

Affymetrix Batch Quality Control using RReportGenerator and R

March 22, 2011

This document was generated by Analysis Type File '*automAffyQC2.Rnw*', Version: 1.2.9,
a protocol for automated QC analysis for Affymetrix expression array data.

The aim of this report is to provide information on multiple QC aspects for a set of microarrays. The interpretation of QC-plots should be done with care since this may lead to delicate decisions. For further information about the various plots in this report please look at the references section. Analysis protocol written by wolfgang.raffelsberger (at) igbmc.fr, LGBI, IGBMC (Strasbourg, France).

QC analysis of 38 cel-files of type 'HG-U133.Plus.2' with 54675 probesets each, found in path :

/data2/g3/wrTest/RNG209/

page 3

Both plots give a resume of the original PM (perfect match) values, either as boxplot or as (kernel) density estimate for the signal distribution (similar to Bioconductor-packages simpleaffy and affyQCReport).

page 4

This figure shows the QC plot from the simpleaffy package. For this purpose the 3' to 5' ratio for spiked-in and control genes (triangles for b-Actin and squares for GAPDH) are considered. The filled (blue) dot with the vertical line (heading to 0) represents the scaling factor. Finally, the percentage of 'present calls' and value of average background are shown in red on the left side. Note that the order of arrays is inversed (compared to other plots in this report).

page 5

This pair of graphs shows the RLE (top) and NUSE (bottom) plots from the affyPLM package. The RLE compares for each probeset (across all arrays) its expression value against its median across all arrays. The NUSE plot shows the standard errors for each probe-set (variability intra), standardized across all arrays.

page 6

This page shows false color images for the RLE residuals of all probes (from the affyPLM package, with red intensities corresponding to positive residuals and blue to negative residuals). Note, that smaller deviating zones have little impact on the overall results. For further information, examples and details see e.g. chapter 3 by Bolstad in book by Gentleman *et al*.

page 7

MA plots (affyPLM package) of each array against a synthetic median array (probe-wise medians of all arrays in project). The red line represents a lowess fit to the scatter plot and may help indicating non-linear relationships.

page 8

Top: RNA degradation plot shows the average intensity with respect to the sorted 5' to 3' position of the probes in the target-sequence. Depending on the type of microarray specific patterns can be observed (see also references Bolstad, Gentleman *et al*). Boxplots (center) and distribution estimates (bottom) for data resulting from GCRMA normalization.

page 9

Top: Similarity between GCRMA summarized samples measured as Euclidean distance : Heatmap of distance values.
Bottom: Dendrogram from hierarchical clustering (with p-values estimated by bootstrapping (reference Suzuki 2004)

page 10

Principal components of GCRMA summarized samples. The main graphic shows the values of the first and second principal component for each of the samples. The smaller plots show the 2nd vs. 3rd as well as the 1st vs. 3rd component, the bargraph at the bottom indicates the amount of total variance captured by the individual principal components.

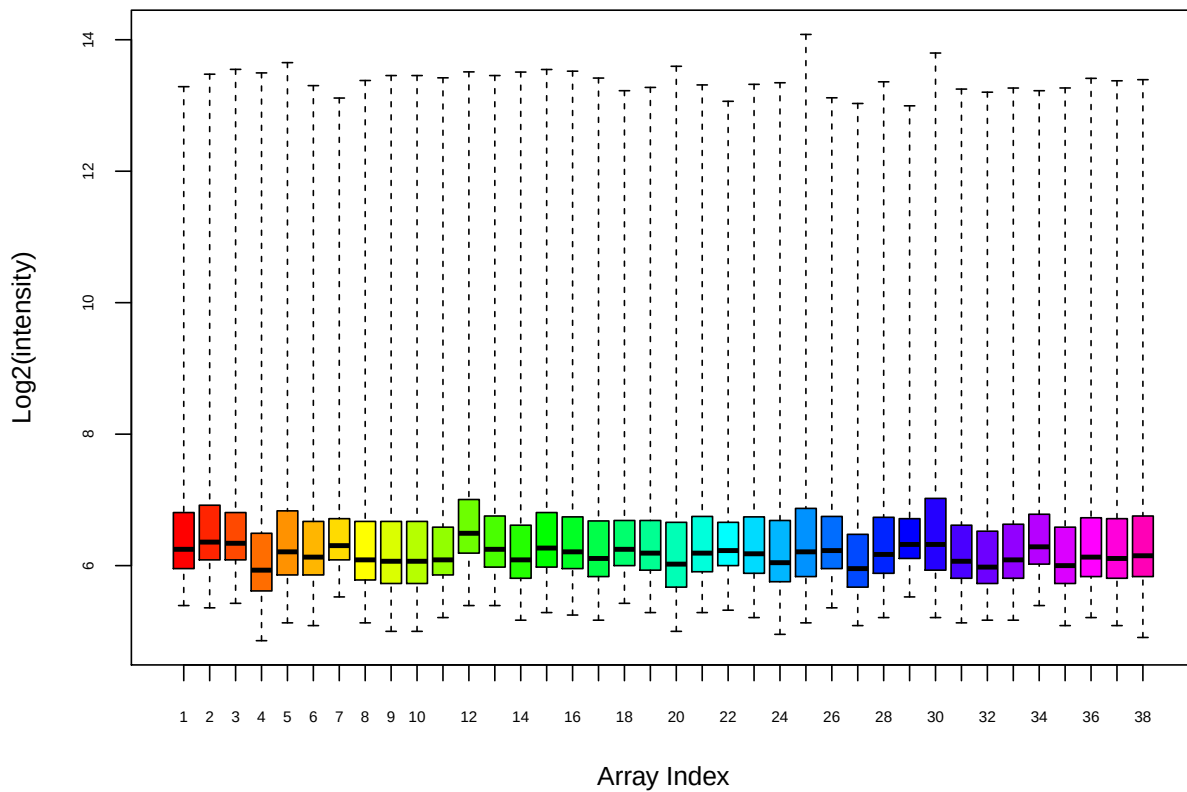
page 11 : References

Overview of Arrays Used

array index	(short) sample names	/ initial file-name
1 ...CTR10	.. CTR10_U133_2.CEL.gz	
2 ...CTR11	.. CTR11_U133_2.CEL.gz	
3 ...CTR12	.. CTR12_U133_2.CEL.gz	
4 ...CTR13	.. CTR13_U133_2.CEL.gz	
5 ...CTR14	.. CTR14_U133_2.CEL.gz	
6 ...CTR15	.. CTR15_U133_2.CEL.gz	
7 ...CTR16	.. CTR16_U133_2.CEL.gz	
8 ...CTR17	.. CTR17_U133_2.CEL.gz	
9 ...CTR18	.. CTR18_U133_2.CEL.gz	
10 ...CTR19	.. CTR19_U133_2.CEL.gz	
11 ...CTR1	.. CTR1_U133_2.CEL.gz	
12 ...CTR2	.. CTR2_U133_2.CEL.gz	
13 ...CTR3	.. CTR3_U133_2.CEL.gz	
14 ...CTR4	.. CTR4_U133_2.CEL.gz	
15 ...CTR5	.. CTR5_U133_2.CEL.gz	
16 ...CTR6	.. CTR6_U133_2.CEL.gz	
17 ...CTR7	.. CTR7_U133_2.CEL.gz	
18 ...CTR8	.. CTR8_U133_2.CEL.gz	
19 ...CTR9	.. CTR9_U133_2.CEL.gz	
20 ...RD10	.. RD10_U133_2.CEL.gz	
21 ...RD11	.. RD11_U133_2.CEL.gz	
22 ...RD12	.. RD12_U133_2.CEL.gz	
23 ...RD13	.. RD13_U133_2.CEL.gz	
24 ...RD14	.. RD14_U133_2.CEL.gz	
25 ...RD15	.. RD15_U133_2.CEL.gz	
26 ...RD16	.. RD16_U133_2.CEL.gz	
27 ...RD17	.. RD17_U133_2.CEL.gz	
28 ...RD18	.. RD18_U133_2.CEL.gz	
29 ...RD19	.. RD19_U133_2.CEL.gz	
30 ...RD1	.. RD1_U133_2.CEL.gz	
31 ...RD2	.. RD2_U133_2.CEL.gz	
32 ...RD3	.. RD3_U133_2.CEL.gz	
33 ...RD4	.. RD4_U133_2.CEL.gz	
34 ...RD5	.. RD5_U133_2.CEL.gz	
35 ...RD6	.. RD6_U133_2.CEL.gz	
36 ...RD7	.. RD7_U133_2.CEL.gz	
37 ...RD8	.. RD8_U133_2.CEL.gz	
38 ...RD9	.. RD9_U133_2.CEL.gz	

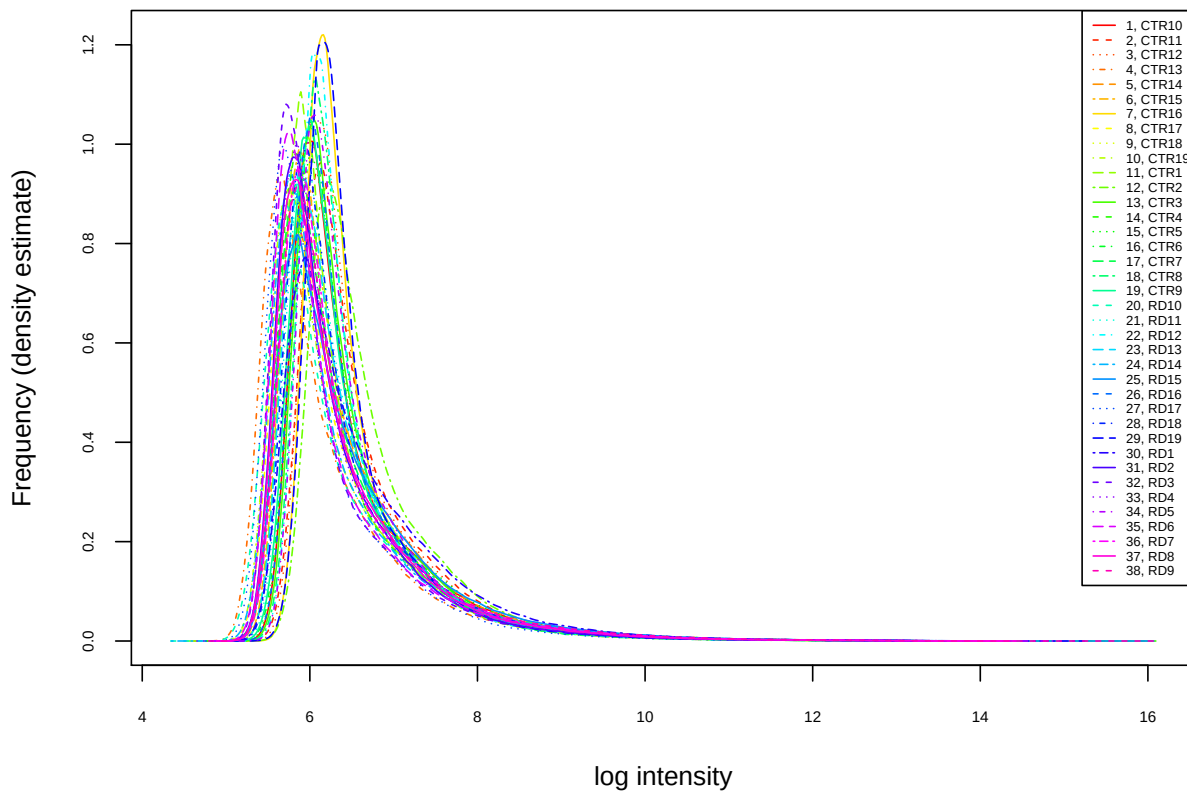
Boxplots for PM Values

data from: /data2/g3/wrTest/RNG209/



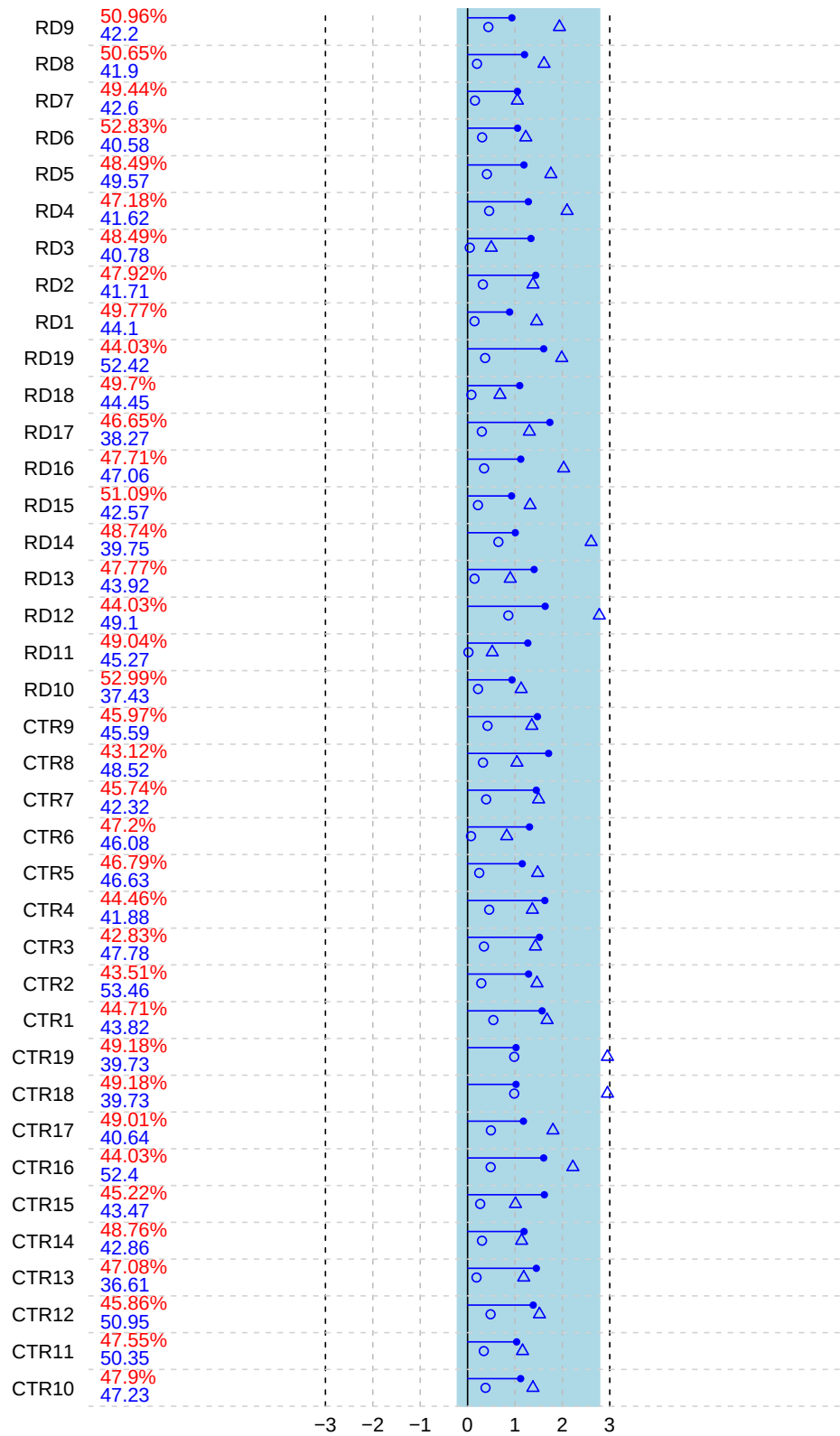
Histogram for PM Values (kernel density estimate)

data from: /data2/g3/wrTest/RNG209/



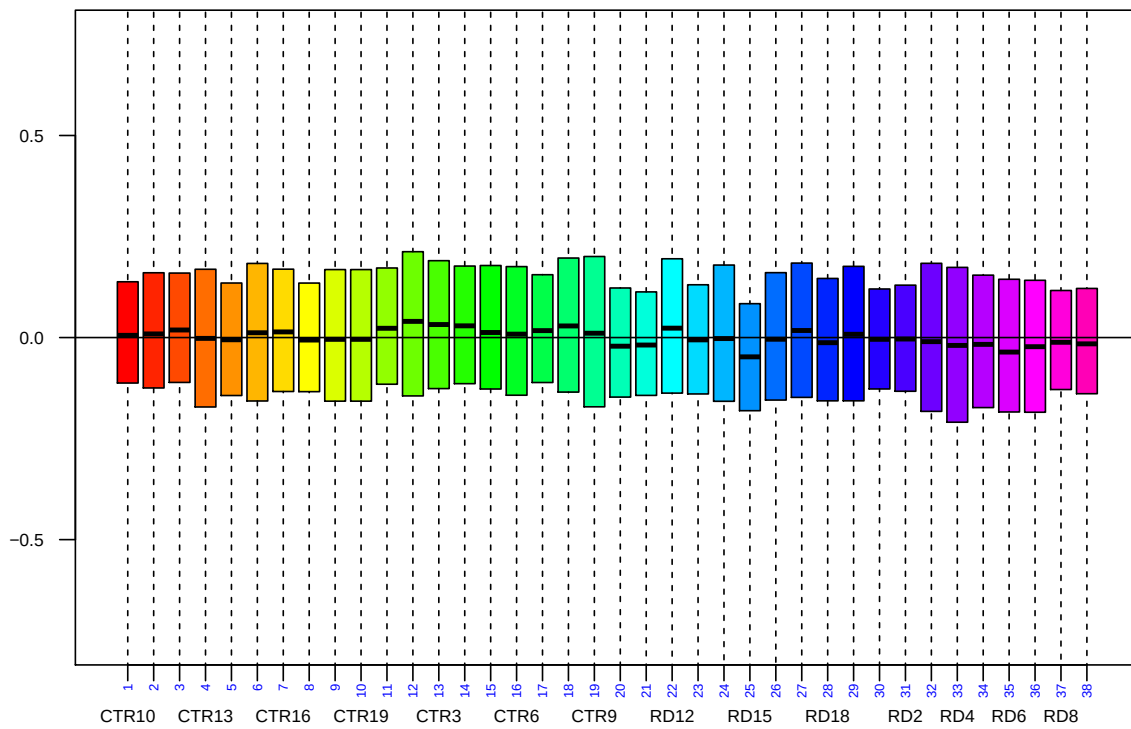
- △ actin3/actin5
- gapdh3/gapdh5

QC Stats

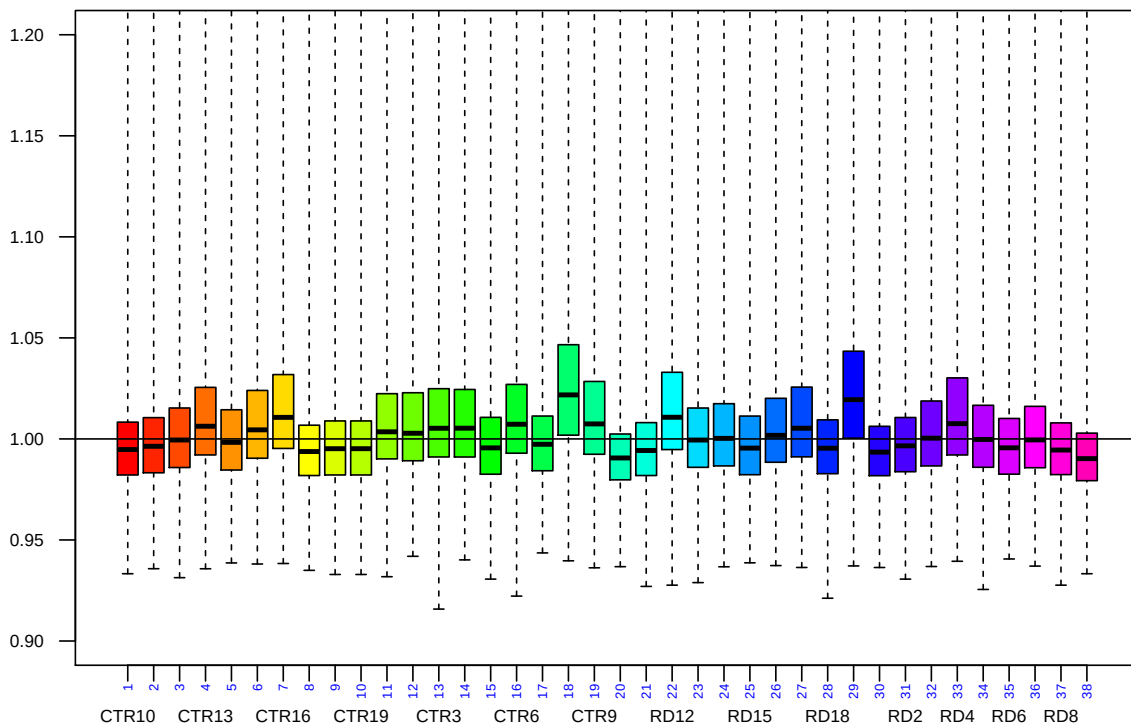


Relative Log Expression (RLE) values

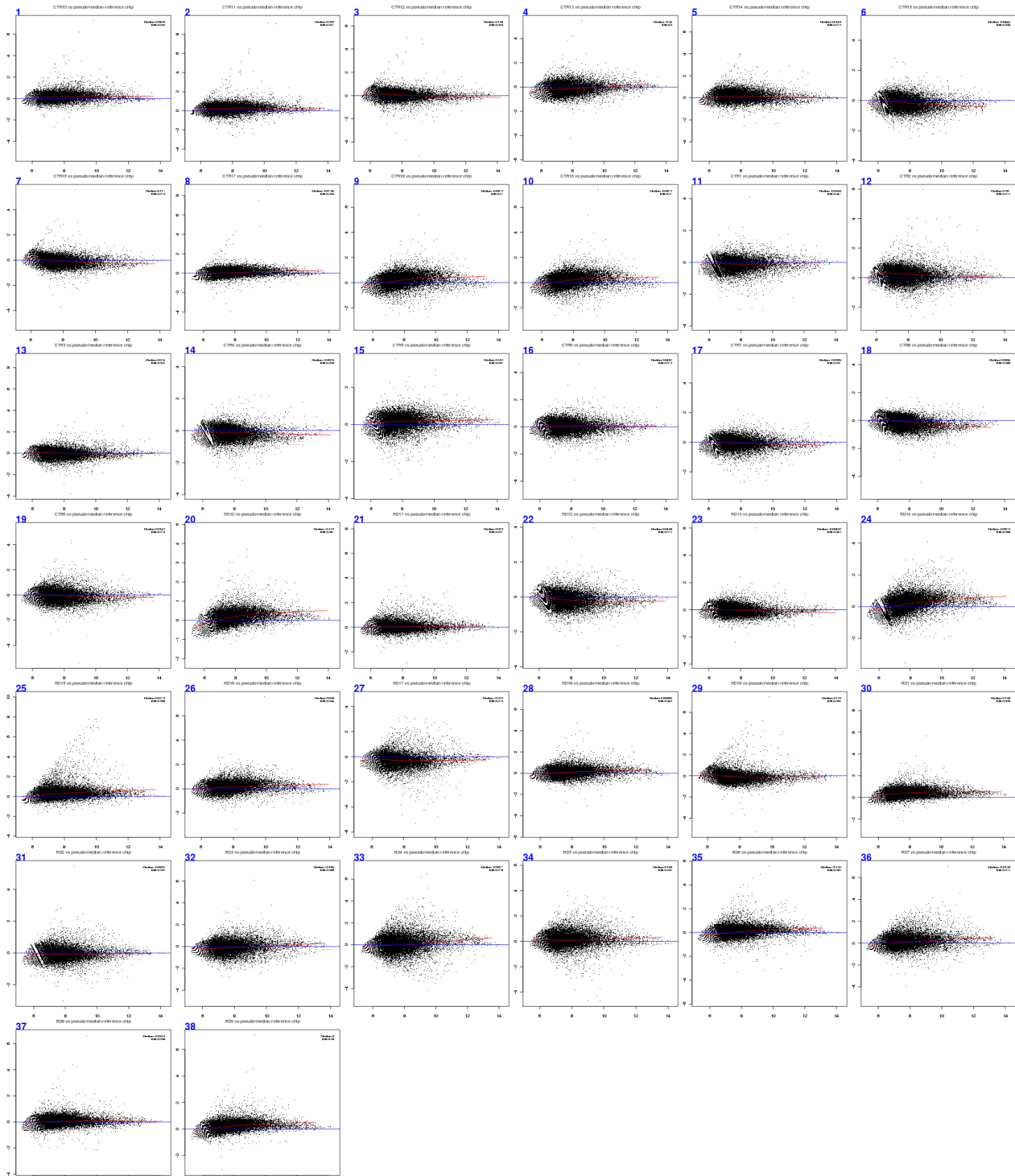
data from: /data2/g3/wrTest/RNG209/



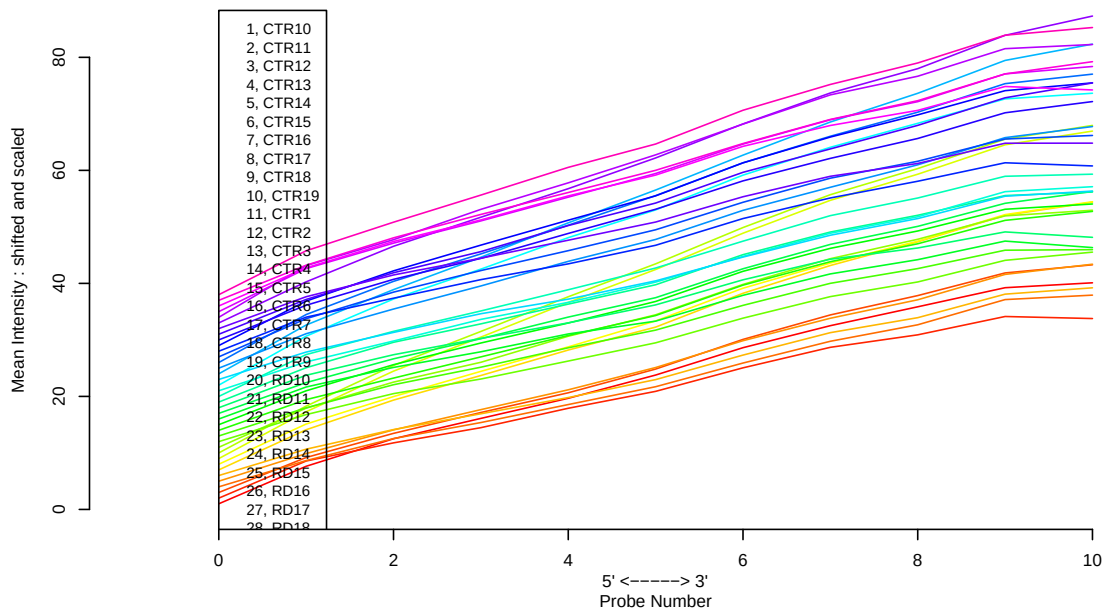
Normalized Unscaled Standard Errors (NUSE) values





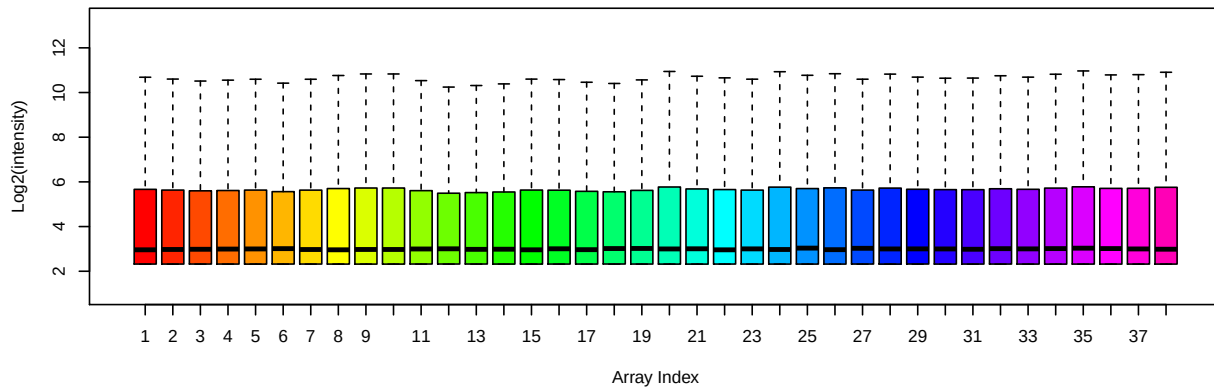


RNA degradation plot

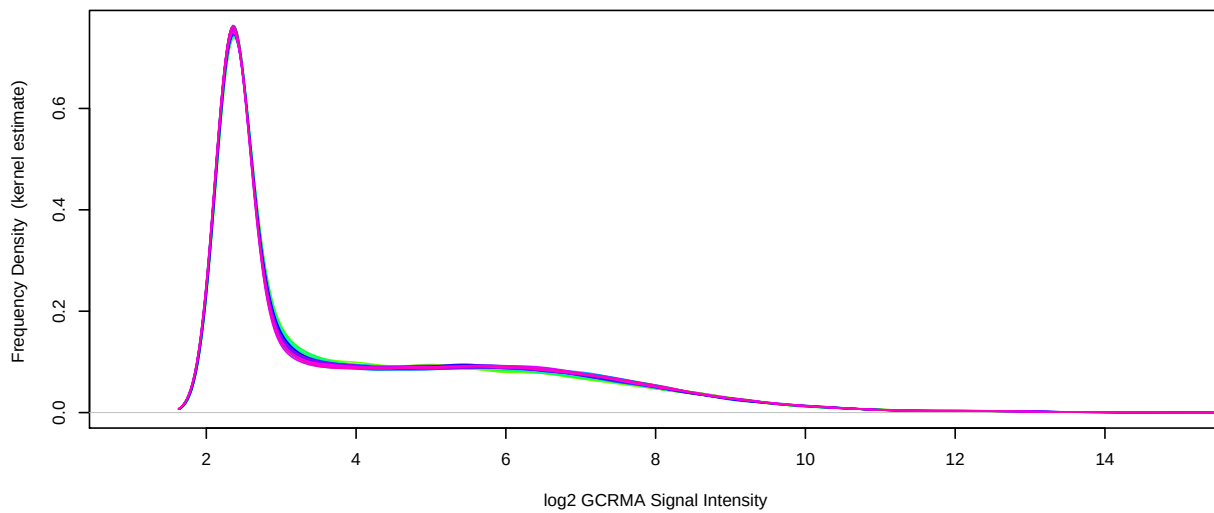


Boxplots for GCRMA Processed Signal Intensity

data from: /data2/g3/wrTest/RNG209/

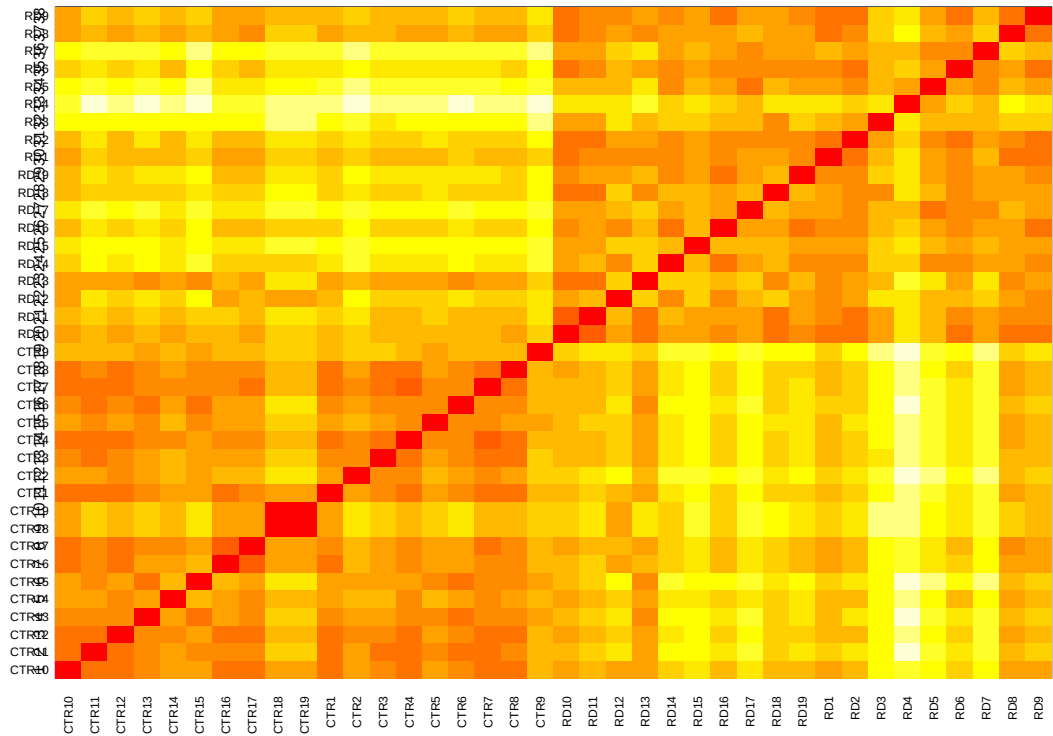


Distribution of GCRMA Processed Signal Intensity



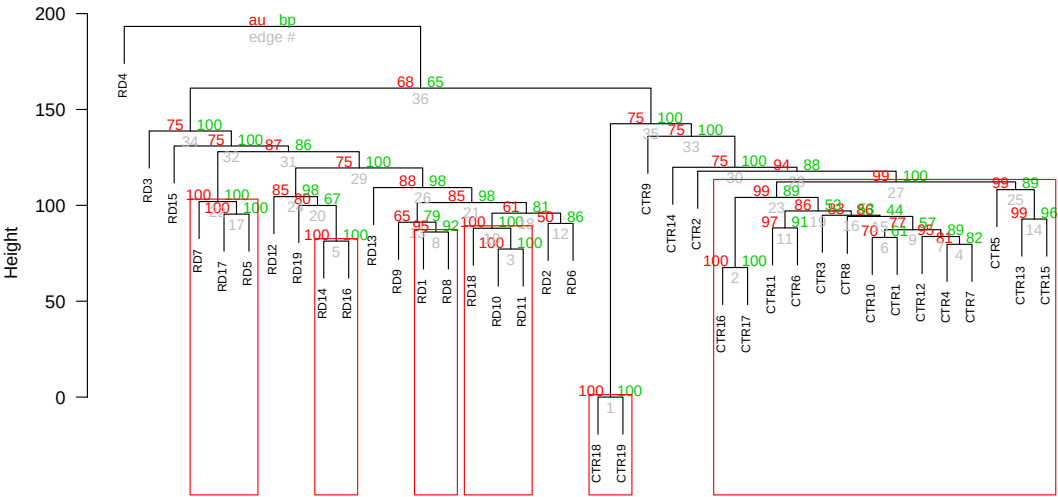
Similarity by GCRMA Summarized Samples (as Euclidean Distances)

using all probesets; n.bootstr=199; p values for AU (Approximately Unbiased) are shown in red, BP (Bootstrap Probability) in green



Dendrogram of Hierarchical Clustering of GCRMA data with Bootstrap p Values

using all probesets; n.bootstr=199; p values for AU (Approximately Unbiased) are shown in red, BP (Bootstrap Probability) in green

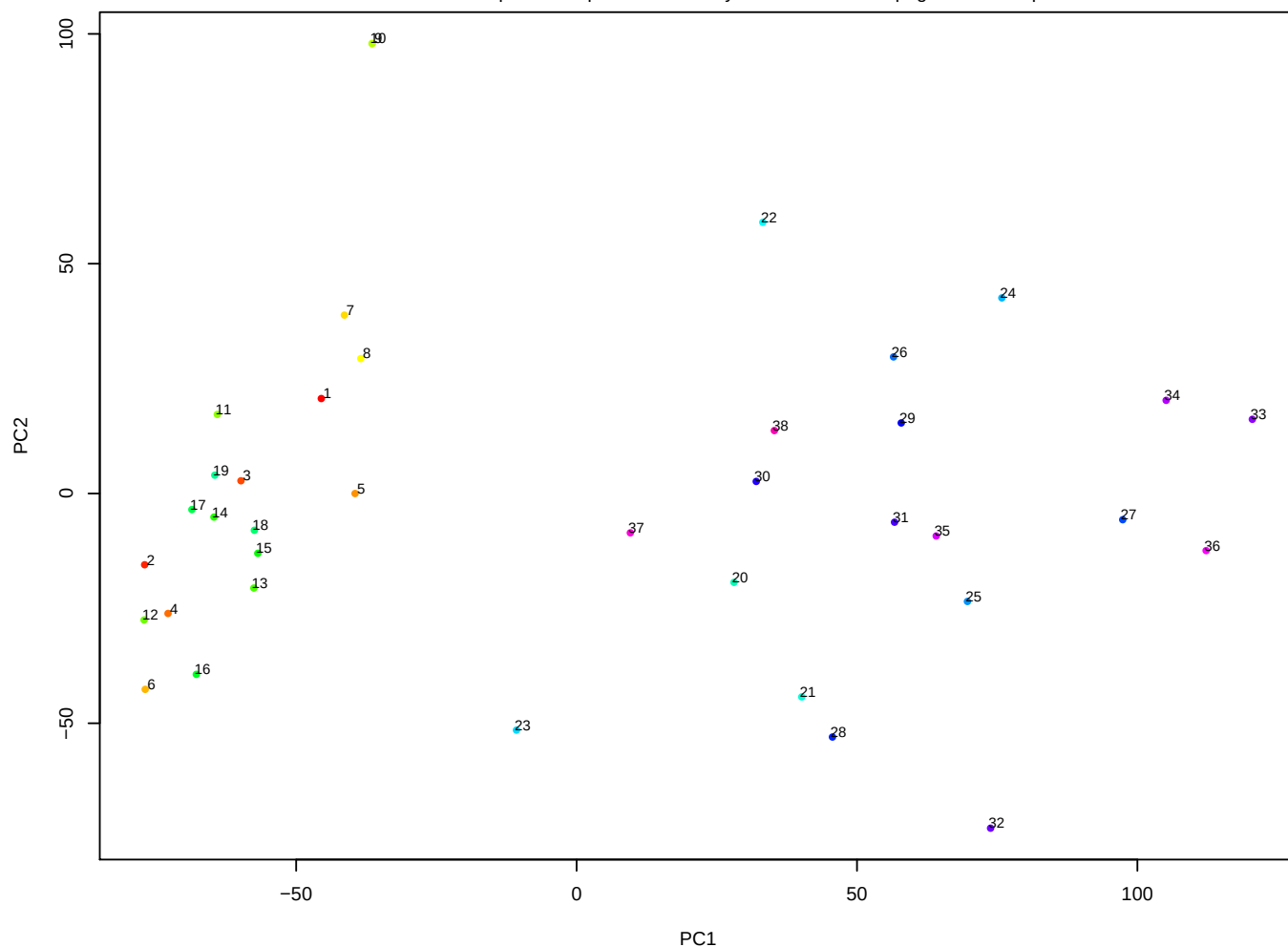


Clusters that 'seem to exist' with p>0.95 are highlighted by red boxes

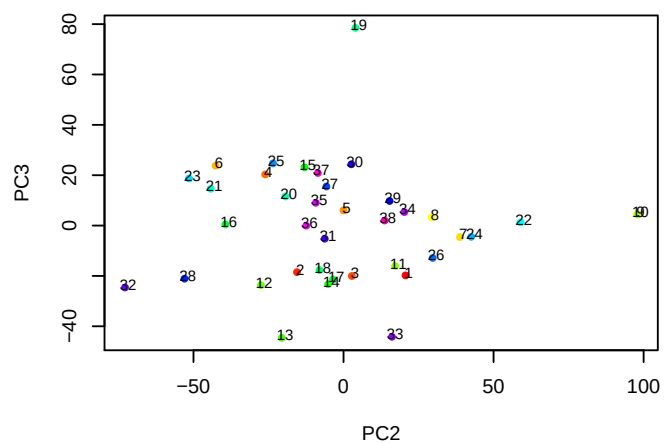
Distance: euclidean
Cluster method: average

Principal Components Plot : 1st and 2nd Component

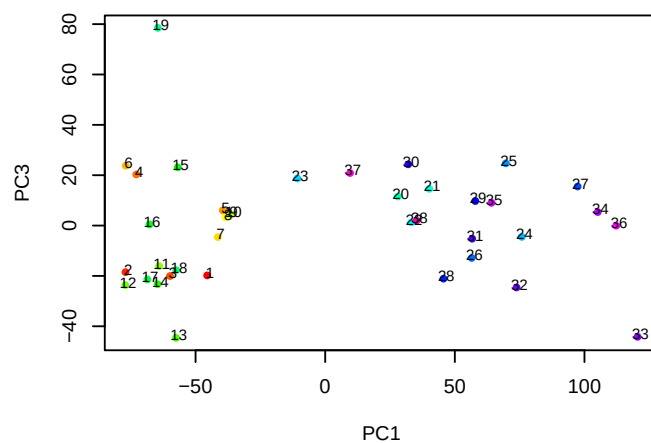
The numbers in the plot correspond to the 'array index' as listed on page 2 of this report



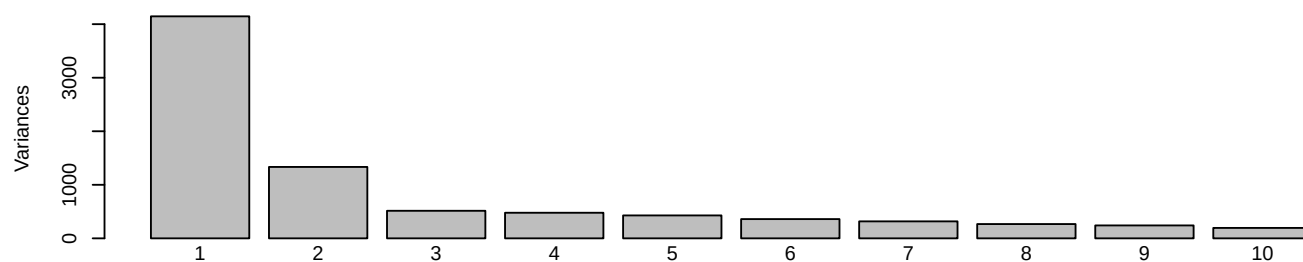
PCA : 2nd and 3rd Component



PCA : 1st and 3rd Component



Screeplot on Variance Captured by the Principal Components



..total computing time : 335.3 min

Using R version 2.12.1 on a x86-64-unknown-linux-gnu system; dedicated R packages :

	Version No.
AnnotationDbi	1.12.0
Biobase	2.10.0
RColorBrewer	1.0-2
affy	1.28.0
affyPLM	1.26.0
affyQCReport	1.28.1
gcrma	2.22.0
genefilter	1.32.0
hgu133plus2cdf	2.7.0
hgu133plus2probe	2.7.0
lattice	0.19-17
preprocessCore	1.12.0
pvclust	1.2-1
simpleaffy	2.26.1
xtable	1.5-6

References :

- Raffelsberger W, Krause Y, Moulinier L, Kieffer D, Morand AL, Brino L and Poch O : RReportGenerator : Automatic reports from routine statistical analysis using R, Bioinformatics. 2008 Jan 15; 24(2): 276-278. and <http://www-bio3d-igbmc.u-strasbg.fr/~wraff>
- R Development Core Team : R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2005 <http://www.R-project.org>.
- Leisch F : Sweave: Dynamic generation of statistical reports using literate data analysis. In: Härdle,W. and Rönz,B. (eds), Proceedings in Computational Statistics, Physica Verlag, Heidelberg, 2002 pp.575-580
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. : Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 2004, 5, R80.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP : A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185-93.
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP : Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.
- Gautier L, Cope L, Bolstad BM, Irizarry RA : affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004 Feb 12;20(3):307-15.
- Irizarry RA, Gautier L, Bolstad BM, Miller C, with contributions from Astrand M, Cope LM, Gentleman R, Gentry J, Halling C, Huber W, MacDonald J, Rubinstein B, Workman C and Zhang J : affy: Methods for Affymetrix Oligonucleotide Arrays. R package.
- Wilson CL, Miller CJ : Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics. 2005 Sep 15;21(18):3683-5.
- Bolstad B : affyPLM: Methods for fitting probe-level models. R package. (2007) <http://bmbolstad.com>
- Parman C, Halling C, Gentleman R: affyQCReport: QC Report Generation for affyBatch objects. R package
- Reimers M, Weinstein JN : Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. BMC Bioinformatics. 2005 Jul 1;6:166.
- Smyth GK : Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005, pages 397-420.
- Gentleman RC, Carey VJ, Huber W, Irizarry R, Dudoit S (editors): Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, 2005. ISBN: 0-387-25146-42005.
- <http://www.affymetrix.com>