oneChannelGUI Package: NGS secondary data analysis

Raffaele A Calogero, Francesca Cordero, Remo Sanges, Cristina Della Beffa

March 23, 2010

1 Introduction

OneChannelGUI started providing a new set of functions extending the capability of affylmGUI package. oneChannelGUI was designed specifically for life scientists who are not familiar with R language but do wish to capitalize on the vast analysis opportunities of Bioconductor. oneChannelGUI offers a comprehensive microarray analysis for single channel pplatforms. Since Next Generation Sequencing (NGS) is becoming more and more used in the genomic area, Bioconductor is extending the number of tools for NGS analysis. Therefore, we are extending the functionalities of oneChannelGUI to handle RNA-seq data with a specific focus on non-coding RNA quantitative and qualitative analysis. We focus our attention on secondary analysis of NGS data, i.e. data mapped on reference genome.

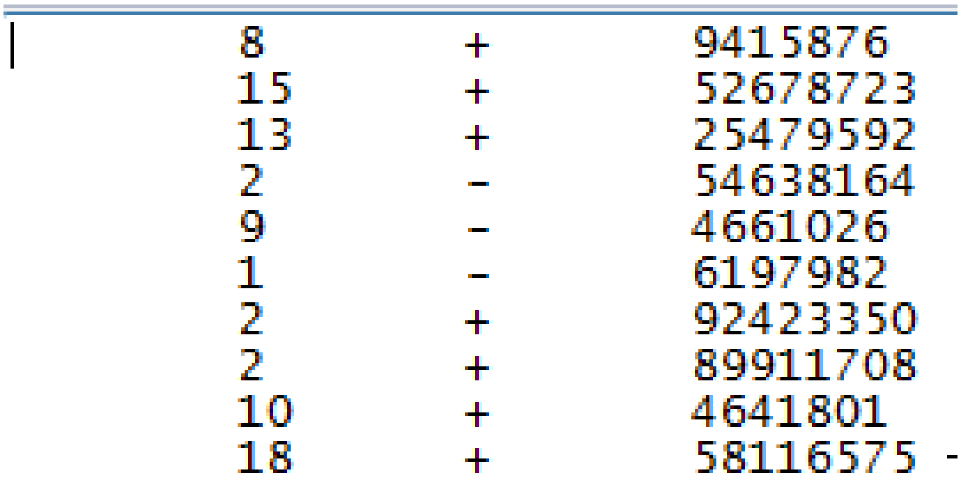
2 Non-coding RNA-seq

The present goal of oneChannelGUI is to provide a graphical interface for the secondary analysis of ncRNA. Since short mature ncRNA, i.e. microRNA, have a length in the order of 18-25 nts the direct mapping of NGS reads over the whole genome produces an enormous amount of spurious results, since in 3.3 Gbases is quite frequente to detect, by chance, a string of 18-25 nts. For this reason oneChannelGUI produces ncRNA-specific reference fasta files that can be used for efficiente mapping of mature short ncRNAs. Thhese fasta files can be retrieved using the function *Export non-coding RNA fasta reference file for ncRNA-seq quantitative analysis* available in the General Tool Menu of oneChannelGUI. These fasta file can be directly retrieved via biomarRt package using the oneChannelGUI standalone function *ncScaffold*. For more information on this function please refer to the stand alone function vignette.

The number of erroneous mapping is reduced using a specifically devoted subset of fasta sequences for ncRNA mapping and it can be done on a common 32 bits laptop. You can perform primary mapping of ncRNAs using the function provided in NGS mapping menu, UNDER DEVELOPMENT. The mapping actually uses SHRIMP with a specific implementation for microRNA detection. The output of this function are a set of files that after post-processing can be loaded in oneChannelGUI using a target file as descriptor of the loaded data.

3 File menu

Mapping tools for Illumina and SOLiD data are many and constantly increasing in number and performances. Since there is not a standard output for mapping data, their outputs can be very different and it is challenging to implement a generally applyable loading procedure. For this reason we have decided for a quite simple data structure, a tab delimited file containing three columns: chromosome number, strand and start position of the mapping. An example of the file structure is given in figure 1. Data are loaded in oneChannelGUI using the *New function* in the File Menu selecting from the menu of the available platforms NGS, figure 2. oneChannelGUI will require a target file which is a tab delimited file with three columns: Names, FileNames and Target, figure 3. The FileName column must contain the names of the files, each one belonging to a mapping experiment, with the structure previously indicated. IMPORTANT: target column contains, togeter with the covariate, also the total number of reads and the total number of mapped reads separated by an underscore.

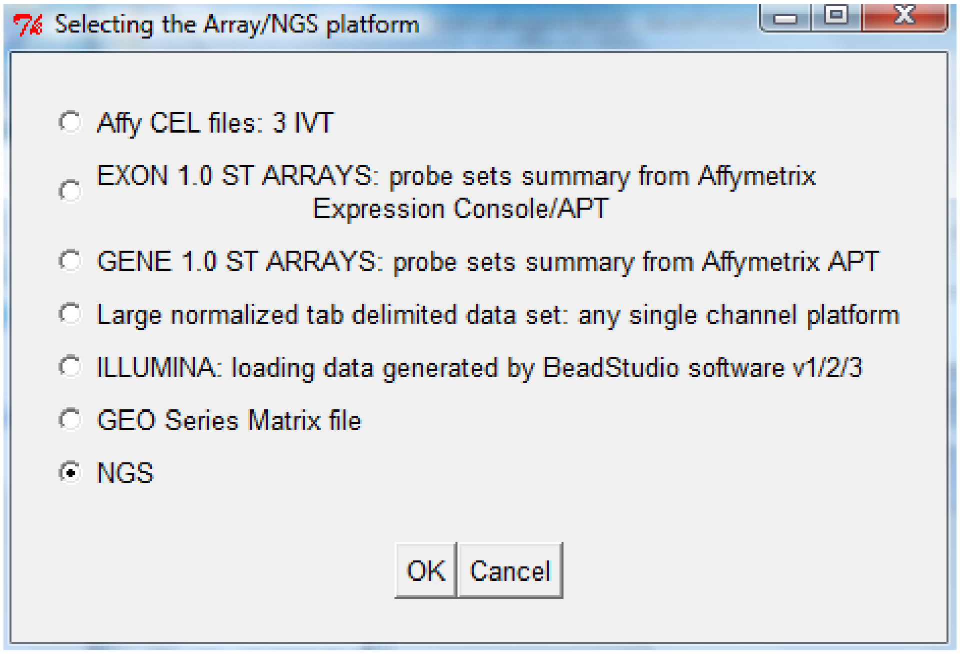


```
8	+	9415876
15	+	52678723
13	+	25479592
2	-	54638164
9	-	4661026
1	-	6197982
2	+	92423350
2	+	89911708
10	+	4641801
18	+	58116575
```

Figure 1: Structure of mapping data that can be imported in oneChannelGUI.

3.1 Reformatting NGS data

The raw data derived from a NGS experiment cannot be used directly for statistical analysis. When user loads the set of NGS runs, described by the target file, those



Selecting the Array/NGS platform

○ Affy CEL files: 3 IVT

○ EXON 1.0 ST ARRAYS: probe sets summary from Affymetrix Expression Console/APT

○ GENE 1.0 ST ARRAYS: probe sets summary from Affymetrix APT

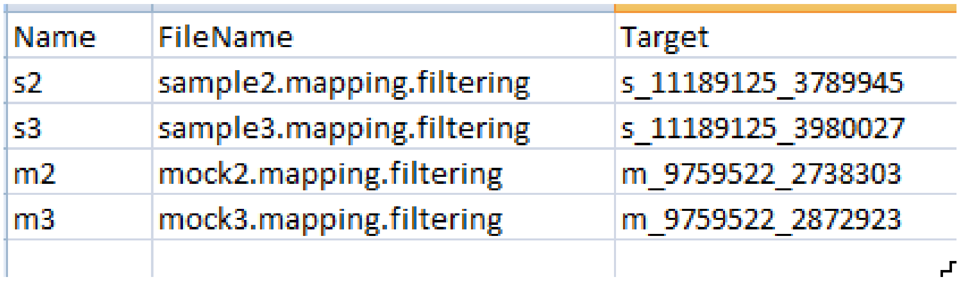
○ Large normalized tab delimited data set: any single channel platform

○ ILLUMINA: loading data generated by BeadStudio software v1/2/3

○ GEO Series Matrix file

● NGS

Figure 2: Data loading menu.



Name	FileName	Target
s2	sample2.mapping.filtering	s_11189125_3789945
s3	sample3.mapping.filtering	s_11189125_3980027
m2	mock2.mapping.filtering	m_9759522_2738303
m3	mock3.mapping.filtering	m_9759522_2872923

Figure 3: Target file structure.

data are reoganized in an ExpressionSet format, which is quite common for microarray data. To reduce RAM consumption and to allow the analysis also on conventional 32 bit computers, the raw data are loaded and stored in files on the bases of their belonging to a specific chromosome. Subsequently, the union of all counts over the all runs for a specific chromosome are used to define the peaks where the reads are clustering on the plus and minus strand. To define chromosome peaks user need to indicate the size of extension of the reads. We usually extend the reads to the real length, e.g. 35. However, extension size for ncRNA quantification values are also between 100 and 200 nts, if the size of cDNA library is considered. Library size is between 108 and 130 nts if a fractionation of small RNAs (10-40 nts) was used. In case cDNA was prepared from total RNA, which is better for quantification, the size is between 150 and 200 nts. For quantification the preparation of the library from total RNA is preferred since it is characterized by a lower inter-experiment variability. User needs also to define the number of reads that are mapped as random event, for a library between 10-20 milion 35-mer tags this value is about 8. It is important to remember that using long extension value an high number of peaks will be created since more peaks will be characterized by having a number of mapped reads greater than 8. The name of each peak is made by:

```
chr name.strand.start position-end position
e.g. chr1.plus.100000-105000
     chr2.minus.500000-500630
```

After reformatting the counts are saved in an ExpressionSet object which is stored in the affylmGUIenvironment and it is ready to be further analyzed.

4 Reformatting-Normalizing NGS data menu

The NGS data stored in the ExpressionSet can be log2 transformed. Since it is possible that some peaks are subset of the same peak, e.g. two peaks located less then 50 bases to each other, the function Refining peaks allows to merge peaks located near to each other given a user define threshold. In figure 4

5 QC menu

The section QC allows to visualized the box plot of the NGS samples after reformatting in peaks as well as samples PCA and hierarchical clustering.

6 Filtering menu

The section filtering allows remove those peaks that are little informatives, i.e. those with too little counts or too little changes between experimental conditions. Furthermore, tab delimited files containing the loaded couts can be exported.

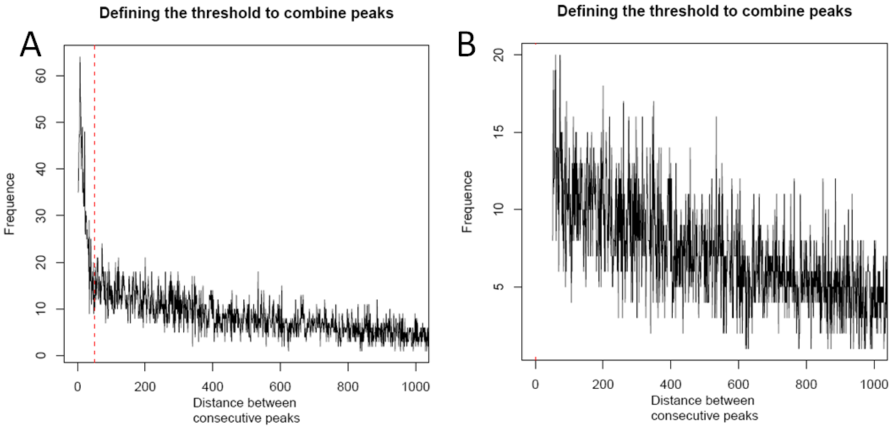


Figure 4: A) Plot of frequence of two nearby peaks versus inter-peaks distance. The dashed lines indicate a max distance defined by user, in this case 50 nts. B) Plot of the refined peaks after recursive merging of nearby peaks.

7 Statistics menu

In this section it is implemented the Rank Product method. Although this method was not specifically developed for NGS data, since it is a non parametric method, it might be useful to detect differential expression in NGS data, expecially when limited number of samples are available, as in the case of NGS data. The RP method implementation assumes a log2 transformation of data signals.

Instead, edgerInterface function provides an interface to edgeR package. edgeR provides statistical routines for determining differential expression in digital gene expression data.

8 Biological Interpretation menu

This section need to be expanded. At the ptesent time is only possible to create a tab delimited file which links ENSEMBL annotation to the peaks present in the dataset loaded in oneChannelGUI. For this annotation it is possible to select Transcription Start Site, or Exon, or miRNA.