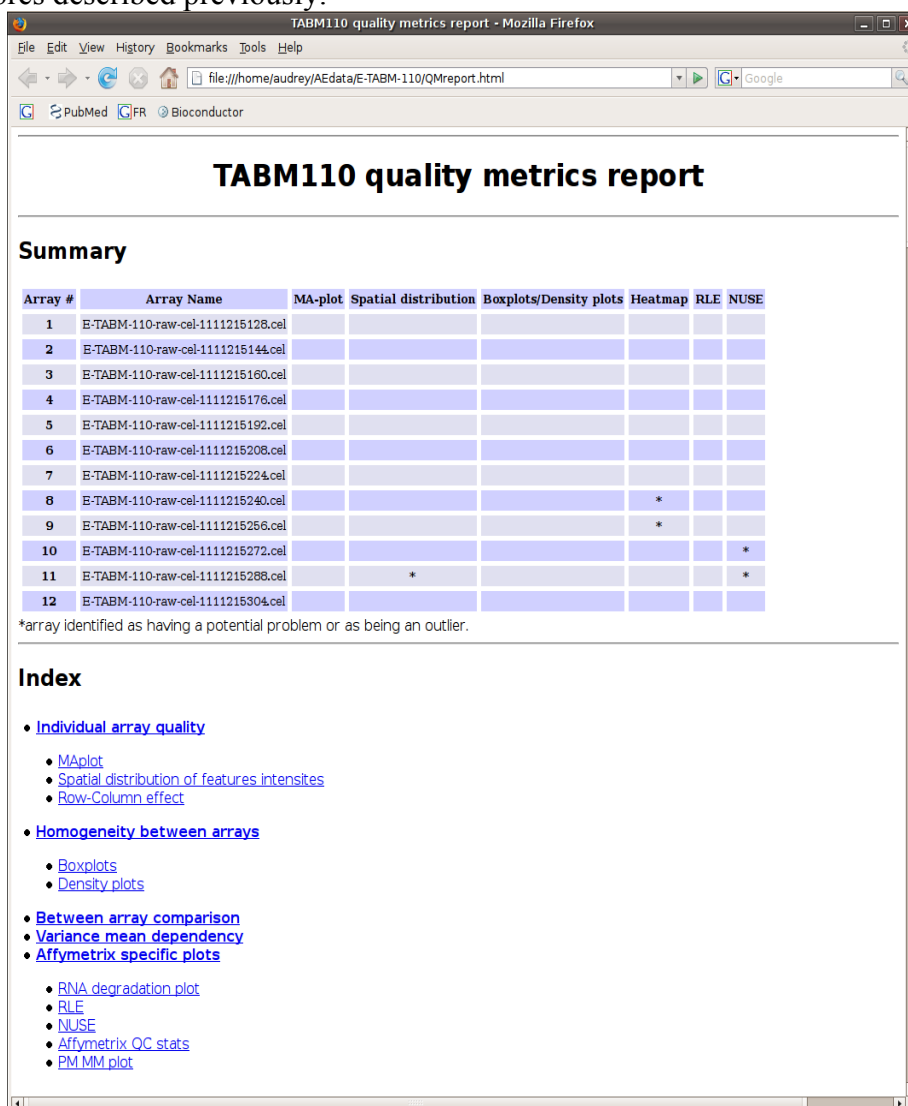


Figure 15: Snapshot of the top of a quality report obtained using arrayQualityMetrics. The report starts with a table summarizing the outliers that have been detected, then an index allowing the user to access different sections is shown. Then, the report contains MA plot, spatial distribution, boxplots of intensities by print-tips, boxplots and density plots per array on the intensities, heatmap, variance versus mean plot, RNA degradation, RLE boxplot, NUSE boxplot, QCStats and PM versus MM density plots.

1.5.3. Future plans

We produced a freely available tool for quality assessment of any kind of microarray data. This tool includes a help for the user to identify outlier arrays thanks to the scores. With the outlier detection, we assess the relative quality of one array compared to the other arrays of the experiment. It would be very useful to assess the absolute quality of each array. The NUSE plot for Affymetrix is an absolute quality metric. We can probably apply this method for other arrays with multiple probes per target molecule but not for other kind of arrays. We will try to develop more absolute quality metrics that will be suitable for most of microarray types.

References

Pozhitkov A.E., Tautz D., Noble P.A. (2007) Oligonucleotide microarrays: widely applied--poorly understood. *Brief Funct Genomic Proteomic* **6(2)**:141-8

Gautier L., Cope L., Bolstad B.M., Irizarry RA. (2004) affy -- analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20(3)**:307-15

Bolstad B.M. affyPLM: Methods for fitting probe-level models. R package version 1.17.0

Wilson C.L. and Miller C.J. (2005) Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* **21(18)**:3683-5

Bioconductor Case Studies (Use R) book by Florian Hahne, Wolfgang Huber, Robert Gentleman and Seth Falcon