Comprehensive Pipeline for Analyzing and Visualizing Agilent and Affymetrix Array-Based CGH Data

Frederic Commo *
Inserm U981, Bioinformatics Group, Gustave Roussy, France

October 28, 2015

1 Introduction

Genomic profiling using array-based comparative genomic hybridization (aCGH) is widely used within precision medicine programs, in combination with DNA sequencing, to match specific molecular alterations (amplifications or deletions) with therapeutic orientations.
We present *rCGH*, a comprehensive array-based CGH analysis workflow, integrating functionalities specifically designed for precision medicine. *rCGH* ensures a full traceability by saving all the process parameters , and facilitates genomic profiles interpretation and decision-making through interactive visualizations.
rCGH supports Agilent (from 44K to 400K arrays), as well as Affymetrix, SNP6 and cytoScanHD arrays.

2 rCGH object structure

In order to store (or update) data, sample information, and the workflow parameters all along a genomic profile analysis process, `rCGH` objects are structured as follow:

- info: the sample information.
- cnSet: the full by-probe dataset.
- param: the workflow parameters, for traceability.
- segTable: the segmentation data.

All these slots are accessible through specific functions, as described in the next sections.
Notice that *rCGH* is a superclass designed for calling common methods. Depending on the type of array and the *read* functions used, the resulting objects will be assigned to classes *rCGH-Agilent*, *rCGH-SNP6*, or *rCGH-cytoScan*. These classes inherit from the superclass, and allow array-specific pre-parametrizations.

*frederic.commo@gustaveroussy.fr

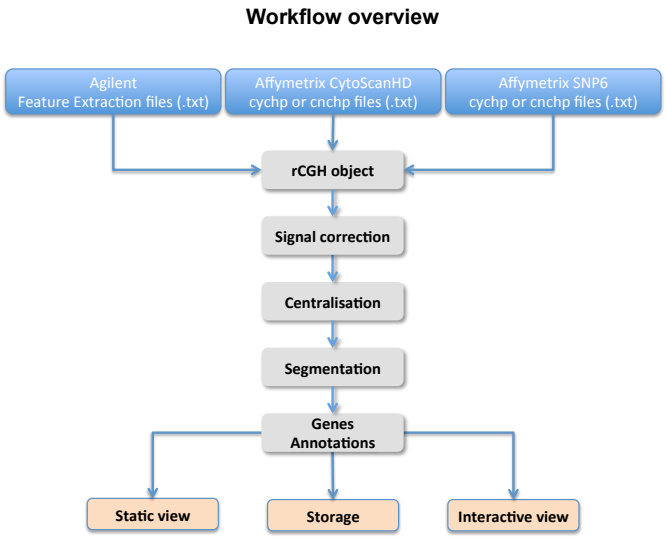

Figure 1: **rCGH workflow.** The global rCGH analysis workflow

3 rCGH functions

rCGH provides functions for each of the analysis steps, from reading files to visualizing genomic profiles. Several *get* functions allow the user to get access to specific results and workflow parameters, saved and stored at each step.

3.1 Reading files

Both Agilent Feature Exraction files (from 44K to 400K arrays), and Affymetrix SNP6 and cytoScanHD, data are supported.
However, and to keep more flexibility, Affymetrix CEL files have to be first read using ChAS or Affymetrix Power Tools (APT) [1], and then exported as cychp.txt or cnchp.txt files.
Notice that cnchp.txt files contain Allelic differences, that allow the loss of heterozygosity (LOH) to be estimated, while cychp.txt files do not.

Due to specific files structures, and since preambles may be missing (depending on ChAS and APT versions), *rCGH* as 3 specific read/build-object functions:

- `readAgilent()`: 44K to 400K FE (.txt) files.
- `readAffySNP6()`: cychp, cnchp and probeset (.txt) files, exported from SNP6.0 CEL.
- `readAffyCytoScan()`: cychp, cnchp and probeset (.txt) files, exported from CytoScanHD CEL.

Each of these functions take the file's path as the unique mandatory argument.
Optional arguments allow the user to save the following information: *sampleName, labName*:

```
> library(rCGH)
> filePath <- system.file("extdata", "Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2",
```

```
+     package = "rCGH")
> cgh <- readAffyCytoScan(filePath, sampleName = "CSc-Example",
+                         labName = "myLab")
```

```
> cgh

                                                    info
fileName          Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2
sampleName                                  CSc-Example
labName                                           myLab
usedProbes                                          snp
platform                                CytoScanHD_Array
barCode             @52082500958167113016424803602715
gridName               CytoScanHD_Array.na33.annot.db
scanDate                                    2015-01-22
programVersion                                   5.0.0
gridGenomicBuild                           hg19/GRCh37
reference          CytoScanHD_Array.na33.r1.REF_MODEL
analyseDate                                 2015-10-28
rCGH_version                                     1.0.2
```

In complement, any kind of useful annotation (logical, string or numeric) can be added, with `setInfo()`:

```
> setInfo(cgh, "item1") <- 35
> setInfo(cgh, "item2") <- TRUE
> setInfo(cgh, "item3") <- "someComment"
```

At any time, the full (or specific) annotations stored can be accessed:

```
> getInfo(cgh)

                                                    info
fileName          Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2
sampleName                                  CSc-Example
labName                                           myLab
usedProbes                                          snp
platform                                CytoScanHD_Array
barCode             @52082500958167113016424803602715
gridName               CytoScanHD_Array.na33.annot.db
scanDate                                    2015-01-22
programVersion                                   5.0.0
gridGenomicBuild                           hg19/GRCh37
reference          CytoScanHD_Array.na33.r1.REF_MODEL
analyseDate                                 2015-10-28
rCGH_version                                     1.0.2
```

```
item1                                                  35
item2                                                TRUE
item3                                          someComment
```

```
> getInfo(cgh, c("item1", "item3"))
```

```
       item1         item3
        "35" "someComment"
```

3.2 Adjusting signals

When Agilent dual-color hybridization are used, GC content and the cy3/cy5 bias are necessary adjustments. `adjustSignal()` handle these steps before computing the $\log_2(RelativeRatios)$ (LRR). In both cases, a local regression (`loessFit`, R package *limma*) is used [2].
Note that by default, the cyanine3 signal is used as the reference. Use `Ref=cy5` if cyanine5 signal has to be used as the reference.
When Affymetrix cychp or cnchp files are used, these steps have already been done, and `adjustSignal()` simply rescale the LRR, when `Scale=TRUE` (default). As for Agilent data, some useful quality scores: the derivative Log Ratio Spread (dLRs) and the LRR Median Absolute Deviation (MAD), are stored in the object.

```
> cgh <- adjustSignal(cgh, nCores=1)
```

```
Log2Ratios QCs:
dLRs:  0.199
MAD: 0.24

Scaling...
Signal filtering...
Modeling allelic Difference...
```

3.3 Centering LRR

Centering LRR is a key step in the genomic analysis process since it defines the base line (the expected 2-copies level) from where gains ad losses are estimated. To do so, LRRs are considered as a mixture of several gaussian populations , and an expectation-maximization (EM) algorithm is used to estimate their parameters.
The centralization value is chosen according to the user specification: the mean of the sub-population with a density peak higher than a given proportion of the highest density peak [3]. The default value is 0.5. Setting `peakThresh = 1` leads to choose the highest density peak.
The `plotDensity()` function gives access to a graphical check on how the centralization step worked, and what LRR population has been chosen for centering the profile:

```
> # Resricted to 3 groups for the purpose of that demo.
> cgh <- EMnormalize(cgh, G = 3)

Smoothing param:  73
Analyzing mixture...
Merging peaks closer than 0.1 ...
Gaussian mixture estimation:
n.peaks =  3

Group parameters:
Grp 1:
prop:  0.396, mean:  0.026, Sd:  0.086, peak height:  1.84
Grp 2:
prop:  0.039, mean:  0.753, Sd:  0.388, peak height:  0.04
Grp 3:
prop:  0.565, mean:  1.341, Sd:  0.075, peak height:  3

Correction value:  0.026
Use plotDensity() to visualize the LRR densities.
```

```
> plotDensity(cgh)
```

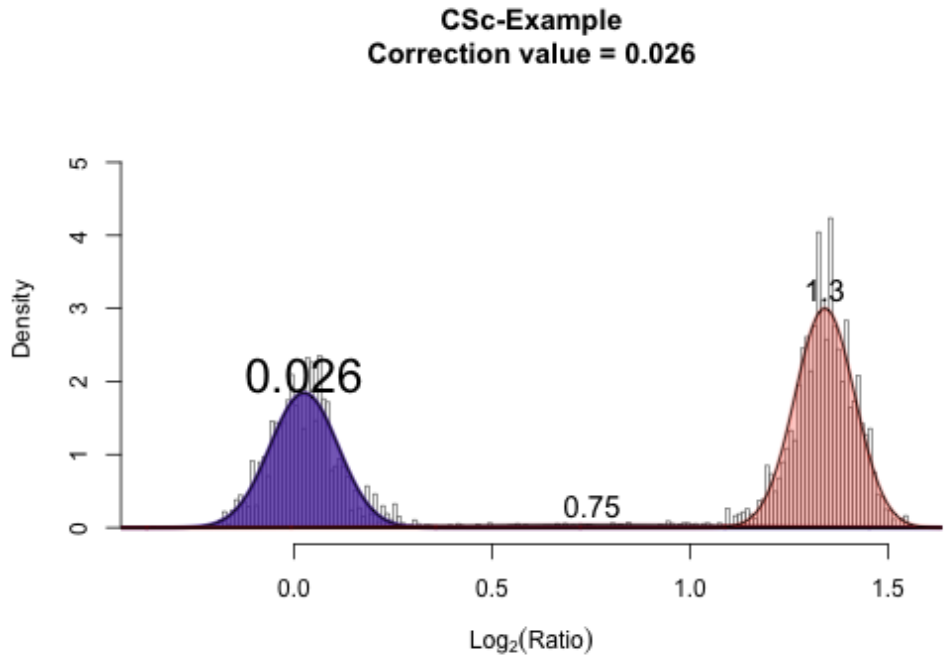

Figure 2: **plotDensity.** `plotDensity()` shows how *EM* models the *LRR* distribution, and what peak is chosen for centralizing the profile (in bold).

3.4 Segmenting

One possible strategy for segmenting the genome profile consists in identifying breakpoints all along the genome, when exist. These breakpoints define the DNA segments start and end positions. To do so, *rCGH* uses the Circular Binary Segmentation algorithm (*CBS*) [4] from the *DNAcopy* package [5]. All the steps are wrapped into one unique easy-to-use function, `segmentCGH()`. In order to faclitate its use, all the parameters but one are predefined: `UndoSD` is kept free. When this parameter is set to `NULL` (default), its optimal value is estimated directly from the values. However, the user can specify its own value, generaly from 0.5 to 1.5.
The resulting segmentation table is of the form of a standard *DNAcopy* output, plus additional columns:

- ID : sample Id.
- chrom : chromosome number.
- loc.start : segment start position.
- loc.end : segment end position.
- num.mark : number of markers within each segment.
- seg.mean : the mean LRR along each segment.
- seg.med : the median LRR along each segment.
- probes.Sd : the LRR standard deviation along each segment.

```
> cgh <- segmentCGH(cgh, nCores=1)

Computing LRR segmentation using UnodSD: 0.245
Merging segments shorter than 10Kb.
Number of segments:  26

> segTable <- getSegTable(cgh)
> head(segTable)

           ID chrom loc.start   loc.end num.mark seg.mean  seg.med probes.Sd
1 CSc.Example     1    882803 120345101      617   0.1135 -0.01885 0.6849012
2 CSc.Example     1 121155528 249198692      591   1.1900  1.27280 0.7150160
3 CSc.Example     2     15703 242775910     1316   1.3409  1.33040 0.4764931
4 CSc.Example     3     62614 197851260     1099  -0.0243 -0.01885 0.5046325
5 CSc.Example     4     46691 190921709     1041   1.3143  1.33040 0.4850337
6 CSc.Example     5    113577 180692833      985   1.3431  1.33040 0.4934583
```

Note that such data format allows GISTIC-compatible inputs to be exported [6].

3.5 Parallelization

rCGH allows parallelization within `EMnormalise()` and `segmentCGH()`, through `mclapply()` from R package *parallel*.
By default, `nCores` will be set to half of the available cores, but any value, from 1 to `detectCores()`, is allowed. However, this feature is currently only available on `Linux` and `OSX`: `nCores` will be auto-

matically set to 1 when a `Windows` system is detected.

3.6 Getting the by-gene table

The next step consists in getting access to the potentially altered genes. `byGeneTable()` extracts the list of genes included in each segment, and constructs a dataset, easy to export and to manipulate outside R. The final genes' list combines position information from *TxDb.Hsapiens.UCSC.hg19.knownGene*, and annotations from *org.Hs.eg.db*.

```
> #geneTable <- byGeneTable(cgh)
> geneTable <- byGeneTable(segTable)
```

```
Creating byGene table...
```

```
> head(geneTable, n=3)
```

```
  entrezid   symbol                       fullName cytoband chr chrStart
1        1     A1BG         alpha-1-B glycoprotein  19q13.4  19 58858172
2   503538 A1BG-AS1           A1BG antisense RNA 1  19q13.4  19 58859117
3    29974     A1CF APOBEC1 complementation factor 10q11.23  10 52559169
    chrEnd width strand Log2Ratio num.mark segNum segLength(kb) relativeLog
1 58874214 16043      -    1.3304      230     22      58836.84           0
2 58866549  7433      +    1.3304      230     22      58836.84           0
3 52645435 86267      -    1.2728      750     11     135330.87           0
  genomeStart
1  2718302494
2  2718303439
3  1732932312
```

3.7 Accessing the analysis parameters

For traceability and reproducibility, it may be useful to keep track to a profile analysis parameters. At each step, the workflow parameters, defined by default or specified by the user, are stored in a `params` slot. They are accessible at any time using `getParam()`.

```
> getParam(cgh)[1:3]
```

```
$ksmooth
[1] 73

$Kmax
[1] 20

$Nmin
```

```
[1] 160
```

3.8 Visualizing the genomic profile

In a context of Precision Medicine, visualizing and manipulating a genomic profile is crucial to interpret imbalances, to identify targetable genes, and to make decisions regarding a potential therapeutic orientation. In many situations, considering LOH can also help to better interpret imbalances.
rCGH provides 2 ways for visualizing a genomic profile: `plotProfile()`, `plotLOH()` and `multiplot()` are simple static ways to visualize a profile, possibly with some tagged gene, while `view()` is a more sophisticated and interactive visualization method, build on top of shiny. A control panel allows the user to interact with the profile, and to export the results.
Notice that `plotLOH()` and `multiplot()` are relevant only in case the allelic difference is available, namely when Affymetrix cnchp.txt files are used.

3.8.1 Static profile visualizations

`plotProfile()` allows the genomic profile visualization. Any gene(s) of interest can be added to the plot by passing a valid HUGO symbol. Other arguments can be used to color the segments according to specified gain/loss thresholds, or to change the plot title.
Two other static functions can be useful for reporting alterations: `plotLOH()` to analyse the LOH, and `multiplot()` to build a full report, including both the genomic profile and the LOH.

```
> multiplot(cgh, c("egfr", "erbb2"))
```

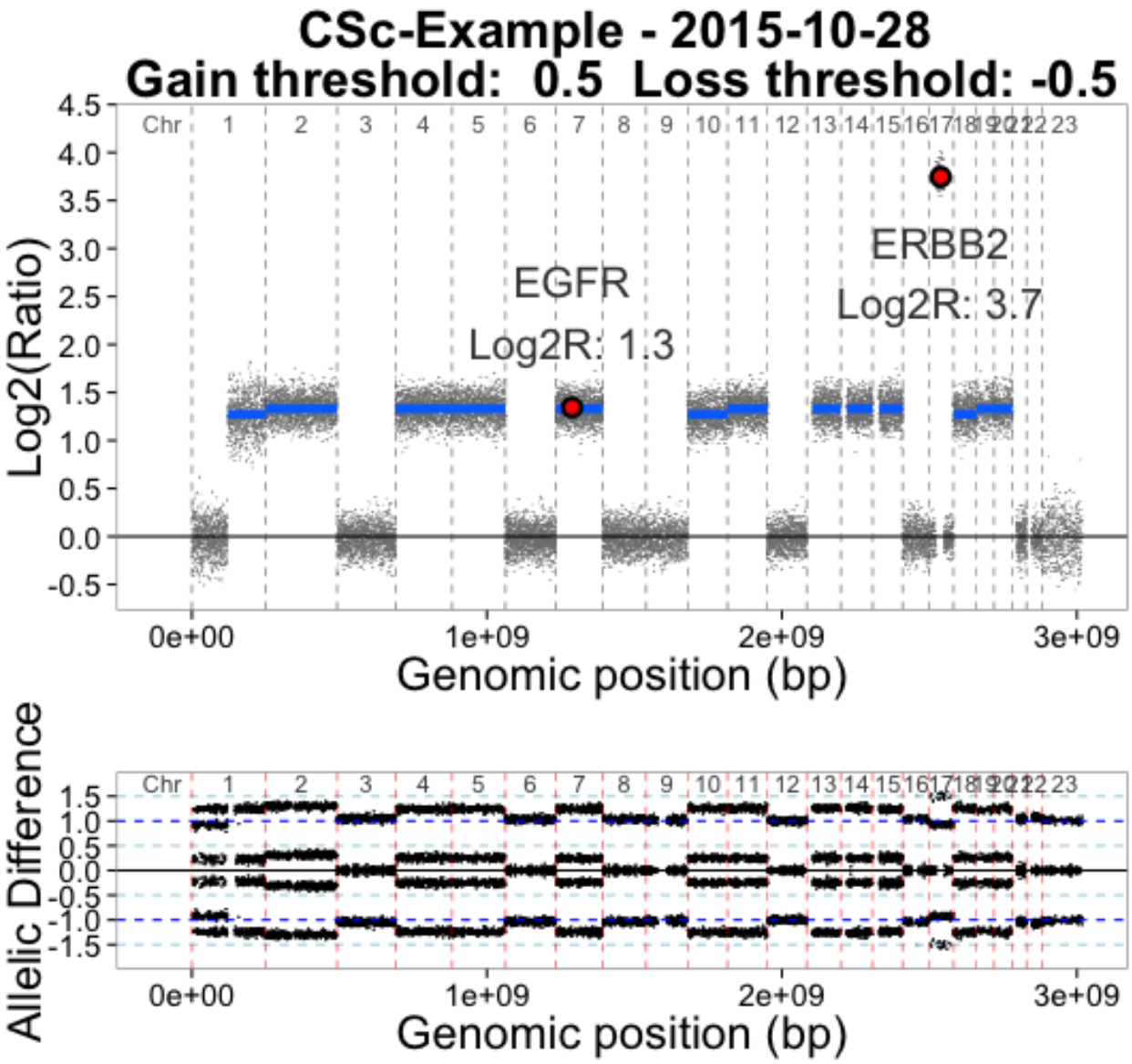

Figure 3: **Static views.** `multiplot()` provides static visualisations combining the genomic profile and the LOH.

3.8.2 Recentering

When the profile centering doesn't seem appropriate, `recenter()` allows the user to choose another centralization value. The new choice has to be specified as the peak index to use: peaks are indexed, from 1 to k (from left to right) as they appear on the density plot.

```
> # Recentering on peak #3
> recenter(cgh) <- 3

Profile recentered on:  1.34

> plotProfile(cgh, "erbb2")
```

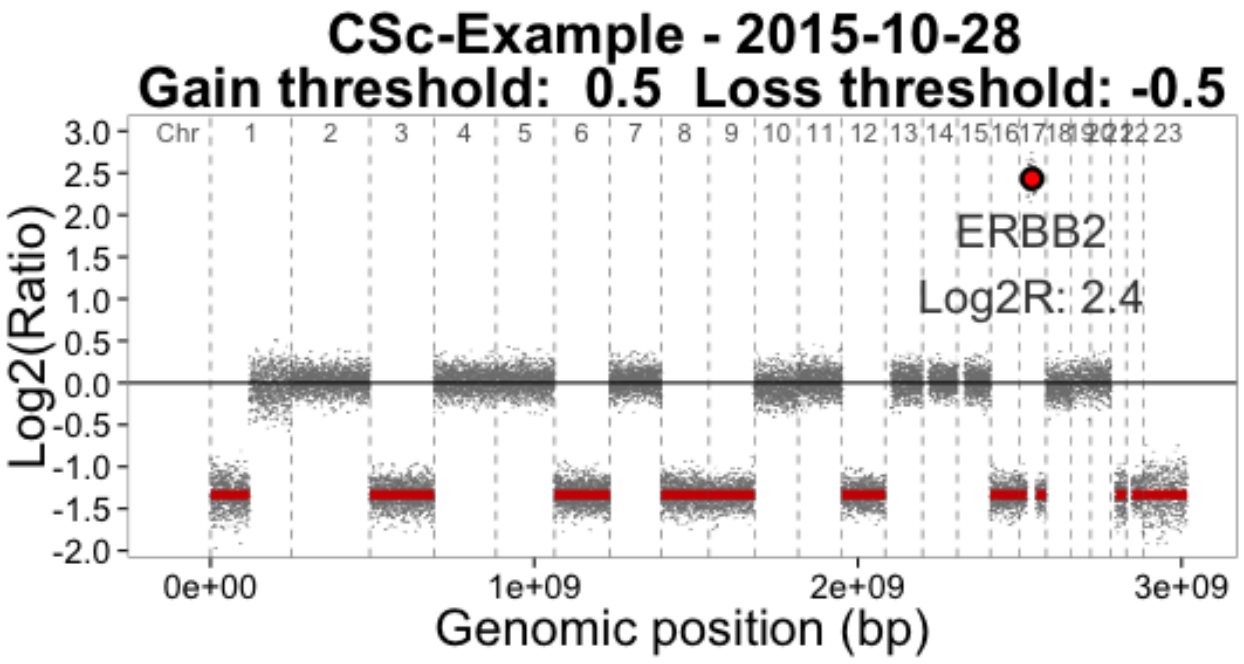

Figure 4: **Recentering.** By default, the EM-based normalization choose a possibly optimal peak to center the profile, but any other peak can be chosen, using `recenter()`.

3.8.3 Interactive visualization

The *view()* function provides a more flexible way for analyzing a genomic profile, and allows interactive graph manipulations through a control panel: defining the gain/loss thresholds, displaying a gene, resizing the y-axis, selecting one unique chromosome, and recentering the entire profile. Note that the *Genes table* is updated whenever changes are made through that control panel, e.g. selecting one unique chromosome on the graph filters the *Genes table* on that chromosome, simultaneously.
The Download buttons, *Plot*, *LOH* and *Table*, allow plots and gene table to be exported, as they have been modified.

The `view()` control panel:

- Gene Symbol : display any existing gene, providing its official HUGO symbol.
- Show chromosome : display the entire profile (default is 'All'), or one specific chromosome.
- Gain/Loss colors : choose blue/red or red/blue.
- Recenter profile : recenter the profile on-the-fly. Gene values are updated in the 'Genes table'.
- Merge segments... : merge segments shorter than the specified value, in Kb. Gene values are updated in the 'Genes table'.
- Recenter profile : recenter the profile on-the-fly. Gene values are updated in the 'Genes table'.
- Rescale max(y) : adjust the top y-axis (0¡y) using a proportion of the maximum value.
- Rescale min(y) : adjust the bottom y-axis (y¡0) using a proportion of the minimum value.
- Gain threshold (Log2ratio) : define the gain threshold. Segments higher than this value are colored according to the chosen color code, and the 'Genes table' is filtered, consequently.
- Loss threshold (Log2ratio) : same as 'Gain threshold' but for losses.
- Download - Profile : download the profile as it is displayed on the screen, including modifications.
- Download - LOH : download the LOH plot as it is displayed on the screen, including modifications.
- Download - Table : download the 'Genes table', including modifications.

```
> view(cgh)
```

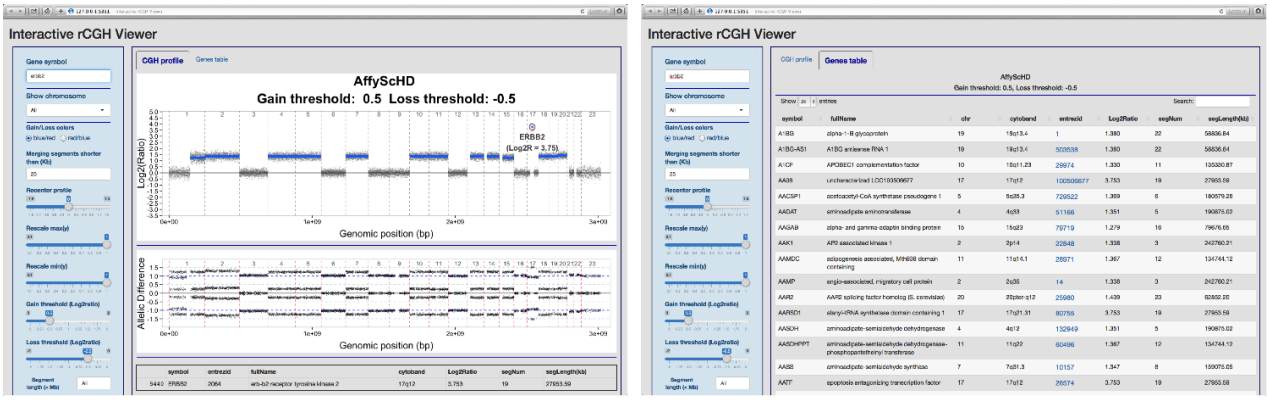

Figure 5: **Interactive profile.** The genomic profile is displayed in the first *CGH profile* tab (left). Several changes can be applied using the control panel (in blue). The list of genes is accessible through the *Genes table* tab (right). Both are updated simultaneously and can be exported, after modifications are applied.

4 Notes regarding the example files

In order to reduce the computation time, we provide subsets of real data for the 3 supported platforms:

```
> list.files(system.file("extdata", package = "rCGH"))
```

```
[1] "Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2"
[2] "Affy_snp6_cnchp.txt.bz2"
[3] "Agilent4x180K.txt.bz2"
```

comment:
In order to speed up demos, the provided example files contain only a subset of the original probes. Affymetrix example files (cytoScan and SNP6) only contain SNP probes. Setting `useProbes = "cn"` in `readAffy` functions should return an error.

5 Server version

A web browser version of the interactive visualization is available at
https://fredcommo.shinyapps.io/aCGH_viewer
As inputs, this application support the *rCGH* segmentation tables, or any segmentation table in the same format as the *CBS* outputs.
For more details about this application, or to install it on your own server, please visit
https://github.com/fredcommo/aCGH_viewer.

6 Session information

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

```
\begin{itemize}\raggedright
  \item R version 3.2.2 (2015-08-14), \verb|x86_64-apple-darwin13.4.0|
  \item Locale: \verb|C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8|
  \item Base packages: base, datasets, grDevices, graphics, methods,
    stats, utils
  \item Other packages: DBI~0.3.1, RSQLite~1.0.0, knitr~1.11,
    rCGH~1.0.2
  \item Loaded via a namespace (and not attached):
    AnnotationDbi~1.32.0, Biobase~2.30.0, BiocGenerics~0.16.0,
    BiocInstaller~1.20.0, BiocParallel~1.4.0, BiocStyle~1.8.0,
    Biostrings~2.38.0, DNAcopy~1.44.0, GenomeInfoDb~1.6.0,
    GenomicAlignments~1.6.1, GenomicFeatures~1.22.0,
    GenomicRanges~1.22.0, IRanges~2.4.1, MASS~7.3-44, R6~2.1.1,
```

```
    RCurl~1.95-4.7, Rcpp~0.12.1, Rsamtools~1.22.0, S4Vectors~0.8.0,
    SummarizedExperiment~1.0.0,
    TxDb.Hsapiens.UCSC.hg19.knownGene~3.2.2, XML~3.98-1.3,
    XVector~0.10.0, aCGH~1.48.0, affy~1.48.0, affyio~1.40.0,
    biomaRt~2.26.0, bitops~1.0-6, cluster~2.0.3, colorspace~1.2-6,
    digest~0.6.8, evaluate~0.8, formatR~1.2.1, futile.logger~1.4.1,
    futile.options~1.0.0, ggplot2~1.0.1, grid~3.2.2, gtable~0.1.2,
    highr~0.5.1, htmltools~0.2.6, httpuv~1.3.3, labeling~0.3,
    lambda.r~1.1.7, lattice~0.20-33, limma~3.26.0, magrittr~1.5,
    mclust~5.1, mime~0.4, multtest~2.26.0, munsell~0.4.2,
    org.Hs.eg.db~3.2.3, parallel~3.2.2, plyr~1.8.3,
    preprocessCore~1.32.0, proto~0.3-10, reshape2~1.4.1,
    rtracklayer~1.30.1, scales~0.3.0, shiny~0.12.2, splines~3.2.2,
    stats4~3.2.2, stringi~1.0-1, stringr~1.0.0, survival~2.38-3,
    tools~3.2.2, xtable~1.7-4, zlibbioc~1.16.0
\end{itemize}
```

References

[1] URL: http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx.

[2] Smyth GK and Speed TP. Normalization of cdna microarray data. *Methods*, 31:265–273, 2003. URL: http://www.statsci.org/smyth/pubs/normalize.pdf.

[3] Commo F, Ferte C, Soria JC, Friend SH, Andre F, and Guinney J. Impact of centralization on acgh-based genomic profiles for precision medicine in oncology. *Ann Oncol.*, 2014.

[4] Venkatraman ES and Olshen AB. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 15(23):657–663, 2007.

[5] Venkatraman E. Seshan and Adam Olshen. *DNAcopy: DNA copy number data analysis*. R package version 1.40.0.

[6] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, and Getz G. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.