Analyzing RNA-seq data with RNASeqGUI

Francesco Russo and Claudia Angelini
Istituto per le Applicazioni del Calcolo–CNR
Laboratory of Statistics and Computational Tools for Bioinformatics,
Napoli, Italy
http://bioinfo.na.iac.cnr.it/BioinfoLab/

Motivations

- **RNA-seq** has quickly become one of the preferred and most widely used approaches for whole transcriptome analysis
- Analyzing RNA-seq data usually requires to carry out several steps, to use different methods and to compare their outputs to obtain more reliable and less biased results.
- Automatic pipelines are very useful, but they are not interactive, neither are easily customizable by non expert users
- Many tools are available in the literature → most of them require the user to be familiar with command-lines and/or some programming languages
- **R** and **Bioconductor** constitute precious resources in terms of open software tools for RNA-seq data analysis

RNASeqGUI

RNASeqGUI is a graphical user interface that facilitates RNA-seq data exploration and analysis. More in detail,

- It is a novel **R-package** (open-souce) that implements a graphical user interface for the **DE** analysis of RNA-Seq data.
- It includes several well known RNA-Seq tools, available at **www.bioconductor.org**. (RNASeqGUI will be soon available there).
- In the spirit of **reproducible research,** it generates a text report of all steps performed during the analysis of a specific project.
- RNASeqGUI has the following features: **High usability, portability, customizability**.

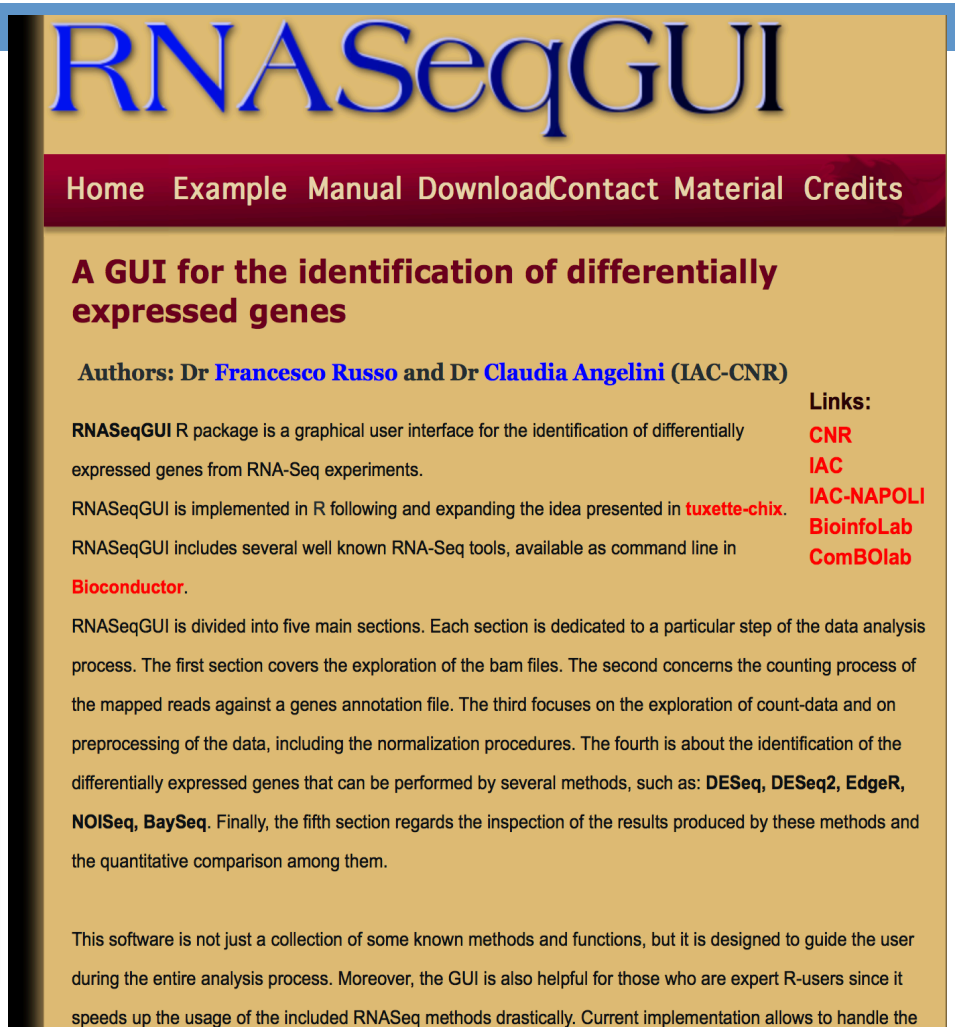
Moreover, the interface is not just a collection of some known methods, but it is designed to guide the user during the entire analysis process.

RNASeqGUI

- RNASeqGUI is implemented in **R**.
- It requires the **RGTK2** graphical Library to run
- It uses **BiocParallel** to speed up the computations.

RNASeqGUI can be downloaded from http://bioinfo.na.iac.cnr.it/RNASeqGUI/ (soon also from Bioconductor)

F. Russo and C. Angelini. **RNASeqGUI: A GUI for analysing RNA-Seq data,** To appear *Bioinformatics*, (2014)


RNASeqGUI

Home Example Manual Download Contact Material Credits

A GUI for the identification of differentially expressed genes

Authors: Dr Francesco Russo and Dr Claudia Angelini (IAC-CNR)

Links:
CNR
IAC
IAC-NAPOLI
BioinfoLab
ComBOlab

RNASeqGUI R package is a graphical user interface for the identification of differentially expressed genes from RNA-Seq experiments.

RNASeqGUI is implemented in R following and expanding the idea presented in tuxette-chix.

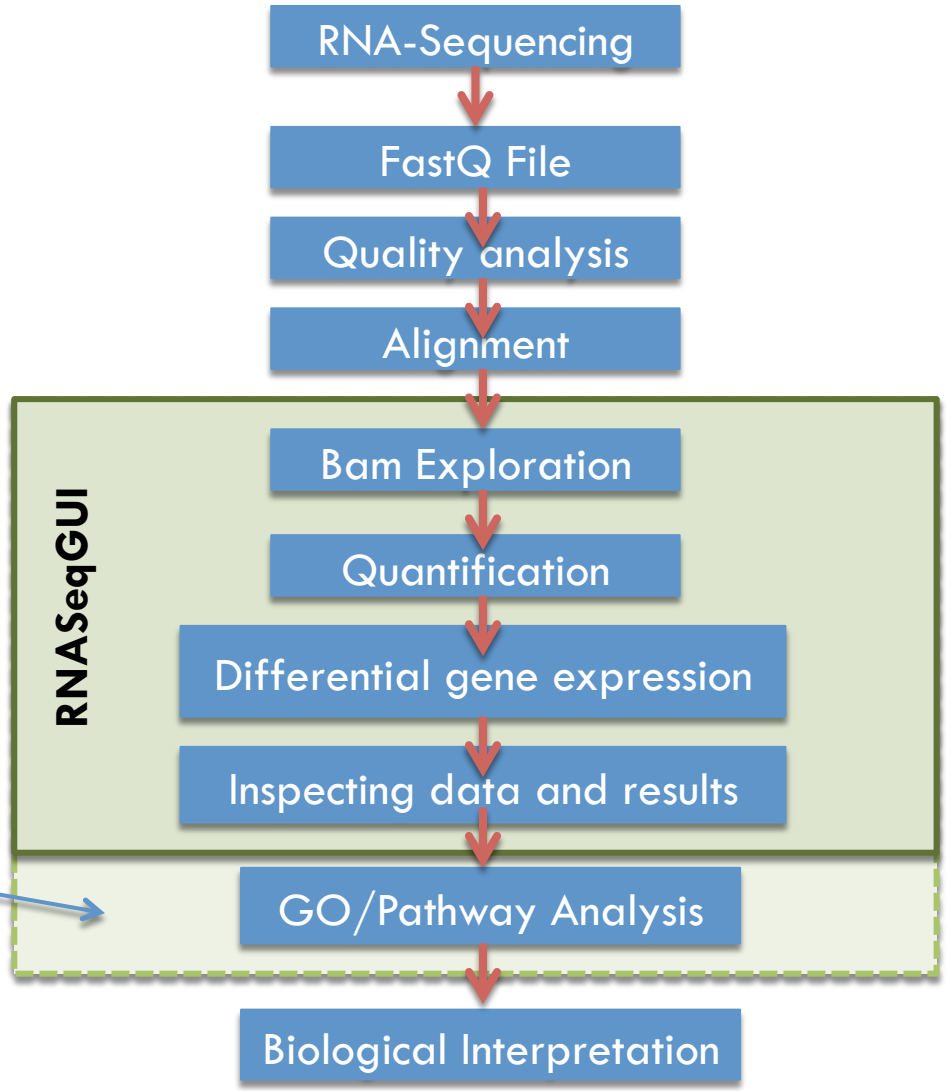
RNASeqGUI includes several well known RNA-Seq tools, available as command line in Bioconductor.

RNASeqGUI is divided into five main sections. Each section is dedicated to a particular step of the data analysis process. The first section covers the exploration of the bam files. The second concerns the counting process of the mapped reads against a genes annotation file. The third focuses on the exploration of count-data and on preprocessing of the data, including the normalization procedures. The fourth is about the identification of the differentially expressed genes that can be performed by several methods, such as: **DESeq, DESeq2, EdgeR, NOISeq, BaySeq**. Finally, the fifth section regards the inspection of the results produced by these methods and the quantitative comparison among them.

This software is not just a collection of some known methods and functions, but it is designed to guide the user during the entire analysis process. Moreover, the GUI is also helpful for those who are expert R-users since it speeds up the usage of the included RNASeq methods drastically. Current implementation allows to handle the

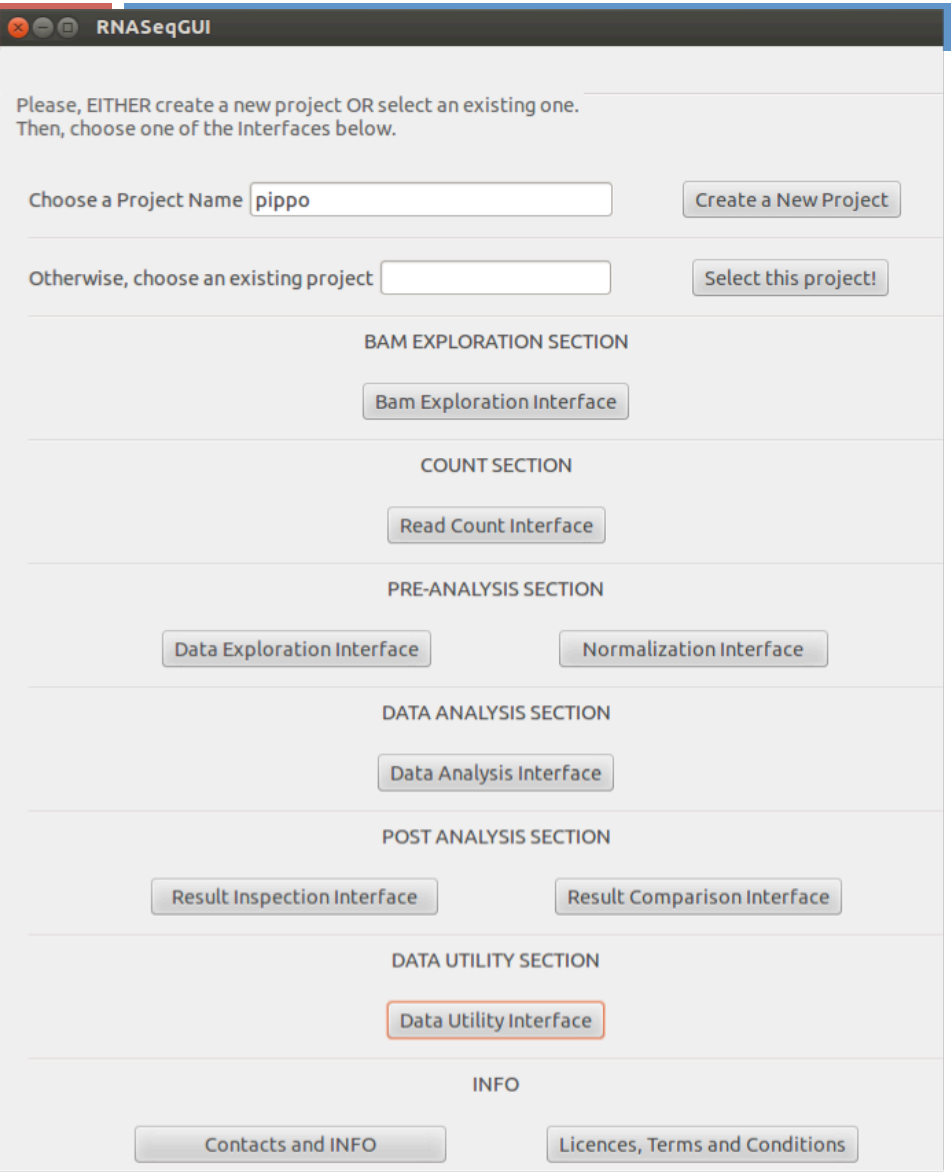
For information about pre-releases or if you require specific functionality, please contact us.

RNASeqGUI workflow



Available in next release (June 2014)

RNASeqGUI Main Interface



The GUI is divided into several sections. Each section is dedicated to a particular step of the data analysis process.

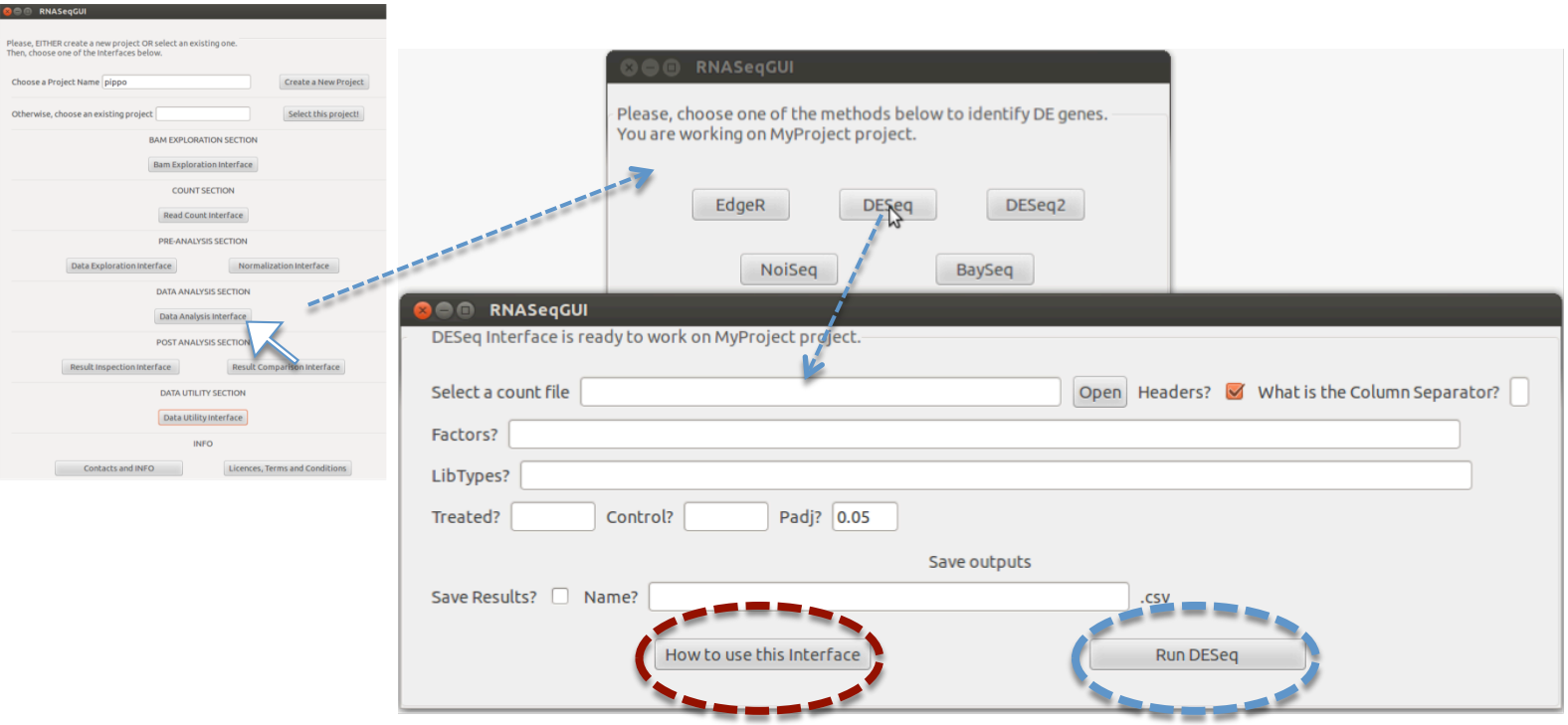
The analysis starts by creating a project or opening an existing project.

Then, the user can access any of RNASeqGUI sections.

Data Analysis Section is the core of RNASeqGUI and contains several methods to identify differentially expressed genes (DE).

Navigating RNASeqGUI

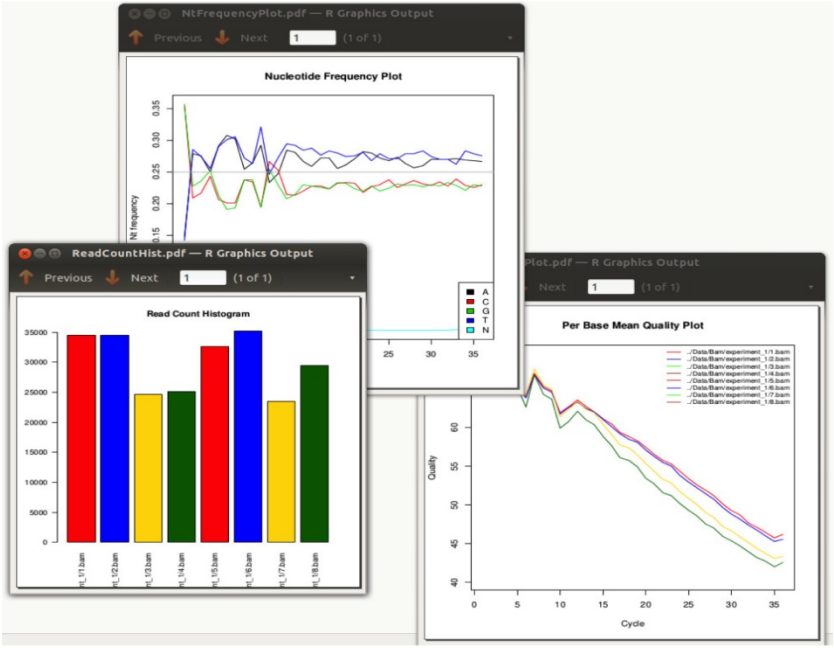
- By clicking to any specific section a new interface, that contains more specific functions, will open.



- The "*how to use*" button in each interface will guide the user in the choice of the best options and parameters.

BAM Exploration Section

- This interface includes five different methods to explore the alignment files (in bam format): Show Read Counts, Mean Quality of the Reads, Quality of Reads, Reads Per Chromosome, Nucleotide Frequencies.



- This section is important to discover possible errors that may have occurred during the alignment step or during the experimental steps.

Count Section

- This interface gives the possibility to perform the **gene quantification** process (against an annotation file in GTF format).
- It works similarly to HTSeq. i.e., it can be used in three different modes (Union, IntersectionStrict and IntersectionNotEmpty)

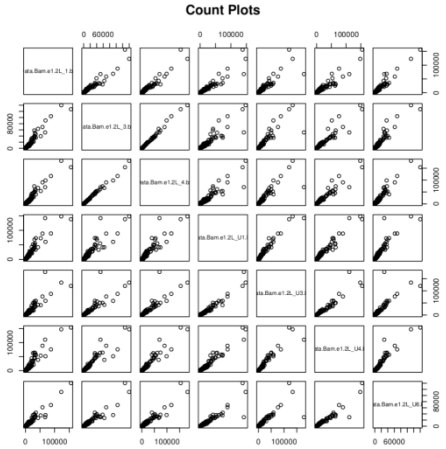
Gene Id	control_1	control_2	treated_1	treated_2
ENSG00000000003	455	463	583	598
ENSG00000000005	0	0	0	1
ENSG00000000419	1174	1210	1545	1533
ENSG00000000457	260	256	305	349
ENSG00000000460	550	607	709	741
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

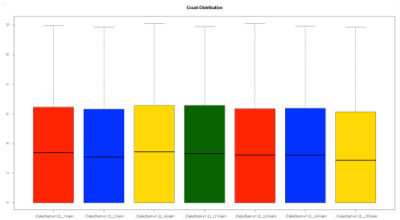
- The "counting" process is realized using the R function **summarizeOverlaps** in the GenomicRanges package that is relatively slow. Current pre-release also includes **featureCounts** that is much more fast and less computationally demanding

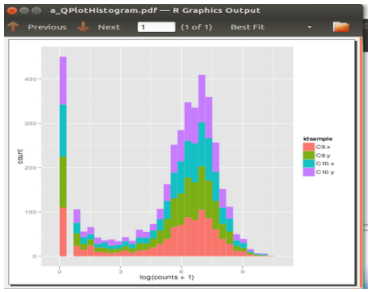
Pre-Analysis Section

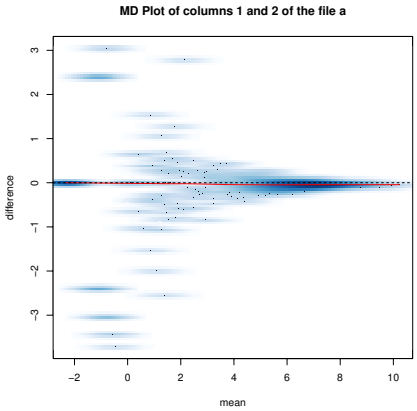
- It is divided in two panels: **Data exploration** and **Data Normalization**

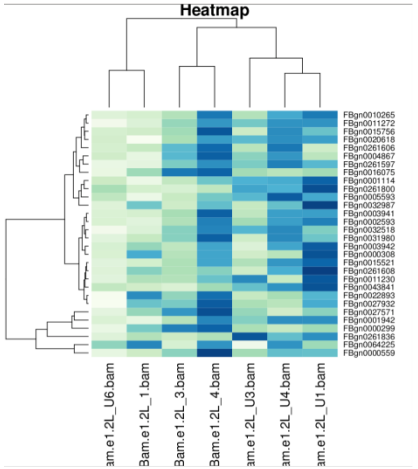
Data exploration





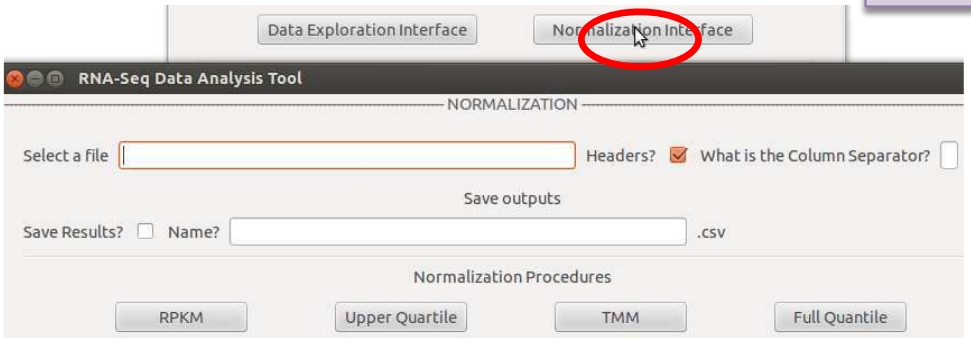






Plots can be customized in terms of colors/ scale/… etc

Data normalization

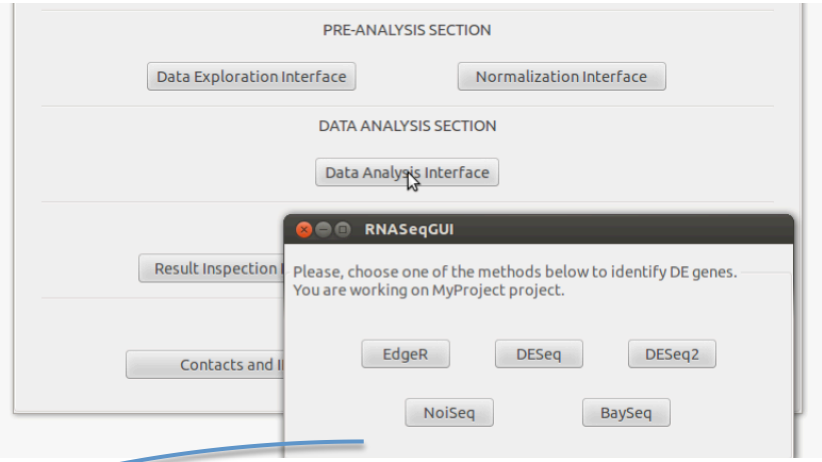


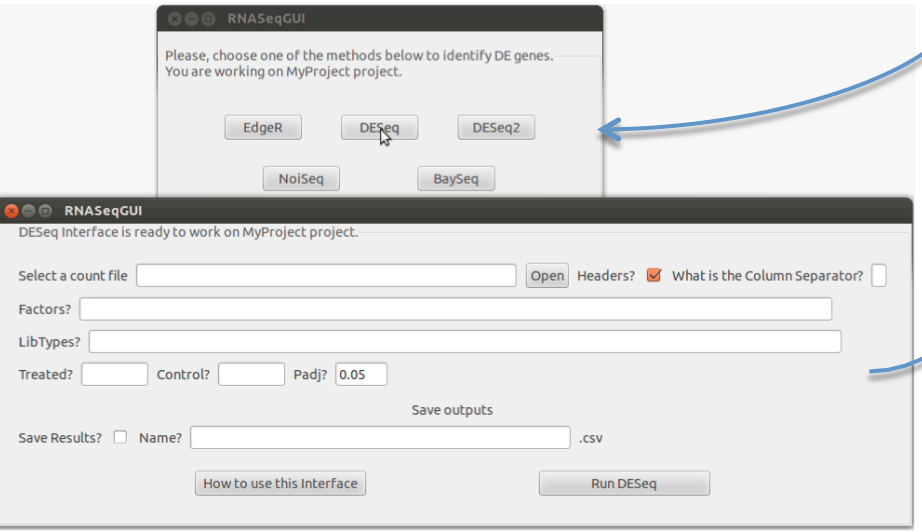
- ✓ Current normalization methods include **RPKM, Upper Quantile, TMM, Full Quantile.**
- ✓ Future release will also include those ones from **TCC R Package**

Data Analysis Section

The Data analysis interface contains several methods for detecting differentially expressed genes

- ✓ Current release includes EdgeR, DESeq, DESeq2, NOISeq, BaySeq
- ✓ Future release will include also EBSeq; Characteristic Direction DEG, as well as others





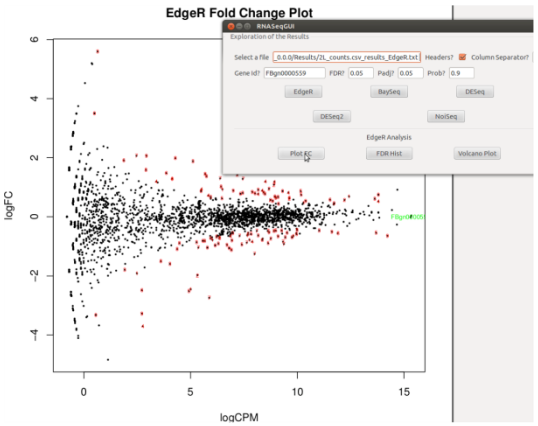
id	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
ENSG...0003	625.025	630.902	619.147	0.981	-0.027	0.774	1
ENSG...0005	0.264	0.528	0	0	-Inf	0.985	1
ENSG...0419	1106.882	1136.118	1077.646	0.948	-0.076	0.297	0.935
ENSG...0457	367.367	362.361	372.374	1.027	0.039	0.744	1
ENSG...0460	617.493	618.055	616.931	0.998	-0.002	0.982	1
....	.....	.....	.....	.....	.....	.....	...
....	.....	.....	.....	.....	.....	.....	...

Tab delimited tables of all genes under investigation and subset of differentially expressed genes are automatically generated

Post-Analysis Section

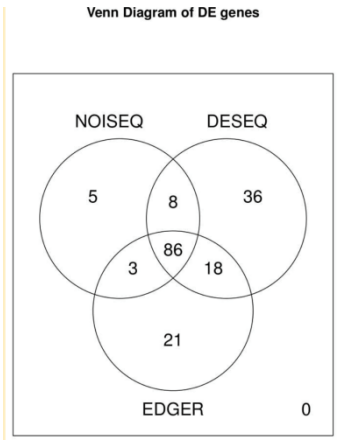
- It is divided in two panels: **Result Exploration** and **Result Comparison**

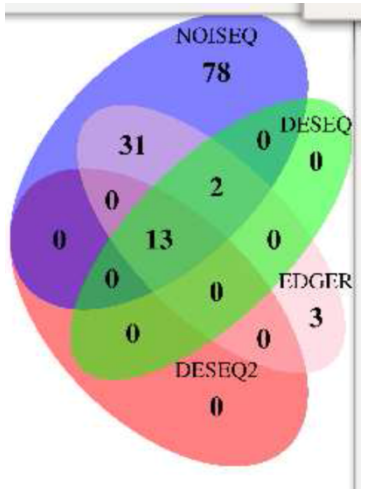
Result exploration



Plots can be customized in terms of colors/scale/… etc

Result comparison





Utility Section

It is a novel section devoted to **general purpose function**.

- ✓ Current pre-release contains functions for combining and for filtering tables
- ✓ Future releases will include also the possibility to incorporate results from **Cufflinks** (2.2.0). → It will be then possible to use functions for exploring and comparing results, including those obtained from Cuffdiff.

Adding new functions in RNASeqGUI

- RNASeqGUI is easily highly customizable→ It is possible to add a new button to the interface in only **3** steps.

RNASeqGUI & Reproducible Research

- Reporting results and steps of data analyses in a reproducible manner is of fundamental importance, although often neglected in many research paper.
- The need for **RR** increases dramatically as data analyses become more complex, involves larger datasets and more sophisticated computations, such as in the case of RNA-seq data analysis
- GUIs do not easily facilitate RR, since results are obtained after clicking several buttons → difficult to keep track of all performed steps
- To this purpose RNASeqGUI automatically generates a txt report of all analysis carried out on a given Project.
- The report includes all versions of the R packages used, all steps, input/output parameters, file names and so on.
- Future releases will make use of **R markdown** documents for RR

Conclusion and Future Work

- **RNASeqGUI** is graphical user interface (GUI) for the identification of DE genes across multiple biological conditions
- It mainly devoted to user that have limited experience with command-line software
- It is designed to work with standard PCs or small workstations – to the analysis of small/ moderate scale projects
- It is also helpful for those who are expert R-users, since it speeds up the usage of the included RNASeq methods drastically.
- It is very easy to add new functions
- For each project it provides automatically the text report of all actions performed on the dataset (in the spirit of RR)

Future versions of **RNASeqGUI** will include

- new methods for DE
- new normalization procedures
- Complex experimental designs
- **Pathway analysis**
- **Gene Ontology.**

Preliminary version available also for based on Shiny (do not require GTK2)
http://www.rstudio.com/shiny/

ACKNOWLEDGEMENTS

Francesco Russo, PhD

The main RNASeqGUI developer

Thanks to all members of

For useful discussions, debugging and suggestions

Supporting Projects

THANK YOU FOR THE ATTENTION

Questions?