

# Textual description of affy

Laurent Gautier, Rafael Irizarry, Leslie Cope, and Ben Bolstad

February 6, 2004

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Changes in Version 1.3.x</b>	<b>3</b>
<b>3</b>	<b>Getting Started: From probe level data to expression values</b>	<b>3</b>
3.1	Quick start . . . . .	4
3.2	Reading CEL file information . . . . .	5
3.3	Expression measures . . . . .	9
3.3.1	expresso . . . . .	10
3.3.2	MAS 5.0 . . . . .	11
3.3.3	Li and Wong's MBEI (dchip) . . . . .	11
3.3.4	C implementation of RMA . . . . .	12
<b>4</b>	<b>Quality Control through Data Exploration</b>	<b>13</b>
4.1	Accessing <i>PM</i> and <i>MM</i> Data . . . . .	14
4.2	Histograms, Images, and Boxplots . . . . .	15
4.3	RNA degradation plots . . . . .	19
<b>5</b>	<b>Normalization</b>	<b>21</b>
<b>6</b>	<b>Classes</b>	<b>21</b>
6.1	AffyBatch . . . . .	21
6.2	ProbeSet . . . . .	22
<b>7</b>	<b>Location to ProbeSet Mapping</b>	<b>23</b>
<b>8</b>	<b>Configuring the package options</b>	<b>28</b>
<b>9</b>	<b>Where can I get more information?</b>	<b>29</b>
<b>A</b>	<b>Previous Release Notes</b>	<b>29</b>

<b>B Changes in Version 1.2.x</b>	<b>29</b>
B.1 Changes in Version 1.1.x . . . . .	30

# 1 Introduction

The *affy* package is part of the Bioconductor<sup>1</sup> project. It is meant to be an extensible, interactive environment for data analysis and exploration of Affymetrix oligonucleotide array probe level data.

The software utilities provided with the Affymetrix software suite summarizes the probe set intensities to form one *expression measure* for each gene. The expression measure is the data available for analysis. However, as pointed out by Li and Wong (2001), much can be learned from studying the individual probe intensities, or as we call them, the *probe level data*. This is why we developed this package. The package includes plotting functions for the probe level data useful for quality control, RNA degradation assessments, different probe level normalization and background correction procedures, and flexible functions that permit the user to convert probe level data to expression measures. The package includes utilities for computing expression measures similar to MAS 4.0's AvDiff (Affymetrix, 1999), MAS 5.0's signal (Affymetrix, 2001), DChip's MBEI (Li and Wong, 2001), and RMA (Irizarry et al., 2003b).

We assume that the reader is already familiar with oligonucleotide arrays and with the design of the Affymetrix GeneChip arrays. If you are not, we recommend the Appendix of the Affymetrix MAS manual Affymetrix (1999, 2001).

The following terms are used throughout this document:

**probe** oligonucleotides of 25 base pair length used to probe RNA targets.

**perfect match** probes intended to match perfectly the target sequence.

*PM* intensity value read from the perfect matches.

**mismatch** the probes having one base mismatch with the target sequence intended to account for non-specific binding.

*MM* intensity value read from the mis-matches.

**probe pair** a unit composed of a perfect match and its mismatch.

**affyID** an identification for a probe set (which can be a gene or a fraction of a gene) represented on the array.

**probe pair set** *PMs* and *MMs* related to a common *affyID*.

**CEL files** contain measured intensities and locations for an array that has been hybridized.

---

<sup>1</sup><http://www.bioconductor.org/>

*CDF* file contain the information relating probe pair sets to locations on the array.

Section 2 describes the main differences between version 1.2 and this version (1.3). Section 3 describes a quick way of getting started and getting expression measures. Section 4 describes some quality control tools. Section 5 describes normalization routines. Section 6 describes the different classes in the package. 7 describes our strategy to map probe locations to probe set membership. Section 8 describes how to change the package's default options. Section ?? describes earlier changes.

**Note:** If you use this package please cite [Gautier et al. \(2003\)](#) and/or [Irizarry et al. \(2003a\)](#).

## 2 Changes in Version 1.3.x

What's new?

- `mas5calls` method added to get Affymetrix's P/M/A calls.
- `Cel` and `Cdf` classes no longer supported. Function, `read.celfile` and other `Cel` related methods and functions removed. Most `Cdf` related functions have moved to the *makecdfenv* package.
- Big speed and memory improvement of `ReadAffy`, `read.affybatch`, and `justRMA`.
- Function `read.probematrix` added. It reads CEL files and returns a matrix of PM, MM, or both. This function is more memory efficient than `read.affybatch`.
- Package no longer depends on `affydata` package. For this reason some examples have been moved from this vignette to the `affydata` vignette.
- The previously deprecated `express` function has been completely removed.
- Most normalization routines for `AffyBatches` can now be called with the parameter `type` which specifies whether the normalization should be applied as a PM-only, MM-only, both PM and MM together or PM and MM separately.

There are some minor differences in what you can do but little functionality has disappeared. Memory efficiency and speed have improved.

## 3 Getting Started: From probe level data to expression values

The first thing you need to do is **load the package**.

```
R> library(affy) ##load the affy package
```

This release of the *affy* package will automatically download the appropriate cdf environment when you require it. However, if you wish you may download and install the cdf environment you need from <http://www.bioconductor.org/data/metaData.html> manually. If there is no cdf environment currently built for your particular chip and you have access to the CDF file then you may use the *makecdfenv* package to create one yourself. To make the cdf packages, Microsoft Windows users will need to use the tools described here: <http://www.stats.ox.ac.uk/pub/R/rw-FAQ.html>.

### 3.1 Quick start

If all you want is to go from probe level data (*Cel* files) to expression measures here are some quick ways.

If you want is RMA, the quickest way of reading in data and getting expression measures is the following:

1. Create a directory, move all the relevant *CEL* files to that directory
2. If using linux/unix, start R in that directory.
3. If using the Rgui for Microsoft Windows make sure your working directory contains the *Cel* files (use “File -> Change Dir” menu item).
4. Load the library.

```
R> library(affy) ##load the affy package
```

5. Read in the data and create an expression, using RMA for example.

```
R> Data <- ReadAffy() ##read data in working directory
R> eset <- rma(Data)
```

Depending on the size of your dataset and on the memory available to your system, you might experience errors like ‘Cannot allocate vector ...’. An obvious option is to increase the memory available to your R process (by adding memory and/or closing external applications<sup>2</sup>). Another option is to use the function *justRMA*.

```
R> eset <- justRMA(Data)
```

This reads the data and performs the ‘RMA’ way to preprocess them at the *C* level. One does not need to call *ReadAffy*, probe level data is never stored in an *AffyBatch*. *rma* continues to be the recommended function for computing RMA.

The *rma* function was written in C for speed and efficiency. It uses the expression measure described in [Irizarry et al. \(2003b\)](#).

---

<sup>2</sup>UNIX-like systems users might also want to check *ulimit* and/or compile **R** and the package for 64 bits when possible.

For other popular methods use **expresso** instead of **rma** (see Section 3.3.1). For example for our version of MAS 5.0 signal uses **expresso** (see code). To get mas 5.0 you can use

```
R> eset <- mas5(Data)
```

which will also normalize the expression values. The normalization can be turned off through the **normalize** argument.

In all the above examples, the variable **eset** is an object of class **exprSet** described in the Biobase vignette. Many of the packages in Bioconductor work on objects of this class. See the *genefilter* and *geneplotter* packages for some examples.

If you want to use some other analysis package you can write out the expression values to file using the following command:

```
R> write.exprs(eset, file="mydata.txt")
```

or if on Microsoft Windows and interested in reading your data into excel

```
R> exprs2excel(eset, file="mydata.csv")
```

## 3.2 Reading CEL file information

The function **ReadAffy** is quite flexible. It lets you specify the filenames, phenotype, and MIAME information. You can enter them by reading files (see the help file) or widgets (you need to have the **tkWidgets** package installed and working)

```
R> Data <- ReadAffy(widget=TRUE) ##read data in working directory
```

This function call will pop-up a file browser widget, see Figure 1, that provides an easy way of choosing cel files.

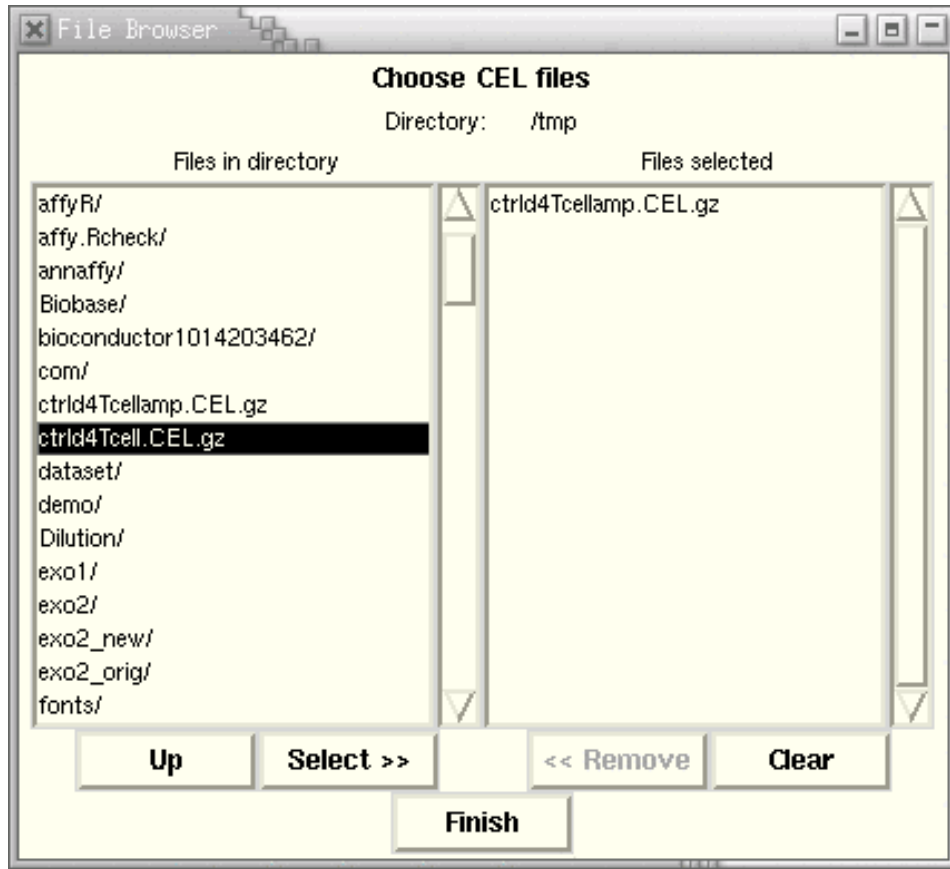


Figure 1: Graphical display for selecting *CEL* files. This widget is part of the *tkWidgets* package.

Next, a widget (not shown) permits the user to enter sample names for the arrays and descriptions for these samples. In the next step a series of three widgets permits the user to enter phenotype or covariate information. See Figure 2.

Finally the a widget is presented for the user to enter MIAME information. Seen in Figure 3.

Notice that it is not necessary to use widgets to enter this information. Please read the help file for more information on how to read it from flat files or to enter it programmatically.

The function `ReadAffy` is a wrapper for the functions `read.affybatch`, `tkSampleNames`, `read.phenoData`, and `read.MIAME`. The function `read.affybatch` has some nice feature that make it quite flexible. For example, the `compression` argument permit the user to read compressed *CEL* files. The argument `compress` set to `TRUE` will inform the readers that your files are compressed and let you read them while they remain compressed. The compression formats *zip* and *gzip* are known to be recognized.

A comprehensive description of all these options is found in the help file:

The figure shows three stacked Tkinter windows for entering phenodata. The first window, titled 'Number of Covariates', has a label 'Enter the Number Of Covariates' and a text input field. The second window, titled 'Covariate Names', has a table with two columns: 'Covariate Names' and 'Description'. It contains two rows: 'Cov 1' with 'treated' and 'drug Z', and 'Cov 2' with 'gender'. The third window, titled 'Pheno Data', has a table with two columns: 'Sample Names' and two sub-columns 'treated' and 'gender'. It contains two rows: 'Array 1' with 'file1', 'TRUE', and 'MALE', and 'Array 2' with 'file2', 'FALSE', and 'MALE'. Each window has a 'Continue' button, and the last window has a 'Finish' button.

**Number of Covariates**

Enter the Number Of Covariates

Continue

**Covariate Names**

	Covariate Names	Description
Cov 1	treated	drug Z
Cov 2	gender	

Back Continue

**Pheno Data**

	Sample Names	treated	gender
Array 1	file1	TRUE	MALE
Array 2	file2	FALSE	MALE

Back Finish

Figure 2: Graphical display for entering phenoData This widget is part of the *tkWidgets* package.

The image shows a window titled "MIAME Information" with a standard Mac OS-style title bar (close, zoom, and scroll buttons). The window contains the following elements:

- Experimenter's Name:** A single-line text input field.
- Laboratory:** A single-line text input field.
- Contact Information:** A multi-line text input area.
- Experiment Title:** A single-line text input field.
- Experiment Description:** A large multi-line text input area.
- URL:** A single-line text input field.
- Exit:** A button located in the bottom right corner of the window.

Figure 3: Graphical display for entering *MIAME* informations. This widget is part of the *tkWidgets* package.

```
R> ?read.affybatch
R> ?read.phenoData
R> ?read.MIAME
```

### 3.3 Expression measures

The most common operation is certainly to convert probe level data to expression values. Typically this is achieved through the following sequence:

1. reading in probe level data.
2. background correction.
3. normalization.
4. probe specific background correction, e.g. subtracting *MM*.
5. summarizing the probe set values into one expression measure and, in some cases, a standard error for this summary.

We detail what we believe is a good way to proceed below. As mentioned the function `expresso` provides many options. For example,

```
R> eset <- expresso(affybatch, normalize.method="qspline", bg.method="rma", pmcorrect.
```

This will store expression values, in the object `eset`, as an object of class `exprSet` (see the *Biobase* package). You can either use R and the Bioconductor packages to analyze your expression data or if you rather use another package you can write it out to a tab delimited file like this

```
R> write.exprs(eset, file="mydata.txt")
```

In the `mydata.txt` file, row will represent genes and columns will represent samples/arrays. The first row will be a header describing the columns. The first column will have the *affyIDs*. The `write.exprs` function is quite flexible on what it writes (see the help file).

For users of Microsoft Windows, who wish to use Excel, the convenient function `exprs2excel` will write out a comma delimited file of expression values. You should be able to open this file by double clicking in Windows (use a `.csv` file extension).

```
R> exprs2excel(eset, file="mydata.csv")
```

### 3.3.1 `expresso`

The function `expresso` performs the steps background correction, normalization, probe specific correction, and summary value computation. We now show this using an `AffyBatch` included in the package for examples. The command `data(affybatch.example)` is used to load these data.

Important parameters for the `expresso` function are:

**`bgcorrect.method`** . The background correction method to use. The available methods are

```
> bgcorrect.methods  
  
[1] "mas" "none" "rma" "rma2"
```

**`normalize.method`** . The normalization method to use. The available methods can be queried by using `normalize.methods`.

```
> data(affybatch.example)  
> normalize.methods(affybatch.example)  
  
[1] "constant"          "contrasts"          "invariantset"       "loess"  
[5] "qspline"           "quantiles"          "quantiles.robust"
```

**`pmcorrect.method`** The method for probe specific correction. The available methods are

```
> pmcorrect.methods  
  
[1] "mas"          "pmonly"       "subtractmm"
```

**`summary.method`** . The summary method to use. The available methods are

```
> express.summary.stat.methods  
  
[1] "avgdiff"      "liwong"        "mas"           "medianpolish"  "playerout"
```

Here we use `mas` to refer to the methods described in the Affymetrix manual version 5.0.

**`widget`** Making the `widget` argument `TRUE`, will let you select missing parameters (like the normalization method, the background correction method or the summary method). Figure 4 shows the widget for the selection of preprocessing methods for each of the steps.

```
R> expresso(affybatch.example, widget=TRUE)
```

There is a separate vignette **`affy: Built-in Processing Methods`** which explains in more detail what each of the preprocessing options does.

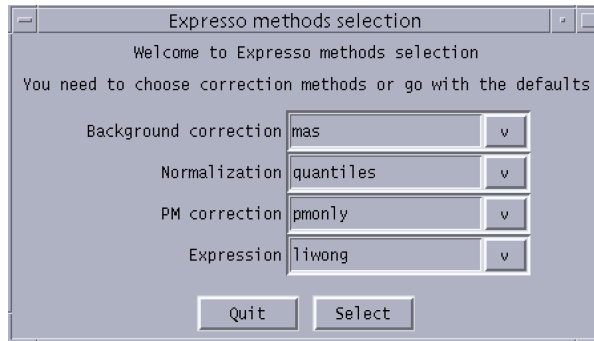


Figure 4: Graphical display for selecting expresso methods.

### 3.3.2 MAS 5.0

To obtain expression values that correspond to those from MAS 5.0, use `mas5`, which wraps `expresso` and `affy.scalevalue.exprSet`.

```
> eset <- mas5(affybatch.example)
```

```
background correction: mas
PM/MM correction : mas
expression values: mas
background correcting...done.
150 ids to be processed
.....
```

A detailed comparison between the MAS 5.0 values that are computed by *affy* and by *Affymetrix's* software can be found at <http://stat-www.berkeley.edu/~bolstad/MAS5diff/Mas5difference.html>.

To obtain MAS 5.0 presence calls you can use the `mas5calls` method.

```
> Calls <- mas5calls(affybatch.example)
```

```
150 ids to be processed
.....
```

This returns an `exprSet` with P/M/A calls in the `exprs` slot and the wilcoxon p-values in the `se.exprs` slot.

### 3.3.3 Li and Wong's MBEI (dchip)

To obtain our version of Li and Wong's MBEI one can use

```
R> eset <- expresso(affybatch.example, normalize.method="invariantset",  
                    bg.correct=FALSE,  
                    pmcorrect.method="pmonly",summary.method="liwong")
```

This gives the current *PM*-only default. The reduced model (previous default) can be obtained using `pmcorrect.method="subtractmm"`.

### 3.3.4 C implementation of RMA

One of the quickest ways to compute expression using the *affy* package is to use the `rma` function. We have found that this method allows a user to compute the RMA expression measure in a matter of minutes for datasets that may have taken hours in previous versions of *affy*. The function serves as an interface to a hard coded C implementation of the RMA method ([Irizarry et al., 2003b](#)). Generally, the following would be sufficient to compute RMA expression measures:

```
> eset <- rma(affybatch.example)
```

```
Background correcting  
Normalizing  
Calculating Expression
```

Currently the `rma` function implements RMA in the following manner

1. Probe specific correction of the PM probes using a model based on observed intensity being the sum of signal and noise
2. Normalization of corrected PM probes using quantile normalization ([Bolstad et al., 2003](#))
3. Calculation of Expression measure using median polish.

The `rma` function is likely to be improved and extended in the future as the RMA method is fine-tuned.

## 4 Quality Control through Data Exploration

Several of the functions for plotting summarized probe level data are useful for diagnosing problems with the data. The plotting functions `boxplot` and `hist` have methods for `AffyBatch` objects. Each of these functions presents side-by-side graphical summaries of intensity information from each array. Important differences in the distribution of intensities are often evident in these plots. The function `mva.pairs` (applied, for example, to `pm(affybatch.example)`), offers pairwise graphical comparison of intensity data. These plots can be particularly useful in diagnosing problems in replicate sets of arrays.

For the users convenience we have included a sample data set:

```
> data(affybatch.example)
```

```
> affybatch.example
```

```
AffyBatch object
size of arrays=100x100 features (237 kb)
cdf=cdfenv.example (150 affyids)
number of samples=3
number of genes=150
annotation=
```

This will create the `affybatch.example` object of class `AffyBatch`. `print` (or `show`) will display summary information. These objects represent data from one experiment. The `AffyBatch` class combines the information of various *CEL* files with a common *CDF* file. This class is designed to keep information of one experiment. The probe level data is contained in this object.

The data in `affybatch.example` is a small sample of probe sets from 2 sets of duplicate arrays hybridized with different concentrations of the same RNA. This information is part of the `AffyBatch` and can be accessed with the `phenoData` and `pData` methods:

```
> phenoData(affybatch.example)
```

```
phenoData object with 1 variables and 3 cases
varLabels
      sample: arbitrary numbering
```

```
> pData(affybatch.example)
```

```
      sample
20A      1
20B      2
10A      3
```

## 4.1 Accessing *PM* and *MM* Data

The *PM* and *MM* intensities and corresponding *affyID* can be accessed with the `pm`, `mm`, and `probeNames` methods. These will be matrices with rows representing probe pairs and columns representing arrays. The gene name associated with the probe pair in row *i* can be found in the *i*th entry of the vector returned by `probeNames`.

```
> Index <- c(1, 2, 3, 100, 1000, 2000)
> pm(affybatch.example)[Index, ]
```

	20A	20B	10A
A28102_at1	149.0	118.0	124.0
A28102_at2	143.5	124.8	116.5
A28102_at3	132.0	111.0	105.0
AB002318_at4	122.3	90.5	111.3
D31815_at8	121.0	89.3	98.0
D87024_at16	120.8	80.3	103.3

```
> mm(affybatch.example)[Index, ]
```

	20A	20B	10A
A28102_at1	847.0	694.0	999.0
A28102_at2	860.3	667.3	1084.8
A28102_at3	815.3	650.0	1057.0
AB002318_at4	847.0	615.0	842.0
D31815_at8	206.0	95.3	154.3
D87024_at16	120.0	86.5	105.0

```
> probeNames(affybatch.example)[Index]
```

```
[1] "A28102_at" "A28102_at" "A28102_at" "AB002318_at" "D31815_at"
[6] "D87024_at"
```

`Index` contains six arbitrary probe positions.

Notice that the column names of *PM* and *MM* matrices are the sample names and the row names are the *affyID*, e.g. AB000114\_at and AB000115\_at together with the probe number (related to position in the target sequence).

```
> sampleNames(affybatch.example)
```

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

**Quick example:** To see what percentage of the *MM* are larger than the *PM* simply type

```
> mean(mm(affybatch.example) > pm(affybatch.example))

[1] 0.5425
```

The `pm` and `mm` functions can be used to extract specific probe set intensities.

```
> gn <- geneNames(affybatch.example)
> pm(affybatch.example, gn[100])
```

	20A	20B	10A
D78156_at1	223.3	148.0	203.0
D78156_at2	149.8	155.5	131.0
D78156_at3	147.3	133.8	145.0
D78156_at4	162.3	131.8	134.0
D78156_at5	459.3	451.0	345.0
D78156_at6	711.0	526.8	601.8
D78156_at7	158.8	142.8	109.0
D78156_at8	219.0	211.3	167.0
D78156_at9	196.0	251.0	182.3
D78156_at10	1715.0	1291.8	1341.3
D78156_at11	710.5	506.3	553.0
D78156_at12	438.0	346.0	282.8
D78156_at13	439.0	311.3	387.0
D78156_at14	114.8	78.0	101.5
D78156_at15	114.0	87.0	89.5
D78156_at16	181.0	122.0	160.0

The method `geneNames` extracts the unique *affyIDs*. Also notice that the 100th probe set is different from the 100th probe! The 100th probe is not part of the the 100th probe set.

The methods `boxplot`, `hist`, and `image` are useful for quality control. Figure 5 shows kernel density estimates (rather than histograms) of *PM* intensities for the 1st and 2nd array of the `affybatch.example` also included in the package

## 4.2 Histograms, Images, and Boxplots

As seen in the previous example, the sub-setting method `[]` can be used to extract specific arrays. **NOTE: Sub-setting is different in this version. One can no longer subset by gene. We can only define subsets by one dimension: the columns, i.e. the arrays. Because the `Cel` class is no longer available `[[` is no longer available.**

The method `image()` can be used to detect spatial artifacts. By default we look at log transformed intensities. This can be changed through the `transfo` argument.

```
> hist(affybatch.example[1:2])
```

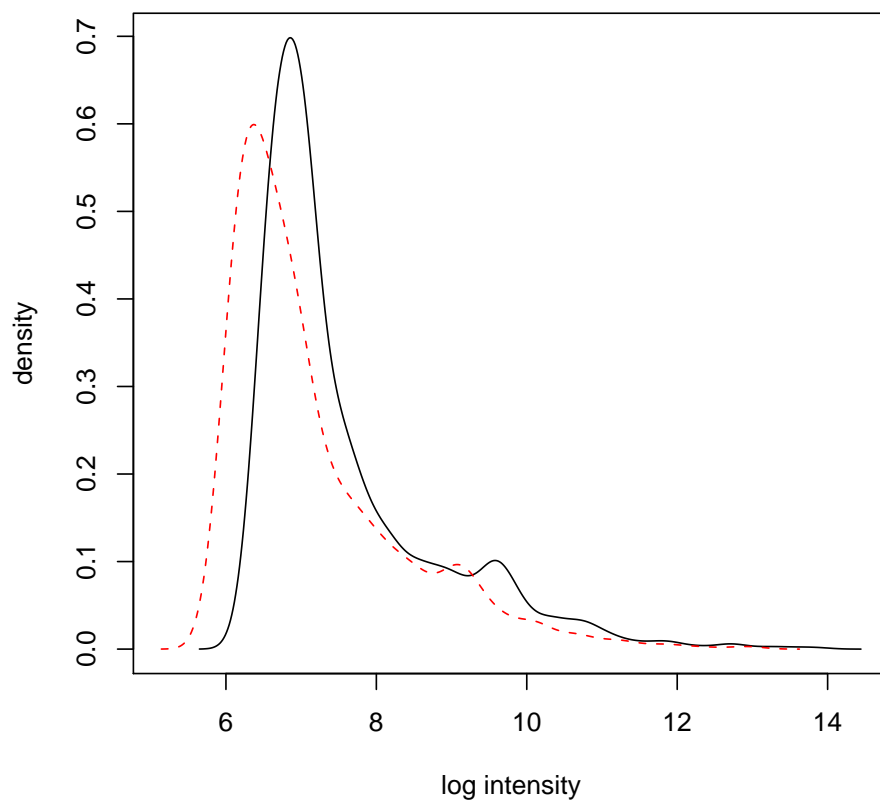


Figure 5: Histogram of  $PM$  intensities for 1st and 2nd array

```
> par(mfrow = c(2, 2))  
> image(affybatch.example)
```

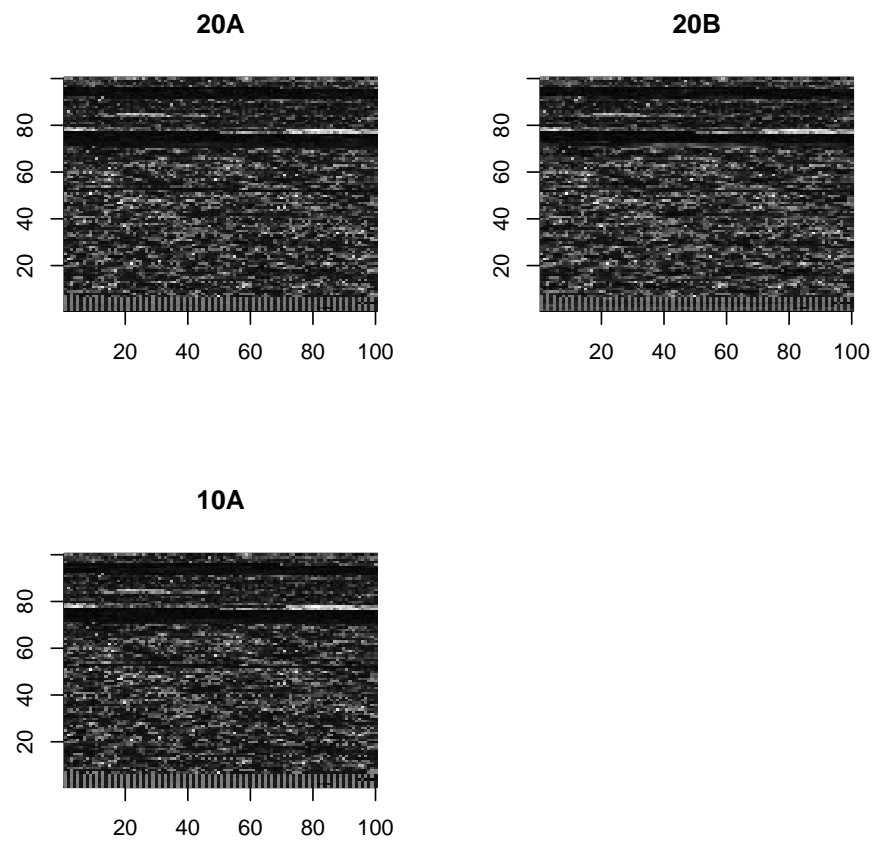


Figure 6: Image of the log intensities.

```
> par(mfrow = c(1, 1))  
> boxplot(affybatch.example, col = c(2, 3, 4))
```

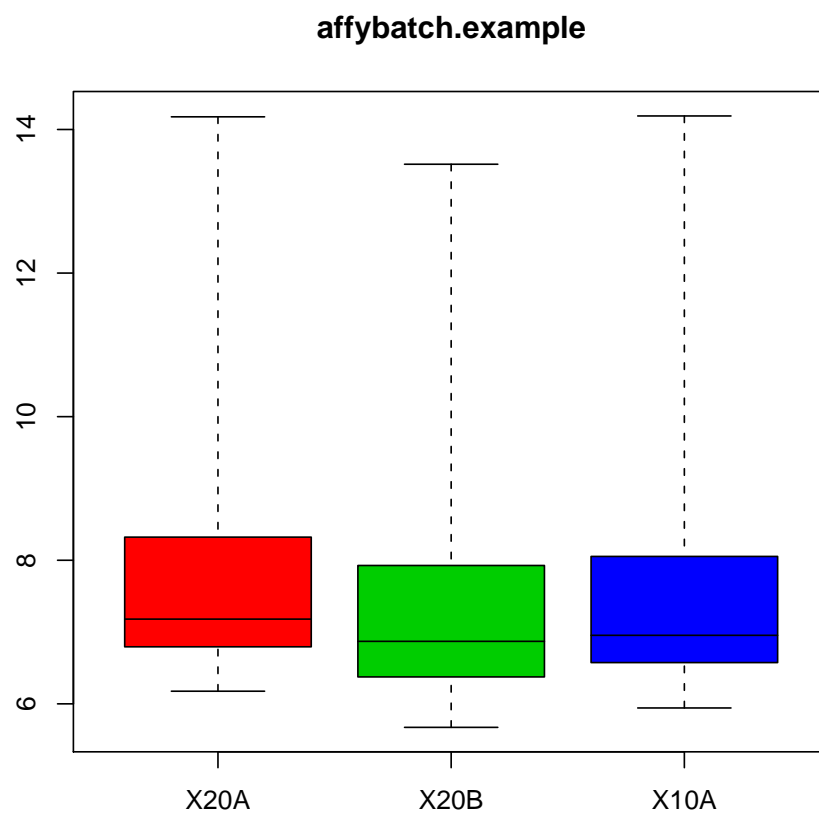


Figure 7: Boxplot of arrays in affybatch.example data.

These images are quite useful for quality control. We recommend examining these images as a first step in data exploration.

The method `boxplot` can be used to show  $PM$ ,  $MM$  or both intensities. As discussed in the next section this plot shows that we need to normalize these arrays.

### 4.3 RNA degradation plots

The functions `AffyRNAdeg`, `summaryAffyRNAdeg`, and `plotAffyRNAdeg` aid in assessment of RNA quality. Individual probes in a probeset are ordered by location relative to the 5' end of the targeted RNA molecule. [Affymetrix \(1999\)](#) Since RNA degradation typically starts from the 5' end of the molecule, we would expect probe intensities to be systematically lowered at that end of a probeset when compared to the 3' end. On each chip, probe intensities are averaged by location in probeset, with the average taken over probesets. The function `plotAffyRNAdeg` produces a side-by-side plots of these means, making it easy to notice any 5' to 3' trend. The function `summaryAffyRNAdeg` produces a single summary statistic for each array in the batch, offering a convenient measure of the severity of degradation and significance level. For an example

```
> deg <- AffyRNAdeg(affybatch.example)
> names(deg)

[1] "N"                "sample.names"    "means.by.number" "ses"
[5] "slope"            "pvalue"
```

does the degradation analysis and returns a list with various components. A summary can be obtained using

```
> summaryAffyRNAdeg(deg)

X20A X20B X10A
slope 0.0767 0.063 0.0842
pvalue 0.136 0.212 0.0911
```

Finally a plot can be created using `plotAffyRNAdeg`, see [Figure 8](#).

```
> plotAffyRNAdeg(deg)
```

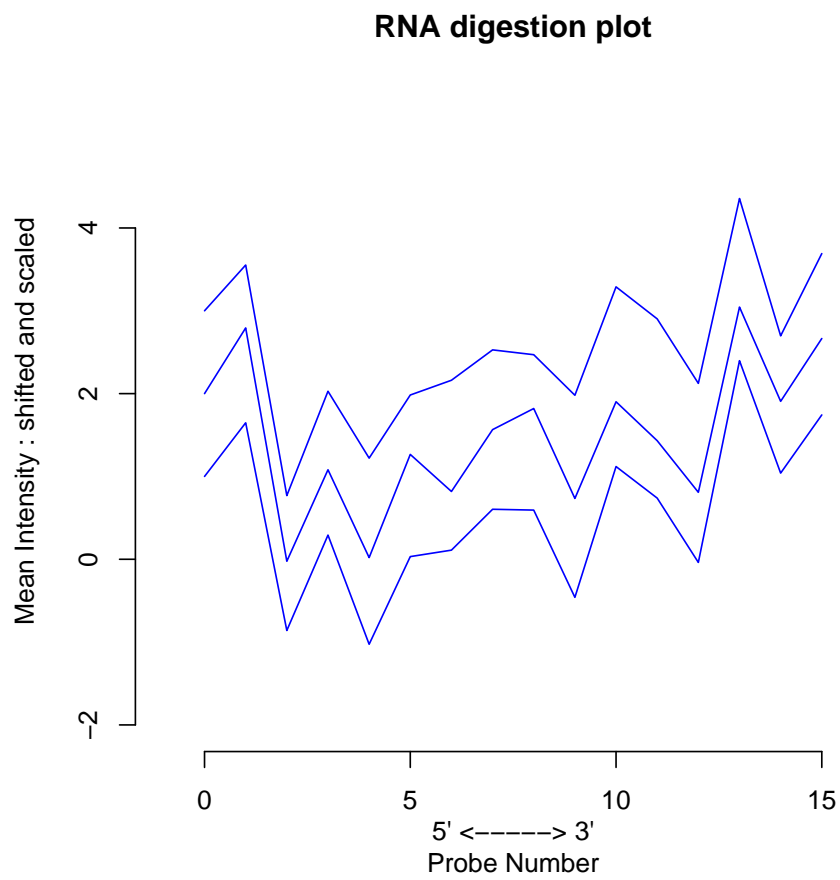


Figure 8: Side-by-side plot produced by `plotAffyRNAdeg`.

## 5 Normalization

Various researchers have pointed out the need for normalization of Affymetrix arrays. See for example [Bolstad et al. \(2003\)](#). The method `normalize` lets one normalize at the probe level

```
> affybatch.example.normalized <- normalize(affybatch.example)
```

For an extended example on normalization please refer to the vignette in the `affydata` package.

## 6 Classes

`AffyBatch` is the main class in this package. There are three other auxiliary classes that we also describe in this Section.

### 6.1 AffyBatch

The `AffyBatch` class has slots to keep all the probe level information for a batch of *Cel* files, which usually represent an experiment. It also stores phenotypic and MIAME information as does the `exprSet` class in the `Biobase` package (the base package for Bioconductor). In fact, `AffyBatch` extends `exprSet`.

The `exprs` slot contains the a matrix with the columns representing the intensities read from the different arrays. The rows represent the *cel* intensities for all position on the array. The cel intensity with physical coordinates<sup>3</sup> ( $x, y$ ) will be in row

$$i = x + \text{nrow} \times (y - 1)$$

. The `ncol` and `nrow` slots contain the physical rows of the array. Notice that this is different from the dimensions of the `exprs` matrix. The number of row of the `exprs` matrix is equal to `ncol` × `nrow`. We advice the use of the functions `xy2indices` and `indices2xy` to shuttle from X/Y coordinates to indices.

For compatibility with previous versions the accessor method `intensity` exists for obtaining the `exprs` slot.

The `cdfName` slot contains the necessary information for the package to find the locations of the probes for each probe set. See Section 7 for more on this.

---

<sup>3</sup>Note that in the *.CEL* files the indexing starts at zero while it starts at 1 in the package (as indexing starts at 1 in **R**).

## 6.2 ProbeSet

The `ProbeSet` class holds the information of all the probes related to an *affyID*. The components are `pm` and `mm`.

The method `probeset` extracts probe sets from `AffyBatch` objects. It takes as arguments an `AffyBatch` object and a vector of *affyIDs* and returns a list of objects of class `ProbeSet`

```
> gn <- geneNames(affybatch.example)
> ps <- probeset(affybatch.example, gn[1:2])
> show(ps[[1]])
```

ProbeSet object:

```
id=A28102_at
pm= 16 probes x 3 chips
```

The `pm` and `mm` methods can be used to extract these matrices (see below).

This function is general in the way it defines a probe set. The default is to use the definition of a probe set given by Affymetrix in the CDF file. However, the user can define arbitrary probe sets. The argument `locations` lets the user decide the row numbers in the `intensity` that define a probe set. For example, if we are interested in redefining the `AB000114_at` and `AB000115_at` probe sets, we could do the following:

First, define the locations of the *PM* and *MM* on the array of the `AB000114_at` and `AB000115_at` probe sets

```
> mylocation <- list("AB000114_at" = cbind(pm = c(1, 2, 3), mm = c(4,
+      5, 6)), "AB000115_at" = cbind(pm = c(4, 5, 6), mm = c(1,
+      2, 3)))
```

The first column of the matrix defines the location of the *PMs* and the second column the *MMs*.

Now we are ready to extract the `ProbeSets` using the `probeset` function:

```
> ps <- probeset(affybatch.example, genenames = c("AB000114_at",
+      "AB000115_at"), locations = mylocation)
```

Now, `ps` is list of `ProbeSets`. We can see the *PMs* and *MMs* of each component using the `pm` and `mm` accessor methods.

```
> pm(ps[[1]])

      20A   20B   10A
[1,] 987.3 603.5 841.8
[2,] 127.3 202.0 118.0
[3,] 1048.8 668.0 958.0
```

```

> mm(ps[[1]])

      20A   20B 10A
[1,] 127.0 164.8 109
[2,] 1050.8 560.0 872
[3,] 130.5 99.0 105

> pm(ps[[2]])

      20A   20B 10A
[1,] 127.0 164.8 109
[2,] 1050.8 560.0 872
[3,] 130.5 99.0 105

> mm(ps[[2]])

      20A   20B   10A
[1,] 987.3 603.5 841.8
[2,] 127.3 202.0 118.0
[3,] 1048.8 668.0 958.0

```

This can be useful in situations where the user wants to determine if leaving out certain probes improves performance at the expression level. It can also be useful to combine probes from different human chips, for example by considering only probes common to both arrays.

Users can also define their own environment for probe set location mapping. More on this in Section 7.

An example of a **ProbeSet** is included in the package. A spike in data set is included in the package in the form of a list of **ProbeSets**. The help file describes the data set. Figure 9 uses this data set to demonstrate that the *MM* also detect transcript signal.

## 7 Location to ProbeSet Mapping

On Affymetrix GeneChip arrays, several probes are used to represent genes in the form of probe sets. From a *CEL* file we get for each physical location, or *cel*, (defined by  $x$  and  $y$  coordinates) an intensity. The *CEL* file also contains the name of the *CDF* file needed for the location-probe-set mapping. The *CDF* files store the probe set related to each location on the array. The computation of a summary expression values from the probe intensities requires a fast way to map an *affyid* to corresponding probes. We store this mapping information in **R** environments<sup>4</sup>. They only contain a part of the information that can be found in the *CDF* files. The *cdfenvs* are sufficient to perform the numerical

---

<sup>4</sup>Please refer to the **R** documentation to know more about environments.

```

> data(SpikeIn)
> pms <- pm(SpikeIn)
> mms <- mm(SpikeIn)
> par(mfrow = c(1, 2))
> concentrations <- matrix(as.numeric(sampleNames(SpikeIn)), 20,
+   12, byrow = TRUE)
> matplot(concentrations, pms, log = "xy", main = "PM", ylim = c(30,
+   20000))
> lines(concentrations[1, ], apply(pms, 2, mean), lwd = 3)
> matplot(concentrations, mms, log = "xy", main = "MM", ylim = c(30,
+   20000))
> lines(concentrations[1, ], apply(mms, 2, mean), lwd = 3)

```

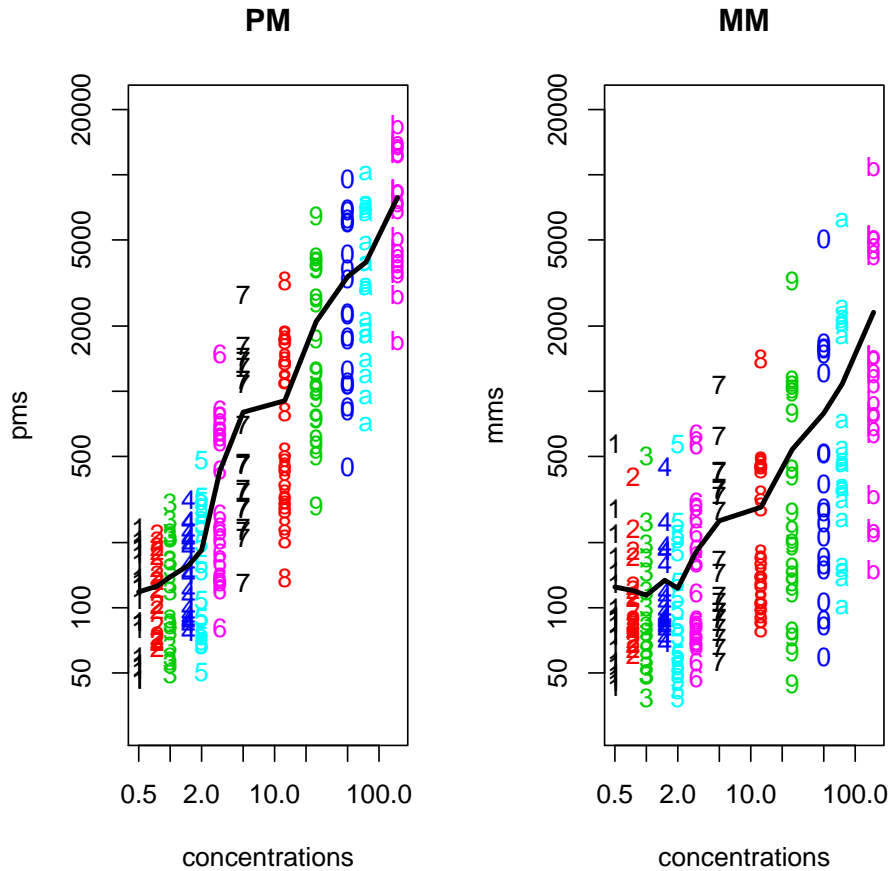


Figure 9: PM and MM intensities plotted against SpikeIn concentration

processing methods included in the package. For each *CDF* file there is package, available from <http://www.bioconductor.org/data/metaData.html>, that contains exactly one of these environments. The *cdfenvs* we store the *x* and *y* coordinates as one number (see above).

In instances of *AffyBatch*, the *cdfName* slot gives the name of the appropriate *CDF* file for arrays represented in the *intensity* slot. The functions `read.celfile`, `read.affybatch`, and `ReadAffy` extract the *CDF* filename from the *CEL* files being read. Each *CDF* file corresponds to exactly one environment. The function `cleancdfname` converts the Affymetrix given *CDF* name to a Bioconductor environment and annotation name. Here are two examples:

These give environment names:

```
> cat("HG_U95Av2 is", cleancdfname("HG_U95Av2"), "\n")
```

```
HG_U95Av2 is hgu95av2cdf
```

```
> cat("HG-133A is", cleancdfname("HG-133A"), "\n")
```

```
HG-133A is hg133acdf
```

This gives annotation name:

```
> cat("HG_U95Av2 is", cleancdfname("HG_U95Av2", addcdf = FALSE),  
+     "\n")
```

```
HG_U95Av2 is hgu95av2
```

An environment representing the corner of an Hu6800 array is available with the package. In the following, we load the environment, look at the names for the first 5 objects defined in the environment, and finally look at the first object in the environment:

```
> data(cdfenv.example)
```

```
> ls(cdfenv.example)[1:5]
```

```
[1] "A28102_at" "AB000114_at" "AB000115_at" "AB000220_at" "AB002314_at"
```

```
> get(ls(cdfenv.example)[1], cdfenv.example)
```

```
      pm  mm  
[1,] 102 203  
[2,] 104 205  
[3,] 106 207  
[4,] 108 209  
[5,] 110 211  
[6,] 112 213
```

```

[7,] 114 215
[8,] 116 217
[9,] 118 219
[10,] 120 221
[11,] 122 223
[12,] 124 225
[13,] 126 227
[14,] 128 229
[15,] 130 231
[16,] 132 233

```

The package needs to know what locations correspond to which probe sets. The `cdfName` slot contains the necessary information to find the environment with this location information. The method `getCdfInfo` takes as an argument an `AffyBatch` and returns the necessary environment. If `x` is an `AffyBatch`, this function will look for an environment with name `cleancdfname(x@cdfName)`. For example:

The call to `data` loads an `AffyBatch` containing an artificial dataset.

```
> print(affybatch.example@cdfName)
```

```
[1] "cdfenv.example"
```

```
> myenv <- getCdfInfo(affybatch.example)
```

```
> ls(myenv)[1:5]
```

```
[1] "A28102_at" "AB000114_at" "AB000115_at" "AB000220_at" "AB002314_at"
```

Notice `affybatch.example` must be loaded (see above). Now lets look at `affybatch.example`

```
> print(affybatch.example@cdfName)
```

```
[1] "cdfenv.example"
```

```
> myenv <- getCdfInfo(affybatch.example)
```

```
> ls(myenv)[1:5]
```

```
[1] "A28102_at" "AB000114_at" "AB000115_at" "AB000220_at" "AB002314_at"
```

Notice `affybatch.example` should be loaded already as above.

By default we search for the environment first in the global environment, then in a package named `cleancdfname(x@cdfName)`, and finally in the `data` directory of the `affy` package. This order can be changed through the options (see Section 8).

Various methods exist to obtain locations of probes as demonstrated in the following examples:

```
> Index <- pmindex(affybatch.example)
> names(Index)[1:2]
```

```
[1] "A28102_at"    "AB000114_at"
```

```
> Index[1:2]
```

```
$"A28102_at"
```

```
[1] 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132
```

```
$"AB000114_at"
```

```
[1] 134 136 138 140 142 144 146 148 150 152 154 156 158 160 162 164
```

pmindex returns a list with probe set names as names and locations in the components. We can also get specific probe sets:

```
> pmindex(affybatch.example, genenames = c("AB000114_at", "AB000115_at"))
```

```
$"AB000114_at"
```

```
[1] 134 136 138 140 142 144 146 148 150 152 154 156 158 160 162 164
```

```
$"AB000115_at"
```

```
[1] 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194 196
```

The locations are ordered from 5' to 3' on the target transcript. The function mmindex performs in a similar way:

```
> mmindex(affybatch.example, genenames = c("AB000114_at", "AB000115_at"))
```

```
$"AB000114_at"
```

```
[1] 235 237 239 241 243 245 247 249 251 253 255 257 259 261 263 265
```

```
$"AB000115_at"
```

```
[1] 267 269 271 273 275 277 279 281 283 285 287 289 291 293 295 297
```

They both use the method indexProbes

```
> indexProbes(affybatch.example, which = "pm")[1]
```

```
$"A28102_at"
```

```
[1] 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132
```

```
> indexProbes(affybatch.example, which = "mm")[1]
```

```
$"A28102_at"
```

```
[1] 203 205 207 209 211 213 215 217 219 221 223 225 227 229 231 233
```

```
> indexProbes(affybatch.example, which = "both")[1]
$"A28102_at"
[1] 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132 203 205 207
[20] 209 211 213 215 217 219 221 223 225 227 229 231 233
```

The `which="both"` options returns the location of the *PMs* followed by the *MMs*.

## 8 Configuring the package options

Package-wide options can be configured, as shown below through examples.

- Getting the names for the options:

```
> opt <- getOption("BioC")
> affy.opt <- opt$affy
> print(names(affy.opt))

[1] "compress.cdf"      "compress.cel"      "use.widgets"       "probesloc"
[5] "bgcorrect.method" "normalize.method"  "pmcorrect.method"  "summary.method"
```

- Default processing methods:

```
> opt <- getOption("BioC")
> affy.opt <- opt$affy
> affy.opt$normalize.method <- "constant"
> opt$affy <- affy.opt
> options(BioC = opt)
```

- Compression of files: if you are always compressing your CEL files, you might find annoying to specify it each time you call a reading function. It can be specified once for all in the options.

```
> opt <- getOption("BioC")
> affy.opt <- opt$affy
> affy.opt$compress.cel <- TRUE
> opt$affy <- affy.opt
> options(BioC = opt)
```

- Priority rule for the use of a cdf environment: The option *probesloc* is a list. Each element of the list is itself a list with two elements *what* and *where*. When looking for the information related to the locations of the probes on the array, the elements in the list will be looked at sequentially. The first one leading to the information is used (an error message is returned if none permits to find the information). The element *what* can be one of *package*, *environment*.

## 9 Where can I get more information?

There are several other vignettes addressing more specialised topics related to the **affy** package.

- **affy: Custom Processing Methods (HowTo)**: A description of how to use custom preprocessing methods with the package. This document gives examples of how you might write your own preprocessing method and use it with the package.
- **affy: Built-in Processing Methods**: A document giving fuller descriptions of each of the preprocessing methods that are available within the **affy** package.
- **affy: Import Methods (HowTo)**: A discussion of the data structures used and how you might import non standard data into the package.
- **affy: Loading Affymetrix Data (HowTo)**: A quick guide to loading Affymetrix data into R.
- **affy: Automatic downloading of cdfenvs (HowTo)**: How you can configure the automatic downloading of the appropriate *cdfenv* for your analysis.

## A Previous Release Notes

## B Changes in Version 1.2.x

What's new?

- slot 'preprocessing' of the MIAME attribute used to store normalization step information [list returned; more complex but organised structures (like a class) are under evaluation.]
- tuning of the implementations of the MAS5.0 methods (`bgcorrect.mas`, ...). <sup>5</sup>
- method `plot.ProbeSet`, an alternative to `barplot`, to plot probe level information.
- parameter 'scale' in the method `barplot` for `ProbeSet`. All the barplots are scaled to each other.
- New functions 'xy2indices' and 'indices2xy' to shuttle from x/y pos to indices (like the ones in `cdfenvs`) (and reverse).
- The documentation for normalization has been improved.

---

<sup>5</sup>A comparison between the implementations of algorithms in MAS5.0 and the ones in **affy** can be found at <http://stat-www.berkeley.edu/~bolstad/MAS5diff/Mas5difference.html>.

- Due to some new protocols `?AffyBatch` no longer will give you the help file. One needs to type `help("AffyBatch-class")`. Same is true for other classes.
- The function `justRMA` added for those who want to use `rma` and are having memory problems.

### What's different?

- Some of the large example datasets have been moved to a new package *affydata*.
- Autoload of `cdfenvs` on demand (uses `reposTools`). Can be configured through the options.
- default methods for normalization, bg correction, pm correction and summary now in the package options [options exist for all, but only used by `normalize` for the moment].
- The default background on the `rma` function has been changed. Now the results from `rma` and `expresso` should agree completely.
- The function `express` is deprecated. It still functions normally but gives warning message. It will be removed in a future release. The function `expresso` should be used as a replacement.
- bug in the parser fixed (infinite loop reported with apparently non-standard CEL files).
- bug in the parser fixed (the 'sd' data returned were not correct).
- missing slot in the dataset `SpikeIn` fixed.
- bug in `normalize.AffyBatch.qspline` fixed (thanks to people at Insightful). The expression data matrix sent to `normalize.qspline` was mistakenly transposed.
- `barplot.ProbeSet` scales plots to each others by default.

## B.1 Changes in Version 1.1.x

### What's new?

- Faster reading functions (type `?read.affybatch`)
- Widgets for reading phenotypic and MIAME information and choices of preprocessing when computing expression measures. (`?read.phenoData`, `?read.MIAME`, `?expresso`)
- No need to read in *CDF* files.

- More efficient expression measure functions. (`?expresso`, `?express`).
- Very fast RMA (`?rma`).
- Our version of MAS 5.0 available (`?expresso`).
- RNA degradation assessment. (`?AffyRNAdeg`)

**What's different?** The new version is not backwards compatible! Unfortunately the changes we had to make to gain efficiency has resulted in some lack of backwards compatibility. Here are some important ones:

- Unless you are using the HG-U95Av2 chip or the HGU133A chip you need to download and install a package for each chip type. They can be obtained from <http://www.bioconductor.org/data/metaData.html>
- You need the latest version of Biobase
- To get RMA you no longer use `express`, you use `rma`
- `ReadAffy` uses the argument `filenames` instead of `CELS` for denoting cel files.
- You can no longer subset probe level objects (now `AffyBatch`) by probe set name

The main difference between Version 1.0 and this version (1.1) is that the user no longer needs to provide the *CDF* files. We now provide a more efficient way of obtaining this information. Data packages containing the necessary CDF information can be obtained from <http://www.bioconductor.org/data/metaData.html>. Simply download as many of these *cdf environments* as you need and install them. The affy package will know where to look. If you are using the **HGU95Av2** or **HGU133A** chip the information is included in the affy package and you do not need to download further packages. You can also create your own cdf environments. See Section 7 for information on how the environments work. A cdf environment making package is available from the Bioconductor web site [www.bioconductor.org](http://www.bioconductor.org).

Version 1.1 provides a unified approach to working with probe level data. *AffyBatch* is the main class the user will manipulate. We believe it combines the simplicity of the former *Plob* with the flexibility of the former *Cel.container*. As before, it bundles the data from a *batch* of experiments. The classes *Cdf* contains the information of *CDF* file, the class *ProbeSet* contain *PM* and *MM* intensities for a particular probe set. Beginners need do not understand these classes. However, they are briefly described in Section 6.

There are some minor differences in what you can do but little functionality has disappeared.

## References

- Affymetrix. *Affymetrix Microarray Suite User Guide*. Affymetrix, Santa Clara, CA, version 4 edition, 1999. [2](#), [19](#)
- Affymetrix. *Affymetrix Microarray Suite User Guide*. Affymetrix, Santa Clara, CA, version 5 edition, 2001. [2](#)
- B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003. [12](#), [21](#)
- Laurent Gautier, Leslie Cope, Benjamin Milo Bolstad, and Rafael A. Irizarry. affy - an r package for the analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 2003. In press. [3](#)
- Rafael A. Irizarry, Laurent Gautier, and Leslie M. Cope. *The Analysis of Gene Expression Data: Methods and Software*, chapter 4. Springer Verlag, 2003a. [3](#)
- Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003b. To appear. [2](#), [4](#), [12](#)
- C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A*, 98:31–36, 2001. [2](#)