yaqcaffy: Affymetrix expression GeneChips quality control and reproducibility with MAQC datasets

Laurent Gatto

October 28, 2009

Contents

1 Introduction

The *yaqcaffy* package is part of the Bioconductor[1] project. It was written to automate the analysis of Affymetrix expression arrays and test in-house Human Whole Genome GeneChips array reproducibility against (a subset of) the Microarray Quality Consortium (MAQC) reference datasets. It is based

[1] `http://www.bioconductor.org/`

on the *affy* and, in particular, *simpleaffy* packages, which do all the hard work. The *simpleaffy* package provides a variety of functions for high-level analysis of Affymetrix data as well as methods to assess some quality metrics of the arrays.

Since *yaqcaffy* is based on the *simpleaffy* (for example, it creates an `YAQCStats` object which is a subclass of *simpleaffy*'s `QCStats`), a basic understanding of the library, its vignette and the simpleaffy QC capabilities described in *QC and Affymetrix data*[2] is welcome.

2 The Affymetrix quality metrics

The **scale factor** (`scale.factors` slot[3]) is an array specific value that is used by Affymetrix software to adjust array intensities towards a user defined target value (default tgt=100 in *simpleaffy* and *yaqcaffy*) based on the (trimmed) mean array intensities. If there are no biases of labeling or hybridization across arrays, the highest value for the scale factor should be less than three times the smallest value.

The **background** and **noise averages** (`average.background`[3] and `average.noise` slots) assume that the hybridization occurred with the similar background and noise. Affymetrix suggests that arrays being compared should ideally have comparable background and noise values.

The **percentage of present calls** (`percent.present`[3] slot) assumes that the number of probe sets called present relative to the total number of probe sets remains similar across arrays. Nevertheless, variability in the percentage of present calls might also represent biological variability.

The internal probe calls **AFFX-r2-Ec-bioB** (M', 3' ,5'), **bioC** (5', 3') and **bioD** (5', 3') (`morespikes` and `bio.calls` slots) are *E. coli* genes that are used as internal hybridization controls and must always be present (P)[4]. Furthermore, the overall signal AFFX-r2-Ec-bioB (All), AFFX-r2-Ec-bioC (All) and AFFX-r2-Ec-bioD (All) for these spikes are present in increasing concentration (1.5 pM, 5 pM and 25 pM for bioB, bioC and bioD respectively).

[2] `http://bioinf.picr.man.ac.uk/simpleaffy/QCandSimpleaffy.pdf`

[3] defined in the *simpleaffy*'s QCStats object

[4] Note that bioB is at the level of array sensitivity and might be absent (A) in less then 50% calls.

The ploy-A controls **AFFX-r2-Bs-Dap**, **AFFX-r2-Bs-Thr**, **AFFX-r2-Bs-Phe** and **AFFX-r2-Bs-Lys** (`morespikes` slot) are modified *B. subtilis* genes and should be called present at a decreasing intensity, to verify that there was no bias during the retro-transcription between highly expressed genes and low expressed genes. Note that the linearity for lys, phe and thr (dap is present at a much higher concentration) is affected by a double amplification.

Note that Affymetrix provides two sets of internal *bio* and *poly-A* controls. If we take as an example the bioB spike control, two similar probe sets IDs are present on some GeneChips: `AFFX-BioB-3_at` and `AFFX-r2-Ec-bioB-3_at`. These two probe sets target the same gene, but the individual probes are slightly shifted. The *r2* probe sets include less probes (11 for each control spike) than the older non-*r2* sets (20 probes per set). The *yaqcaffy* package uses the *r2* probe sets unless these are not available (as in older GeneChips).

The **GAPDH** and **β-Actin** 3'/5' signal ratios are RNA degradation controls (see slot `gcos.probes`). These values should generally be smaller then 3. Nevertheless, double amplification is known to have a significant impact on these two parameters.

More information regarding the Affymetrix internal controls can be found in the *GeneChip Expression Analysis* and *Data Analysis Fundamentals* manuals [5].

To assess the quality of the samples to analyses, we suggest that qc metrics should lie within 2 standard deviations of one another across the entire set of arrays. We apply this rule to the above mentioned metrics. For the scale factor, we define the upper and lower limits as the $mean/2$ and $mean * 1.5$ respectively to stick to Affymetrix's three-fold rule.

3 The MAQC reference datasets

The Microarray Quality Consortium (MAQC) project[6] provides a set of reference datasets for a set of platforms (see *Summary of the MAQC Data Sets*[7] for more details). Regarding the Affymetrix platform (AFX prefix), a total

[5] `http://www.affymetrix.com/support/technical/manual/expression_manual.affx`

[6] `http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/`

[7] `http://edkb.fda.gov/MAQC/MainStudy/upload/Summary_MAQC_DataSets.pdf`

of 120 Human Genome U133 Plus 2.0 GeneChips have been generated. Four different reference RNAs have been used: (A) 100% of Stratagene's *Universal Human Reference RNA*, (B) 100% of Ambion's Human Brain Reference RNA, (C) 75% of A and 25% of B and (D) 25% of A and 75% of B. Each reference has been repeated 5 times (noted _A1_ to _A5_) on six different test sites (noted _1_ to _6_). As an example, the .CEL result file for the first replicate of test site 2, for the reference ARN C is named `AFX_2_C1.CEL`.

These datasets are freely available and allow researchers, among other things, to compare the reproducibility of their own Human Genome U133 Plus 2.0 arrays with a set of high quality .CEL files. Nevertheless, using all the 30 available .CEL files (per reference RNA) is memory consuming and further reproducibility calculations time consuming. We randomly chose 6 .CEL file for each reference RNA, one for each test site as reference to compare the user's data to. These 6 .CEL files are distributed with the *MAQCsubsetAFX* package as associated data (respectively called `refA.RData`, `refB.RData`, `refC.RData` and `refD.RData`). These subsets are used to compute the Pearson correlation factors and draw scatterplots with the users data (see section 6).

4 Generating an `YAQCStats` objects

As an example, we will use *affydata*'s `Dilution` dataset. We will modify the raw probe intensities of the first sample to illustrate some of *yaqcaffy*'s functions below.

```
> library("yaqcaffy")
> library("affydata")
> data(Dilution)
> tmp <- exprs(Dilution)
> tmp[, 1] <- tmp[, 1] * 2
> exprs(Dilution) <- tmp
```

The next step is the creation of the `YAQCStats` object that will hold the data that will subsequently be used to assess the quality of the arrays (see section 5). The `YAQCStats` object is a subclass of the `QCStats` object, defined in the *simpleaffy* package.

The function `yaqc` computes the following values that are used for quality assignment:

1. the scale factors, percent of present calls, average background and noise that are tested as described above;

2. the bioB, bioC and bioD calls;

3. the intensity values for the bioB, bioC, bioD and dap, lys, phe and thr probes, as computed by the Affymetrix GCOS software;

4. the intensity values for GAPDH and β-actin probes as computed by the Affymetrix GCOS software.

The newly created object can then be visualized as a `data frame` with the `show()` function.

```
> yqc <- yaqc(Dilution)
> show(yqc)
```

	20A	20B
scale.factors	"0.446700673288299"	"1.26536267137974"
average.background	"188.506453894315"	"63.6385483408235"
average.noise	"5.99634613568846"	"2.05635393118791"
percent.present	"48.6732673267327"	"49.7029702970297"
AFFX-BioB-3_at	"3193.34776384752"	"2403.61326911566"
AFFX-BioB-5_at	"664.475559859502"	"529.820083206257"
AFFX-BioB-M_at	"3405.73144678145"	"2848.81587661467"
AFFX-BioC-3_at	"252.436293058404"	"205.437770825853"
AFFX-BioC-5_at	"140.223227877387"	"102.335999707106"
AFFX-BioDn-3_at	"3877.1350169243"	"3394.51773110207"
AFFX-BioDn-5_at	"1.51769065090198"	"0.845418666059053"
AFFX-LysX-3_at	"4.80101892491885"	"5.19664179741686"
AFFX-LysX-5_at	"5.5833031150604"	"2.33298400139847"
AFFX-LysX-M_at	"9.18049463242504"	"5.52469246504482"
AFFX-PheX-3_at	"7.08179780076799"	"6.16637709724212"
AFFX-PheX-5_at	"1.44519375614278"	"0.670665049720503"
AFFX-PheX-M_at	"0.674714424858515"	"1.67990083326996"
AFFX-ThrX-3_at	"3.24359573960642"	"1.69876702933264"
AFFX-ThrX-5_at	"6.10352070691629"	"14.2456808606269"
AFFX-ThrX-M_at	"9.21704036082897"	"4.98124090682067"
AFFX-DapX-3_at	"47.623445854456"	"47.0670722097149"

AFFX-DapX-5_at	"70.7879945420667"	"77.6307445901374"
AFFX-DapX-M_at	"142.108734911279"	"157.471548891886"
AFFX-HSAC07/X00351_3_at	"5545.08639240761"	"5016.28641670242"
AFFX-HSAC07/X00351_M_at	"5076.6312266744"	"4429.0480858254"
AFFX-HSAC07/X00351_5_at	"3422.54064237669"	"3043.59824738725"
AFFX-HUMGAPDH/M33197_3_at	"4453.70276234178"	"3975.28201930944"
AFFX-HUMGAPDH/M33197_M_at	"4643.6097792756"	"4013.10843566896"
AFFX-HUMGAPDH/M33197_5_at	"3276.2164923853"	"3112.48782158316"
AFFX-BioB-5_at_call	"A"	"A"
AFFX-BioB-3_at_call	"A"	"A"
AFFX-BioC-5_at_call	"P"	"P"
AFFX-BioC-3_at_call	"A"	"A"
AFFX-BioDn-5_at_call	"A"	"A"
AFFX-BioDn-3_at_call	"A"	"A"
	10A	10B
scale.factors	"1.14484301856915"	"1.84540671835491"
average.background	"80.0943568071944"	"54.2582973752169"
average.noise	"2.41104721237696"	"1.53954177290118"
percent.present	"49.2514851485149"	"49.639603960396"
AFFX-BioB-3_at	"3803.05793263888"	"4651.34002639248"
AFFX-BioB-5_at	"782.380763943258"	"868.048189730846"
AFFX-BioB-M_at	"4099.20419046011"	"4708.37272584289"
AFFX-BioC-3_at	"292.390728738772"	"400.137460869786"
AFFX-BioC-5_at	"134.208345100585"	"172.144437858793"
AFFX-BioDn-3_at	"4730.57421072258"	"5815.90360991049"
AFFX-BioDn-5_at	"1.66675953810322"	"1.05867006723521"
AFFX-LysX-3_at	"7.43964195242547"	"3.09333383020266"
AFFX-LysX-5_at	"2.01530814419427"	"1.0358874831702"
AFFX-LysX-M_at	"10.1816741622856"	"1.25190837897068"
AFFX-PheX-3_at	"8.83419969356381"	"6.15585638526382"
AFFX-PheX-5_at	"1.79541277725376"	"1.0836801640999"
AFFX-PheX-M_at	"2.06622009441486"	"1.53941417497261"
AFFX-ThrX-3_at	"3.88468657759839"	"1.89381748154357"
AFFX-ThrX-5_at	"2.00185328136356"	"9.39203341871418"
AFFX-ThrX-M_at	"4.50047460899868"	"1.90547577524205"
AFFX-DapX-3_at	"83.1313670261053"	"82.2030556006241"
AFFX-DapX-5_at	"109.999093608403"	"133.380317244418"
AFFX-DapX-M_at	"224.636335945386"	"231.292608780094"

```
AFFX-HSAC07/X00351_3_at   "7047.69553069851" "6690.4076575177"
AFFX-HSAC07/X00351_M_at   "6087.54358313079" "5538.67447202837"
AFFX-HSAC07/X00351_5_at   "3852.89986182388" "3508.18103953567"
AFFX-HUMGAPDH/M33197_3_at "4937.95681357743" "5074.631527531"
AFFX-HUMGAPDH/M33197_M_at "3681.16271724012" "4693.68394083838"
AFFX-HUMGAPDH/M33197_5_at "3658.51361772065" "3412.0483110118"
AFFX-BioB-5_at_call       "A"                "A"
AFFX-BioB-3_at_call       "A"                "A"
AFFX-BioC-5_at_call       "P"                "P"
AFFX-BioC-3_at_call       "A"                "A"
AFFX-BioDn-5_at_call      "A"                "A"
AFFX-BioDn-3_at_call      "A"                "A"
```

In the above examples, the data given as input is of class `AffyBatch` object. An `YAQCStats` object can also be created by providing an `ExpressionSet`, in which case some of the qc metrics cannot be computed: only the intensity values for the bioB, bioC, bioD and dap, lys, phe and thr probes and GAPDH and β-actin probes are used.

5 Quality control analysis

The quality metrics in the `YAQCStats` object can be plotted out to allow an easy and rapid overview, as shown on figure 1:

- the scale factors for the different arrays are plotted with the upper and lower limits as a dotchart;
- boxplots for the average background and noise, the percentage of present calls and GAPDH and β-actin $\frac{3'}{5'}$ ratios.
- boxplots of the control probes *biob*, *bioc*, *biod* and *dap*, *thr*, *phe*, *lys* intensities respectively

The mean (longdashed line), upper and lower 2 standard deviations (dotted lines) are also plotted on the graphs. The upper and lower limits may however not appear when they are outside of the boxplot y-axis. For the internal probes, a grey rectangle represents the mean (middle segment) and the +/- 2 stdev range.

```
> plot(yqc)
```

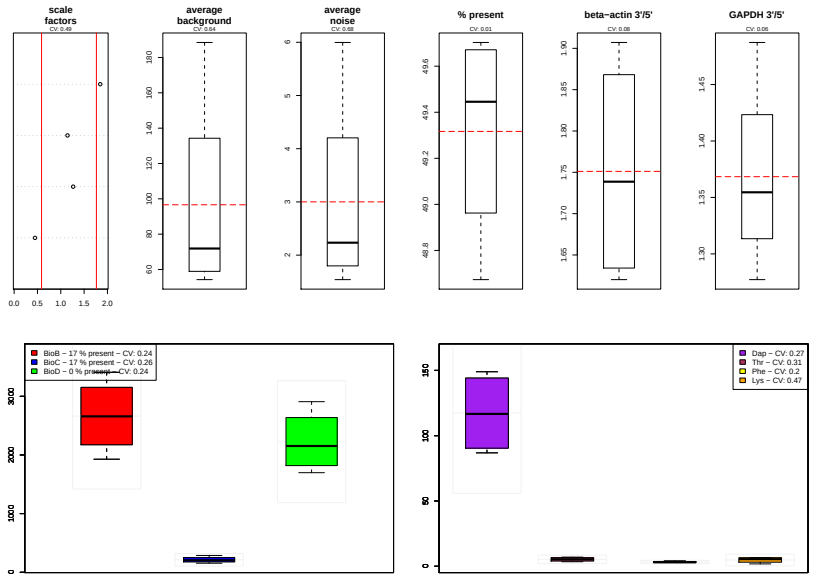


Figure 1: Graphical representation of the `YAQCStats` object.

The outliers (i.e. the data points the lie outside the mean +/- 2 stdev) can be queried and listed for each qc metrics using the `getOutliers()` function. The arguments are the `YAQCStats` object and a string describing the metrics that should be queried. In the above example, we can see that the scale factors of the fourth samples (counting from the botton) is out of range and not even present on the dotchart. We can retrieve the name of the sample and its scale factor value by typing:

```
> getOutliers(yqc, "sfs")

      20A       10B
0.4467007 1.8454067
```

The qc metrics strings are respectively sfs, avbg, avns, pp, actin, gapdh, biob, bioc, biod, dap, thr, phe, lys (listed in their order of apperance on the qc plot). Individual plots can also be generated with the `which` argument: 'sfs' for the scale factor, 'avbg' and 'avns' for the average background and noise, 'pp' for the percentage of present calls, 'gapdh' and 'actin' for the GAPDH and β-actin ratios, 'bio' for the hybridization controls and 'spikes' for the retro-transciption spiked controls. In addition, the coefficient of variation is calculated for each qc metric and indicated on the qc plot. The outliers can be summerized in a data frame calling the `summary()` function on a `YAQCStats` object.

It is also possible to combine two `YAQCStats` object into one with the `merge()` function. To illustration this function, we will use the `arrays()` function that outputs the arrays names of the `YAQCStats` provided as parameter.

```
> yqc2 <- yaqc(Dilution[, 2:3])
> arrays(yqc)

[1] "20A" "20B" "10A" "10B"

> arrays(yqc2)

[1] "20B" "10A"

> yqc3 <- merge(yqc, yqc2)
> arrays(yqc3)

[1] "20A" "20B" "10A" "10B" "20B" "10A"
```

6 Human genome U133 Plus 2.0 reproducibility

To illustrate this section, we will compare the first array of the RNA B reference dataset (`AFX_1_B1.CEL`) to the RNA A reference dataset[8].

```
> library(MAQCsubsetAFX)
> data(refB)
> d <- refB[, 1]
> sampleNames(d)
```

```
[1] "AFX_1_B1.CEL"
```

We will compare this CEL file to the `refA` dataset using the `reprodPlot` function. The name of the `AffyBatch` object to be tested is given as first argument and the reference data is specified as a character provided as second parameter (respectively `"refA"`, `"refB"`, `"refC"` or `"refD"`). The reference dataset is automatically loaded and merged with the user's `AffyBatch` object, normalized and results are plotted. The intensities used for the statistics are normalized using the RMA algorithm implemented in the *affy* package (`normalize="rma"`, default). It is also possible the use GCRMA (as implemented in the *gcrma* package, `normalize="gcrma"`), MAS5 (as implemented in *affy*, `normalize="mas5"`)) or no normalization (`normalize="none"`).

The `reprodPlot` function draws a 6 by 6 matrix showing scatterplots (below the diagonal) and the Pearson correlation factors (above the diagonal) for all comparisons. The sample names are given on the diagonal. The gray lines on the scatterplots represent respectively 2, 4 and 8 fold change differences.

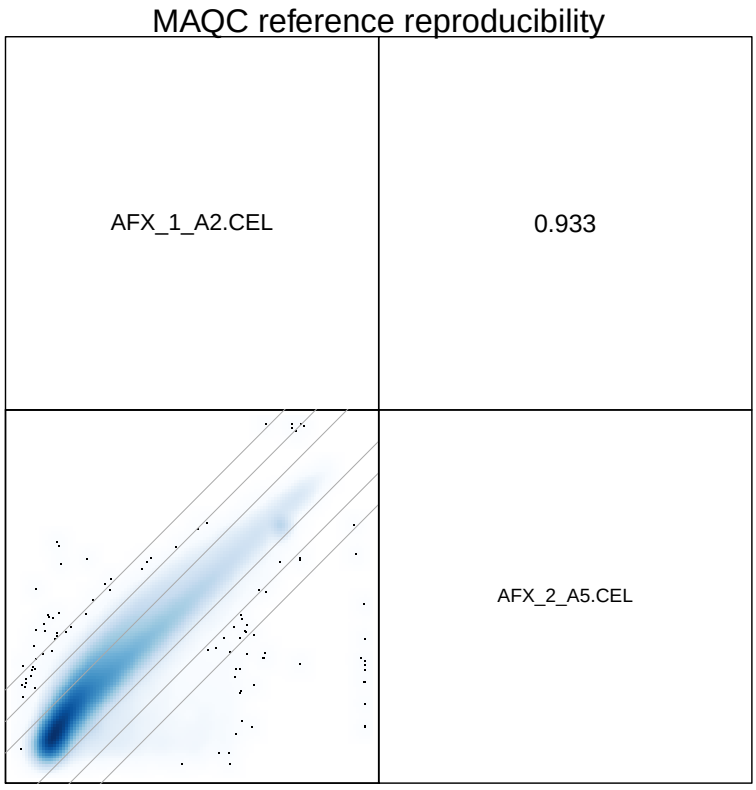
```
> reprodPlot(d, "refA", normalize = "rma")
```

The figure below is an example of the `reprodPlot` for 2 unnormalized samples [9].

```
> reprodPlot(d, "test", normalize = "none")
```

[8] Note that the reproducibility statistics will *de facto* be low, as the conditions to be compared are different.

[9] This `test` plot is used instead of the 6 by 6 plot to reduce time and size requirements to build the vignette.



7 Session information

```
> sessionInfo()

R version 2.10.0 (2009-10-26)
x86_64-unknown-linux-gnu

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=C              LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] MAQCsubsetAFX_1.0.3 hgu95av2cdf_2.5.0   affydata_1.11.10
[4] yaqcaffy_1.6.0      simpleaffy_2.22.0   gcrma_2.18.0
[7] genefilter_1.28.0   affy_1.24.0         Biobase_2.6.0

loaded via a namespace (and not attached):
 [1] affyio_1.14.0        annotate_1.24.0      AnnotationDbi_1.8.0
 [4] Biostrings_2.14.0    DBI_0.2-4            IRanges_1.4.0
 [7] KernSmooth_2.23-3    preprocessCore_1.8.0 RSQLite_0.7-3
[10] splines_2.10.0       survival_2.35-7      xtable_1.5-5
```