Introduction to the Codelink package

Diego Diez

October 17, 2016

1 Introduction

This package implements methods to facilitate the preprocessing and analysis of Codelink microarrays. Codelink is a microarray platform for the analysis of gene expression that uses 30 base long oligonucleotides. Codelink is currently owned by Applied Microarrays, Inc. (previously was GE Healthcare and before that Amersham). There is a proprietary software for reading scanned images, perform spot intensity quantification and diagnostics. A Codelink microarray consists of a number of species-specific probes to measure gene expression, as well as some other control probes (see Table 1). The Codelink software assigns quality flags to each spot (see Table 2) on the basis of a signal to noise ratio (SNR) computation (Eq: 1) and other morphological characteristics as irregular shape of the spots, saturation of the signal or manufacturer spots removed. By default, the software performs background correction (subtract) followed by median normalization. The results can be exported in several formats as XML, Excel, plain text, etc.

The codelink package enables loading Codelink data into R, and stores it as a CodelinkSet object. The CodelinkSet-class inherits from ExpressionSet all methods, and enables straightforward interfacing with other Bioconductor structures, and packages.

NOTE: the old Codelink-class infrastructure is maintained for backbard compatibility, and information about its use can be found in the vignette Codelink_Legacy.pdf.

<table>
<thead>
<tr>
<th>Probe type</th>
<th>Description</th>
<th>Default weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISCOVERY</td>
<td>Measure gene expression</td>
<td>1</td>
</tr>
<tr>
<td>POSITIVE</td>
<td>Positive control</td>
<td>0</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>Negative control</td>
<td>0</td>
</tr>
<tr>
<td>FIDUCIAL</td>
<td>Grid alignment</td>
<td>0</td>
</tr>
<tr>
<td>OTHER</td>
<td>Other controls and housekeeping genes</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Probe types for Codelink arrays.

<table>
<thead>
<tr>
<th>Flag</th>
<th>Description</th>
<th>Default value set</th>
<th>Default weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>Good signal (SNR ≥ 1)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>Limit signal (SNR &lt; 1)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>Saturated signal</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>Irregular shape</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>MSR spot (-9999)</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>Background contaminated</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>User excluded spots</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Quality Flag description. SNR: Signal to Noise Ratio.
\[ SNR = \frac{S_{mean}}{(B_{median} + 1.5 \times B_{stdev})} \]  

2 Reading data

Only Codelink data exported as plain text from the Codelink software is supported. Unfortunately the Codelink exported text format can have arbitrary columns and header fields so depending of what has been exported, reading it into a `CodelinkSet` object may be more or less complicated. As a rule of thumb it is recommended to include in the exported files at least `Spot_mean` and `Bkgd_median` values so that background correction and normalization can be performed within R. In addition, `Bkgd_stdev` will be needed to compute the SNR. If `Raw_intensity` or `Normalized_intensity` columns are present then it is possible to avoid background correction and/or normalization, and use the ones performed by the Codelink software. The `Feature_id` column will be use to assign unique identifiers to each spot, so that CodelinkSet object can be read appropriately (or else will try to guess those). To read codelink data:

```r
> # NOT RUN #
> library(codelink)
> # to read data as CodelinkSet object:
> f = list.files(pattern = "TXT")
> codset = readCodelinkSet(filename = f)
> # NOT RUN #
```

This assumes that the files have the extension "TXT" (uppercase) and are in the working directory. You can prepare a targets file with each file’s name and additional phenotypic information, then pass this information to `readCodelinkSet()` so that it is stored in the CodelinkSet object.

```r
> # NOT RUN #
> pdata = read.AnnotatedDataFrame("targets.txt")
> codset = readCodelinkSet(filename = pdata$FileName, phenoData = pdata)
> # NOT RUN #
```

```r
> # sample dataset.
> data(codset)
> codset
CodelinkSet (storageMode: lockedEnvironment)
assayData: 35129 features, 4 samples
  element names: background, exprs, flag, snr, weight
protocolData: none
phenoData
  sampleNames: Sample-1 Sample-2 Sample-3 Sample-4
  varLabels: sample
  varMetadata: labelDescription
featureData
  featureNames: 1001 1002 ... 328112 (35129 total)
  fvarLabels: probeName probeType ... meanSNR (5 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: rwgcod
```

To convert old Codelink objects into the new CodelinkSet the handy function `Codelink2CodelinkSet()` can be used:

```r
```
> data(codelink.example)
> print(is(codelink.example))
[1] "Codelink"  "list"   "vector"  "AssayData"
[5] "vectorORfactor"
> tmp = Codelink2CodelinkSet(codelink.example)
> tmp
CodelinkSet (storageMode: lockedEnvironment)
  assayData: 20469 features, 2 samples
    element names: background, exprs, flag, snr, weight
  protocolData: none
  phenoData
    rowNames: 1 2
    varLabels: sample
    varMetadata: labelDescription
  featureData
    featureNames: 1 2 ... 20469 (20469 total)
    fvarLabels: probeName probeType ... meanSNR (5 total)
    fvarMetadata: labelDescription
  experimentData: use 'experimentData(object)'
Annotation: h10kcod.db

2.1 Flags and weights

Traditionally the codelink package has used flag information to assign NAs to values. This behavior has been changed since the version released with Bioconductor 2.13 (October, 2013). To reproduce the old behavior call readCodelinkSet() with argument old=TRUE.

In the current implementation, only probes flagged as MSR spots (flag 'M' - which have an intensity value assigned of -9999), will be automatically converted to NA. This value cannot be adjusted since the value of the probes itself does not represent any measure of signal.

In addition to this, probe weights will be computed by default, based on the conversion table shown in tables 1 and 2. The weight computation follows this process. First, weights are assigned based on type, with DISCOVERY probes being assigned \texttt{weight}=1 and other probes \texttt{weight}=0. Then, weights are adjusted based on flags. The worst weight (type or flag weights when multiple) is assigned to each probe. The weights assigned can be controled by the \texttt{type.weights} and \texttt{flag.weights} argument to readCodelinkSet(). It is possible also to reassign weights after reading with the function createWeights(). Weights can be used during preprocessing (background correction and normalization) and linear modeling.

```r
> w = createWeights(codset)
> ## NOTE: a proper replacement function will be provided later:
> assayDataElement(codset, "weight") = w
```

2.2 Accessing data

Data stored in a CodelinkSet object can be accessed using several accessor functions:

```r
> # get signal intensities. alias: getInt()
> head(exprs(codset))
   Sample-1 Sample-2 Sample-3 Sample-4
1001  1645.4359 1175.0750 1191.2703 1127.0000
```
> # get background intensities.
> head(getBkg(codset))

Sample-1 Sample-2 Sample-3 Sample-4
1001 31 31 31 31
1002 31 31 31 31
1004 31 31 32 31
1005 31 30 31 32
1006 32 31 30 32
1007 31 30 31 32

> # get SNR values.
> head(getSNR(codset))

Sample-1 Sample-2 Sample-3 Sample-4
1001 37.79071 27.61049 28.06214 26.18738
1002 1.10020 0.92866 0.97113 0.85644
1004 1.00796 0.92988 1.01469 0.84757
1005 22.49587 11.48678 14.98805 10.91259
1006 2.57643 1.75660 2.04631 1.80408
1007 36.99935 25.40197 25.93054 24.35704

> # get flags.
> head(getFlag(codset))

Sample-1 Sample-2 Sample-3 Sample-4
1001 "G" "G" "G" "G"
1002 "G" "L" "L" "L"
1004 "G" "L" "G" "L"
1005 "G" "G" "G" "G"
1006 "G" "G" "G" "G"
1007 "G" "G" "G" "G"

> # get weights.
> head(getWeight(codset))

Sample-1 Sample-2 Sample-3 Sample-4
1001 0 0 0 0
1002 1 1 1 1
1004 1 1 1 1
1005 1 1 1 1
1006 1 1 1 1
1007 0 0 0 0

> # get phenoData:
> head(pData(codset))

sample
Sample-1 T3-5(3)
Sample-2 TX-1(3)
Sample-3 T3-2(3)
Sample-4 T3-4(1)
3 Background correction

If Spot_mean and Bkgd_median values are available then background correction can be performed with `codCorrect()`. Background correction methods are borrowed from the limma package, including methods `none`, `subtract`, `half` and `normexp`. The default is set to `half`, because it is very fast. However, more sensitive (although slower) methods like `normexp` are recommended. It is possible to assign an offset to avoid low intensity probes to have high M variances.

```r
> codset = codCorrect(codset, method = "half", offset = 0)
```

4 Normalization

Normalization of the background corrected intensities is done by the wrapper function `normalize` (or the alias `codNormalize()`). Here again, normalization is borrowed from the limma package. Methods `median`, `quantile` (the default) and `loess` are available.

```r
> codset = codNormalize(codset, method = "quantile")
```

Method `loess` performs CyclicLoess normalization and accepts weights. Weights are used in a per-probe fashion (that is, one weight for one probe, not different weights for each sample). When weights are used for normalization the minimum of all the weights for each probe along all the samples will be used. This is to ensure that for each array there is an equal contribution of each probe in the normalization process.

```r
> # NOT RUN
> codset = codNormalize(codset, method = "loess", weights = getWeight(codset),
+                  loess.method = "fast")
> # NOT RUN
```

5 Diagnostic plots

There are some plot facilities to help diagnose the effect of background correction and normalization, as well as identify putative faulty arrays. The most commonly used functions are MA plots, density plots and array images. All these functions can be accessed through the function `codPlot()`. The parameter `what` specifies the type of plot: `ma` (default), `density`, `scatter` and `image` are valid choices. Figures 1 and 2 below show examples of these plotting functions.

```r
> codPlot(codset) # by default MA plot.
> codPlot(codset, what = "density")
```

When the columns `Logical_row` and `Logical_col` are present in the original data files, this information is used to assign the physical location of each probe in the array to plot a pseudo image. It is possible to plot the background intensities (default), the spot mean, raw and normalized intensities and the SNR values. This images are useful to identify spatial artifact that may be affecting the analysis.

```r
> codPlot(codset, what = "image")
```
Figure 1: MA plot (left) and density plot (right).

Figure 2: Pseudo image plot of an array
6 Fitting linear models

A typical analysis include the testing for differentially expressed probes between two populations. This can be performed using a variety of different R/Bioconductor packages, but the `limma` package is one of the most popular options. Limma can readily use `CodelinkSet` objects, and can take advantage of weights generated during data reading. In this case, weights will be use probe-wise (i.e. different weights for the same probe in different samples will be considered).

```r
> fit = lmFit(codset, design = c(1, 1, 2, 2), weights = getWeight(codset))
> fit2 = eBayes(fit)
> topTable(fit2)
```

<table>
<thead>
<tr>
<th>probeName</th>
<th>probeType</th>
<th>logicalRow</th>
<th>logicalCol</th>
<th>meanSNR</th>
<th>logFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>255020</td>
<td>GE1262775</td>
<td>DISCOVERY</td>
<td>255</td>
<td>224.6663</td>
<td>7.985204</td>
</tr>
<tr>
<td>311011</td>
<td>GE21204</td>
<td>DISCOVERY</td>
<td>311</td>
<td>197.73053</td>
<td>7.967831</td>
</tr>
<tr>
<td>31103</td>
<td>GE1152142</td>
<td>DISCOVERY</td>
<td>31</td>
<td>222.18072</td>
<td>7.863261</td>
</tr>
<tr>
<td>106027</td>
<td>GE1204034</td>
<td>DISCOVERY</td>
<td>106</td>
<td>130.05434</td>
<td>7.468283</td>
</tr>
<tr>
<td>24019</td>
<td>GE1126416</td>
<td>DISCOVERY</td>
<td>24</td>
<td>122.54245</td>
<td>7.312816</td>
</tr>
<tr>
<td>150024</td>
<td>GE20195</td>
<td>DISCOVERY</td>
<td>150</td>
<td>105.40705</td>
<td>7.265655</td>
</tr>
<tr>
<td>321107</td>
<td>GE1181836</td>
<td>DISCOVERY</td>
<td>321</td>
<td>98.68877</td>
<td>7.259977</td>
</tr>
<tr>
<td>242033</td>
<td>GE1221296</td>
<td>DISCOVERY</td>
<td>242</td>
<td>102.22355</td>
<td>7.224388</td>
</tr>
<tr>
<td>114068</td>
<td>GE22145</td>
<td>DISCOVERY</td>
<td>114</td>
<td>102.59456</td>
<td>7.199294</td>
</tr>
<tr>
<td>312012</td>
<td>GE19692</td>
<td>DISCOVERY</td>
<td>312</td>
<td>93.38577</td>
<td>7.192154</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AveExpr</th>
<th>t</th>
<th>P.Value</th>
<th>adj.P.Val</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.33260</td>
<td>12.70075</td>
<td>6.257377e-37</td>
<td>1.505759e-32</td>
<td>72.17457</td>
</tr>
<tr>
<td>13.21738</td>
<td>12.67312</td>
<td>8.898759e-37</td>
<td>1.505759e-32</td>
<td>71.83290</td>
</tr>
<tr>
<td>13.14444</td>
<td>12.50680</td>
<td>7.293671e-36</td>
<td>8.227747e-32</td>
<td>69.79209</td>
</tr>
<tr>
<td>12.44730</td>
<td>11.87587</td>
<td>1.609255e-32</td>
<td>1.361511e-28</td>
<td>62.32685</td>
</tr>
<tr>
<td>12.23010</td>
<td>11.63130</td>
<td>2.993858e-31</td>
<td>2.028363e-27</td>
<td>59.49396</td>
</tr>
<tr>
<td>12.14916</td>
<td>11.55629</td>
<td>7.180842e-31</td>
<td>3.855796e-27</td>
<td>58.64639</td>
</tr>
<tr>
<td>12.06931</td>
<td>11.54725</td>
<td>7.954666e-31</td>
<td>3.855796e-27</td>
<td>58.54471</td>
</tr>
<tr>
<td>12.07106</td>
<td>11.49065</td>
<td>1.536931e-30</td>
<td>6.501604e-27</td>
<td>57.90924</td>
</tr>
<tr>
<td>11.98362</td>
<td>11.43938</td>
<td>2.776572e-30</td>
<td>9.396475e-27</td>
<td>57.33636</td>
</tr>
</tbody>
</table>

7 Citation

```r
> citation(package = "codelink")
```

To cite codelink in publications use:

Diego Diez, Rebeca Alvarez and Ana Dopazo. codelink: An R package for analysis of GE Healthcare Gene Expression Bioarrays. 2007, Bioinformatics

A BibTeX entry for LaTeX users is

```latex
@Article{},
  title = {codelink: an R package for analysis of GE Healthcare Gene Expression Bioarrays},
  author = {{Diego Diez} and {Rebeca Alvarez} and {Ana Dopazo}}},
```
8 Session info

> sessionInfo()
R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

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

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

other attached packages:
[1] knitr_1.14 codelink_1.42.0 limma_3.30.0
[4] Biobase_2.34.0 BiocGenerics_0.20.0

loaded via a namespace (and not attached):
[1] IRanges_2.8.0 XML_3.98-1.4 bitops_1.0-6
[4] DBI_0.5-1 xtable_1.8-2 stats4_3.3.1
[7] formatR_1.4 magrittr_1.5 RSQLite_1.0.0
[10] evaluate_0.10 highr_0.6 stringi_1.1.2
[13] annotate_1.52.0 S4Vectors_0.12.0 tools_3.3.1
[16] stringr_1.1.0 RCurl_1.95-4.8 AnnotationDbi_1.36.0