oneChannelGUI Package Vignette

Raffaele A Calogero, Francesca Cordero, Remo Sanges

April 14, 2007

1 Introduction

This package is a GUI for "mouse-click" based QC, statistical analysis and data mining for one channel microarray data. It is designed for Bioconductor beginners having limited or no experience in interacting with Bioconductor line commands. OneChannelGUI is a set of functions extending the affylmGUI capabilities. The basic GUI of affylmGUI has been rearranged and extended. A detailed description of the usage of the available functions can be found in the help pages associated to each function. This package allows to perform, in a graphical environment, the analysis pipe-line shown in figure 1, green box.

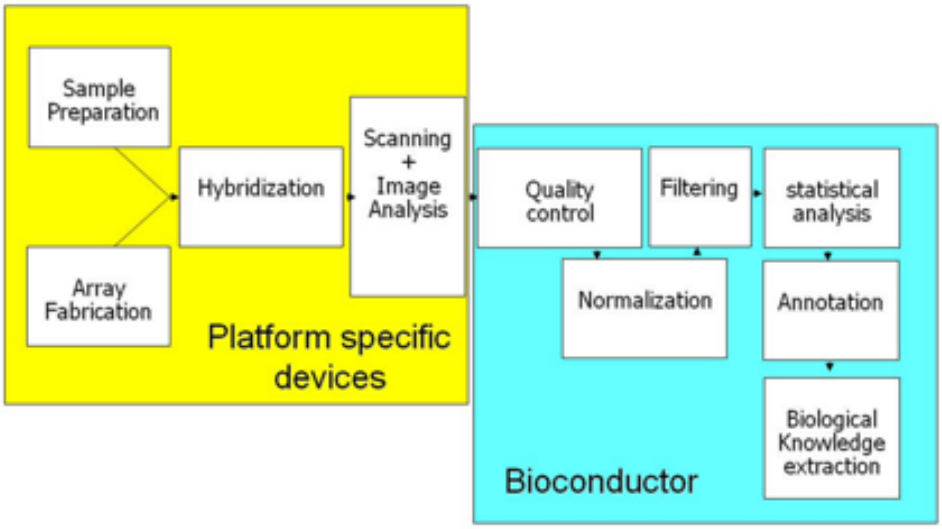


Figure 1: Microarray analysis pipe-line.

This vignette gives a general view of the available graphical tools present in oneChannelGUI. An extended description of oneChannelGUI usage with exercises and test data

can be found at: http://www.bioinformatica.unito.it/oneChannelGUI/

The oneChannelGUI Documentation can be accessed online or by installing the oneChannelGUI package locally. Then at the R prompt you can type

```
> if (interactive()) {
+     browseURL("http://www.bioinformatica.unito.it/downloads/oneChannelGUI.pdf")
+ }
```

2 Inizialization

When the oneChannelGUI library is loaded in R it loads the affylmGUI library and all the libraries needed for the microarray analysis.

```
> assign("affylmGUIenvironment", new.env(), .GlobalEnv)
> if (.Platform$OS.type != "windows") {
+     library("oneChannelGUI")
+     oneChannelGUI()
+ }

To begin, type oneChannelGUI()

Searching for user-defined affylmGUI commands in /home/biocbuild/bbs-2.0-bioc/R/librar
[1] ""

> con <- url("http://www.bioinformatica.unito.it/downloads/exon.sample.txt")
> load(con, envir = affylmGUIenvironment)
> close(con)
```

Ones a new project is started, on the basis of the type of microarray data under analysis, a specific menu is layered over the affylmGUI ones. Actually three menus are available: 3'IVT Affymetrix arrays, Affymetrix Exon arrays, Large data set (i.e. any kind of single channel array data can be loaded). Two other menus are under development for AB1700 and Illumina arrays.

All data uploaded in oneChannelGUI as well as results derived from their analysis are saved in affylmGUIenvironment, which is an R environment

```
> ls(affylmGUIenvironment)

 [1] "affylmGUIVersion"
 [2] "ArraysLoaded"
 [3] "cdf"
 [4] "CDFFile"
 [5] "cdfName"
```

```
 [6] "connectedSubGraphs"
 [7] "ContrastParameterizationList"
 [8] "ContrastParameterizationNamesVec"
 [9] "ContrastParameterizationTREEIndexVec"
[10] "exonAffyData"
[11] "exonAffyData.Available"
[12] "exprConsoleLibs"
[13] "exprConsoleLibs.Available"
[14] "FileNameText"
[15] "geneNames"
[16] "geneSymbols"
[17] "intronicBg"
[18] "intronicBg.available"
[19] "limmaDataSetNameText"
[20] "LimmaFileName"
[21] "limmaVersion"
[22] "LinearModelFit.Available"
[23] "maSigProSigs.Available"
[24] "MLdesign"
[25] "MLdesign.Available"
[26] "NormalizedAffyData"
[27] "NormalizedAffyData.Available"
[28] "NormMethod"
[29] "numConnectedSubGraphs"
[30] "NumContrastParameterizations"
[31] "NumParameters"
[32] "NumRNATypes"
[33] "NumSlides"
[34] "Pset"
[35] "Pset.Available"
[36] "PsetData.Available"
[37] "PvalExon"
[38] "RawAffyData"
[39] "RawAffyData.Available"
[40] "SlideNamesVec"
[41] "spliceIndexData"
[42] "spliceIndexData.Available"
[43] "Targets"
[44] "TargetsFile"
[45] "testAffyData.available"
[46] "ThresholdFC"
[47] "trainAffyData.available"
```

```
[48] "Try"
[49] "weightsPLM"
[50] "whichArrayPlatform"
```

2.1 CEL files: 3'IVT, Exon and Gene Affymetrix arrays

The GUI allows the manipulation of .CEL files from 3'IVT as well as of the new human Gene 1.0 ST arrays using the cdf file hugene10stv1cdf available at `http://www.bioinformatica.unito.it/oneChannelGUI`. Also exon .CEL can be loaded on oneChannelGUI, using the cdf huex10stv2cdf available at `http://www.bioinformatica.unito.it/oneChannelGUI` Since Exon arrays are very big they cannot be handled on winXP even if on systems with 2 Gb RAM. However, they can be handled without any problem on linux if enough RAM is available. In our lab, we analyze exon arrays using a SUN work station with 24Gb RAM, running SUSE linux. Having loaded raw .CEL files the oneChannelGUI allows primary (probe level QC, probe set summary and normalization), and secondary analysis (replicates QC by PCA, sample size evaluation, filtering, Limma and SAM analysis, maSigPro time course analysis) and data mining (GO enrichment) is provided Furthermore, a tool to generate template A files to upload on Ingenuity database `http://www.ingenuity.com` is also implemented. The menu functions are summarized in figure 2

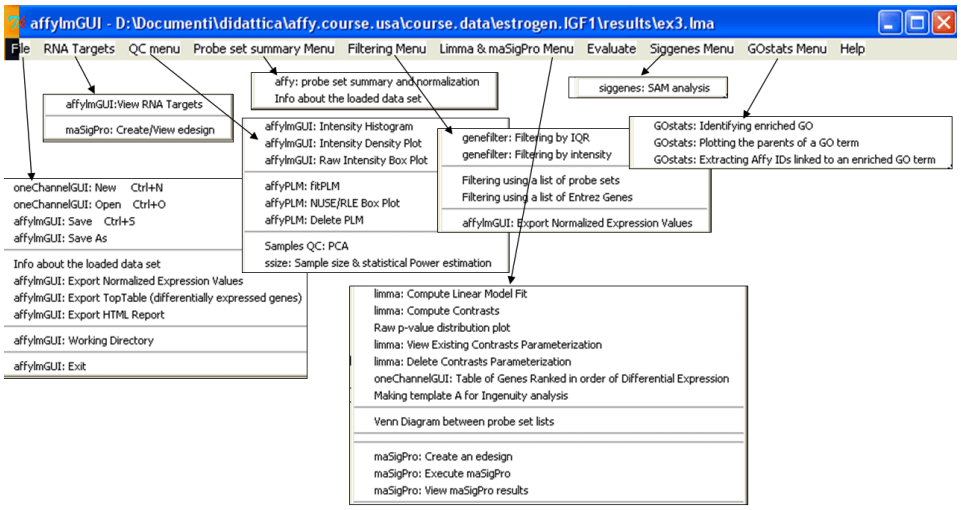
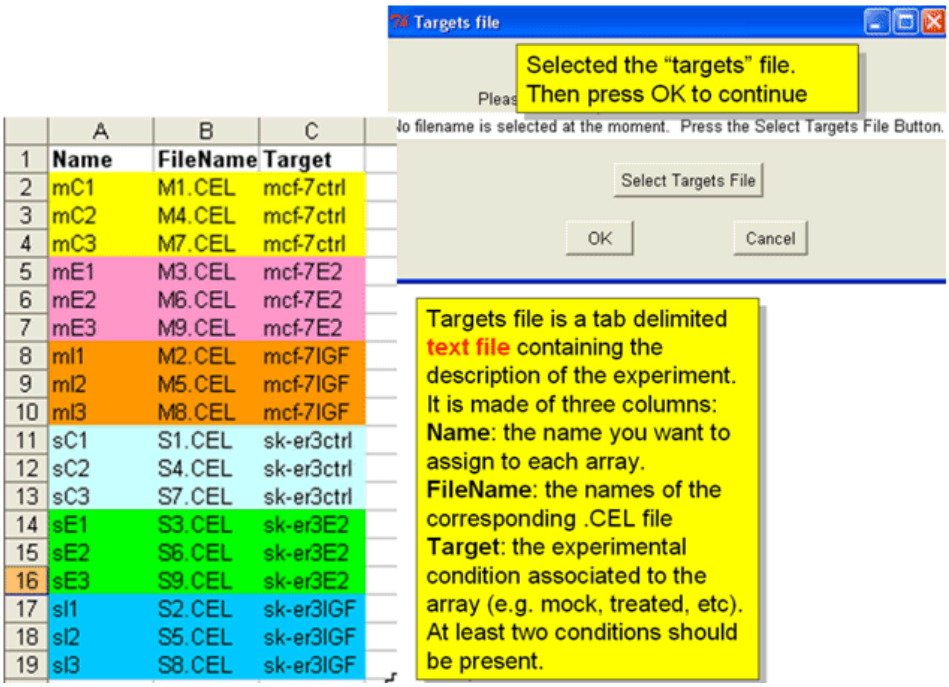


Figure 2: 3'IVT Affymetrix microarray menu.

Each item in the menu is simply a graphical implementation of a function present in a specific Bioconductor library (e.g. ssize: sample size and statistical power estimation). To get more information on those libraries please refer to their specific vignettes.

2.1.1 Target file structure

Any time a new analysis project is started, user will be asked to select the working folder, i.e. the one in which are present the .CEL files or the tab delimited file containing the expression data. In this folder should be present a tab delimited file, target file, containing all information needed to upload in the affylmGUI environment the .CEL files. Target file has a fix header and its structure is shown in figure 3. For For 3'IVT/Gene Affymetrix



	A	B	C
1	**Name**	**FileName**	**Target**
2	mC1	M1.CEL	mcf-7ctrl
3	mC2	M4.CEL	mcf-7ctrl
4	mC3	M7.CEL	mcf-7ctrl
5	mE1	M3.CEL	mcf-7E2
6	mE2	M6.CEL	mcf-7E2
7	mE3	M9.CEL	mcf-7E2
8	mI1	M2.CEL	mcf-7IGF
9	mI2	M5.CEL	mcf-7IGF
10	mI3	M8.CEL	mcf-7IGF
11	sC1	S1.CEL	sk-er3ctrl
12	sC2	S4.CEL	sk-er3ctrl
13	sC3	S7.CEL	sk-er3ctrl
14	sE1	S3.CEL	sk-er3E2
15	sE2	S6.CEL	sk-er3E2
16	sE3	S9.CEL	sk-er3E2
17	sI1	S2.CEL	sk-er3IGF
18	sI2	S5.CEL	sk-er3IGF
19	sI3	S8.CEL	sk-er3IGF

Figure 3: Basic target file.

arrays it is also available a widget to create a target file in File menu: oneChannelGUI: Create a Target file.

For time course experiments a specific target file is needed. Target file has a fix header and its structure is shown in figure 4

2.2 Affymetrix Exon arrays: probe sets summary derived from Expression Console/APT

The menu functions are summarized in figure 5

Since it is unfeasible to analyse raw exon 1.0 ST arrays data in winXP. This limitation can be overcome if probe set summary and normalization is performed using Affymetrix Expression Console `http://www.affymetrix.com/support/technical/software_downloads.affx` or APT `http://www.affymetrix.com/support/developer/powertools/index.`

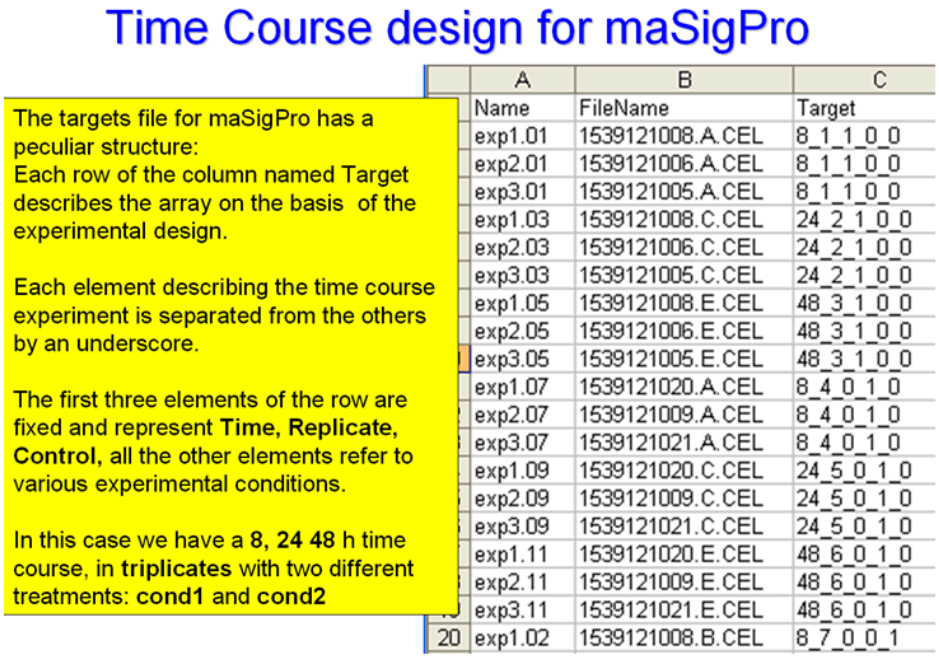


Figure 4: Target file for time course analysis.

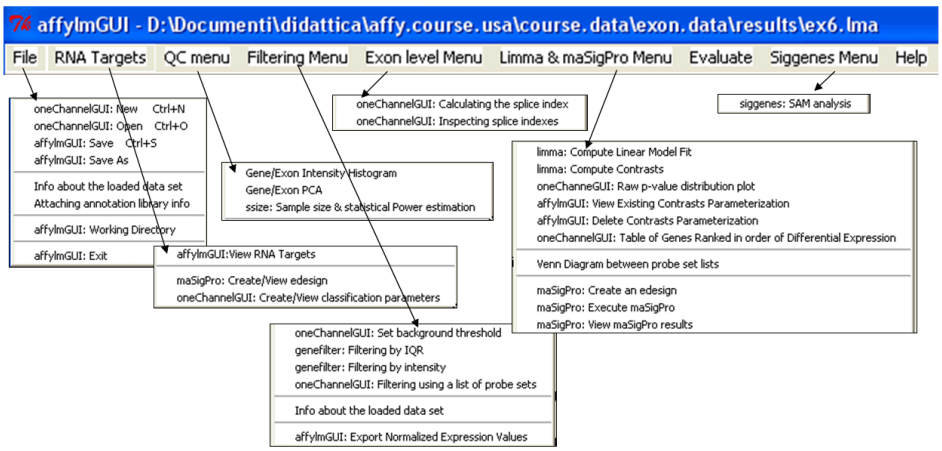


Figure 5: Exon Affymetrix microarray menu.

`affx`, for more info see oneChannelGUI user guide. Using this approach only gene/exon level secondary analysis (replicates QC by PCA, sample size evaluation, filtering, Limma and SAM analysis, maSigPro time course analysis, basic Splice Index inspection) is available. To load exon array data a target file with the same structure shown in figure 3 is needed together with two tab delimited files, with exon and gene level data, exported from the Expression Console/APT. Furthermore, the directory where exon libraries are located (.mps files) must be also provided.

2.2.1 Samples QC for Affymetrix exon arrays data

We will use a file `exon.sample.txt` which contains gene/exon level data for 291 genes extracted from Affymetrix public brain and breast exon arrays.

```
> con <- url("http://www.bioinformatica.unito.it/downloads/exon.sample.txt")
> load(con, envir = affylmGUIenvironment)
> close(con)
```

Gene expression values are available as an Expression set in:

```
> affylmGUIenvironment$NormalizedAffyData
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 291 features, 6 samples
  element names: exprs
phenoData
  rowNames: huex_wta_breast_A.CEL, huex_wta_breast_B.CEL, ..., huex_wta_cerebellum_C.C
  varLabels and varMetadata:
    Name: Name
    FileName: FileName
    Target: Target
featureData
  rowNames: 2375664, 3080033, ..., 3448088 (291 total)
  varLabels and varMetadata: none
experimentData: use 'experimentData(object)'
Annotation [1] ""
```

Exon expression values are available as an Expression set in:

```
> affylmGUIenvironment$exonAffyData
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 5025 features, 6 samples
  element names: exprs
phenoData
```

```
  rowNames: huex_wta_breast_A.CEL, huex_wta_breast_B.CEL, ..., huex_wta_cerebellum_C.C
  varLabels and varMetadata:
    Name: Name
    FileName: FileName
    Target: Target
featureData
  rowNames: 2322104, 2322105, ..., 4054672 (5025 total)
  varLabels and varMetadata: none
experimentData: use 'experimentData(object)'
Annotation [1] ""
```

Sample quality control can be evaluated using Principal component analysis (QC menu: Gene/exon PCA) or by the line command:

```
> ocPlotPCA()
```

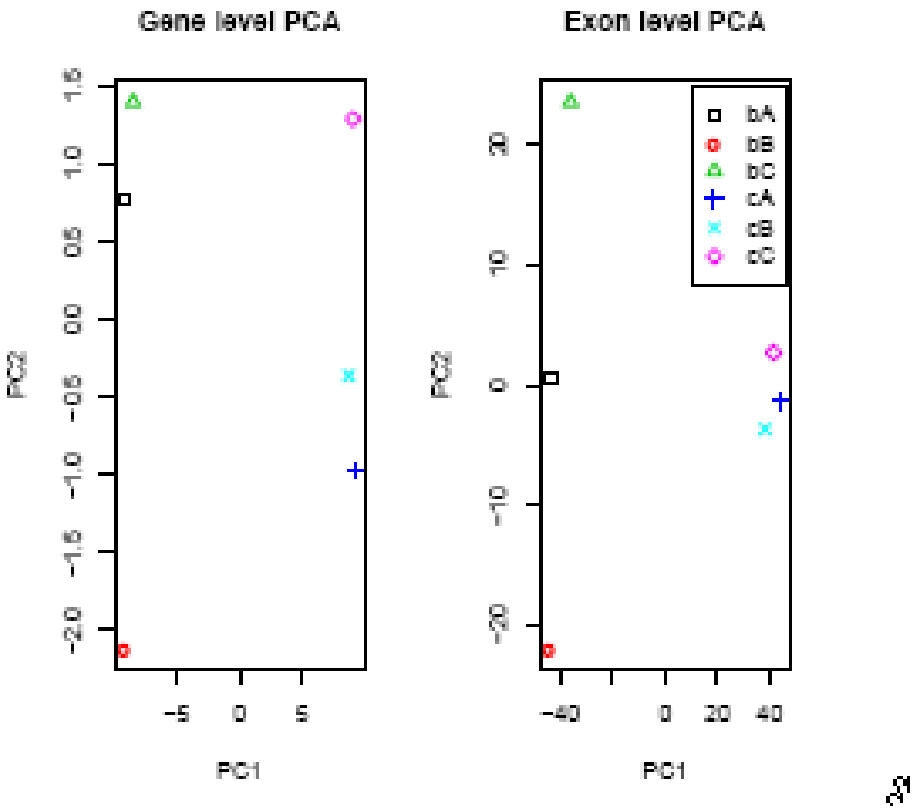


Figure 6: Samples PCA at gene and exon level

It is expectable that replicates derived from the same experimental condition will cluster together as different experimental conditions will cluster in different position of the plot (Figure 6).

2.2.2 Splice index calculation and inspection

This analysis is possible only on two groups experiment. To investigate the presence of alterative splicing, Splicing Index can be calculated using this code:

```
> spliceIndex()
```

Or using the grafical interface Exon level Menu: Calculating the splice index
Splice indexes data are saved as a data frame object in:

```
> dim(affylmGUIenvironment$spliceIndexData)

[1] 5025    6

> affylmGUIenvironment$spliceIndexData[1:3, ]

        huex_wta_breast_A.CEL huex_wta_breast_B.CEL huex_wta_breast_C.CEL
2375665             -0.084928             -0.440942             -0.459955
2375666             -0.560334             -0.135493             -0.581447
2375671             -0.448799             -0.232435             -0.062672
        huex_wta_cerebellum_A.CEL huex_wta_cerebellum_B.CEL
2375665                 -0.626948                 -0.367716
2375666                 -0.641139                 -0.627325
2375671                 -0.412696                 -0.233489
        huex_wta_cerebellum_C.CEL
2375665                 -0.506423
2375666                 -0.191048
2375671                 -0.333573
```

Spice index captures the basic metric for the analysis of alternative splicing. It is a measure of how much exon specific expression (with gene induction factored out) differs between two samples. Putative alternative splicing events are selected evaluating the presence of significant differences between any of the exons by t-test and by a certain amount of difference between splice indexes within the two different experimental conditions under analysis.

To inspect it you can run at the line command:

```
> inspecting.splice.index()
```

Or using the grafical interface Exon level Menu: Inspecting splice indexes. The result of this analysis is a pdf file containing a series of graphs helping the user to evaluate the presence of alternative splicings (see below). Actually this tool is only a gene by gene visual inspection of putative splicing events. It will be expanded as soon as new methods for massive analysis of alternative splicing identification will be available.

The splice Index inspection for a specific transcript cluster probe set can be done by.

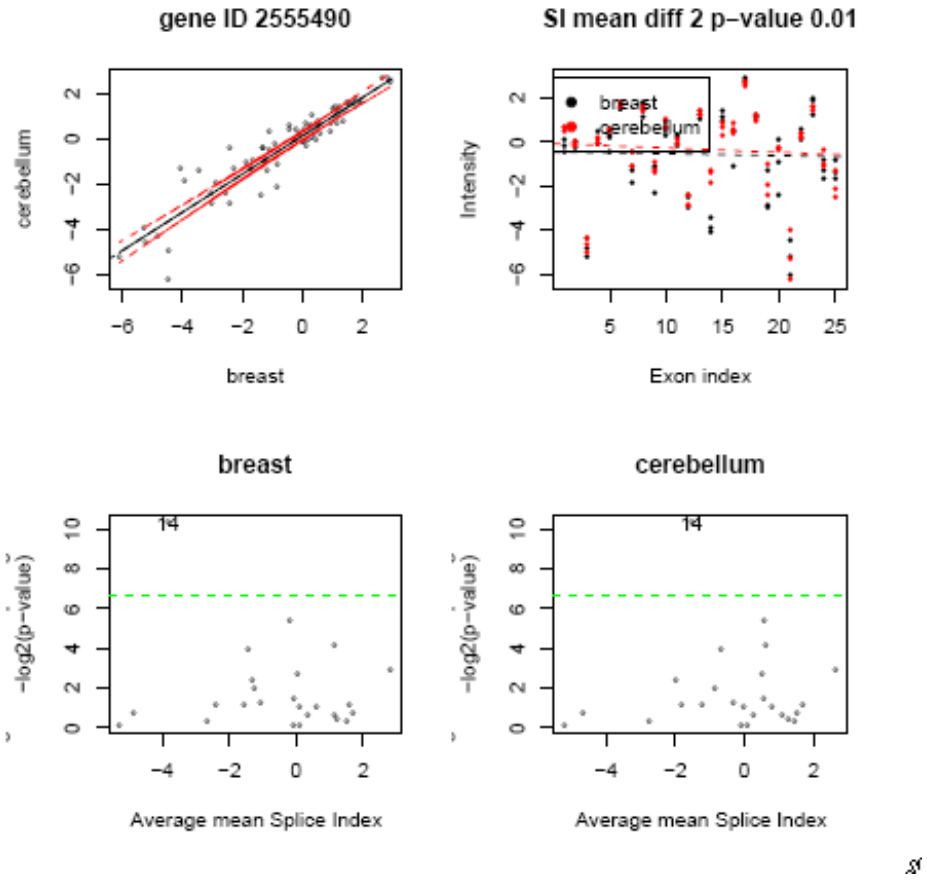


Figure 7: Splice Index inspection

```
> evalRcode()
> inspecting.one.splice.index("2555490", 2, 0.01)
```

Result of this analysis are shown in 7

Upper left panel (7) gives some advise about the scattering levels of the Splice Indexes over the gene under analysis. The basis of this approach is that the overall similarity between gene splice indexes under two different experimental conditions is relatively high unless for few regions affected by alternative splicing. This plot shows the linear model of splice indexes over the two experimental conditions. Red dashed lines indicate the confidence interval of the model. It is clear that similarity between the two condition is quite consistent if exon intensity is similar of greater of the gene level expression summary (Splice Index > 0). Instead, scattering increases if the exon signal is much lower of gene level expression summary (Splice Index < 0) Bottom panels (7) are the plots of significance p-value of the alternative splicing versus the average Splice Index values. In this example only one exon seems to be differentially spliced : exon 14. The right top panel (7) shows exons probe set intensity summaries versus exon index. The header of this panel shows the filtering conditions used (absolute splice indexes difference $= 2$ and p-value $= 0.01$). This plot is useful to check the quality of replicates associated to a detected putative alternative splicing event.

2.3 Large data set

The menu functions are summarized in figure 8

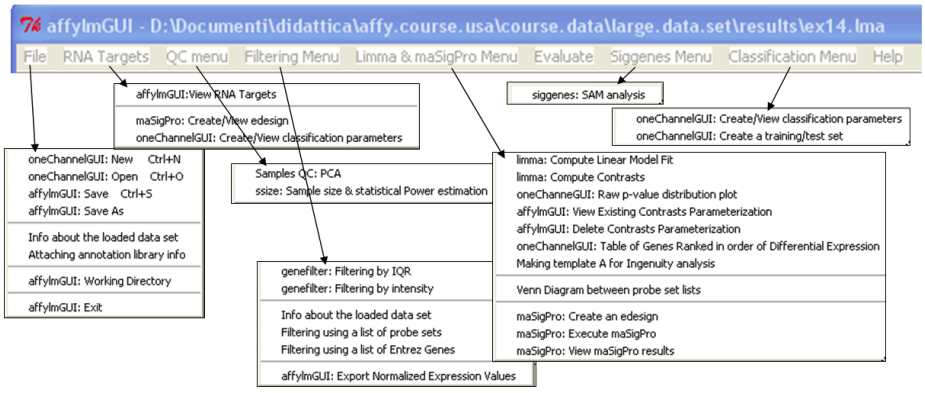


Figure 8: Large data set microarray menu.

This menu is devoted to the analysis of large data sets derived from Affymetrix Expression Console/APT/GEO/ArrayExpress data. This is particularly useful if more than 12 plus2 CEL files are loaded, since under windows environment RAM limitation can be encountered. In this menu it is only available secondary analysis (replicates QC,

filtering, statistical analysis, PAMR classification) To load a large data set a target file like that described in figure 9 is needed together with a tab delimited file containing the expression values of the microarrays.

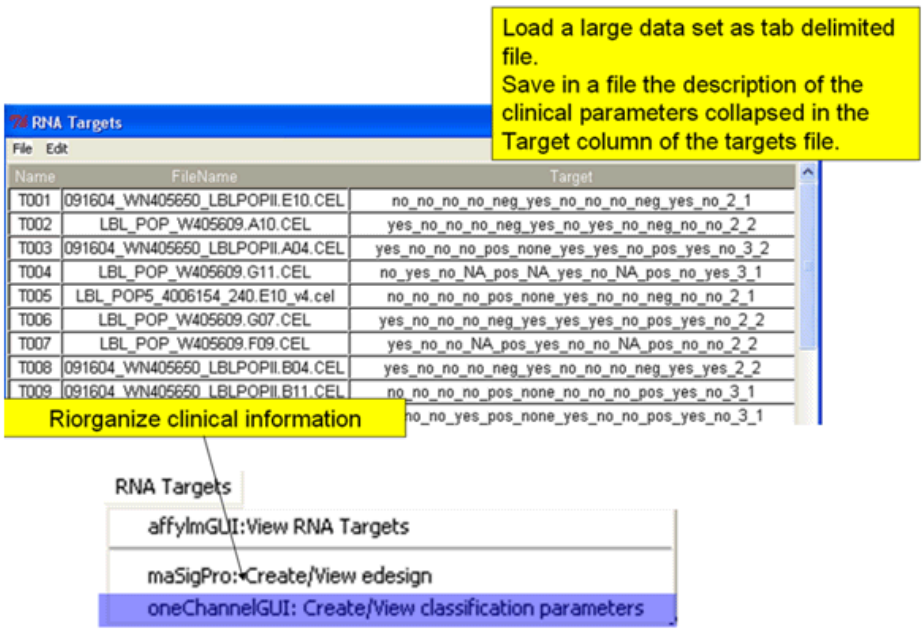


Figure 9: Target file for large data set also associated to clinical information.

More info are available at: `http://www.bioinformatica.unito.it/oneChannelGUI/`