Quality Report for Affymetrix Microarray Experiment CLLbatch

May 11, 2007

Contents

This is a quality assessment report for the dataset *CLLbatch*. The data are comprised of 24 arrays, of type `HG_U95Av2`.

For details on the software packages that were used to produce this report see Section 4.

1 The quality metrics recommended by Affymetrix

Affymetrix recommends a number of quality metrics that can be calculated for each array.

- Average background intensity, scale factors and percent of genes called present. These are shown in Table 1. The values should be similar across arrays. In the presented data, the ratio of the largest to the smallest value of average background is 1.475. Since this ratio is less than 3 there is unlikely to be a problem. Among the scale factors, the ratio of the maximum to the minimum value is 6.241. Since this ratio is larger than 3 there is a potential problem. For the percent present calls, it is 1.755. Since this ratio is less than 3 there is unlikely to be a problem.

- Ratios of hybridization efficiency between probes at the 3' and 5' ends of some control probe sets. These are displayed in Table 2. They should all be less than 3.

- External control probes. The protocols suggest that labelled cRNAs be added during sample preparation. These are BioB, BioC, BioD and CreX and are derived from Bacillus subtiliis. Nothing else should bind to their probesets. The results for these quantities are reported in Table 3. It is intended that BioB be spiked in at the lower limit of detection and that BioC, BioD and CreX be spiked in at higher concentrations. If BioB is routinely absent, then there may be a problem with sensitivity.

These quality metrics are also summarized in Figure 1. Any metric that is shown in red is out of the manufacturer's specified boundaries and suggests a potential problem.

The quality metrics reported in this Section and Figure 1 were generated using the simpleaffy package. For further information, we recommend the documentation and vignettes in the simpleaffy package.

2 Per array intensity distributions

2.1 Before normalization

The quality metrics in this section look at the distribution of the (raw, unnormalized) feature intensities for each array. Figure 2 shows density estimates (histograms), and Figure 3 presents boxplots of the same data. Arrays whose distributions are very different from the others should be considered for possible problems.

2.2 After normalization

MA-plots are useful for pairwise comparisons between arrays. M and A are defined as

$$\begin{aligned} M &= \log_2(X_1) - \log_2(X_2) = \log_2 \frac{X_1}{X_2}, \\ A &= \frac{1}{2}\left(\log_2(X_1) + \log_2(X_2)\right) = \log_2 \sqrt{X_1 X_2}, \end{aligned}$$

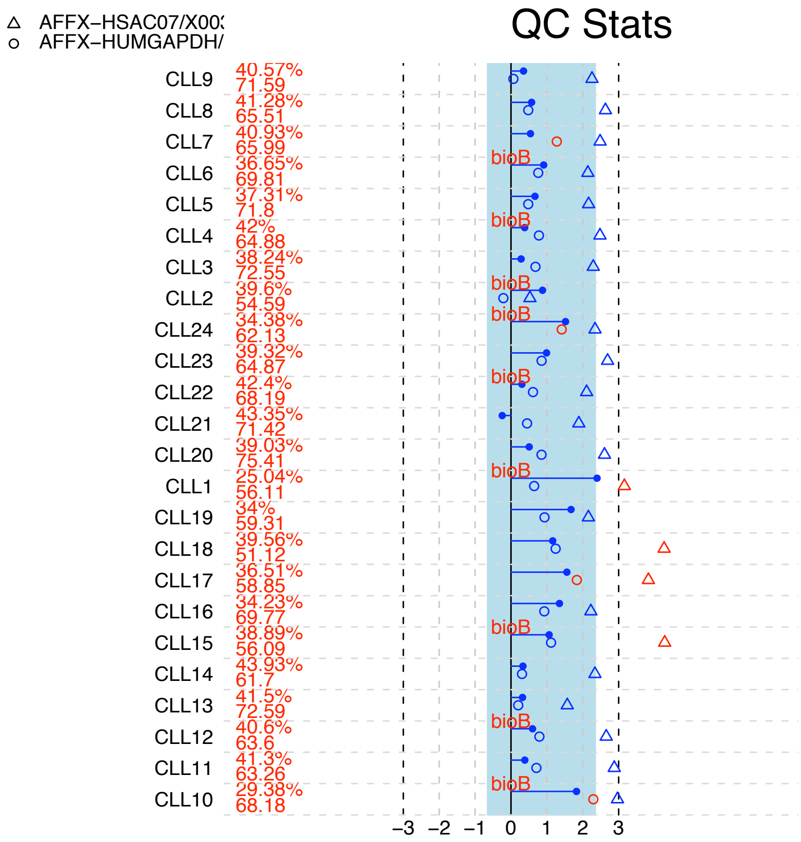

Figure 1: Quality metrics overview diagnostic plot.

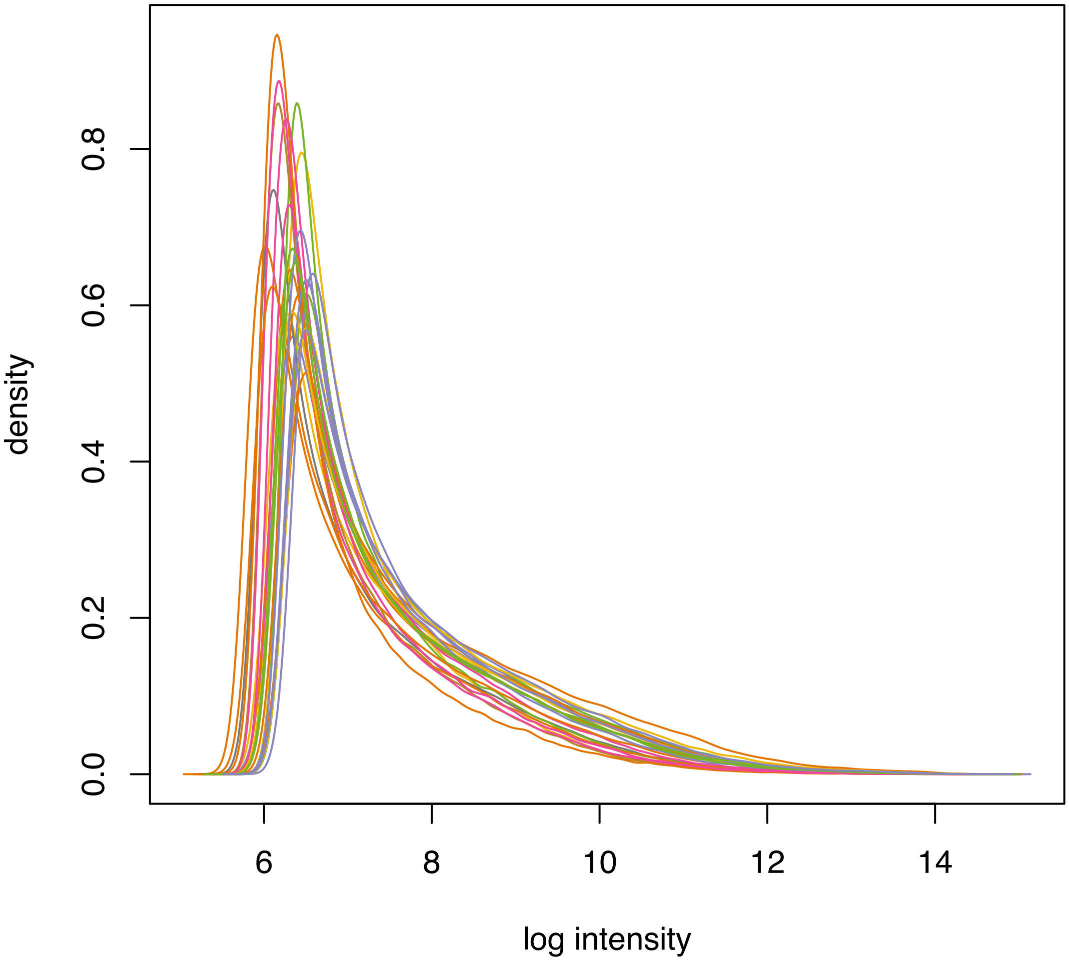

Figure 2: Density estimates (histograms) for arrays CLL10, CLL11, CLL12, CLL13, CLL14, CLL15, CLL16, CLL17, CLL18, CLL19, CLL1, CLL20, CLL21, CLL22, CLL23, CLL24, CLL2, CLL3, CLL4, CLL5, CLL6, CLL7, CLL8, CLL9.

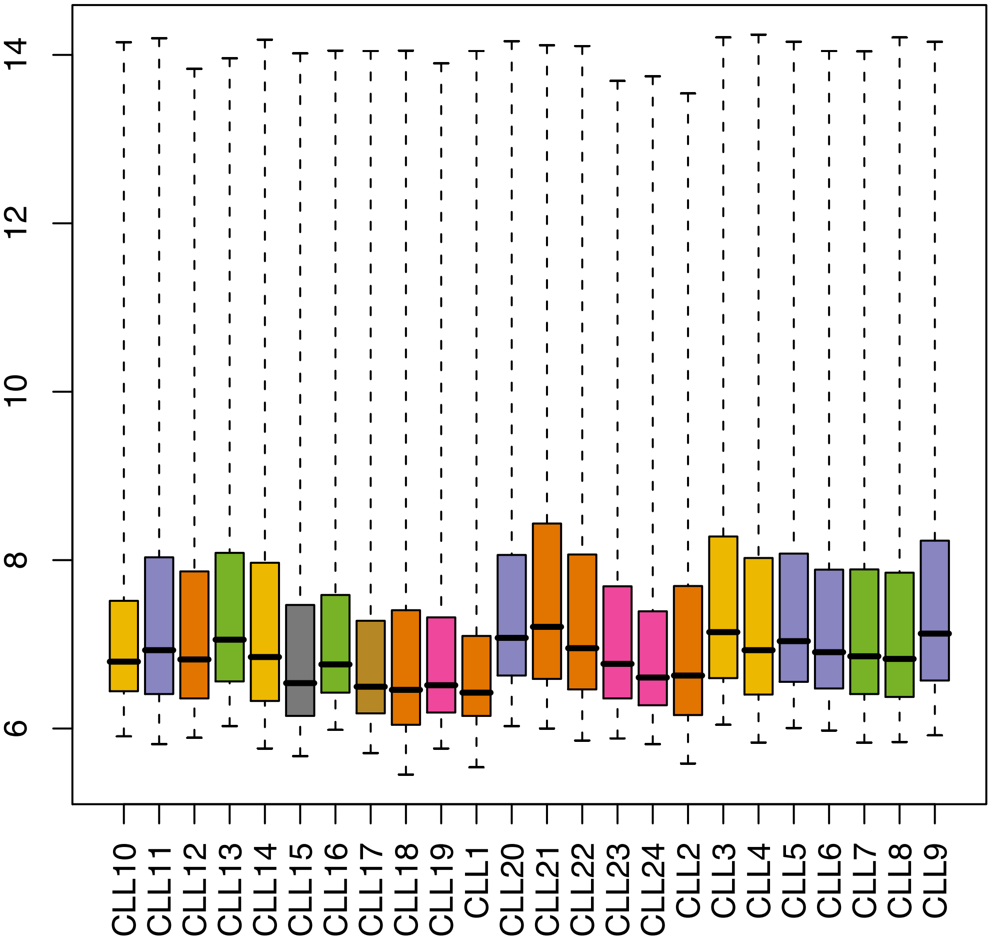

Figure 3: Boxplots for arrays CLL10, CLL11, CLL12, CLL13, CLL14, CLL15, CLL16, CLL17, CLL18, CLL19, CLL1, CLL20, CLL21, CLL22, CLL23, CLL24, CLL2, CLL3, CLL4, CLL5, CLL6, CLL7, CLL8, CLL9.

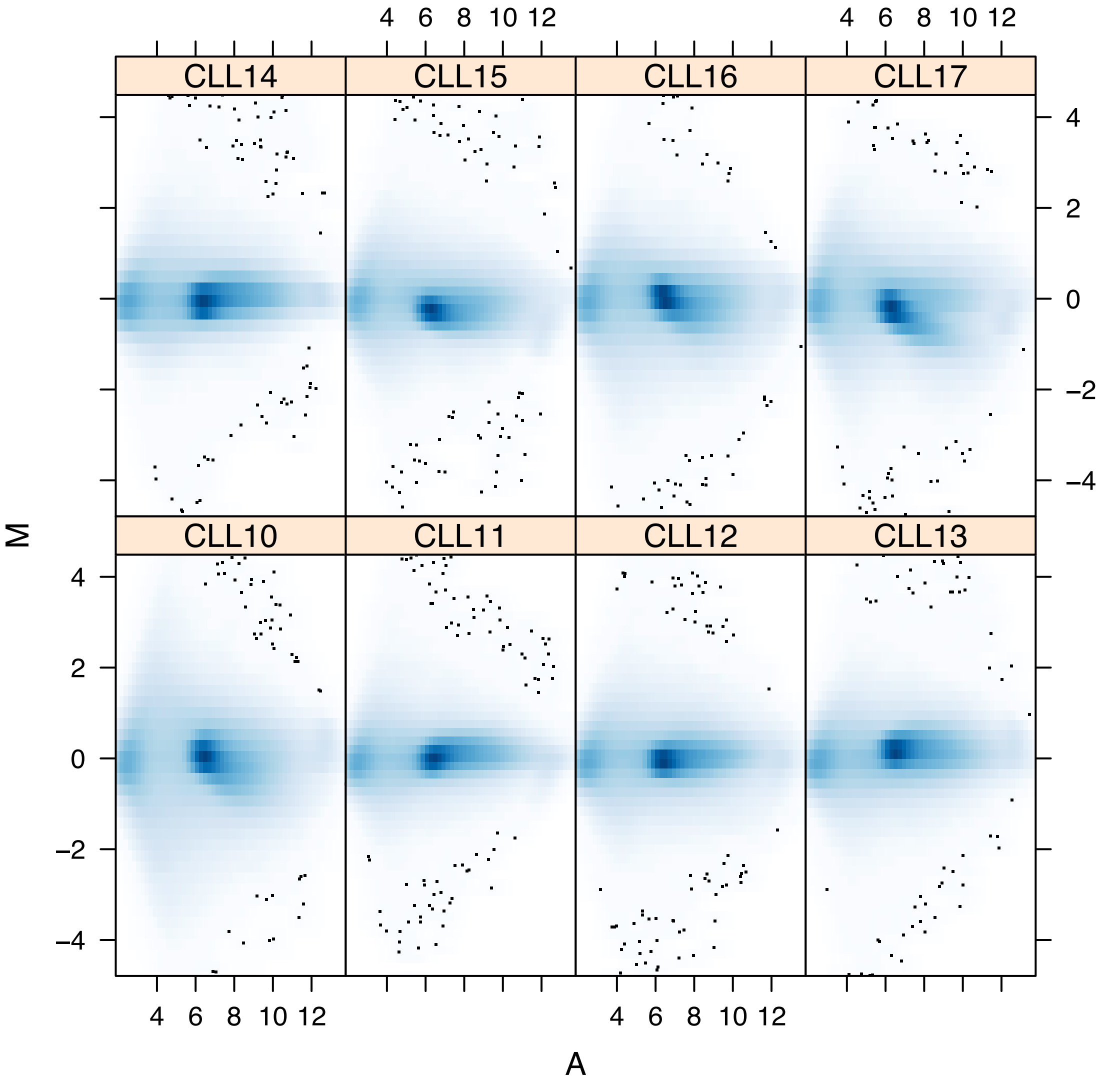

Figure 4: MA plots. A *reference array* array is calculated from the median across arrays, and for each array M and A values are calculated for the comparison to that reference.

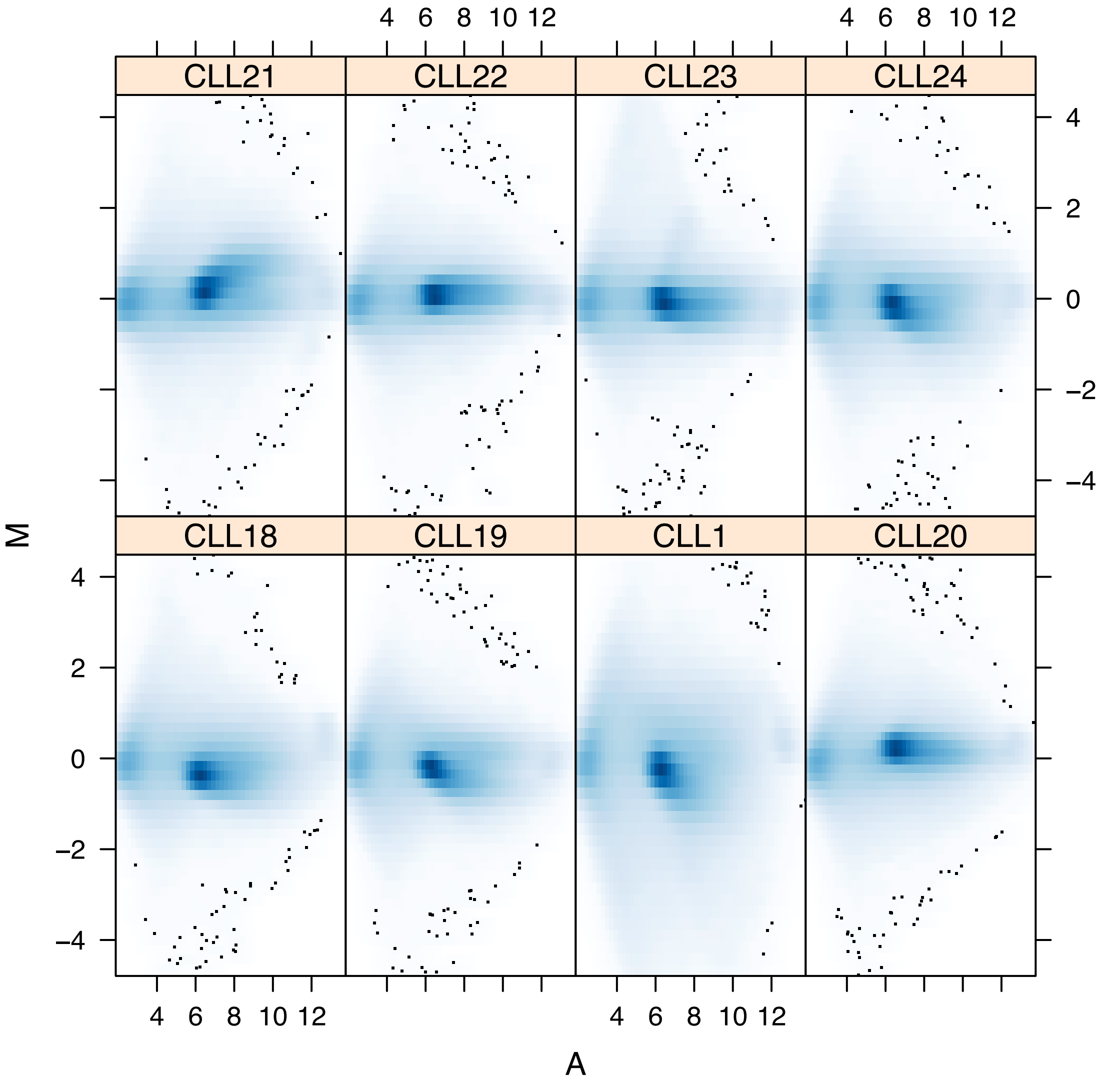

Figure 5: MA plots. A *reference array* array is calculated from the median across arrays, and for each array M and A values are calculated for the comparison to that reference.

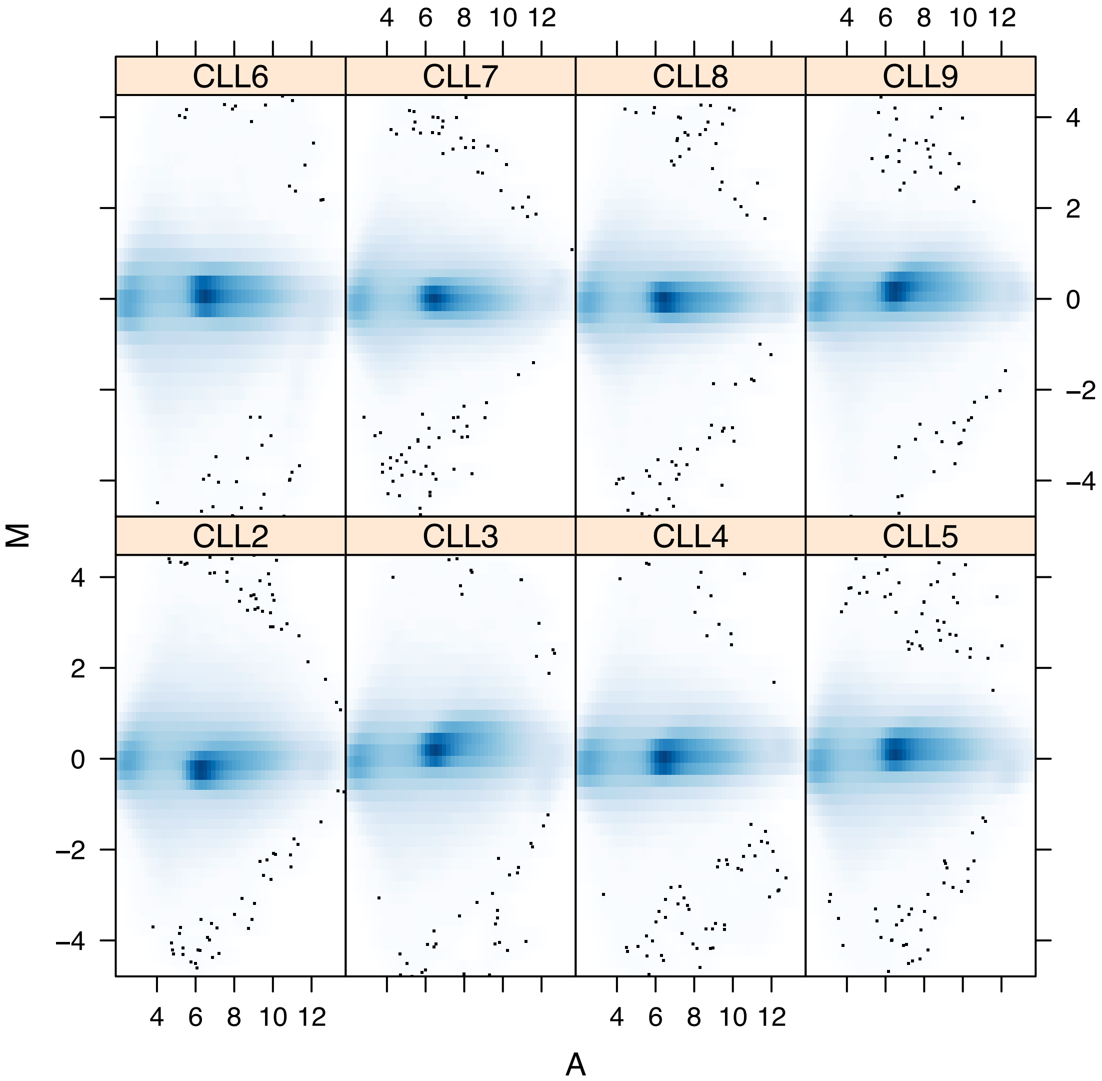

Figure 6: MA plots. A *reference array* array is calculated from the median across arrays, and for each array M and A values are calculated for the comparison to that reference.

	AvBg	ScaleF	PerCPres
CLL10	68.18	3.54	29.38
CLL11	63.26	1.31	41.30
CLL12	63.60	1.52	40.60
CLL13	72.59	1.25	41.50
CLL14	61.70	1.26	43.93
CLL15	56.09	2.09	38.89
CLL16	69.77	2.55	34.23
CLL17	58.85	2.94	36.51
CLL18	51.12	2.24	39.56
CLL19	59.31	3.19	34.00
CLL1	56.11	5.27	25.04
CLL20	75.41	1.42	39.03
CLL21	71.42	0.84	43.35
CLL22	68.19	1.24	42.40
CLL23	64.87	1.99	39.32
CLL24	62.13	2.87	34.38
CLL2	54.59	1.84	39.60
CLL3	72.55	1.22	38.24
CLL4	64.88	1.30	42.00
CLL5	71.80	1.59	37.31
CLL6	69.81	1.88	36.65
CLL7	65.99	1.46	40.93
CLL8	65.51	1.49	41.28
CLL9	71.59	1.27	40.57

Table 1: Average background, scale factor and percent present calls.

where X_1 and X_2 are the vectors of normalized intensities of two arrays, on the original data scale (i. e. not logarithm-transformed).

For the MA-plots shown in Figure 6, the data were background corrected and normalized, but not summarized (so there is one value per probe, not one value per probeset). Rather than comparing each array to every other array, here we compare each array to a single median "pseudo"-array.

Typically, 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.

Note that a bigger width of the plot of the M-distribution at the lower end of the A scale does not necessarily imply that the variance of the M-distribution is larger

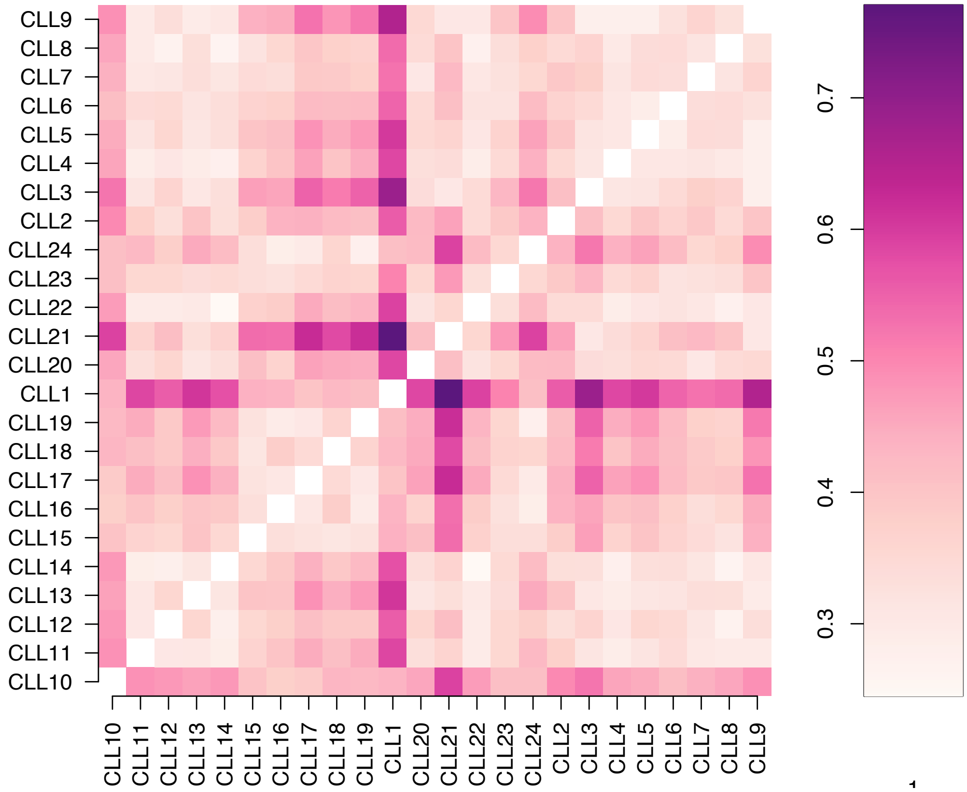

Figure 7: Pairwise differences between arrays, computed as the median absolute deviation (MAD) of the differences of the M-values.

at the lower end of the A scale: the visual impression might simply be caused by the fact that there is more data at the lower end of the A scale. To visualize whether there is a trend in the variance of M as a function of A, consider plotting M versus rank(A).

3 Between array comparisons

Figure 7 shows a false color display of between arrays distances, computed as the MAD of the M-values of each pair of arrays.

$$d_{ij} = c \cdot \underset{m}{\operatorname{median}} \left| x_{mi} - x_{mj} \right| .$$

Here, x_{mi} is the normalized intensity value of the m-th probe on the i-th array, on the original data scale. $c = 1.4826$ is a constant factor that ensures consistency with the empirical variance for Normally distributed data (see manual page of the *mad* function in R).

Figure 7 is an exploratory plot that can help detecting (a) outlier arrays and (b) batch effects. The analysis of this plot is subjective and context-dependent: there are no objective numeric thresholds when to call something an outlier. Consider the following decomposition of x_{mi}:

$$x_{mi} = z_m + \beta_{mi} + \varepsilon_{mi}, \tag{1}$$

where z_m is the probe effect for probe m (the same across all arrays), ε_{mi} are i.i.d. random variables with mean zero and β_{mi} is such that for any array i, the majority of values β_{mi} are negligibly small (i. e. close to zero). β_{mi} represents differential expression effects. In this model, all values d_{ij} are (in expectation) the same, namely $\sqrt{2}$ times the standard deviation of ε_{mi}. Arrays whose distance matrix entries are way different give cause for suspicion.

If there is an outlier array, you will expect to see vertical and horizontal stripes in the plot of darker color. Batch effects that are aligned to the order of the arrays as they are read in can be seen as blocks along the diagonal. If you see neither, you are lucky, and the data passes this quality criterion.

4 Other plots (degradation and affyPLM)

In this section we present diagnostic plots based on tools provided in the affyPLM package.

In Figure 8 a RNA digestion plot is computed. In this plot each array is represented by a single line. It is important to identify any array(s) 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.

Figure 9 is a Normalized Unscaled Standard Error (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.

Figure 10 is a Relative Log Expression (RLE) plot and an array that has problems will either have larger spread, or will not be centered at M = 0, or both.

Acknowledgements

This report was generated using version 1.14.0 of the affyQCReport, written by Craig Parman, Conrad Halling, and Robert Gentleman. It uses functions from the affy package written by R. Irizarry et al, the simpleaffy package, written by C. J.

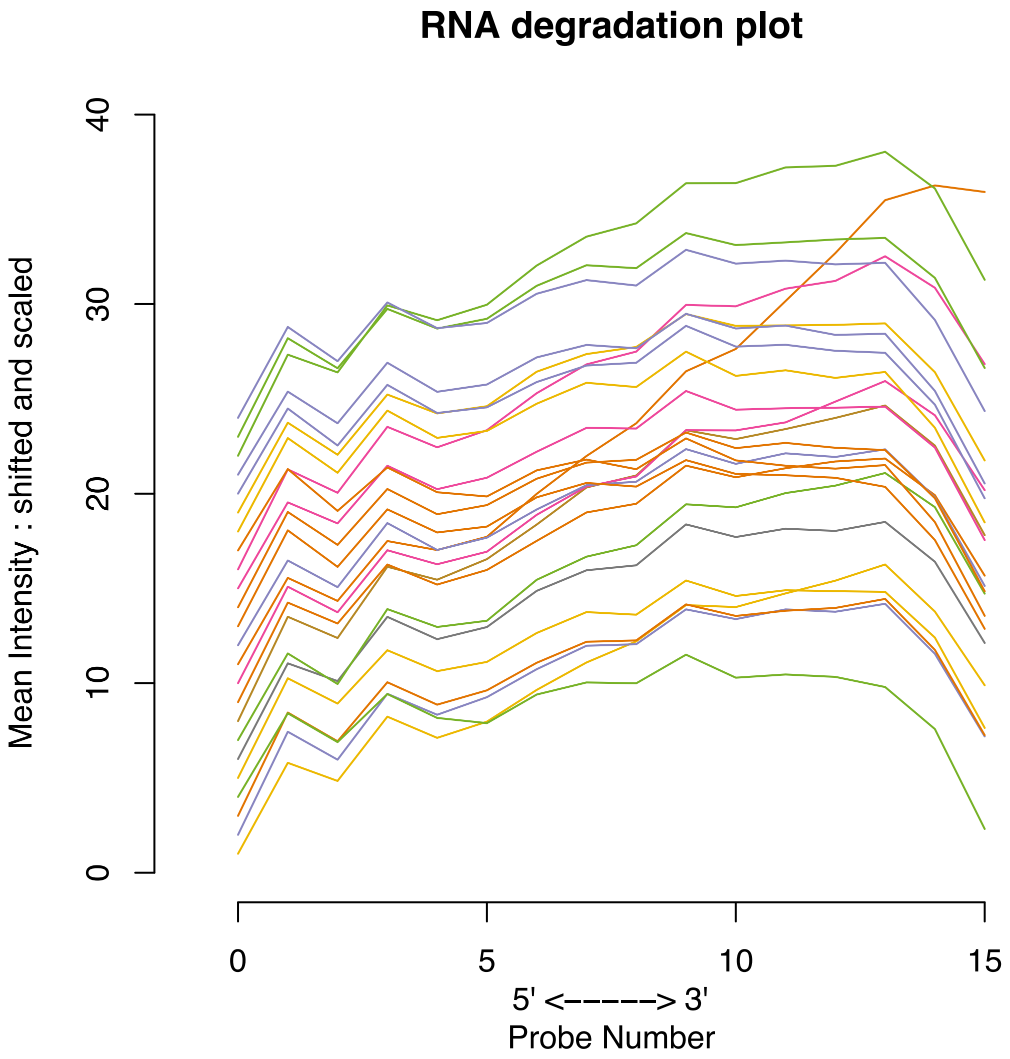

Figure 8: RNA digestion / degradation plots for arrays CLL10, CLL11, CLL12, CLL13, CLL14, CLL15, CLL16, CLL17, CLL18, CLL19, CLL1, CLL20, CLL21, CLL22, CLL23, CLL24, CLL2, CLL3, CLL4, CLL5, CLL6, CLL7, CLL8, CLL9.

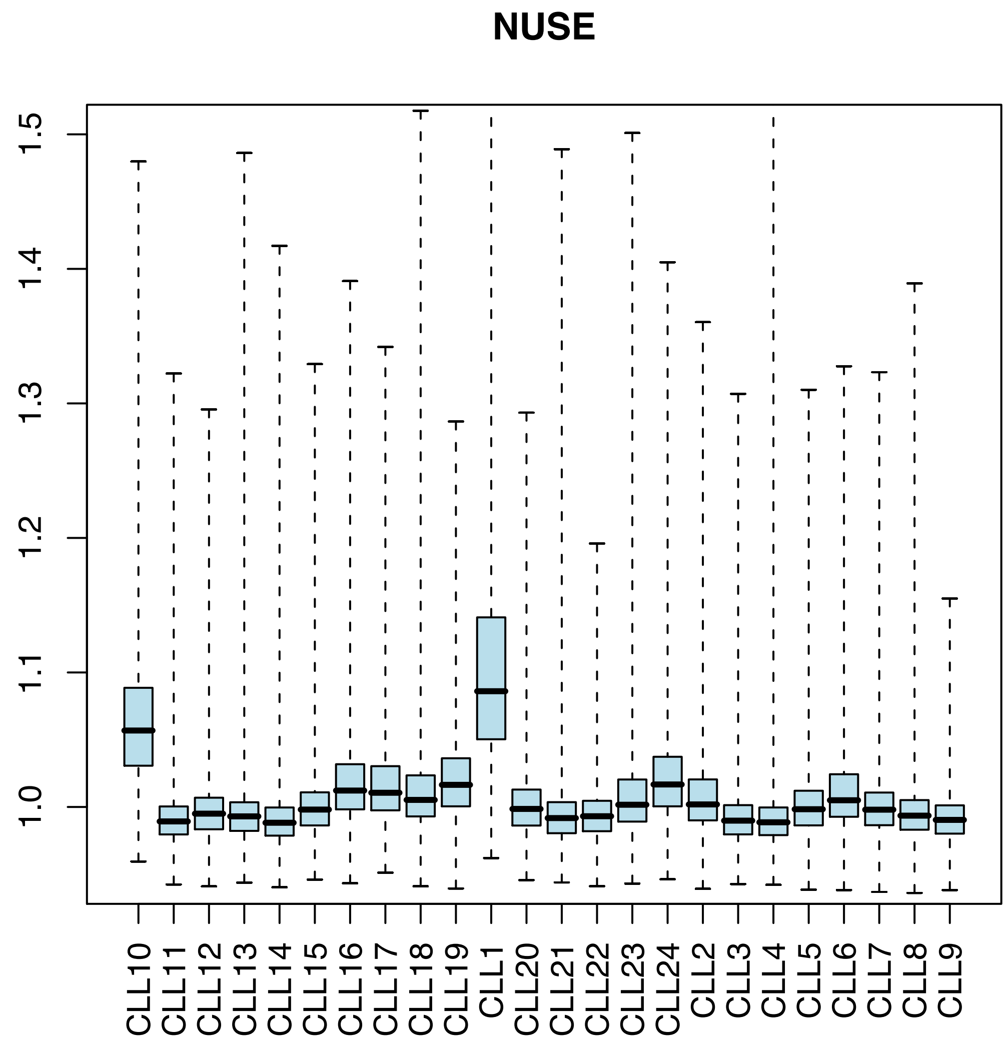

Figure 9: NUSE plot.

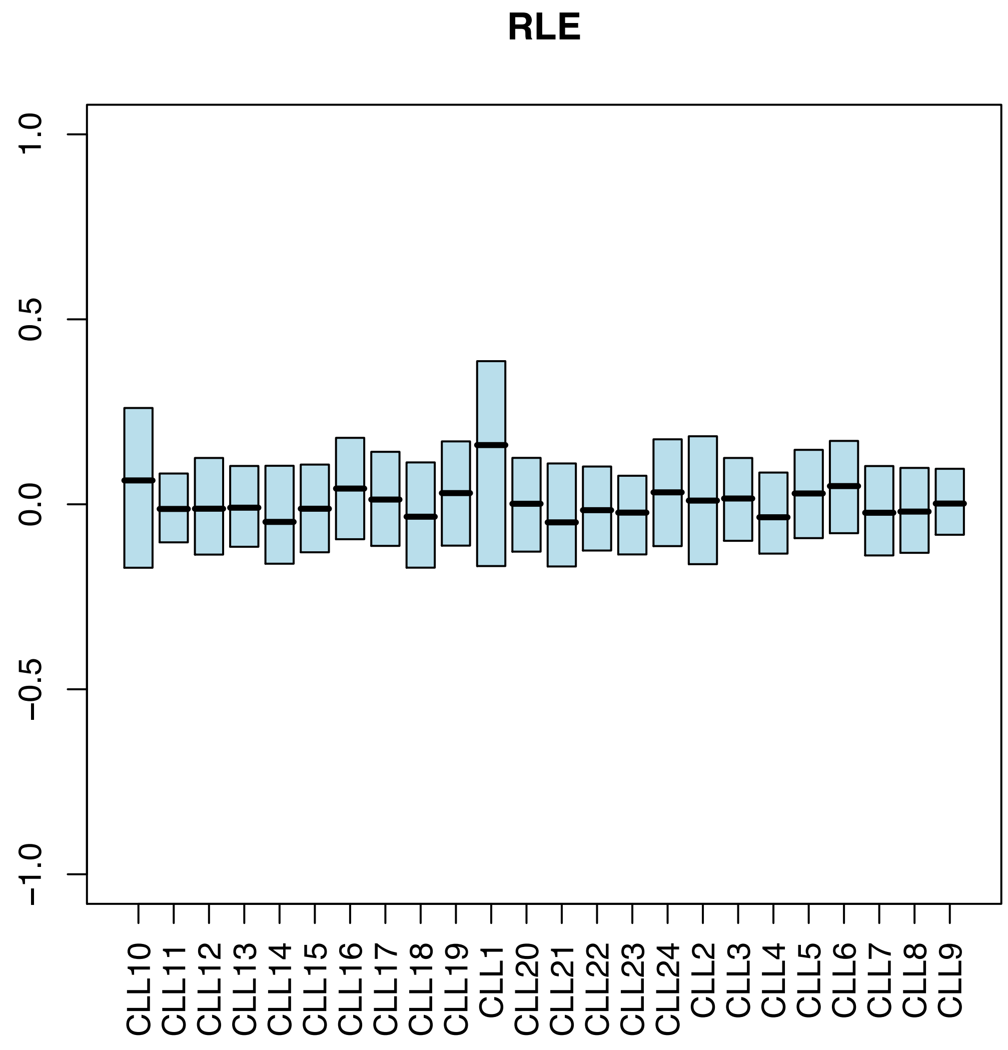

Figure 10: RLE plot.

Miller and the affyPLM package written by B. M. Bolstad. W. Huber contributed substantially to the format and functions. D. Sarkar contributed the lattice graphics for the MA plots.

SessionInformation:

- R version 2.5.0 RC (2007-04-22 r41275), `i386-apple-darwin8.9.1`
- Locale: `C`
- Base packages: base, datasets, grDevices, graphics, methods, splines, stats, tools, utils
- Other packages: Biobase 1.14.0, CLL 1.2.2, RColorBrewer 0.2-3, affy 1.14.0, affyPLM 1.12.0, affyQCReport 1.14.0, affydata 1.11.2, affyio 1.4.0, annotate 1.14.1, gcrma 2.8.0, genefilter 1.14.1, geneplotter 1.14.0, hgu95av2cdf 1.16.0, lattice 0.15-5, matchprobes 1.8.1, simpleaffy 2.11.2, survival 2.31, xtable 1.4-3

	a	b	c	d
CLL10	2.97	2.29	3.13	2.58
CLL11	2.88	0.71	1.36	0.62
CLL12	2.65	0.79	0.81	0.30
CLL13	1.57	0.20	0.34	0.01
CLL14	2.34	0.31	0.71	0.11
CLL15	4.28	1.12	3.15	0.64
CLL16	2.23	0.93	1.72	1.02
CLL17	3.83	1.83	2.75	1.33
CLL18	4.27	1.24	2.71	0.87
CLL19	2.16	0.94	1.03	0.83
CLL1	3.16	0.65	2.38	1.31
CLL20	2.61	0.85	1.76	0.60
CLL21	1.89	0.45	0.43	−0.11
CLL22	2.11	0.61	0.65	0.14
CLL23	2.69	0.85	0.91	0.37
CLL24	2.34	1.41	1.27	1.32
CLL2	0.53	−0.21	0.03	−0.13
CLL3	2.29	0.68	0.94	0.40
CLL4	2.48	0.78	0.94	0.58
CLL5	2.16	0.48	0.96	0.09
CLL6	2.14	0.76	0.61	0.21
CLL7	2.48	1.28	1.11	1.02
CLL8	2.63	0.48	1.02	0.26
CLL9	2.26	0.07	0.92	−0.01

Table 2: 3’/5’ ratios. a) HSAC07/X00351 3’/5’ b) HUMGAPDH/M33197 3’/5’ c) HSAC07/X00351 3’/M d) HUMGAPDH/M33197 3’/M.

	BioBCall	BioB	BioC	BioDn	CreX
CLL10	P	7.807	9.317	11.899	13.232
CLL11	A	3.469	6.727	9.319	10.712
CLL12	P	6.329	8.557	11.065	12.462
CLL13	M	5.134	7.11	9.89	10.998
CLL14	P	6.268	8.072	10.756	12.03
CLL15	P	6.362	7.973	10.574	11.788
CLL16	M	5.823	7.992	10.503	11.483
CLL17	P	5.856	7.781	10.351	11.61
CLL18	P	7.144	8.725	11.32	12.706
CLL19	P	6.738	8.516	11.379	12.655
CLL1	P	7.689	9.167	11.917	13.465
CLL20	A	5.294	7.497	10.255	11.522
CLL21	P	5.494	6.839	8.914	9.838
CLL22	P	5.494	7.263	10.06	11.368
CLL23	M	5.498	7.072	9.833	11.109
CLL24	P	6.309	8.117	10.824	12.087
CLL2	M	5.411	7.361	9.941	11.388
CLL3	A	5.111	6.976	9.536	10.321
CLL4	P	6.26	7.781	10.277	11.802
CLL5	M	5.741	7.33	9.959	11.563
CLL6	P	6.471	7.291	9.854	10.86
CLL7	M	5.103	7.287	10.103	11.338
CLL8	P	6.39	8.49	10.899	12.231
CLL9	P	6.053	8.113	10.492	11.981

Table 3: BioB and friends