RNASeqGUI: A GUI for analysing RNA-Seq data

Francesco Russo 1,*, Claudia Angelini 1

1Istituto per le Applicazioni del Calcolo, CNR, Napoli, Italy.

Associate Editor: Prof. Ivo Hofacker

ABSTRACT
Summary: We present RNASeqGUI R package, a graphical user interface (GUI) for the identification of differentially expressed genes across multiple biological conditions. This R package includes some well known RNA-Seq tools, available at www.bioconductor.org. RNASeqGUI package is not just a collection of some known methods and functions, but it is designed to guide the user during the entire analysis process. RNASeqGUI package is mainly addressed to those users that have little experience with command-line software. Therefore, thanks to RNASeqGUI they can conduct analogous analyses using this simple graphical interface. Moreover, RNASeqGUI is also helpful for those who are expert R-users since it speeds up the usage of the included RNASeq methods drastically.

Availability and Implementation: RNASeqGUI package needs the RGTK2 graphical library (Lawrence et al. (2010)) to run. This package is open source and is freely available under GPL licence at http://bioinfo.na.iac.cnr.it/RNASeqGUI/Download
Contact: masergui@na.iac.cnr.it

Supplementary information: A comprehensive user manual with a usage example is available at http://bioinfo.na.iac.cnr.it/RNASeqGUI

1 INTRODUCTION
There is a plethora of RNA-Seq data analysis tools available to study the difference of the gene expression across multiple biological conditions, see Soneson et al. (2013) for a review. Generally, a complete analysis requires to carry out several steps, to use different methods and to compare their outputs to obtain more reliable and less biased results. RNASeqGUI is a tool that facilitates and speeds up the exploration of the RNA-Seq data, the usage of several RNA-Seq methods and the comparison of different results. Moreover, RNASeqGUI is a modular software. This gives to the user the possibility to customize the software for a specific type of study.

1.1 Other GUIs and objectives of RNASeqGUI
Several bioinformatics tools, such as Wettenhall et al. (2004, 2006); Sanges et al. (2007); Angelini et al. (2008); Lohse et al. (2012); Pramana et al. (2013), have been implemented as user friendly graphical interfaces to provide point and click access to sophisticated data analysis. Among them, RNASeqGUI has a clear focus on the analysis of RNA-seq data. Analogous focus, but with different functionalities, is present in Lohse et al. (2012). More recently, Sanges et al. (2007) has been extended to RNA-Seq data analysis as well. With respect to these interfaces, RNASeqGUI has some overlaps, but also several unique features to be considered a useful and valid alternative.

The design of RNASeqGUI main interface is inspired to that one presented in Angelini et al. (2008). It tries to be intuitive and to guide the user through the RNA-Seq data analysis. In order to meet this goal, the main interface (described in detail in the rest of the paper) is organized into several different sections/interfaces (see Fig. 1), each of them devoted to a specific stage of the analysis. Usability is enhanced thanks to the presence of numerous explanatory vignettes. Moreover, RNASeqGUI is designed to facilitate the extensibility thanks to its software development organization. In fact, it is extremely easy to add new buttons that calls new functionalities. Therefore, a user can customize RNASeqGUI interfaces for his own purposes and benefits by adding the methods he needs mostly (for more details see How to customize RNASeqGUI: Adding a new button in just three steps of the user manual).

2 STRUCTURE OF RNASEQGUI MAIN INTERFACE
RNASeqGUI R package was implemented by following and expanding the idea presented in Villa-Vialaneix et al. (2013). Its main interface is divided into five Sections (see Fig. 1). Each section corresponds to a particular step of the RNA-Seq data analysis workflow and includes one or more Graphical Interfaces. Inside each interface, there are several available functions (also called methods). Each function takes specific inputs that can be numeric ones, strings or both and produces one or more outputs that can be plots, text files or both. Each analysis starts by creating or retrieving a specific project. Despite the fact that each function can be executed by simply clicking the corresponding button, each project a detailed history of all performed steps is automatically saved in a report file.

In the following, we briefly describe all sections of RNASeqGUI main interface.

2.1 Bam Exploration Section
In the first section of RNASeqGUI main interface, we find the Bam Exploration Interface that can be easily called by clicking the corresponding button (see Fig. 1). This interface includes five different methods to explore the alignment files (in bam format), such as: Read Counts, Mean Quality of the Reads, Per Base Quality of Reads, Reads Per Chromosome, Nucleotide Frequencies. Each
of these functions takes as input a folder containing all the bam files that the user wants to explore. Usually, each bam file corresponds to a sample. This section is important to discover possible errors that may have occurred either during the alignment step or during the experimental steps (i.e., sequencing, PCR, extraction of RNA-sequences, etc).

2.2 Count Section
In the second section of RNASeqGUI, we find the Read Count Interface that gives the possibility to perform the quantification process against an annotation file in GTF format. It works similarly to htseq-count (www-huber.embl.de/users/anders/HTSeq). The Count Reads button, inside this interface, calls summarizeOverlaps function from the package GenomicRanges (Lawrence et al. (2013)). It can be used in three different modes (Union, IntersectionStrict and IntersectionNotEmpty) and returns a table of counts, where the first column represents the gene names, while the remaining columns correspond to the names of the bam files. Rows report the number of reads that have hit a particular gene in the given sample. Read counting can be a computationally expensive process, especially for large experiments with several samples and big alignment files. The R environment is not optimized for this particular task. Therefore, this procedure makes use of bplapply function of the BiocParallel package Morgan et al. (2014) to parallelize the code in order to reduce the execution time.

2.3 Pre-Analysis Section
In the third section of RNASeqGUI, there are two interfaces: Data Exploration Interface and Normalization Interface. Both interfaces take an input count file that must be tab-delimited as those provided by Count Reads function as output, with rows representing genes names and the columns representing the samples. Data Exploration Interface: This interface includes twelve methods, such as: Plot Pairs of Counts, Plot all Counts, Count Distr, Density, MDPlot, MeanVarPlot, Heatmap, PCA, PCA3D, Component Histogram, QplotHistogram, Qplot Density. This interface uses several functions defined in Risso et al. (2011) and it is crucial to find out possible biases that could affect the RNASeq-experiment and provide useful diagnostic figures to decide whether a normalization procedure is needed. Normalization Interface: This interface includes four normalization procedures, such as: RPKM (Mortazavi et al. (2008)), Upper Quantile (Bullard et al. (2010)), TMM (Robinson et al. (2010)), Full Quantile (Smyth et al. (2005), Bolstad et al. (2003)).

2.4 Data Analysis Section
This section is the core of RNASeqGUI and contains the Data Analysis Interface. This interface includes five different statistical methods to identify differentially expressed genes, such as: edgeR (Robinson et al. (2010), Robinson et al. (2007), Robinson et al. (2008), McCarthy et al. (2012)), DESeq, DESeq2 (Anders et al. (2010)), NOISeq (Tarazona et al. (2011)), baySeq (Hardcastle et al. (2010)). Each method takes an input count file and returns two text files and one or more plots. The first text file shows the results of the chosen method, while the second text file shows the differentially expressed genes only.

2.5 Post Analysis Section
This section includes two interfaces: Result Inspection Interface and Result Comparison Interface. Result Inspection Interface: This interface includes the possibility to generate volcano plots, fold change plots and histograms of the FDRs or Pvalues for each methods in order for the user to explore the results. It is also possible to display a specific gene of interested inside the volcano or the fold change plot. All generated plots are automatically saved in pdf format. Result Comparison Interface: This interface includes the possibility to generate venn diagrams and text files that show those genes that have been identified as differentially expressed by the chosen method, while the second text file shows the differentially expressed genes only.

3 USAGE EXAMPLE
In this usage example, we start the analysis of the RNASeq data from alignment files and we compare the results of edgeR, DESeq and NOISeq among them. We analysed the dataset published by Brooks et al. (2011) and used in Anders et al. (2013) as a real data working example. We selected the chromosome 2L only to reduce the execution time. Aligned data (bam files) are available at http://bioinfo.na.iac.cnr.it/RNASeqGUI/Example. We analysed the expression of 2986 genes belonging to 2L chromosome. The methods found 128, 148,
102 DE gene respectively. Among these 86 genes were found DE in all the three used methods (for more details, see the user manual).

4 CONCLUSION AND FUTURE WORKS

RNASeqGUI is an R package that allows the quick usage of several methods to identify the differentially expressed genes from RNASeq experiments. In future, we will include new methods, new normalization procedures, the possibility to define more complex experimental designs, the pathway analysis and the Gene Ontology.

ACKNOWLEDGEMENTS

We are thankful to reviewers for corrections and suggestions that improved this work substantially and to M. Franzese, V. Costa and R. Esposito for suggestions and discussions, D. Granata for technical support.

Funding: This work was supported by the Italian Flagship InterOmic Project (PB.P05) and by BMBS COST Action BM1006.

REFERENCES