Advanced topics: Customizing arrayQualityMetrics reports and programmatic processing of the output

Audrey Kauffmann, Wolfgang Huber

April 14, 2011

Contents

1 Introduction

If you are new to this package, then please consult the vignette *Introduction: microarray quality assessment with arrayQualityMetrics*.
This vignette addresses advanced topics. It explains how to customize the report by selecting specific modules and sections, or by adding your own ones. Furthermore, we will see how to (programmatically) postprocess the results of the outlier detection, or how to adapt the detection criteria to your needs.
Terminology: In the documentation of this package, we refer to a *module* as a self-contained element of a report that investigates one particular quality metric. A module consists of a figure and an explanatory text. It may also contain a data structure (an object of class *outlierDetection*) that marks a subset of the arrays in the dataset as outliers according to the particular metric investigated in the module. We refer to a *section* as a collection of one or more modules that are thematically related.
For the following examples, let us load the needed packages and some data.

```
> library("arrayQualityMetrics")
> library("ALLMLL")
> data("MLL.A")
```

2 Data preparation

Some of the computations that are needed for the modules are common between several modules, and thus we perform them once, beforehand. These functions are called `prepdata` and `prepaffy`, and we refer to their manual page for details. For example,

```
> preparedData = prepdata(expressionset = MLL.A,
+                       intgroup = c(),
+                       do.logtransform = TRUE)
```

The arguments `intgroup` and `do.logtransform` are the same as for `arrayQualityMetrics`, but in `prepdata` they have no defaults, so we need to set them explicitely.

3 Module generating functions

The package contains a variety of functions that compute modules, and they are listed on a manual page which you can access by typing:

```
> ?aqm.boxplot
```

Here, let us create a report with only two quality metrics modules: boxplots and density plots.

```
> bo = aqm.boxplot(preparedData)
> de = aqm.density(preparedData)
> qm = list("Boxplot" = bo, "Density" = de)
```

The objects `bo` and `de` are of class *aqmReportModule*; please consult the class' manual page for more information.
If you want to create your own modules, please have a look at the code for the various existing functions for this purpose, and adapt it. The function `aqm.pca` is a good place to start.

4 Outlier detection

Some of the modules perform outlier detection. For instance, in the default report produced by `arrayQualityMetrics`, the module headed *Boxplots* is followed by one headed *Outlier detection for Boxplots*. In the corresponding *aqmReportModule* object, this is reflected by a non-trivial value for the slot named `outliers`:

```
> bo@outliers
```

```
An object of class "outlierDetection"
Slot "statistic":
   JD-ALD009-v5-U133A.CEL    JD-ALD051-v5-U133A.CEL    JD-ALD052-v5-U133A.CEL
                0.0673025                 0.1011325                 0.1569350
   JD-ALD057-v5-U133A.CEL    JD-ALD078-v5-U133A.CEL    JD-ALD180-v5-U133A.CEL
                0.0597350                 0.1181025                 0.2325025
JD-ALD226-NA-v5-U133A.CEL    JD-ALD232-v5-U133A.CEL    JD-ALD237-v5-U133A.CEL
```

```
                0.2221800                 0.0942600                 0.1493875
   JD-ALD244-v5-U133A.CEL    JD-ALD294-v5-U133A.CEL    JD-ALD380-v5-U133A.CEL
                0.1306675                 0.3751150                 0.0500425
   JD-ALD381-v5-U133A.CEL    JD-ALD384-v5-U133A.CEL    JD-ALD385-v5-U133A.CEL
                0.0950500                 0.4391800                 0.1652275
   JD-ALD420-v5-U133A.CEL    JD-ALD421-v5-U133A.CEL    JD-ALD431-v5-U133A.CEL
                0.1007775                 0.1239300                 0.1933900
   JD-ALD433-v5-U133A.CEL    JD-ALD520-v5-U133A.CEL
                0.1489500                 0.0600150

Slot "threshold":
JD-ALD385-v5-U133A.CEL
            0.3062894

Slot "which":
JD-ALD294-v5-U133A.CEL JD-ALD384-v5-U133A.CEL
                    11                     14

Slot "description":
[1] "Kolmogorov-Smirnov statistic <i>K<sub>a</sub></i>"
[2] "data-driven"
```

The slot named `statistic` contains, for each array, a single number based on which outlier detection is performed. For instance, in the case of `bo`, the slot `bo@outliers@statistic` is the Kolmogorov-Smirnov statistic for the comparison between each array's intensity distribution and the distribution of the pooled data. The slot `threshold` contains the threshold against which the valuess of `statistic` were compared. Arrays with a value of `statistic` greather than `threshold` are called outliers. Their indices are listed in the vector `which`. Finally, the slot `description` contains a textual description of the definition of `statistic` and indicates how the `threshold` was chosen.
The actual details of outlier detection are different for each report module, and are documented in the figure caption of the report module. For more information, please look at the code of the report module generating function of interest – for instance, at the first few lines of the `boxplot` function. The code there uses the helper functions `outliers` and `boxplotOutliers`, which are documented in their manual pages.

5 Rendering the report

A report is rendered by calling the function `aqm.writereport` with a list of *aqmReportModule* objects.

```
> outdir = tempdir()
> aqm.writereport(modules = qm, reporttitle = "My example", outdir = outdir,
+                 arrayTable = pData(MLL.A))
> outdir

[1] "/tmp/RtmptuzBPz"
```

Point your browser to the `index.html` file in that directory.

Session Info

```
> toLatex(sessionInfo())
```

- R version 2.13.0 (2011-04-13), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=C`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: ALLMLL 1.2.8, Biobase 2.12.0, affy 1.30.0, arrayQualityMetrics 3.8.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.14.0, Biostrings 2.20.0, Cairo 1.4-9, DBI 0.2-5, Hmisc 3.8-3, IRanges 1.10.0, RColorBrewer 1.0-2, RSQLite 0.9-4, SVGAnnotation 0.9-0, XML 3.2-0, affyPLM 1.28.0, affyio 1.20.0, annotate 1.30.0, beadarray 2.2.0, cluster 1.13.3, genefilter 1.34.0, grid 2.13.0, hwriter 1.3, lattice 0.19-23, latticeExtra 0.6-14, limma 3.8.0, preprocessCore 1.14.0, setRNG 2009.11-1, splines 2.13.0, survival 2.36-5, tools 2.13.0, vsn 3.20.0, xtable 1.5-6