Quality assessment with arrayQualityMetrics

Audrey Kauffmann, Wolfgang Huber

August 7, 2008

Introduction

The function arrayQualityMetrics can be used on AffyBatch for Affymetrix data sets, ExpressionSet in the case of non Affymetrix one colour experiments, NChannelSet for dual colour experiments and BeadLevelList for Illumina bead arrays. arrayQualityMetrics produces a HTML report as an output.

1 Data preparation

1.1 Data import

arrayQualityMetrics can be used on unnormalized or normalized data sets. In this example, we will produce a report on a normalized NChannelSet. We first need to load the data.

```
> library("Biobase")
> library("CCl4")
> data("CCl4")
```

1.2 Normalization

We can normalize the data using the variance stabilization method available in the package vsn.

```
> library("vsn")
> nCCl4 = justvsn(CCl4, subsample=2000)
```

1.3 Data specificities

Some of the quality metrics provided by the package are performed using specific information about the features of the arrays. For an optimal use of the package, the data can be prepared accordingly to the following conventions.

X and Y coordinates of the spots To plot the images of the arrays in the case of ExpressionSet and NChannelSet, arrayQualityMetrics needs the coordinates of the spots on the chip. Two columns corresponding to the row and column numbers of the features are thus required in the featureData. These columns should be named "X" for rows and "Y" for columns. If the arrays are split into blocks, then the function addXYfromGAL should be executed prior to arrayQualityMetrics to convert the rows and columns of the blocks in absolute "X" and "Y" on the array.

```
> featureData(nCCl4)$X = featureData(nCCl4)$Row
> featureData(nCCl4)$Y = featureData(nCCl4)$Column
```

GC content of the reporters If the GC content of the reporters is known, then it is possible to include it as percentages in the featureData of the NChannelSet under the column name "GC". Then a study of the GC content effect on intensities of the arrays can be performed.
Mapping of the reporters  The report can also include a study of the effect of the target mapping of the reporters. Thus a featureData column named "HasTarget" should include logical "TRUE" if the reporter matches for a coding mRNA and "FALSE" if not.

```r
> featureData(nCCl4)$hasTarget = (regexpr("NM", featureData(nCCl4)$Name) > 0)
```

Covariate  The phenoData can also have a column containing a group name (up to 12 groups) for each sample accordingly to a factor of interest. This phenoData will be used to provide side bars on the heatmap. In the case of the CCl4 dataset, the RNA hybridized to the arrays can be of good, medium or poor quality accordingly to its RIN number (see CCl4 vignette).

```r
> datapath = system.file("extdata", package="CCl4")
> p = read.AnnotatedDataFrame("samplesInfo.txt", path=datapath)
> cond = paste(p$RIN.Cy3,p$RIN.Cy5,sep="/")
> poor = grep(cond,pattern="2.5")
> medium = grep(cond,pattern="^5/|/5")
> good = grep(cond,pattern="9.7")
> cov = rep(0, length = nrow(p))
> cov[good] = "Good"
> cov[medium] = "Medium"
> cov[poor] = "Poor"
> phenoData(nCCl4)$RNAintegrity = cov
```

2  Report production

To produce a report, the function arrayQualityMetrics is called with the following arguments:

- **expressionset**: is an object of class ExpressionSet, AffyBatch, NChannelSet or BeadLevelList.
- **outdir**: is the directory in which the result files are created.
- **force**: if TRUE, if outdir already exists, it will be overwritten.
- **do.logtransform**: if TRUE, the data are log transformed before the analysis.
- **split.plots**: if the number of studied arrays is more than 50 it is advised to define a number of experiments to represent on the density plots.
- **intgroup**: is the name of the column in the phenoData that contains the information about a covariate of interest to be shown as a side bar on the heatmap. The default name to this column is 'Covariate'.

```r
> library("arrayQualityMetrics")
> arrayQualityMetrics(expressionset = nCCl4,
+  outdir = "CCl4",
+  force = TRUE,
+  do.logtransform = FALSE,
+  split.plots = FALSE,
+  intgroup = "RNAintegrity")
```

A report named QMreport.html is produced in the subdirectory CCl4. It contains text illustrated by .png files. Each .png is linked to corresponding .pdf files in order to provide high quality images.
Session Info
> `toLatex(sessionInfo())`

- R version 2.7.1 (2008-06-23), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=C;LC_MESSAGES=en_US;LC_PAPER=en_US;LC_NAME=C;LC_ADDRESS=C;LC_TELEPHONE=C;LC_MEASUREMENT=en_US;LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, grid, methods, splines, stats, tools, utils
- Other packages: affy 1.18.2, affyio 1.8.1, affyPLM 1.16.0, annotate 1.18.0, AnnotationDbi 1.2.2, arrayQualityMetrics 1.6.1, beadarray 1.8.0, Biobase 2.0.1, CCl4 1.0.7, DBI 0.2-4, germa 2.12.1, genefilter 1.20.0, geneplotter 1.18.0, lattice 0.17-12, latticeExtra 0.5-1, limma 2.14.5, matchprobes 1.12.0, preprocessCore 1.2.1, RColorBrewer 1.0-2, RSQLite 0.6-9, simpleaffy 2.16.0, survival 2.34-1, vsn 3.6.0, xtable 1.5-2
- Loaded via a namespace (and not attached): KernSmooth 2.22-22