BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2514–2516
doi:10.1093/bioinformatics/btu308

Gene expression

Advance Access publication May 7, 2014

RNASeqGUI: a GUI for analysing RNA-Seq data

Francesco Russo* and Claudia Angelini

Istituto per le Applicazoni del Calcolo, CNR, 80131, Napoli, Italy

Associate Editor: 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 who 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 because it speeds up the usage of the included RNASeq methods drastically.

Availability and implementation: RNASeqGUI package needs the RGTK2 graphical library to run. This package is open source and is freely available under General Public License at http://bioinfo.na.iac.cnr.it/RNASeqGUI/Download.

Contact: f.russo@na.iac.cnr.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 10, 2013; revised on March 27, 2014; accepted on April 27, 2014

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 and Delorenzi (2013) for a review. Generally, a complete analysis requires carrying out several steps, using different methods and comparing 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 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 (Angelini *et al.,* 2008; Lohse *et al.,* 2012; Pramana *et al.,* 2013; Sanges *et al.,* 2007; Wettenhall and Smyth, 2004; Wettenhall *et al.,* 2006) 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. To meet this goal, the main interface (described in detail in the rest of the article) is organized into several different *sections*/*interfaces* (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 call 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 and Leroux (2013) and using the RGTK2 graphical library (Lawrence and Temple Lang, 2010). Its main interface is divided into five *sections* (Fig. 1). Each section corresponds to a particular step of the RNA-Seq data analysis work-flow 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, for 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 (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,

*To whom correspondence should be addressed.

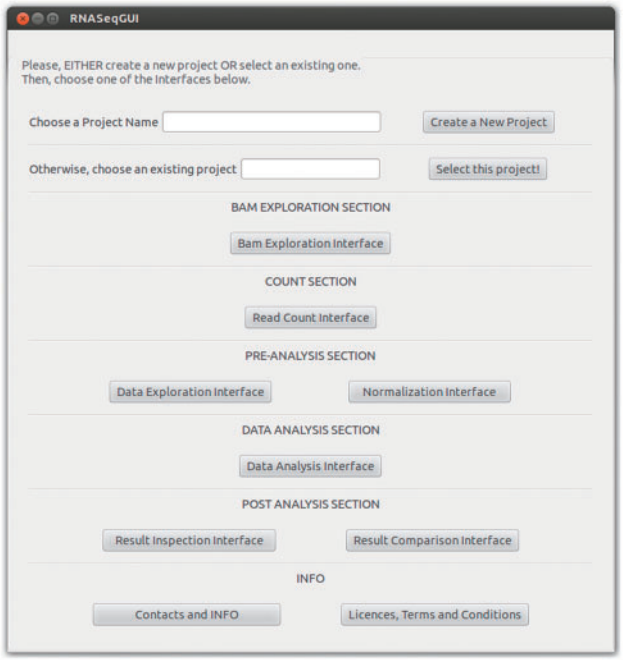

Fig. 1. RNASeqGUI main interface. It appears on the screen after typing library(RNASeqGUI) and RNASeqGUI() in an R environment

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).

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 gene transfer format (*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 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 12 methods, such as Plot Pairs of Counts, Plot all Counts, Count Distr, Density, MDPlot, MeanVarPlot, Heatmap, Principal Component Analysis (PCA), PCA3D, Component Histogram, QplotHistogram and 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 reads per kilobase per million mapped reads (RPKM) (Mortazavi *et al.*, 2008), Upper Quartile (Bullard *et al.*, 2010), trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010), Full Quantile (Bolstad *et al.*, 2003; Smyth, 2005).

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 (DE) genes, such as edgeR (McCarthy *et al.*, 2012; Robinson *et al.*, 2007, 2008, 2010), DESeq, DESeq2 (Anders and Huber, 2010), NOISeq (Tarazona *et al.*, 2011), baySeq (Hardcastle and Kelly, 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, whereas 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 false discovery rates (FDRs) or *P*-values for each method for the user to explore the results. It is also possible to display a specific gene of interest 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 methods used in the Data analysis section.

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 genes, respectively. The intersection of these three gene sets contains 86 genes. Hence, all the three methods used found 86 DE genes in common (for more details, see the Supplementary data).

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

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

Funding: Italian Flagship Project "InterOmics" Project (PB.P05), by BMBS COST Action BM1006 and by PON01_02460 (DIAGEN_SC).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765–1786.

Angelini,C. *et al.* (2008) BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, **9**, 415.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Brooks,A.N. *et al.* (2011) Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, **21**, 193–202.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, **11**, 94.

Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *Bioinformatics*, **11**, 422.

Lawrence,M. and Temple Lang,D. (2010) RGtk2: a Graphical User Interface Toolkit for R. *J. Stat. Softw.*, **37**, 1–52.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**e1003118.

Lohse,M. *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNASeq-based transcriptomics. *Nucleic Acid Res.*, **40**, W622–W627.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

Morgan,M. *et al.* (2014) BiocParallel: bioconductor facilities for parallel evaluation. R package version 0.4.1, http://www.bioconductor.org/packages/release/bioc/html/BiocParallel.html.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

Pramana,S. *et al.* (2013) neaGUI: an R package to perform the network enrichment analysis (NEA). R package version 1.0.0, http://www.bioconductor.org/packages/release/bioc/html/neaGUI.html.

Risso,D. *et al.* (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 1–480.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson,M.D. *et al.* (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Robinson,M.D. *et al.* (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.

Sanges,R. *et al.* (2007) oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. *Bioinformatics*, **23**, 3406–3408.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, e91.

Tarazona,S. *et al.* (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2222.

Villa-Vialaneix,N. and Leroux,D. (2013) sexy-rgtk: a package for programming RGtk2 GUI in a user-friendly manner. In: *Proceedings of: 2mes rencontres R.*

Wettenhall,J.M. *et al.* (2006) affylmGUI: a graphical user interface for linear modeling of single channel microarray data. *Bioinformatics*, **22**, 897–899.

Wettenhall,J.M. and Smyth,G.K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, **20**, 3705–3706.