ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ

ΙΝΣΤΙΤΟΥΤΟ ΧΗΜΙΚΗΣ ΒΙΟΛΟΓΙΑΣ

**ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**

**ΒΙΟΕΠΙΧΕΙΡΕΙΝ**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Ανάπτυξη ροών εργασιών lncRNA στην πλατφόρμα Galaxy

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:**
**Νικόλαος Μπαλατσός, Παν. Θεσσαλίας**

**ΣΥΜΒΟΥΛΟΙ: Αριστοτέλης Χατζηιωάννου, Ελευθέριος Πιλάλης, e-NIOS**

**Πασάτ Μπιάνκα Αλεξάνδρα**

**Αθήνα, 2023**

UNIVERSITY OF THESSALY

SCHOOL OF HEALTH SCIENCES

DEPARTMENT OF BIOCHEMISTRY AND BIOTECHNOLOGY
NATIONAL HELLENIC RESEARCH FOUNDATION

INSTITUTE OF CHEMICAL BIOLOGY

# INTERISTITUTIONAL PROGRAM OF POSTGRADUATE STUDIES

# IN


## MASTER THESIS

# "Development of lncRNA analysis workflows with Galaxy platform"

**Supervisor: Nikolaos Balatsos, University of Thessaly**

**Advisors: Aristotelis Chatziioannou, Eleutherios Pilalis, e-NIOS**

**Bianca Alexandra Pasat**

**Athens, 2023**

**ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ**

| **ΟΝΟΜΑΤΕΠΩΝΥΜΟ** | **ΒΑΘΜΙΔΑ** | **ΥΠΟΓΡΑΦΗ** |
|---|---|---|
| (Επιβλέπων/) Μπαλατσός Νικόλαος | | |
| (Μέλος 1) Γεωργιάδης Παναγιώτης | | |
| (Μέλος 2) Γιακουντής Αντώνιος | | |

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στο

**ΒΙΟΕΠΙΧΕΙΡΕΙΝ**

που απονέμει το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, σε συνεργασία με *χώρος εκπόνησης της διπλωματικής εργασίας* (αν υπάρχει).

Εγκρίθηκε την …….. από την τριμελή εξεταστική επιτροπή:

# Ackownledements

This master thesis was created as a result of my work during my internship period at eNIOS Aplications PC. I would like to acknowledge eNIOS and all its members for their help and constructive collaboration. I would like to thank each of you Chara Mastrokalou, Theodwro Koutsandrea, Elena Gkotsi for the time we spent together and for your great help and especially I would like to thank Aristotelis Chatziioannou and Eleftherios Pilalis who gave me the opportunity to work with them and conduct with them my master's project and conceived the idea of the lncRNA workflow. This work would not have been possible without them. Special thanks to my supervisor, Nikolaos Balatsos, who guided me and helped me throughout the completion of my thesis.

I would also like to thank my family who supported me tirelessly all these years and stood by every decision I made. Thank you!
Special thanks to my grandfather. I will miss you everyday

Finally, I would like to thank my dearest human, Kostas.

Thank you for everything.

**Στον παππού μου**

# LncRNA on Galaxy

## Contents

## Abstract

Undoubtfully, we live in an era with continuous technological advances. This expansion in the biological field has led to the concept of “omics” and methodological advancement in systematically interrogating a biological sample. OMICS technologies are constantly advancing, and new discoveries are made every day. Notably, one of them is lncRNA. It is worth noting that, although OMICS technologies are very powerful, they yield a great amount of data that steadily increases without being used to its full potential. We need to address the issue of aggregating big data that needs to be analyzed to extract critical biological information in an accessible and easy to use manner. To address these problems, we created our instance of Galaxy platformy with many added custom tools and workflows for transcriptomic and genomic analysis.

A novel module-based workflow was created for RNAseq analyses with special focus on lncRNA. LncRNAs are RNA molecules that have many similarities with mRNAs but are not translated into proteins. Although they were considered as junk DNA, recent developments in sequencing technologies have revealed their regulatory potential on gene expression and their consequent involvement in various pathologies. We wanted to create a unified framework to be dedicated to their analysis in order to limit lab to lab discrepancies. The pipeline starts with raw files aligns, assemblies and quantifies the reads into counts, dividing them into three different matrices representing mRNAS, miRNAs and lncRNAs. For the creation of the lncRNA matrix extra steps to ensure the validity of lncRNAs are taken, like estimating coding potential and excluding short reads. Following the matrix creation step is differential expression analysis on these features using nonparametric tests. A sub workflow was created for unraveling the different interaction networks that these features. This sub routine correlates the expression of the different RNAs and then searches for RNA-interaction databases to validate this correlation. If some features are not found in databases, an RNA-RNA interaction prediction is performed to discover potential new pairs. Two different networks are created, one regarding molecules with inverse expression and one regarding sets of three with 1 common target, for example one lncRNA and one miRNA targeting the same mRNA. The analysis ends with pathway enrichment analysis of the two networks and on all differentially expressed molecules. In summary, the analysis starts with reads and ends with prioritized genes and enriched systemic processes that best describe the observed phenotype.

The design of the workflow is module-based. This means that the whole workflow consists of standalone sub workflows and this pattern continues covering all routinely performed tasks in a transcriptomic analysis. For the creation of all modules, publicly available tools were used, as well

as newly custom-made which were created using DOCKER and are available on Galaxy's toolshed. In this thesis, I will present the local Galaxy instance that was created during my internship time at eNIOS, along with the tools and possibilities it offers. Special attention to lncRNA workflow will be given and two case studies will be explored showing the different aspects of our instance. The whole instance is on **http:www.biotranslator.gr:8080**

**KEY WORDS:** Big Data, OMICS, lncRNA, Galaxy, pipelines, Bioinformatics

## Aim of this project

This thesis was part of my work during my internship at eNIOS. Its primary objectives were to build a reproducible, easy to use pipelines for genomic and transcriptomic analyses. Best practices were used in order to assess scalability and easy deployment on cloud technologies such as sevenBridges and Amazon services.

# Introduction

## Big data and omics data: Advantages of omics but drawbacks of analyzing

We are living without a doubt in the information era. The massive explosion of technological advances has affected our lives from the way we communicate to the way we work. In particular, one field of science, Biology, has yielded many gains from these frantic advancements, with some of the most outstanding examples being omics technologies.

Omics technologies are these methods that are primarily aimed at the universal detection of genes (genomics), RNAs (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. The key word here is “universal”. [1] When investigating a disease or a pathology or even two different organisms, what we are really interested in is the underlying mechanism that makes them differ. More often than usual this mechanism consists of a multitude of different proteins, RNAs and metabolites that all together coordinate each other forming a complex network. So understandably, having a universal approach to tackle this issue holds the potential of providing major insights into the biology of interest.

## Transcriptomics

The last few decades have witnessed an explosive growth in genomic sequencing technologies. Transcriptomics is the study of the ‘transcriptome'. The term transcriptome refers the complete set of all the ribonucleic acid (RNA) molecules expressed in some given entity, such as a cell, tissue, or organism. As a result of increased throughput, higher accuracies and lower costs, there has been an exponential growth in genomic sequence databases over the last two decades. [2] However, a major challenge in molecular biology continues to be the complex mapping of the same genome to diverse phenotypes in different tissue types, development stages and environmental conditions. A better understanding of the transcripts and expression of gene regulation is not only non-trivial but lies at the heart of this challenge. Transcriptomics offers important insights on gene structure, expression, and regulation and has been widely studied in many organisms. [3]

### Microarrays

The first attempt to capture global gene expression was with microarrays technology. A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Each DNA spot

contains tiny parts of a specific DNA sequence, known as probes, that will hybridize with a target cDNA sequence. Probe-target hybridization can be detected and quantified to determine the relative abundance of nucleic acid in the target. Scientists use cDNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. [4] An example of its application is in SNPs arrays for polymorphisms in cardiovascular diseases, cancer, pathogens and Genome-wide association studies (GWAS). It is also used for the identification of structural variations and the measurement of gene expression. Because of their high throughput and lower cost, microarrays were very popular throughout the 2000s. However, they are limited to probing to an array of genes that are already known, so a reference genome or transcriptome is a must for microarrays.[5]

### Next Generation Sequencing

Most of the limitations of microarrays are tackled with next generation sequencing (NGS) techniques. Second generation of sequencing (SGS) alternatively referred as next generation sequencing technology, originated in 2005 to support massively parallel sequencing of hundreds of thousands of short DNA strands that are anchored and read through multiple “wash and scan” cycles. [6] For example, Illumina HiSeq platforms can generate upward of 5 billion reads and 1500 Gb per run. Also, this approach is able to generate high read accuracy with much lower cost. Admittedly, NGS has many applications in biological studies, such as DNA-sequencing to assemble a previously unknown genome, and RNA-sequencing to analyze gene expression and to identify the regions of DNA or RNA binding proteins. In addition to gene expression quantification, RNA-Seq is quite effective in detecting alternative splicing events.[2] As a result, it has grown to be the most popular transcript profiling approach over the last decade. Nevertheless, it has several limitations. First, there is the amplification bias and the template sequence errors contributed by the polymerase chain reaction (PCR) amplification step. More importantly, read length is limited to a couple of 100s. The short-read RNA-Seq approach has several inherent limitations. It fails to accurately identify multiple full-length transcripts reconstituted from the short reads [7]. This problem is pervasive particularly when dealing with complex genomes (mostly eukaryotic), which exhibit a large number of isoforms per gene because of alternative splicing and where genes have multiple candidate promoters and 3′ ends [8]. As a result, short-reads RNA-Seq is lacks efficiency when studying gene regulation, the protein-coding potential of the genome and ultimately the phenotypic diversity.

## Third Generation Sequencing

Third Generation Sequencing are the technologies that are primarily distinguished from previous generations in their focus on uninterrupted sequencing of a single DNA or RNA molecule. This makes them highly preferable for a number of use cases such as de novo assembly, improved genome annotations, and epigenome characterization [8]. With long-read sequencing technologies, it has become reality that one read is one transcript, and each transcript can be accurately captured and studied individually since it directly provides full-length cDNA sequences [9]. Techniques such as Oxford Nanopore and PacBio SMS, are designed to avoid the step of assembly and therefore are better suited to comprehensively identify full length transcripts and to profile allele specific expressions. While TGS techniques are optimal for de novo sequencing for small-to-moderate sized genomes (<1 Gbp), they become cost-prohibitive for high coverage of larger genomes.[2]

One issue that arises with these kinds of technologies is the amount and high complexity of the data produced. These data often referred as *Big Data* need to be transferred, stored and more importantly analyzed and transformed into meaningful biological information. This is often laborious work and there are people dedicated to this task. However, the need to rely on someone's expertise slows down if not halting entirely the advancement of the study and the potential benefit from it. Therefore, numerous efforts have been made to produce platforms that would be extremely useful and at the same time user-friendly, in order non – bioinformatics scientists to be able to make immediate sense of their experiments or clinical studies and act accordingly. One of the most famous publicly available platforms is ***www.usegalaxy.org***

## Galaxy platform and its utilities

The Galaxy project started 18 years ago, in 2005, with its key goal being the need to make computational biology accessible to scientists who have no computer experience. Although the initial "target" was genomics it soon broadened its utilities to affect different biological research fields. Their value proposition is accessibility, transparency and reproducibility and the way they reach that goal is by providing either a ready to use platform with all the tools and workflows necessary for basic analysis ready to use in the form of a website, or the infrastructure for someone to build its one Galaxy instance, highly customizable to one's own needs. Recent statistics show that members of Galaxy project are steadily increasing each year and moreover users tend to know a priori how and when they will use Galaxy. Nevertheless, users can always

get help through their various communication channels. [10] This shows that this platform is very popular, helpful and useful and taken together with the fact that we can build our own instance with our custom tools we decided to use this infrastructure for designing our analyses pipelines.

## Our Galaxy instance with focus on lncRNAs

As mentioned above, we were interested in a highly customizable platform. The reason for that was that we sought to design a workflow for analyzing long non-coding RNAs (lncRNAs). LncRNAs are RNA molecules that are transcribed by polymerases, have 5' and 3' UTR regions but –usually- don't encode for proteins. Instead, they seem to have regulatory roles in the network they reside in. Recent studies have shown the critical role of some lncRNAs in various pathologies. [11] LncRNAs are a classic example of how improvements in technologies broaden our understanding of gene expression and regulation and genomics in general, as these molecules were once considered junk DNA. RNA sequencing has revealed that not only are they not junk but together with other non-coding RNAs form a complex network that fine tune gene expression. Since we are only now starting to realize the variety and complexity of these different RNA molecules, a consensus set of tools and workflows has yet to be established. In this thesis we will try to provide such a framework through our local galaxy instance at **http:www.biotranslator.gr:8080**

# Methods

In this chapter we will describe the prerequisites for deploying your own Galaxy instance and refer to the basics of Galaxy architecture.

## Requirements

The most basics requirements for launching a Galaxy instance are UNIX/Linux or Mac operating system a together with python 3.7. It is preferable to have a dedicated server to host the instance rather than a local machine in order to provide 24-hour service without the risk of damaging your computer. This instance of Galaxy runs on a virtual machine hosted on Hypatia infrastructure provided by ELIXIR- Greece. It has 8 GB of RAM and 530GB of memory. In addition to the server, an SQL database needs to be created and connected to the instance. Finally, a system for sending links for user authentication was set up and connected to the instance.

## General Architecture of Galaxy

Galaxy is an open-source software implemented using Python programming language. It is developed by the Galaxy team at Penn State, Johns Hopkins University, Oregon Health & Science University, and the Galaxy Community. [12][13] It is a collection of scripts that define the layout of a website and connects its features with back-end processes. The different back-end and front-end features are defined in a configuration file called galaxy.yml

Inside that file, administrators are defined, as well as the IP address on which you want to host the instance, the IP address of your SQL database and the connection to the smtp server is established. After this basic configuration, the user needs to declare 4 important directories:

- The location of the FTP server that will be used by this instance
- The jobs directory which is the location where results from users' analyses will be stored
- The tools directory

- The tool-data directory which holds the loc files of the reference files required by various tools

## Tools

Another important configuration file is tool_conf.xml. In there, the location of the XMLs of the different tools can be found as well as the order and grouping in which they will be displayed on the website. In Figure 1 its format is shown. All directories are relative to the one declared in galaxy.yml

ubuntu@galaxyncrna: /data/galaxy/config

```
<?xml version='1.0' encoding='utf-8'?>
<toolbox monitor="true">
  <label name="Basic Tools" id="cbt" />
  <section id="getext" name="Get Data">
    <tool file="data_source/upload.xml" />
    <tool file="data_source/ucsc_tablebrowser.xml" />
    <!-- <tool file="data_source/ucsc_tablebrowser_test.xml" /> -->
    <tool file="data_source/ucsc_tablebrowser_archaea.xml" />
    <tool file="data_source/ebi_sra.xml" />
    <tool file="data_source/fly_modencode.xml" />
    <tool file="data_source/intermine.xml" />
    <tool file="data_source/flymine.xml" />
    <!-- <tool file="data_source/flymine_test.xml" /> -->
    <tool file="data_source/modmine.xml" />
    <tool file="data_source/mousemine.xml" />
    <tool file="data_source/ratmine.xml" />
    <tool file="data_source/yeastmine.xml" />
    <tool file="data_source/worm_modencode.xml" />
    <tool file="data_source/wormbase.xml" />
    <!-- <tool file="data_source/wormbase_test.xml" /> -->
    <tool file="data_source/zebrafishmine.xml" />
    <tool file="data_source/eupathdb.xml" />
    <tool file="data_source/hbvar.xml" />
  </section>
  <section id="send" name="Send Data">
    <tool file="cloud/send.xml" />
  </section>
```

*Figure 1 Tool_conf.xml file holds the directories of the XMLs of all tools displayed on the website*

Each xml file represents a specific tool. A closer look at a particular xml is shown in Figure 2

```
ubuntu@galaxyncrna: /data/galaxy/tools/lncRNA
<tool id="discardfasta" name="discardFasta" >
    <description>"discard_fasta_sequences_less_than_specific_length"</description>
    <requirements>
        <container type="docker">bianca7/lncrna:discard_fasta</container>
    </requirements>
    <command><![CDATA[perl /discard_fasta.pl $length $fasta_file > $out
   ]]></command>
   <inputs>
     <param type="data" name="fasta_file" format="fasta" />
     <param type="integer" value="200" name="length" label="length cut off" />
  </inputs>
  <outputs>
         <data name="out" format="fasta" label="discarded ${fasta_file.name}" />
  </outputs>
  <help><![CDATA[
         "Discard fasta sequences less than X nucleotides"
         Inputs: fasta
         Outputs: filtered fasta
         ]]>
  </help>
</tool>
```

*Figure 2 Format of an XML of a tool. In blue the different sections are displayed*

It is apparent that in this file we set the name, inputs, outputs, description and help information of our tool. This file will establish the layout of our specific tool on our instance as shown in Figure 3


*Figure 3 Appearance of the tool on website*

In the section command we set what process or processes our inputs will be put into. For example, in this example we call discard_fasta.pl script together with length parameter and our input and we declare where we want our output. The script is not located on the server but rather inside a Docker container, defined in the requirements section. Docker containers are isolated virtual environments that hold together all the different components of an application, which in our case are the custom R, Python or Perl scripts along with their dependencies. Generally, Docker is a platform designed to help developers build, share, and run modern applications. It takes away repetitive, mundane configuration tasks and is used throughout the development lifecycle for fast,

easy, scalable and portable application development – desktop and cloud. [14] Dockers were created for all tools provided in our instance both for the public and the custom made. All tools can be found on toolshed.

## Reference files

Often tools used for transcriptomic and genomic analyses require reference files or specific databases. These files can be defined in special files that have the format shown in Figure 4. Basically, these are tab delimited files containing the directory of each reference file, one per line. The mapping of location files with their respective tool is defined in a tool_data_table_config.xml file as shown in Figure 5. The directory of loc files is relative to the directory defined in galaxy.yml

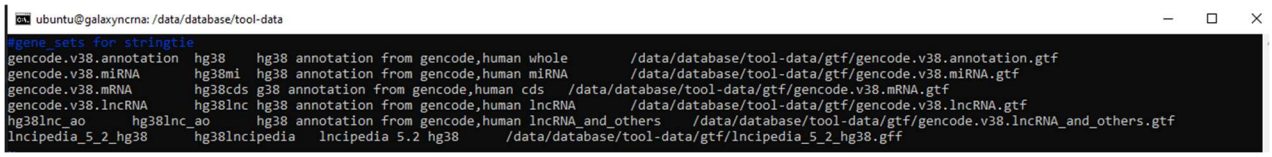


*Figure 4 Format of a loc file. A loc file is a Tab delimited file that holds information about the name and location of reference files, one file per line*

```
<!-- Paths are relative to the value of `tool_data_path` in galaxy.ini -->
<tables>
    <!-- Locations of all fasta files under genome directory -->
    <table name="organisms" comment_char="#" allow_duplicate_entries="False">
        <columns>index, name, abbreviation, value, taxonomy</columns>
        <file path="/data/database/tool-data/Proteostasis_Phylogenetic_Analysis/organisms.loc" />
    </table>
    <table name="all_fasta" comment_char="#" allow_duplicate_entries="False">
        <columns>value, dbkey, name, path</columns>
        <file path="all_fasta.loc" />
    </table>
    <!-- Locations of indexes in the BFAST mapper format -->
    <table name="bfast_indexes" comment_char="#" allow_duplicate_entries="False">
        <columns>value, dbkey, formats, name, path</columns>
        <file path="bfast_indexes.loc" />
    </table>
    <!-- Locations of nucleotide BLAST databases -->
    <table name="blastdb" comment_char="#" allow_duplicate_entries="False">
        <columns>value, name, path</columns>
        <file path="tool-data/blastdb.loc" />
    </table>
```

*Figure 5 Format of a tool_data_table_conf.xml. This file maps the reference files to the tools that use them*

# Results

## General presentation of the platform.

Numerous scientists from different disciplines use Galaxy because of its great variety of tools and analyses you can perform within it. For instance, one can develop an instance that will be more focused on image analysis or phylogenetic analysis etc. Our instance is more focused on transcriptomic and genomic analyses. We also offer a great variety of versatile tools for text manipulations, because these processes are necessary in any kind of analysis. Each tool has "Help" section with details about the inputs, outputs and parameters. In Figure 6 we can see the front page of our instance. One can either navigate to our website though the icons or choose a specific tool from the "Tools" section.

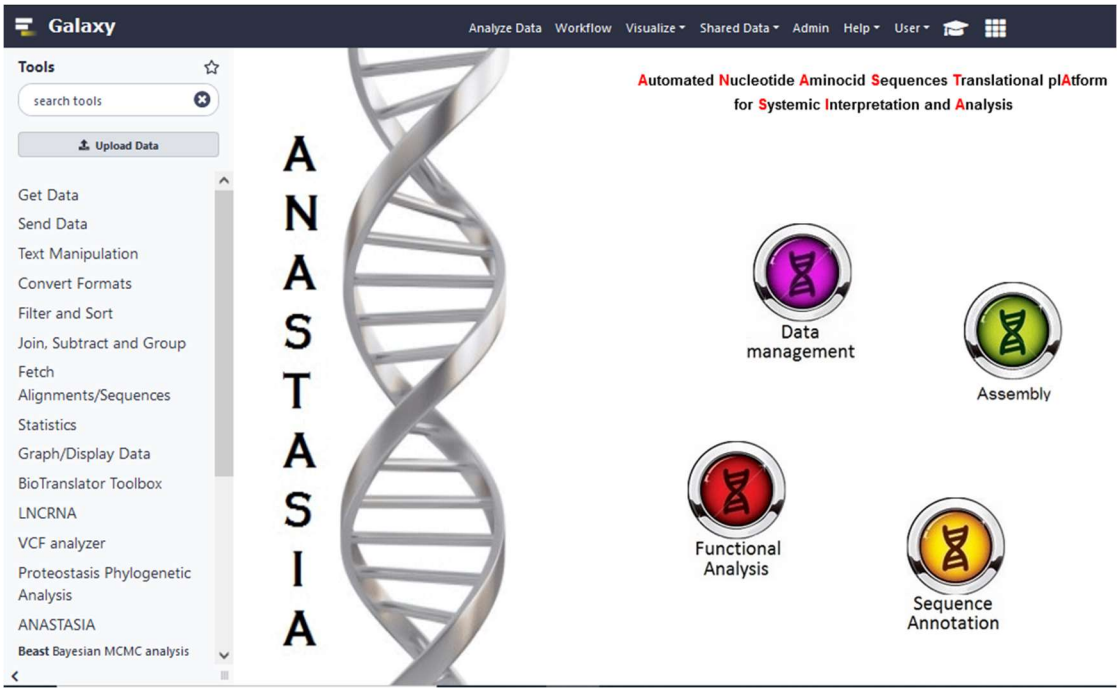


*Figure 6 Appearance of our Galaxy's frontpage*

### *Data libraries*

An important section of Galaxy is “Shared Data” (Figure 7). In this section one can find the tab “Data Libraries”. In there the user can find reference files required for tools and workflows, like the GTF and FASTA of the human genome. Test files for demonstration of the various workflows also exist there.

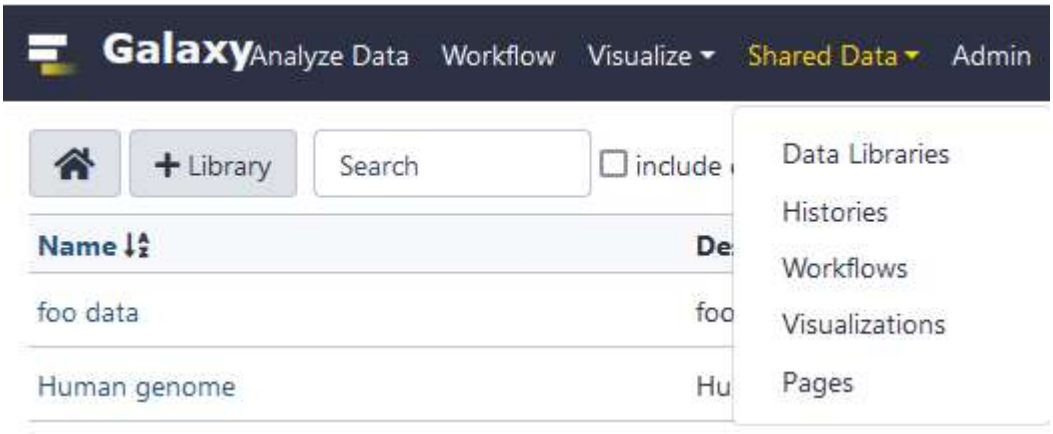


*Figure 7 Libraries under Shared Data section provides with reference files and test files*

### *Workflows picture*

Another tab worth-mentioning in the section “Shared Data” is that of “Workflows”. Here a user can find all the published workflows in our instance. The possibility of uploading and publishing your own workflows is supported. After importing the workflow of interest, one can run it or edit it to suit specific demands.

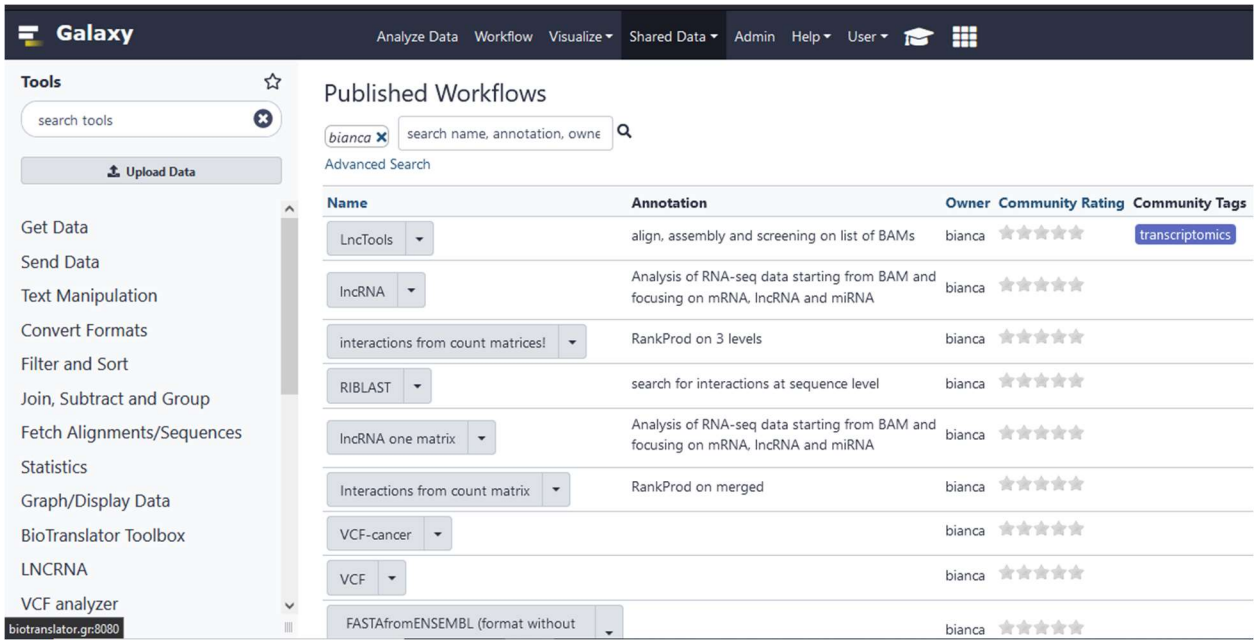


*Figure 8 Workflows tab under Shared Data section provides with all published workflows of our team. The user needs to import them before using them*

The use of the platform is click-based, and no programming language is required. Any analysis starts with the upload of the required files. The files can either be locally or remotely uploaded using FTP or a URL from the internet.

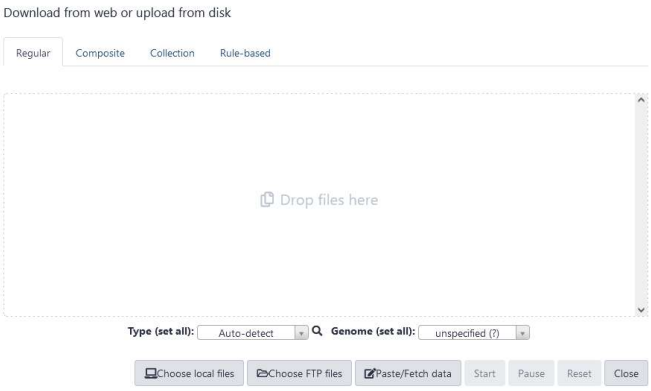


*Figure 9 Upload tab provides the possibility to upload data either from a local device or from an FTP address or from an URL address*

### *Variant Calling File (VCF) analyzer*

Often in biological studies, one of the main questions is key mutations that drive the respective phenotype. To investigate that, scientists perform whole genome sequencing or whole transcriptome sequencing in order to identify these mutations. After the sequencing part, the aligning of the raw files to a reference genome is performed in order to identify the differences between the two, or in other words, identify the variants. The result is a file which is called Variant Calling File (VCF). A workflow was added that takes these VCF files, annotates these mutations using snpeEff, snpSift and COSMIC databases and filters for the most significant ones. [15][16][17][18]. Once we have established which genes have significant mutations, we can perform pathway enrichment analysis on the genes bearing significant mutations. In between the workflow uses several different Python scripts for quality control of each step. These scripts and the backbone of the workflow had been developed by Chara Mastrokalou for SevenBridges platform (https://www.sevenbridges.com/platform/). My task was to adjust it for Galaxy implementation.

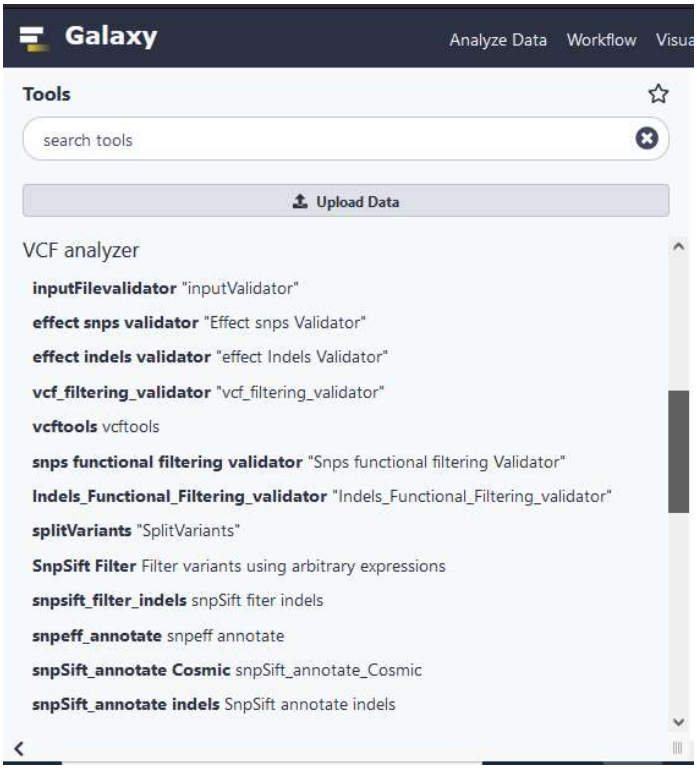


*Figure 10 Tools provided under VCF-analyzer section, under Tools section*

### *Metagenomic analysis - ANASTASIA*

This section is suitable for users from metagenomics field. Metagenomics is the study of the structure and function of entire nucleotide sequences isolated and analyzed from all the organisms (typically microbes) in a bulk sample. Metagenomics is often used to study a specific community of microorganisms originating from diverse environments. Usually, such analyses are performed in order to establish the different groups of microorganisms that exist in a sample. Moreover, their specific genetic signature is studied and how this influences their environment. . ANASTASIA is a suite of algorithms dedicated to such analyses. We provide classic tools for *de novo* assembly, like MEGAHIT and very popular options for the annotation of prokaryotic genomes (PROKKA, KRAKEN2). [19][20][21] In addition, tools like *CD-hit* and *hmmer* are provided to cluster and annotate protein patterns (Figure 11). [22][23] Elena Gkotsi has developed a workflow for routine analyses.

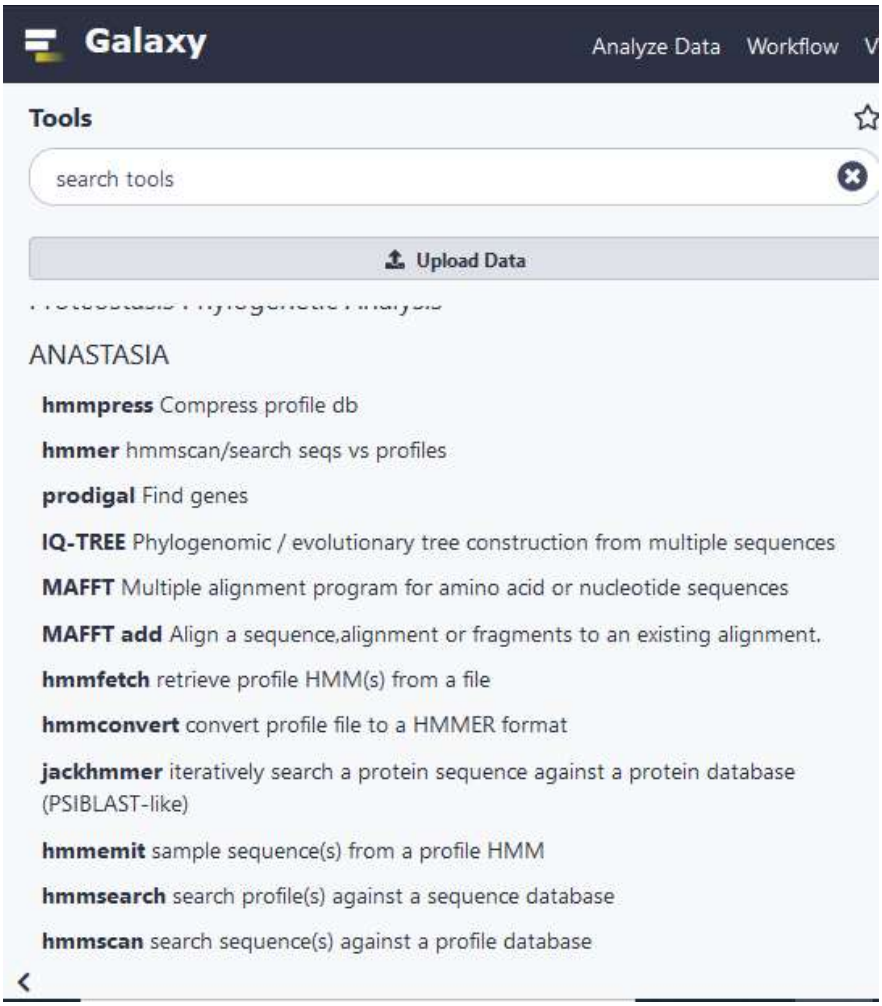


*Figure 11 Some of the tools provided in the ANASTASIA tab, under Tools section*

### *LNCRNA*

In the LNCRNA section a great variety of align and assembly tools for RNA-seq data can be found. In addition, if someone is interested in prediction of RNA-folding they can use RNAfold or RNApifold [24]. Good visualization of data can have a great and immediate impact on the interpretation of your results, so we offer heatmaps and correlation plots to aid this effort. We should note here that each tool used in the lncRNA workflow can be found in this section and can be used in different contexts depending on the question of interest.

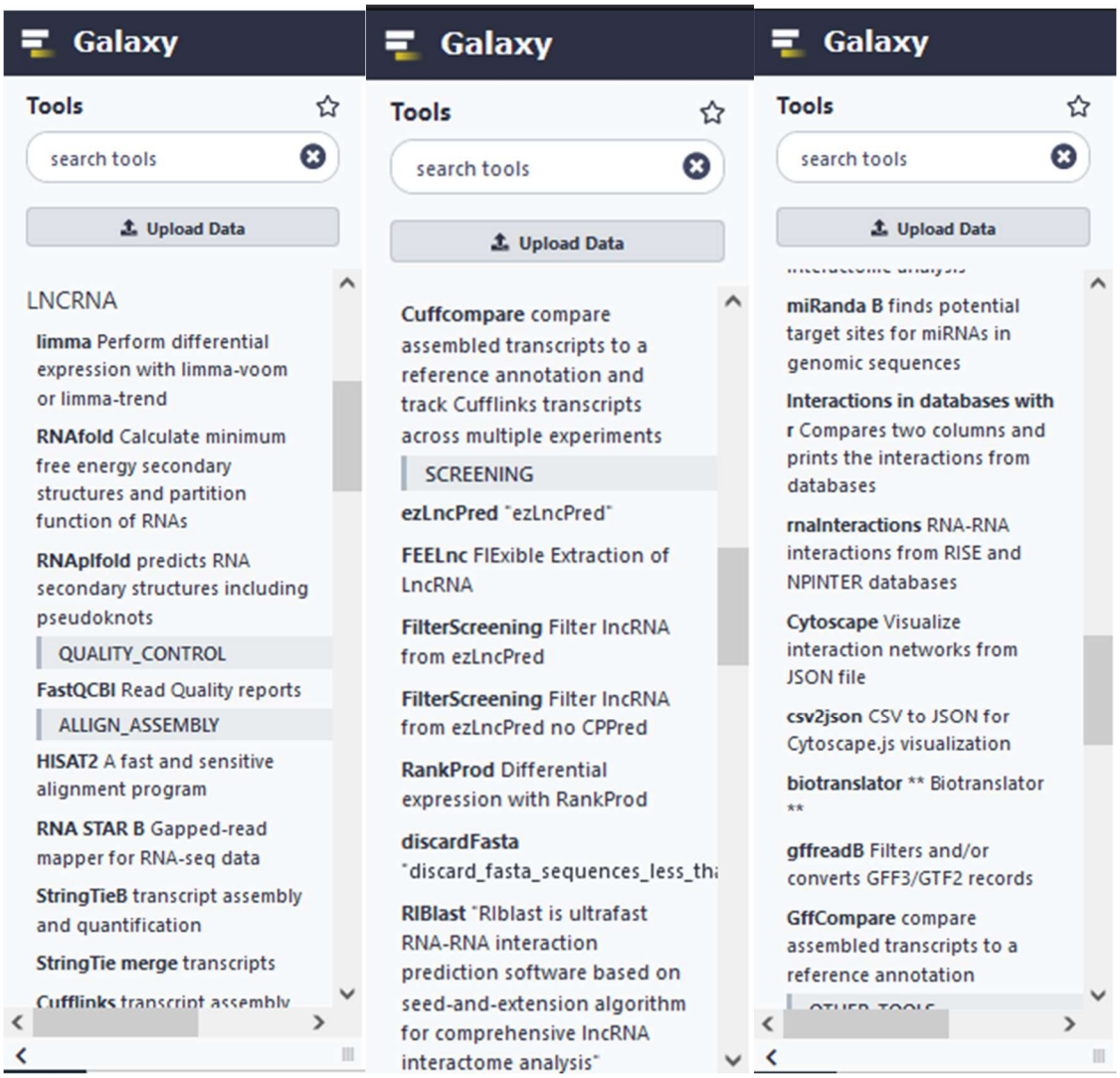


*Figure 12 All tools provided in LNCRNA tab, under Tools section*

## LNCRNA WORKFLOW

The use of transcriptomic experiments has greatly increased in the past years. As sequencing technologies are becoming more accessible the amount of data increases exponentially. In order to utilize them in the best way possible, a unified framework of analysis must be applied. This way will minimize the potential discrepancies observed when statistics are done by different labs and a more robust interpretation can be concluded on. In order to decipher the role of non-coding RNAs and especially of lncRNAs in gene expression we developed a lncRNA workflow. The objective of this workflow is to start directly after the Illumina sequencer (or any kind of sequencer),

i.e., raw files and end with biological insights about the samples of interest. Biological insights in this specific context means key genes and processes that drive the observed phenotype. All workflows discussed here are module-based. This means that they follow a “Russian doll” model in which the big workflow consists of two sub workflows and each consisting of smaller sub-workflows (or modules). Some examples are the RIBLAST workflow, Cytoscape or FASTAfromEnsemblID. They were developed with this design because in this way not only can one make use of all the standalone modules depending on the specific questions imposed but also have the flexibility to combine them into new workflows.

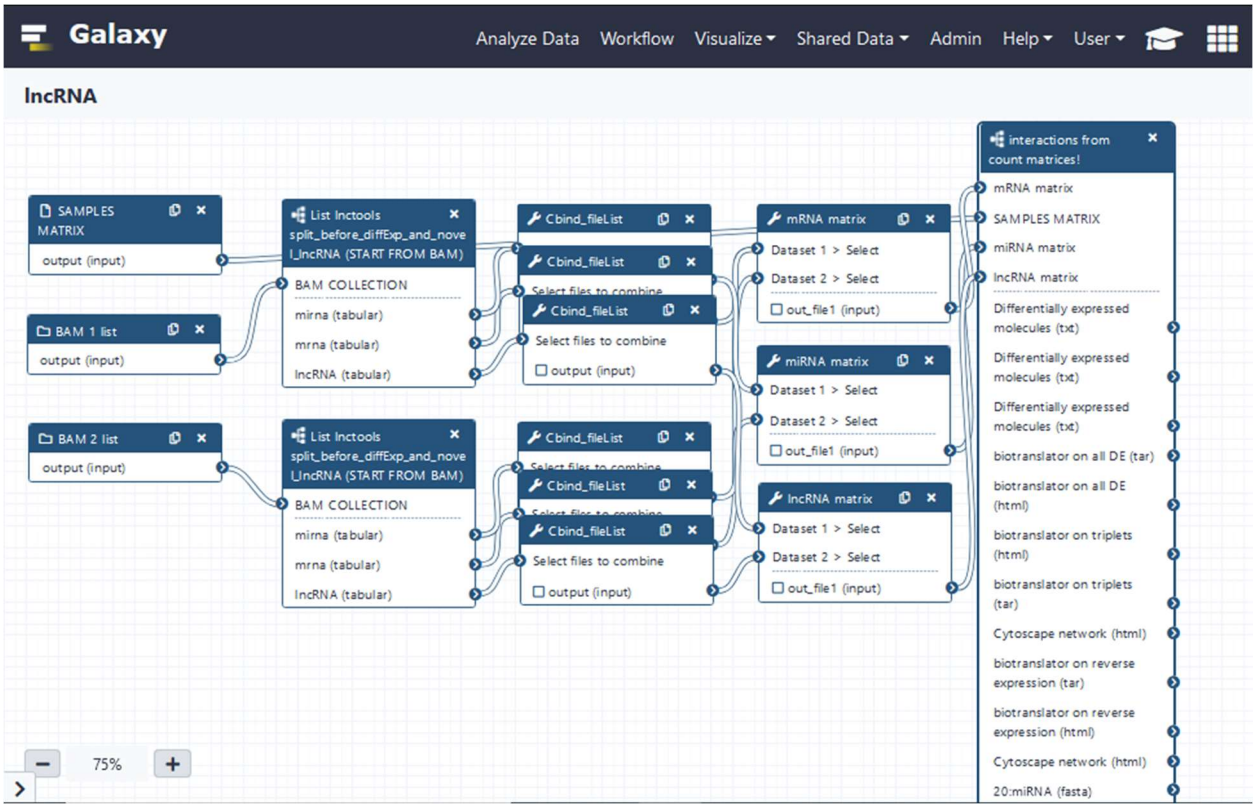


*Figure 13 Format of full lncRNA workflow*

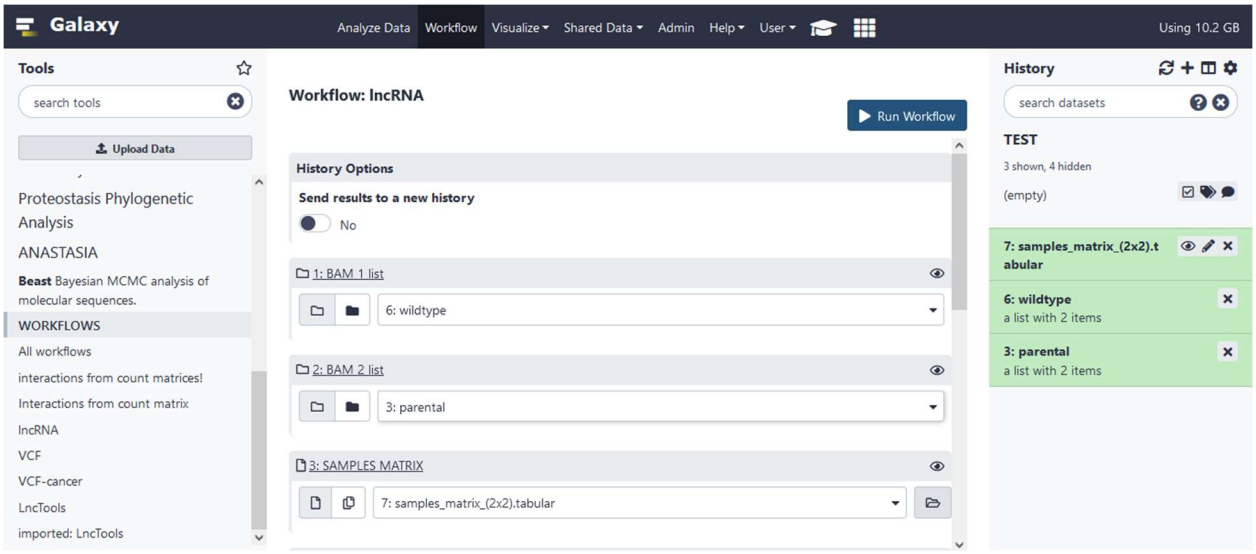


*Figure 14 Appearance of lncRNA workflow on our Galaxy website*

## First part: lncTools

LncTools workflow starts from raw files like FASTA or FASTQ and ends at the step of the generation of 3 matrices. This part is suitable for experienced users who want to do the pre-processing steps, change specific parameters etc.

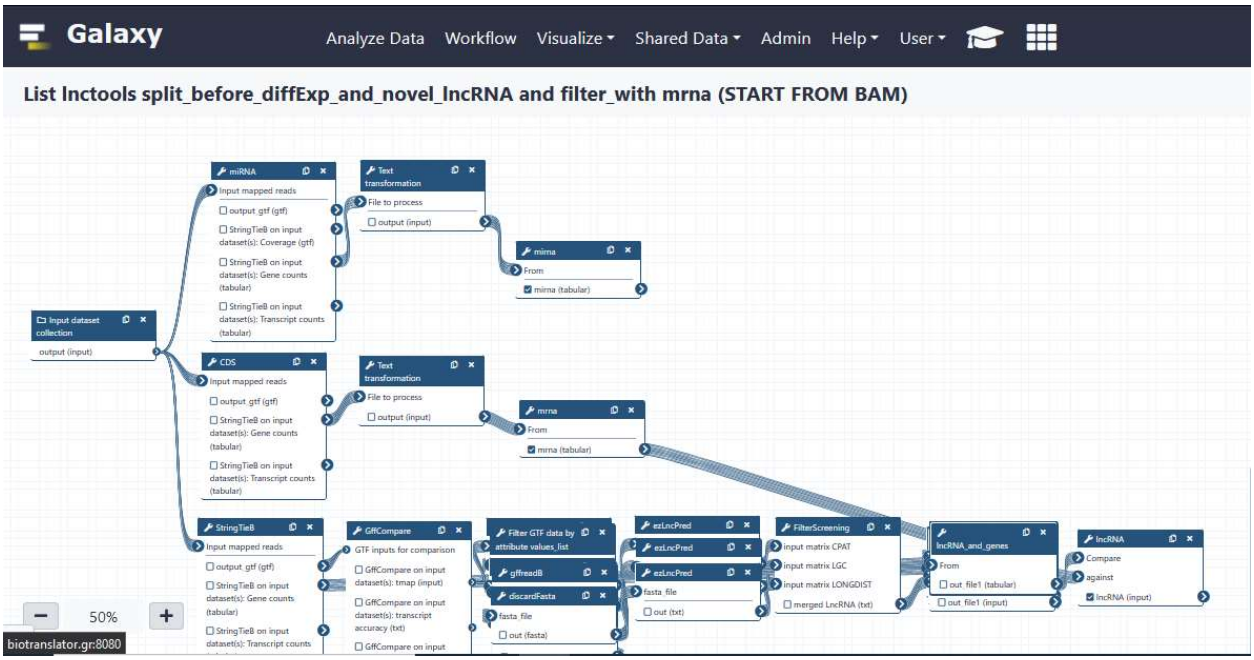


*Figure 15 Format of lncTools, the first part of lncRNA workflow*

### *Inputs*

The lncRNA workflow that starts from FASTQ is performed on demand due to the computational limitations of the server that hosts the website. Nevertheless, a complete workflow starting from BAM format is provided. The only thing the user has to provide is a list of BAM files separated based on condition of interest and a samples info file which describes the condition of each sample. To date we provide pairwise comparisons. We are considering in the future adjusting our workflow for more conditions. These two lists of BAM files are used both for whole lncRNA and for lncTools

### *Outputs*

Once the workflow is completed, three txt are produced representing the lncRNA matrix, mRNA matrix and miRNA matrix. These files are filled with counts and have features in the rows and samples in the columns, in other words each file is an abundance matrix of the separate RNA molecules.

### *Details of workflow*

#### Quality control

The first step of every experiment and every analysis is quality control. In this step one establishes that the experiment worked as planned and that enough useful samples to proceed exist. Part of this process is to look at GC content of your reads, search for contaminants, investigate if a tile was overloaded etc. FASTQ algorithm was used for this step [25] If the reads don't meet the quality criteria they are removed and then one can proceed, if possible, with downstream analysis.

#### Alignment

Another extremely important step in analysis of RNA-seq data is the aligning process. It is a field of active research, and much thought must be put into this in order to reach an acceptable equilibrium between computer power that is needed and the accuracy you want to achieve. Many aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases. Of course, a limiting factor here is also the technologies used for the generation

of these data, but this is a topic exceeds the research of this thesis. We chose the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNAseq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, while at the same time improving alignment sensitivity and precision. In addition to unbiased de novo detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts and is also capable of mapping full-length RNA sequences. [26]

### Assembly

Once the aligning process is complete one can proceed to the assembly step. In this step the reads are assembled into transcripts and then measured to produce counts that depict the expression of the RNA molecule of your interest (mRNA, lncRNA, miRNA etc.) Although STAR offers directly this opportunity, we figured that for our type of analysis we should also include StringTie algorithm. [27] StringTie is a fast and highly efficient assembler of RNASeq alignments into potential transcripts. It uses a novel network flow algorithm as well as an optional de novo assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus. Its input can include not only alignments of short reads that can also be used by other transcript assemblers, but also alignments of longer sequences that have been assembled from those reads. Of note, we provide several assembly algorithms like HTSeq and featureCounts on our platform, to suit the needs of all potential users. [28],[29]

### Detection of new lncRNA, miRNA, mRNA

As discussed previously, our interest was in lncRNAs and their interactions with other RNA molecules like mRNA and miRNA. There are two ways to look at this problem depending on the testable question. One can search for established lncRNAs in a dataset and correlate their expression with other RNA molecules, or one can search for novel lncRNA molecules and proceed from there. Since lncRNAs are a relatively new field we thought to provide both opportunities to the user.

In the first case we aligned the reads to three different reference genomes: the lncRNA subset from GENCODE v29, the miRNA subset from GENCODE v29 and finally the mRNA subset from GENCODE v29. After that reads are assembled into transcripts with StringTie algorithm using as

guide the abovementioned reference genomes. After that, 4 different algorithms are used to establish that the detected lncRNAs are indeed long noncoding and not protein coding.

The opportunity to detect novel lncRNAs is also given to the users that are interested in these kinds of questions. In this case the reads are aligned to the complete GENCODE v29 reference genome. StringTie is used afterwards using as guide the reference genome and with the option to detect novel transcripts. After that, as with the other option, 4 different algorithms to detect the coding potential of each transcript are used: CPAT, CPC2, longdist and LGC.

### CPAT

Coding Potential Assessment Tool (CPAT) is a novel alignment-free method, which rapidly recognizes coding and noncoding transcripts from a large pool of candidates. To this end, CPAT uses a logistic regression model built with four sequence features: open reading frame size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias. CPAT has the advantages of being highly accurate and being quite fast, enabling its users to process thousands of transcripts within seconds. [30]

### CPC2

Coding potential calculator 2 (CPC2) is a fast and accurate algorithm based on sequence intrinsic features like Fickett TESTCODE score, open reading frame (ORF) length, ORF integrity and isoelectric point (pI). While the Fickett TESTCODE score is derived from the weighted nucleotide frequency of the inputted full-length transcript (22), the rest of three features (ORF length, ORF integrity and isoelectric point) are calculated based on the longest putative ORF identified in silico. [31]

### Longdist

Longdist is a Support Vector Machine (SVM) based method to distinguish lncRNAs from PCTs, using features based on frequencies of nucleotide patterns and ORF lengths, in transcripts. The proposed method is based on SVM and uses the 4 features from ORF relative length and 6 groups of frequencies of nucleotide patterns selected by PCA as features. Then based on these features a model is trained to distinguish lncRNAs from PCTs. [32]

### LGC

Finally, LGC characterizes and identifies lncRNAs based on the relationship between open reading frame (ORF) Length and GC content. LGC is able to accurately distinguish lncRNAs from protein-coding RNAs in a cross-species manner without species-specific adjustments and is robustly effective (>90% accuracy) in discriminating lncRNAs from protein-coding RNAs across diverse species that range from plants to mammals. [33]

As mentioned after the assembly step, the transcripts are analyzed with these 4 different algorithms and the consensus of these algorithms is kept for further analysis. At the end of this step the user has three different matrices regarding the three different populations of RNA molecules.

### Second part: interaction from count matrices

The second part of the workflow starts from the three different matrices representing our RNA populations. It focuses on detecting differentially expressed features, correlating their expression and finally establish the network they reside in with the final goal of identifying key features or hub features that hold the most descriptive information of the system

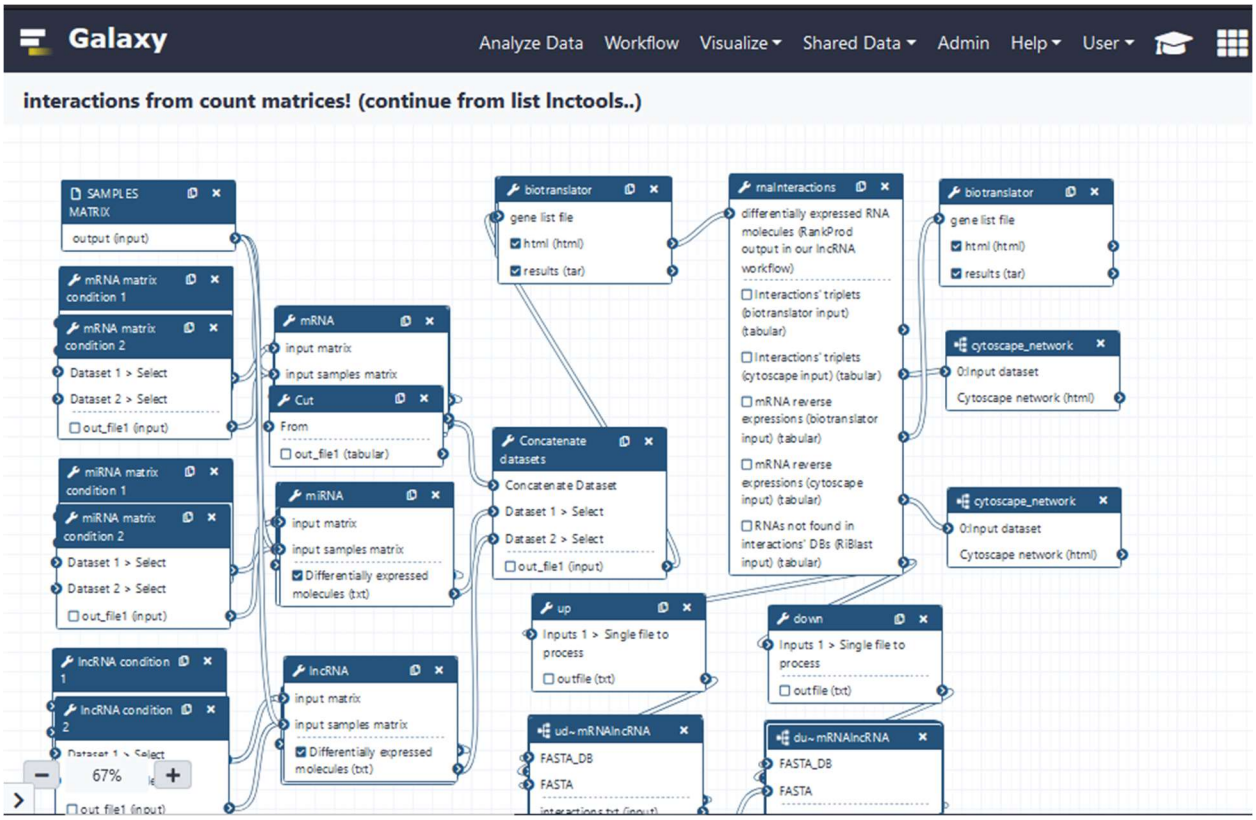


*Figure 16 Interactions from Count matrices. LncRNA workflow starting from count matrices and ending to biomarkers of the condition of interest*

### *Inputs*

"Interaction from count matrices" starts with input three matrices, each representing the abundance of mRNA, lncRNA and miRNA molecules (Figure 16) . These three matrices can be generated by lncTools or uploaded by the user. Often in science the ideal scenario of having access to the abundance of all three different RNA molecules is not realized. For this reason, we developed "Interaction from count matrix", which takes as input a txt matrix file that has all RNA molecules together and divides them in a subsequent step.

### *Outputs*

After the completion of the above-mentioned workflows the user can find various files and plots that can be used for potentially downstream analysis and for interpretation of the data. Specifically, two cytoscape networks are created one named "cytoscape of triplets" and

"cytoscape of reverse expression". *Cytoscape of triplets* refers to the RNA molecules whose expression correlate with each other, whereas in the *cytoscape of reverse expression* we have duals of reverse expression only. Pathway enrichment analysis and identification of hub genes of the network of these sets of molecules is also provided. These files are called "Bioinfominer on triplets" and "Bioinfominer on reverse expression". These files come in two formats, one zip file which contains results in detail and one html file containing heatmaps and dendrograms of enriched systemic processes and hub genes that can be viewed on-spot or in any other browser. In addition, an extra file with novel interactions is produced. This file refers to novel interactions predicted with RIBLAST sub-workflow along with the sequences of the molecules involved in FASTA format.

### *Details of workflow*

#### Differential expression analysis

Finding differentially expressed molecular features when comparing different conditions plays a pivotal role in all kinds of molecular profiling studies ('omics'). Differential expression analysis is referred to the type of analysis that aims to identify distinctive features that characterize the phenotype of interest. Different approaches are used based on the experimental design, the question of interest or the nature of the data. We decided to include packages that include both parametric and non-parametric tests to cover all potential needs. The non-parametric tests are performed with a newly added custom tool. In our workflow differential expression analysis is performed on the three different matrices of RNA molecules separately.

#### Linear Models for Microarray Data, LIMMA

Linear models are a way of describing a response variable in terms of a linear combination of predictor variables. The response should be a continuous variable and be at least approximately normally distributed. Such models find wide application but cannot handle clearly discrete or skewed continuous responses. Naturally we considered offering limma as a tool. [34] Limma is a very popular bioconductor package that was initially developed for microarray analyses, but it soon expanded to also cover high-throughput PCR data. The package contains particularly strong facilities for reading, normalizing and exploring such data. In addition, limma can perform both differential expression and differential splicing analyses of RNA sequencing (RNA-seq) data. Furthermore, the package is now able to go past the traditional gene-wise expression analyses

in a variety of ways, analyzing expression profiles in terms of co-regulated sets of genes or in terms of higher-order expression signatures. This provides enhanced possibilities for biological interpretation of gene expression differences. [34]

### RankProd

Non-parametric tests offer an alternative to parametric tests when the conditions of the latter aren't met. Non-parametric tests are distribution free statistics, that means that no assumption about the probability distribution of the variables being assessed is taken. In this kind of tests, firstly variables are sorted in ascending order and then replaced with a rank. The p-values are then calculated based on these ranks and they tend to be more robust because they are less sensitive to random noise. One very popular R package which performs these tests is RankProd. [35]. The Rank Product (RP) and the Rank Sum (RS) are two non-parametric statistics widely used to detect variables consistently upregulated (or downregulated) in replicate experiments [36]. Originally developed for the analysis of gene expression microarrays, both methods are more accurate and powerful than their usual competitors in several different scenarios (e.g. abnormally distributed noise, heterogeneity of samples, small fraction of changed features, small sample size), as demonstrated in extensive numerical studies. [36] The main identified weakness of the RP method is its sensitivity to variable-specific measurement variance. Nevertheless, this problem has been successfully addressed by a number of variance-stabilizing normalization techniques. [36]

## Interaction analysis

### Correlation analysis of differential expressed molecules

Once differentially expressed features regarding mRNAs, miRNAs and lncRNAs are discovered, correlation analysis can be performed. In correlation analysis, we search for features that are either correlated, e.g., mRNAs with lncRNAs, lncRNAs with miRNAs or inversely correlated, like miRNAs with mRNAs. We developed a custom tool that first uses spearman correlation to search for these correlations and then filters out pairs that didn't meet the 0.9 cut off. After that, it will search in the databases described below to find additional proof of interaction.

### Databases of RNA-RNA interactions

We implemented two databases in the step of interaction analysis, to make our correlation analysis more robust. The databases that we used are NPINTER and RISE.

#### NPINTER

Despite the growing number of databases, few databases have been established to particularly collect ncRNA functional interaction data, and no bio-molecules network with an ncRNA component has been described. NPInter documents functional interactions between noncoding RNAs and biomolecules (proteins, RNAs and DNAs) which are experimentally verified. Interactions are manually collected from publication in peer-reviewed journals, followed by an annotation process against known databases including NONCODE, miRBase and UniProt. NPInter is one of the 42 expert databases chosen by RNAcentral database. [37]

#### RISE

RISE provides a comprehensive collection of RNA-RNA interactions (RRIs) identified in human, mouse and yeast, together with extensive molecular annotations for each RRI. The RRIs were curated from transcriptome-wide sequencing studies and targeted sequencing studies, as well as other public databases and datasets. In addition, annotation of RRIs with extensive molecular and functional information is provided including RBP binding, RNA editing and modification sites, SNPs and pan-cancer mutations, as well as (iv) gene expression levels from various cell and tissue types. [38]

### NOVEL INTERACTIONS

#### RIBLAST

A proportion of the differentially expressed RNA molecules couldn’t be found in databases of RNA interactome. This could be observed because they in fact do not interact or because their interaction hasn’t been previously observed. Therefore, we provided RIBLAST sub workflow which predicts RNA-RNA interactions. The inputs for this module are the sequences of RNA molecules of interest and output is a file with novel predicted interactions. RIBLAST is an ultrafast RNA–RNA interaction prediction method based on the seed-and-extension approach. RIBlast discovers seed regions using suffix arrays and subsequently extends seed regions based on an

RNA secondary structure energy model. Computational experiments indicate that RIBlast achieves a level of prediction accuracy similar to those of existing programs, but at much greater speed than existing programs. [39]

### Visualize interactions

#### Cytoscape

Many times, it is very useful to depict our interactions in the form of a network. To achieve this objective, we developed a sub-workflow that takes input a 2-column file in the form of the dual "feature-target" and outputs the respective network. Chara Mastrokalou has been of great aid in completing this task. [40]

### Functional analysis

The last step of our analysis is the pathway enrichment analysis. Pathway enrichment analysis refers to a set of methods the goal of which is the identification of hub features or important features that play a critical role in the system of question along with the system processes that they reside in. For this purpose, we exploited BioInfoMiner. [41] BioInfoMiner was developed during Theororos Koutsandreas, PhD, as a standalone program, and was incorporated into Galaxy. It is a novel computational approach that was devised for the interpretation of various omics data, based on semantic network analysis, exploiting biomedical ontologies. It relies on an unsupervised, automated, analytical workflow, which was developed for the semantic interpretation of omics data. The workflow combines the execution of pathway analysis and gene prioritization, using the annotation and structure of biomedical ontologies. Its final goal is to transform the distribution of signals of a high-throughput experiment into a semantic network and detect the pivotal regulatory biomolecules into it. [41]

### Survival analysis

Often, after the discovery of a gene signature, the question that comes next is whether this signature can stratify patients and if this stratification correlates with clinical parameters. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects who survived or were saved after that intervention over a period. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time

and the analysis of group data as survival analysis. We added a custom tool for this kind of analysis, using data from TCGA and Kaplan-Meier estimator from TCGAbiolinks package. [42]

### TCGA

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions. Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use. We used transcriptome data based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga."

### Kaplan-Meier estimator

The Kaplan-Meier estimator is the simplest way of computing the survival probability of patients over time in spite of all the difficulties associated with subjects or situations. The survival curve can be created assuming various situations. It involves computing of probabilities of occurrence of event at a certain point of time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate. This can be calculated for two groups of subjects in addition to their statistical difference in the survivals. [43]

## Case Studies

Once the pipeline was developed our next step was to use it for two different projects regarding IRE1 activity. IRE1 is an endoplasmic reticulum (ER) transmembrane protein that acts as a sensor of the accumulation of unfolded proteins in the ER. Its downstream signaling involves phosphorylation of target proteins by its kinase domain and cleavage of a large number of cytosolic RNA molecules by its RNase domain. The best characterized substrate of its RNase activity is *xbp1u* mRNA, from which a 26-nucleotide intron is removed, producing *xbp1s* mRNA

that is translated into a potent transcription factor, XBP1s. Several other cytosolic RNAs, including mRNAs, microRNAs and other ncRNAs, that possess an *xbp1u*-like stem loop sequence can also be cleaved and undergo degradation in a process termed Regulated IRE1-Dependent Decay (RIDD). [44]

The wider literature demonstrates a role of IRE1 in most of the hallmarks of cancer [45]. In addition, inhibition of IRE1 enhances the effect of chemotherapeutic drugs in reducing tumor growth. IRE1 is constitutively active in several cancers and is especially important in Triple-negative breast cancer, TNBC [46][47]. IRE1 signaling has been linked in TNBC to changes in the secretome and in lipid metabolism, both of which impact the ability of TNBC cells to survive [47][48]. Due to its kinase and RNase activity, IRE1 is linked to several layers of biological information, including the transcriptome through XBP1's transcription regulatory activity and RIDD-mediated RNA degradation, the proteome including protein expression, protein phosphorylation, and the secretome. I wished to investigate whether and/or how these different layers are regulated by ncRNAs. The case studies used here are intended for demonstration purposes.

### 1. *IRE1 in Triple Negative Breast Cancer*

Although it is established that IRE1 has a critical role in the progression of many cancers, the processes through which the RNase domain contributes to TNBC are not fully understood. We obtained a study with accession number GSE176454, in which the researchers used MKC8866, a novel small molecule that inhibits the IRE1 RNase activity to find novel targets of IRE1 in TNBC, in an effort to elucidate how it contributes to the disease. The group used MDA-MB-231, a cell line of TNBC origin. The cells were treated with 20 μM MKC8866 or DMSO for 8 and 24 h, then RNA was extracted, and the transcriptomic profile was determined with RNA sequencing (n=3). [48]. We imported the raw data into our pipeline.

After completing the lncRNA workflow we get various outputs regarding our differentially expressed features upon IRE1 inhibition. Firstly, we explore the network that is formed from triplets of features upon MKC8866 treatment at 8 hours (Figure 17). We remind the reader that triplets are defined when the expression of two features is related to an expression of a third one. For example, a miRNA and a lncRNA targeting the same mRNA constitute a triplet. Some of the systemic processes that we observe are expected, like "Cholesterol biosynthetic process via 24,25 dihydrolanesterol", "Regulation of viral life cycle" and processes involved in calcium

signaling as well as inflammatory responses (Figure 18), because they are known to be IRE1-dependent [53]. Interestingly, all prioritized features have been previously implicated with TNBC prognosis or progression [49][50][51][52]. Especially, miR-221 has been suggested to promote tumorigenesis in TNBC.[49] Other interesting results include the identification of the biological process “Regulation of G1/S transition of mitotic cell cycle”, and that miR-221 is required for this process (Figure 19). Processes that involve cell proliferation, differentiation or apoptosis are critical in cancer pathology. Next, we move to explore the features whose expression is inversely correlated (Figure 20). The pathway enrichment analysis for the network of features with reverse expression yielded one prioritized gene CXCL8 (Figure 21). This gene, apart from being involved in “Response to Endoplasmic Reticulum Stress” systemic process (Figure 22), a biological ontology that we know will be deregulated because of the nature of the drug administrated, has been implicated in TNBC brain metastasis and recent studies suggest it could be used as a prognostic biomarker. [50]

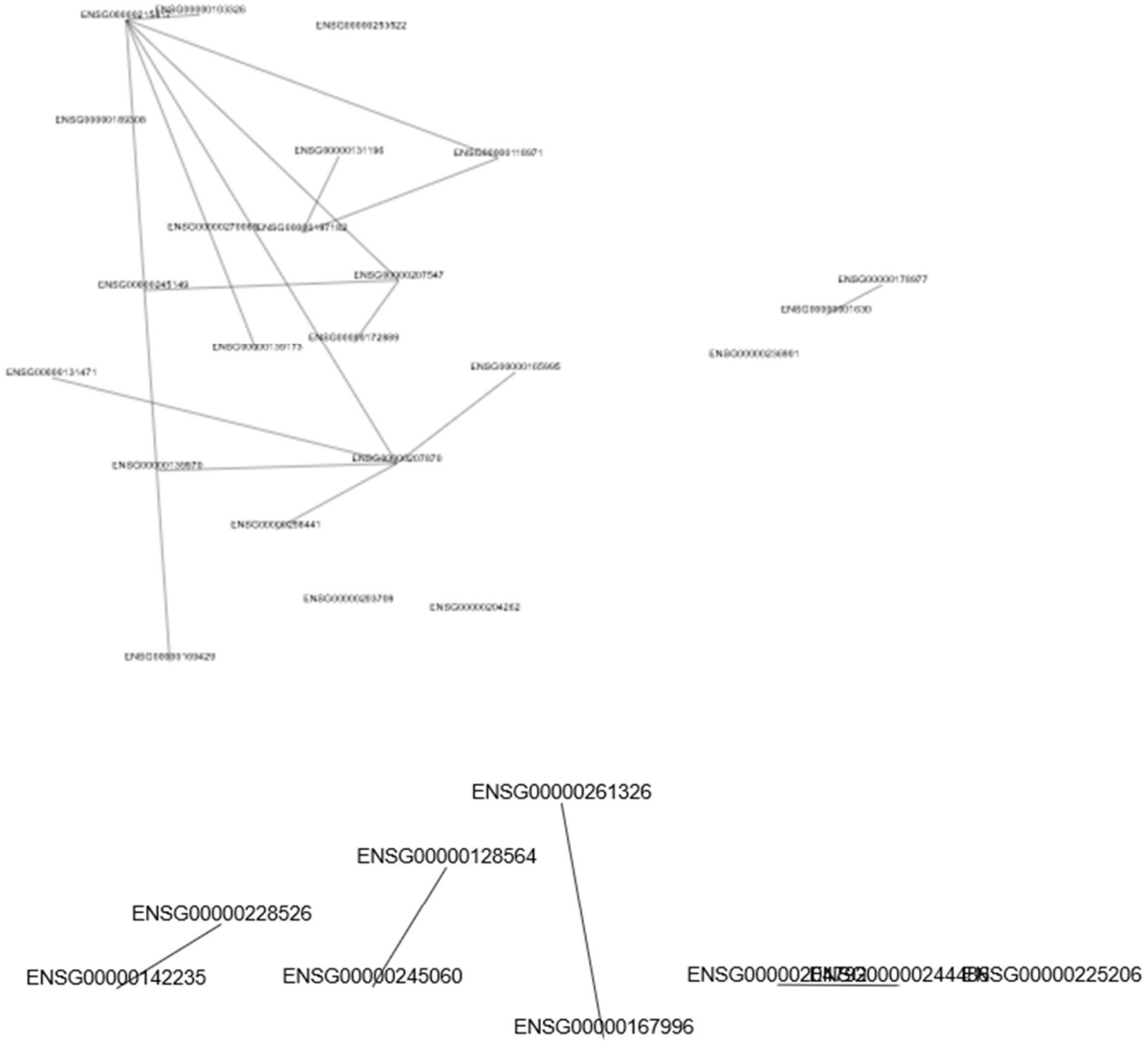


*Figure 17 Network of triplets of interaction at 8-hour time point, upon IRE1 inhibition with MKC8866. Triplets are defined when the expression of two features is related to an expression of a third one. For example, a miRNA and a lncRNA targeting the same mRNA constitutes a triplet.*

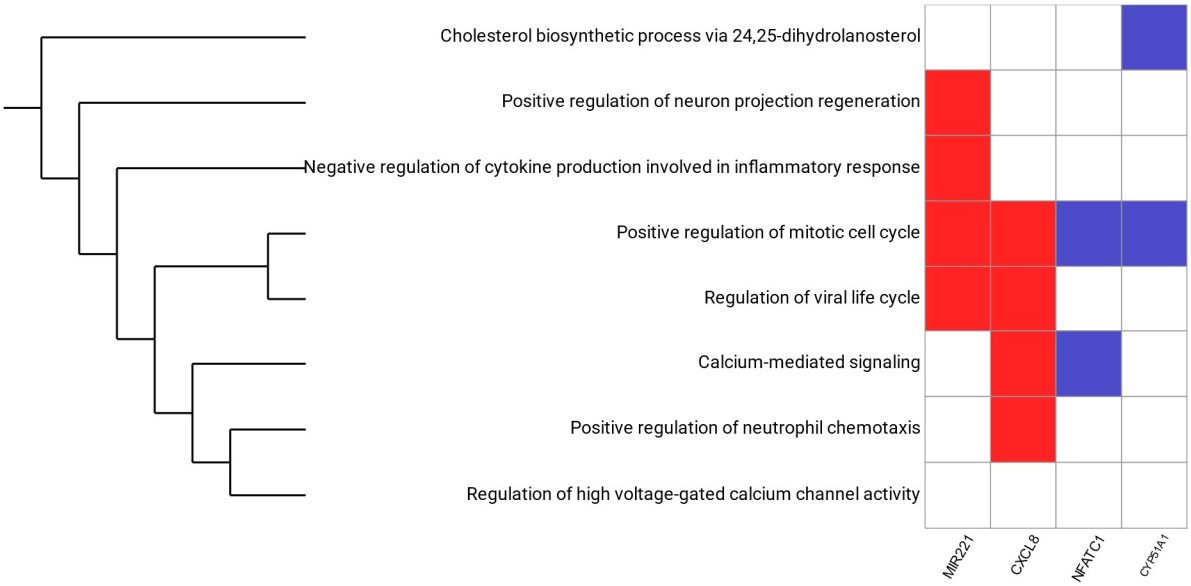


*Figure 18 Heatmap of hub genes and prioritized processes of network of triplets at 8-hour time point, upon IRE1 inhibition with MKC8866. Triplets are defined when the expression of two features is related to an expression of a third one. For example, a miRNA and a lncRNA targeting the same mRNA constitutes a triplet.*

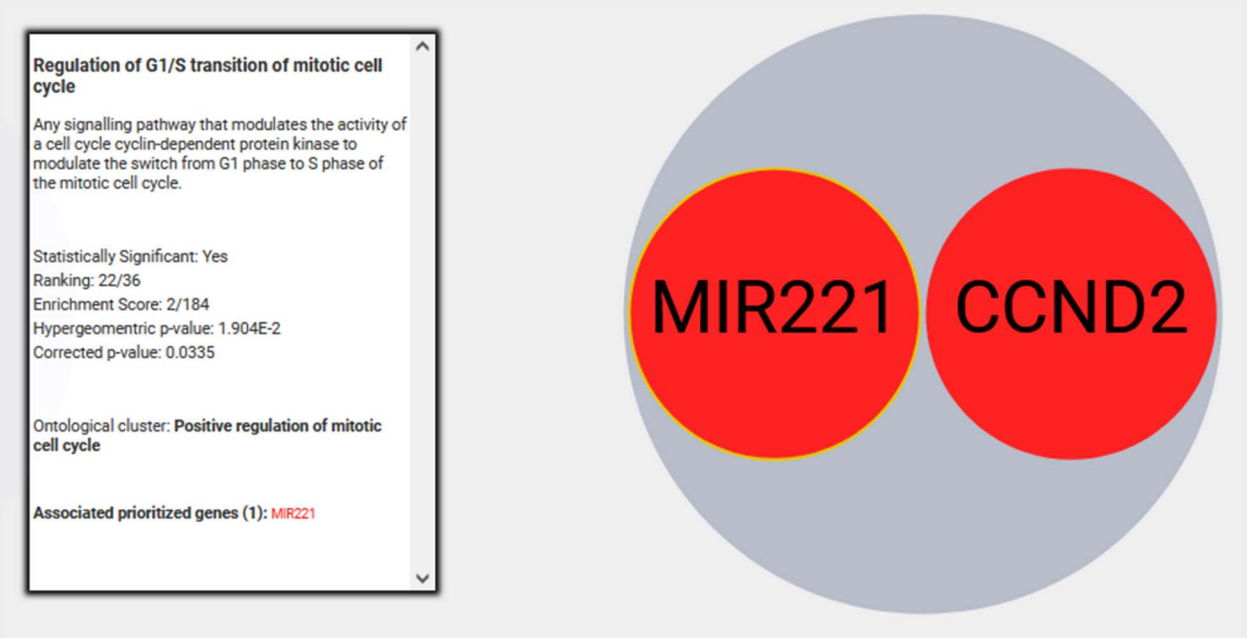


*Figure 19 Focus on one of the prioritized systemic processes enriched from the network of triplets of interaction at 8-hour time point, upon IRE1 inhibition with MKC8866*

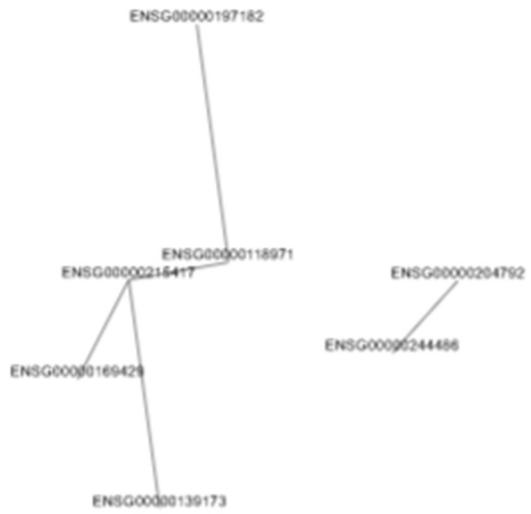


*Figure 20 Network of reverse-expression at 8-hour time point upon IRE1 inhibition with MKC8866*

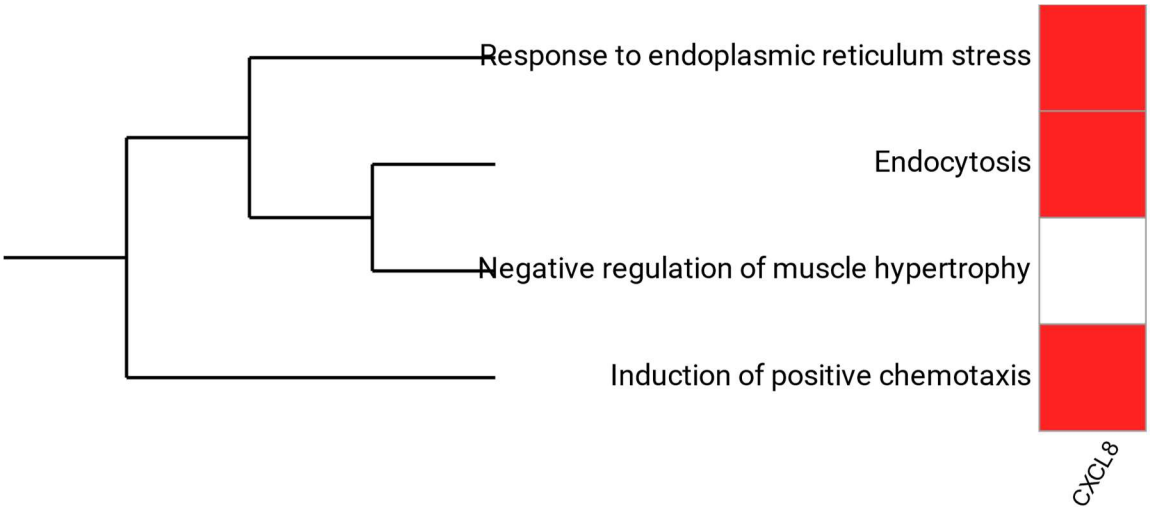


*Figure 21 Heatmap of network of reverse expression at 8-hour time point, upon IRE1 inhibition with MKC8866*

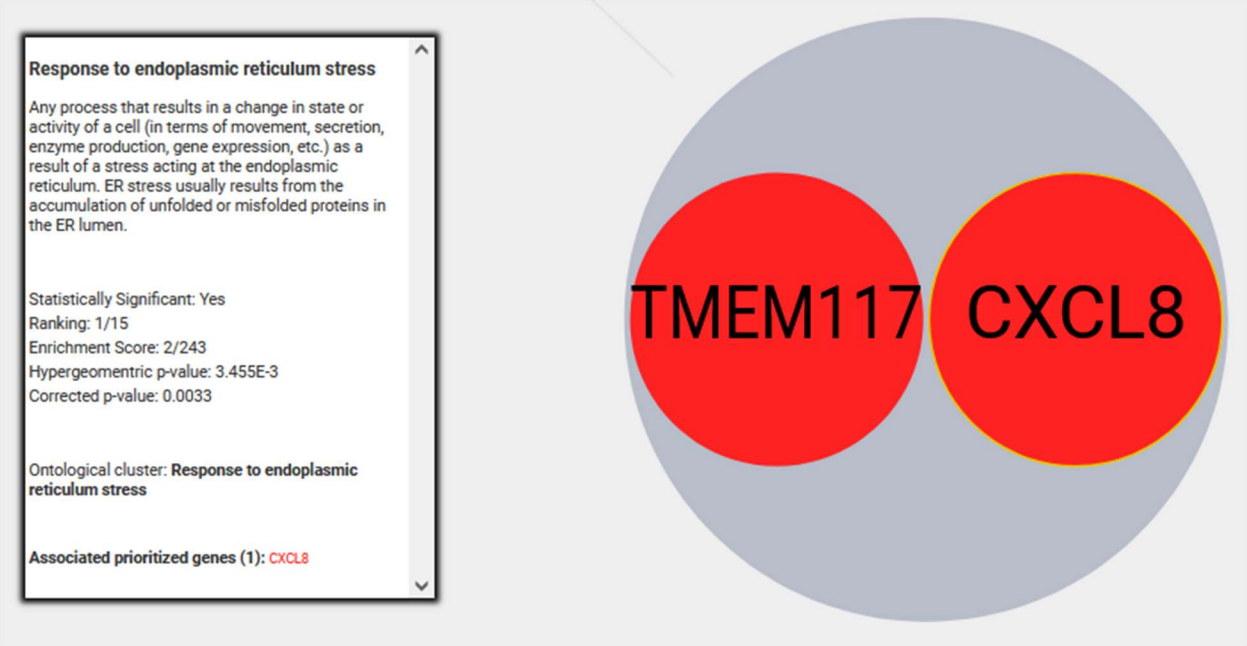


*Figure 22 Focus on one of the prioritized systemic processes enriched from the network of reverse expression at 8-hour time point, upon IRE1 inhibition with MKC8866*

At the 24-hour time point we observed processes involved in differentiation processes, something that was also expected. Continuing the trend of the 8-hour time point, all the enriched features at 24 hours, both for the network or triplets and for the network of reverse expression (Figures 23 and 26) are dictated by the literature to be involved in TNBC progression and pathology [45]. This shows that our method is robust and can capture important biological information. Some of the systemic processes that stand out are the “Regulation of apoptotic signaling pathway” because it is widely known from literature to be IRE1-RIDD regulated and “Gene Silencing by miRNA” due to the focus on non-coding molecules and due to the biological implication of this process, which is cellular proliferation (Figures 24, 25, 27 and 28) .

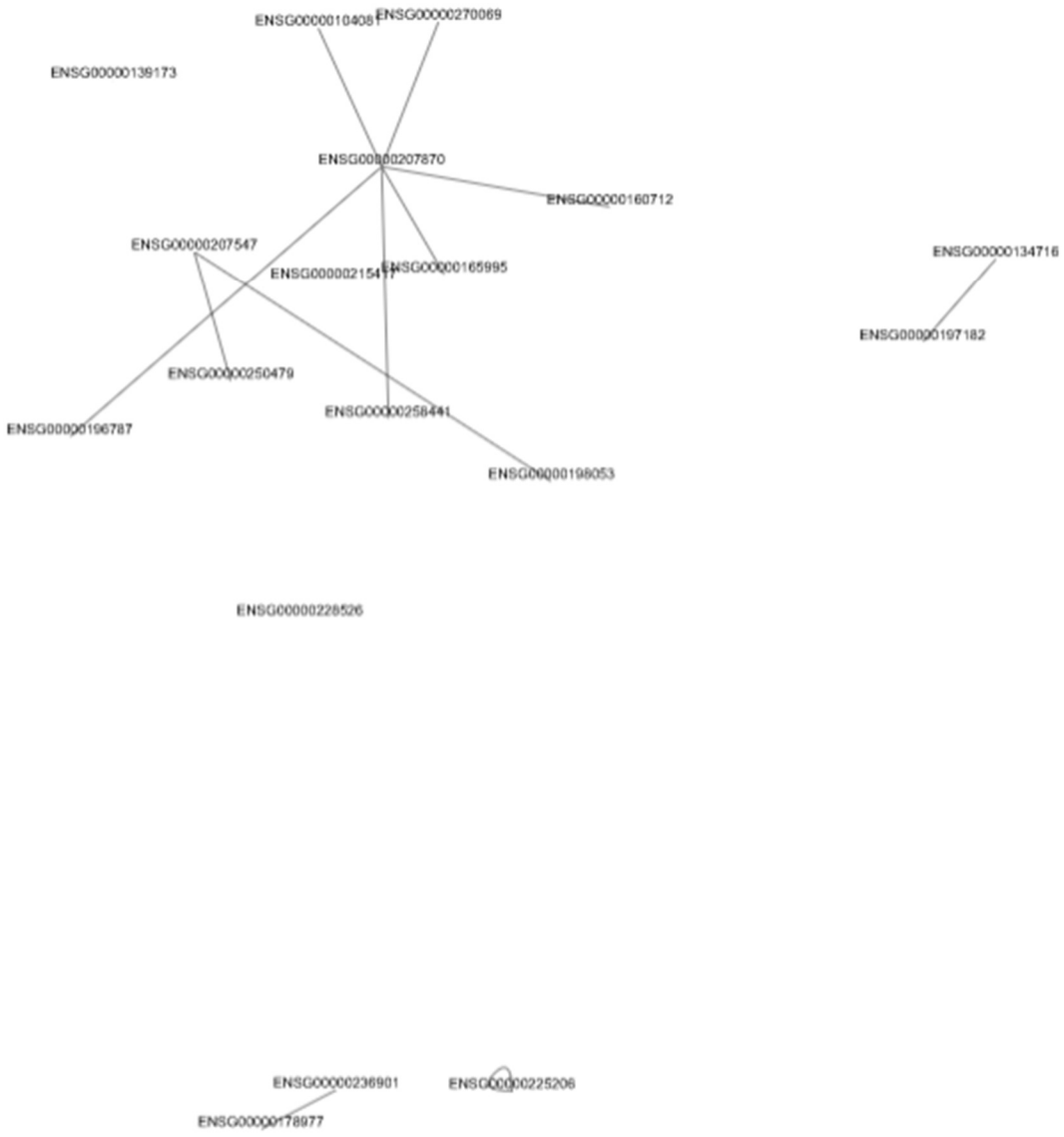


*Figure 23 Network of triplets of interaction at 24-hour time point, upon IRE1 inhibition with MKC8866. Triplets are defined when the expression of two features is related to an expression of a third one. For example, a miRNA and a lncRNA targeting the same mRNA*

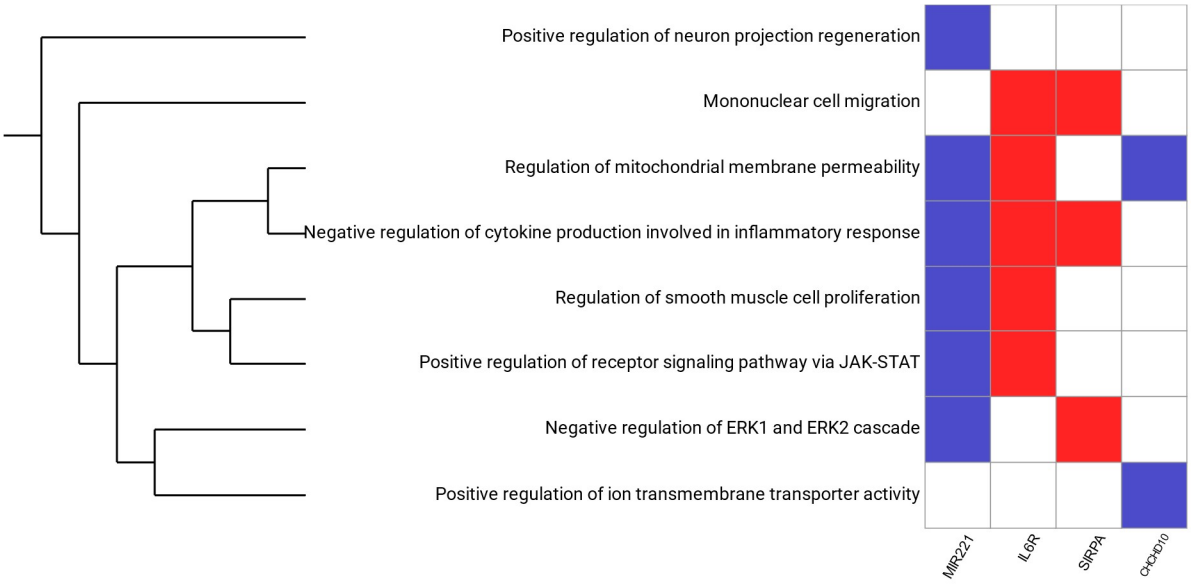


*Figure 24 Heatmap of hub genes and prioritized processes of network of triplets at 24-hour time point, upon IRE1 inhibition with MKC8866. Triplets are defined when the expression of two features is related to an expression of a third one. For example, a miRNA and a lncRNA targeting the same mRNA*

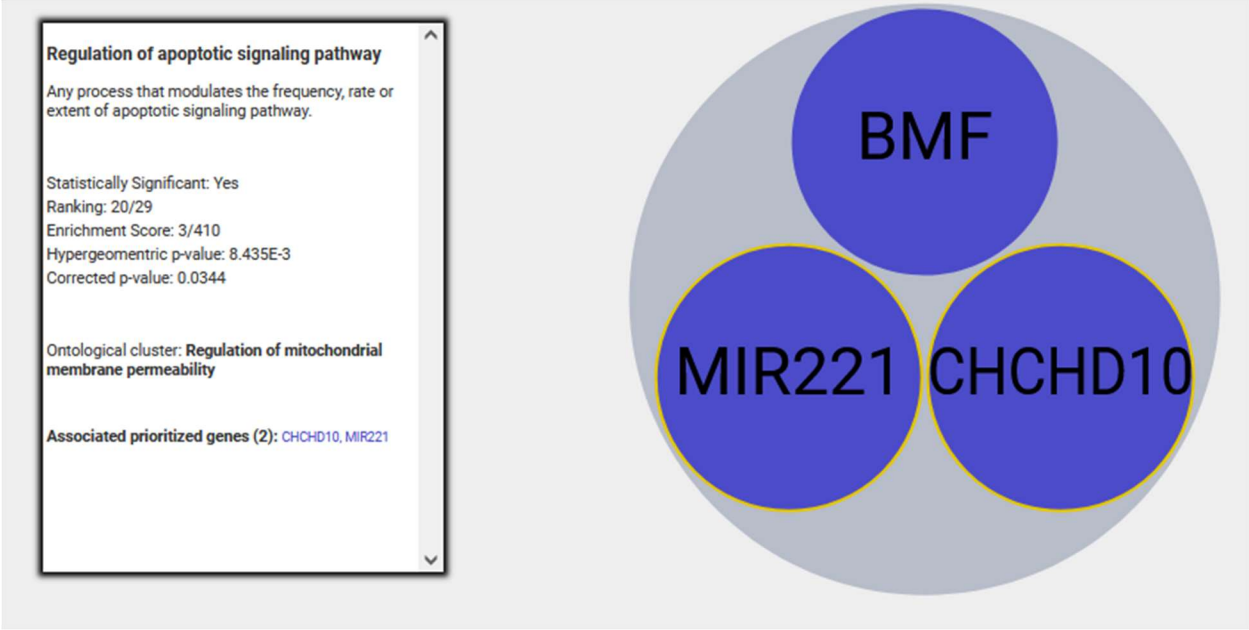


*Figure 25 Focus on one of the prioritized systemic processes enriched from the network of triplets of interaction at 24-hour time point, upon IRE1 inhibition with MKC8866*

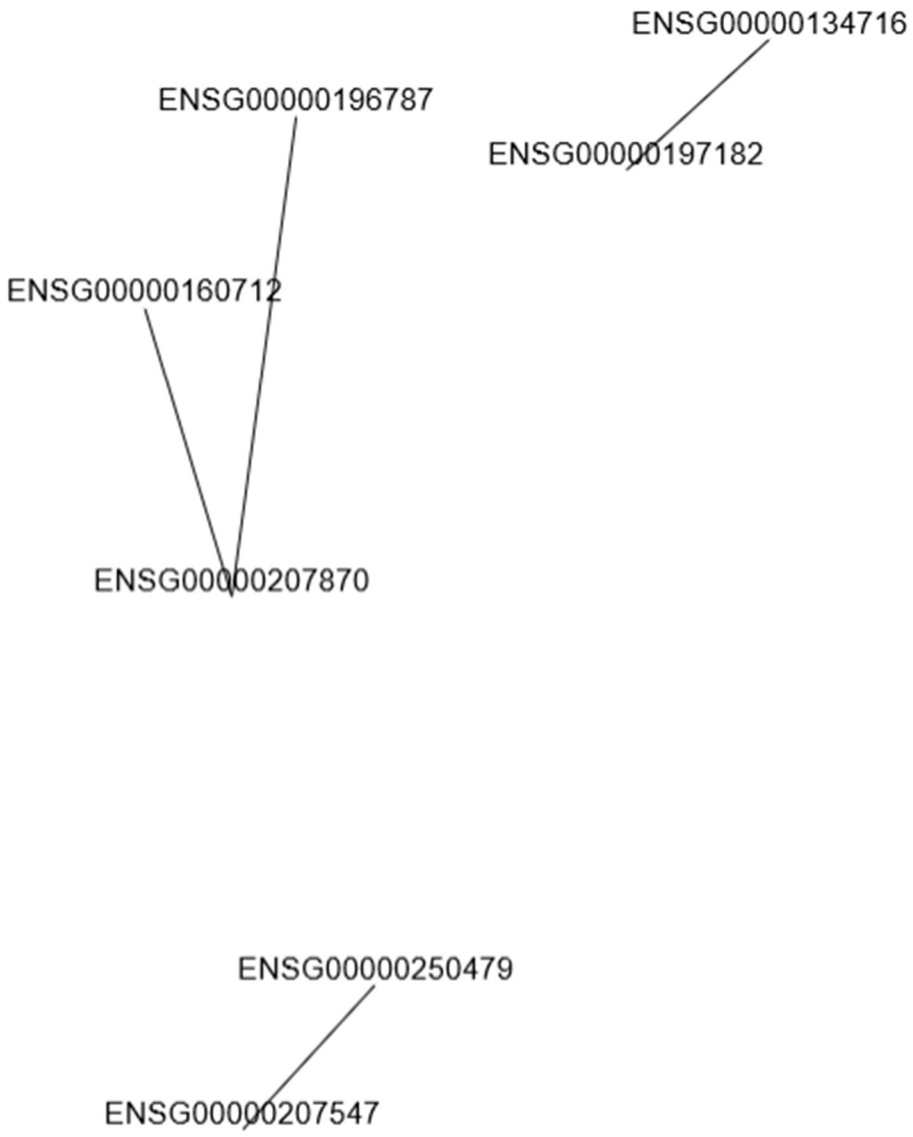


*Figure 26 Network of reverse-expression at 24-hour time point upon IRE1 inhibition with MKC8866*

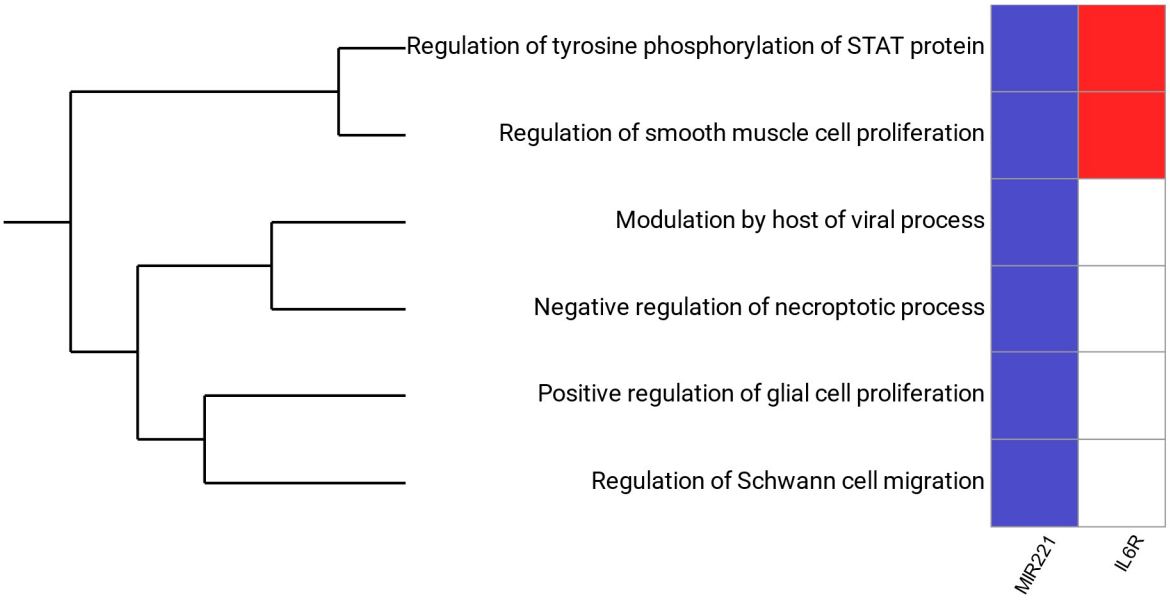


*Figure 27 Heatmap of hub genes and prioritized processes of network of reverse-expression at 24-hour time point, upon IRE1 inhibition with MKC8866*

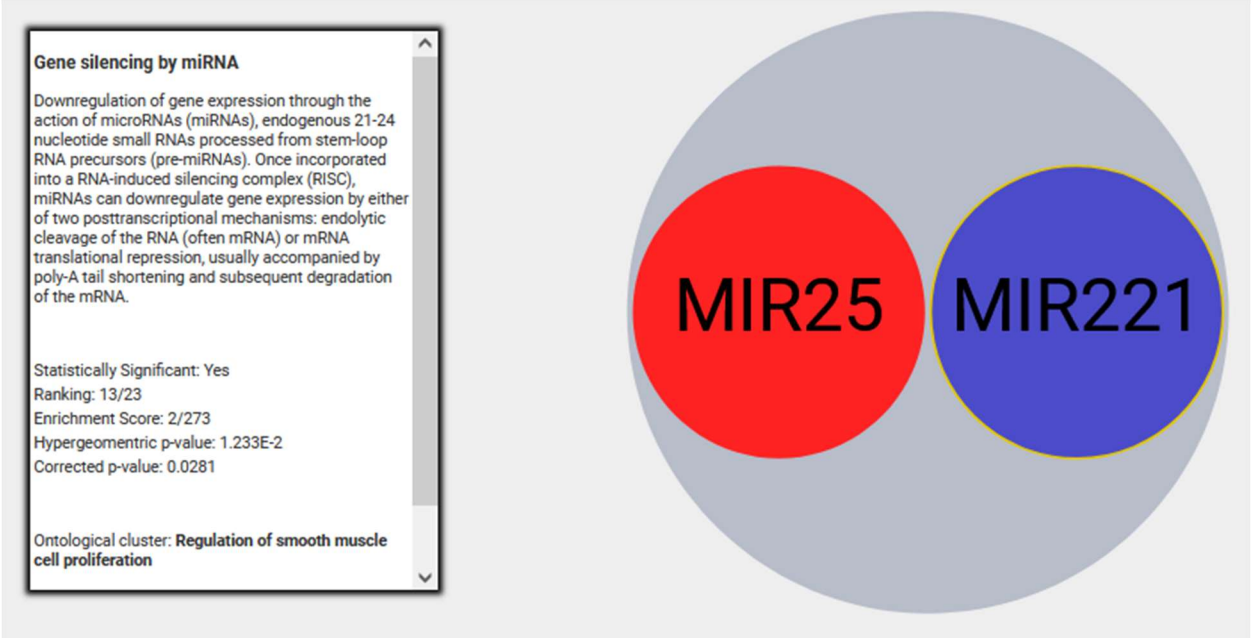


*Figure 28 Focus on one of the prioritized systemic processes enriched from the network of reverse-expression at 24-hour time point, upon IRE1 inhibition with MKC8866*

### *2. IRE1 signature and patients from TCGA*

As previously discussed, IRE1 RNase signaling pathway consists of two arms: XBP1 and RIDD. The cleavage of 26 nucleotides from the mRNA of XBP1 leads to the formation of an active transcription factor XBP1s, which can go into the nucleus and initiate the transcription of various genes that will help alleviate the unfolded protein load. In general, activation of XBP1 arm has pro survival effects on the cell. RIDD arm, on the other hand, has been linked with pro-apoptotic effects. In fact, researchers suggest that the decision of life and death comes from a fine-tuned balance between these two arms. [53] These findings together with the fact that IRE1 seems to be critical in various cancers' progression, shape an intricate molecular landscape with many cross-talks between the various signaling pathways. We considered investigating further this phenomenon with a special interest in non-coding biomarkers.
First, we obtained transcriptomic data for patients with TNBC from the TCGA database. Then, we stratified the patients based on IRE1 activity. To do that, we used a previously published IRE1 signature consisting of 36 genes, 18 regarding XBP1 arm and 18 referring to RIDD arm. [54] Due to the peculiarity of our signature we had to slightly modify our custom tool to meet our unique demands. Specifically, we assessed high IRE1 activity when the 18 genes of XBP1 arm were upregulated and the 18 of RIDD arm were downregulated and vice versa. In this way we produced four subgroups of patients that were divided as we discussed based on their relative IRE1 expression.
Next, we correlated IRE1 high and low activity with the patients' survival probability and saw a significant correlation as shown in Figure 29.

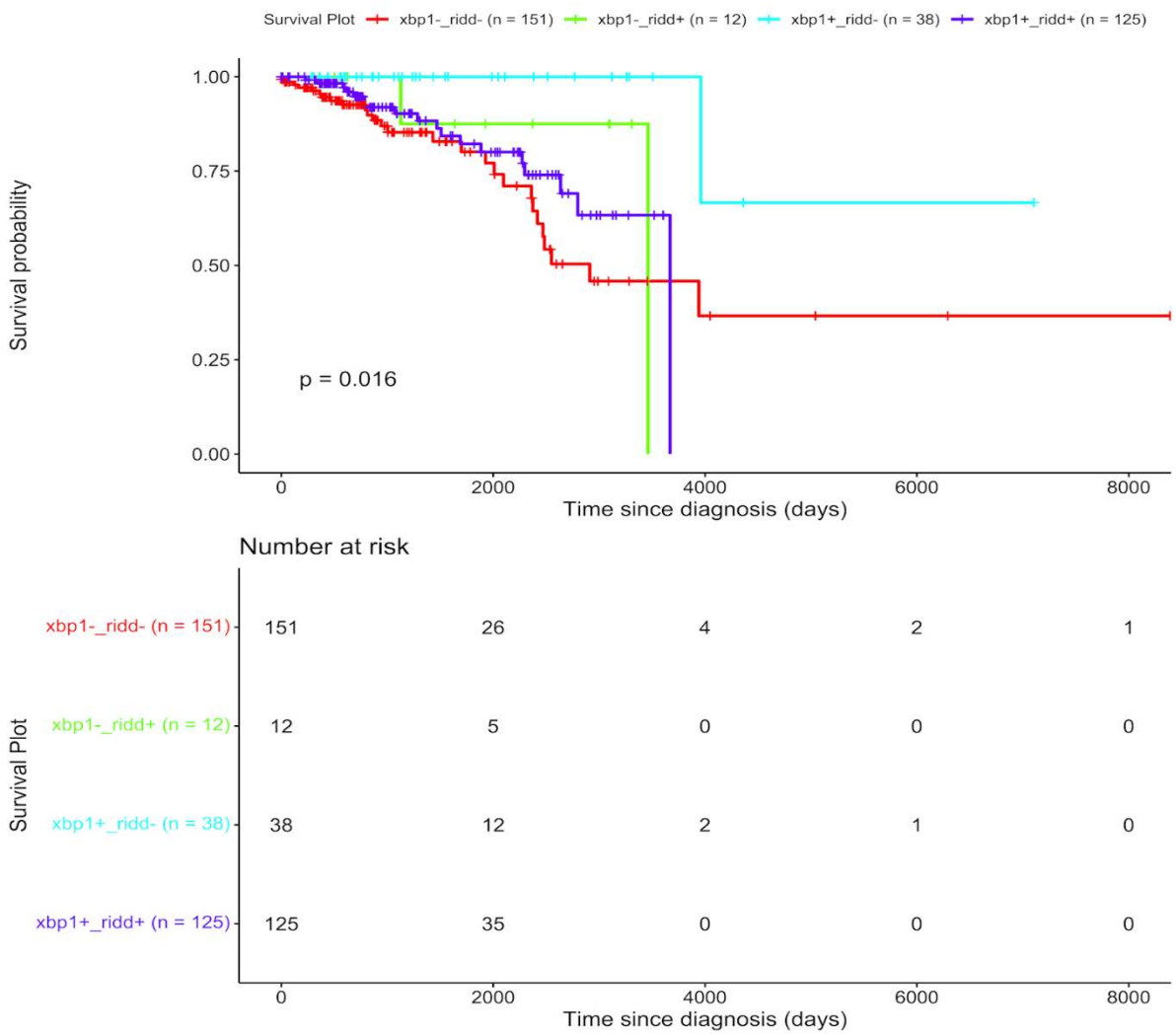


*Figure 29 Survival analysis of the different sub-groups of patients based on IRE1 activity*

After that we wished to investigate the potential implication of non-coding molecules to the observed phenotype. So, we took transcriptomic data for these 2 groups of patients (IRE1 high and low) and imported them into our Galaxy platform. We analyzed the data using the standalone second part of our workflow, starting from count matrices.

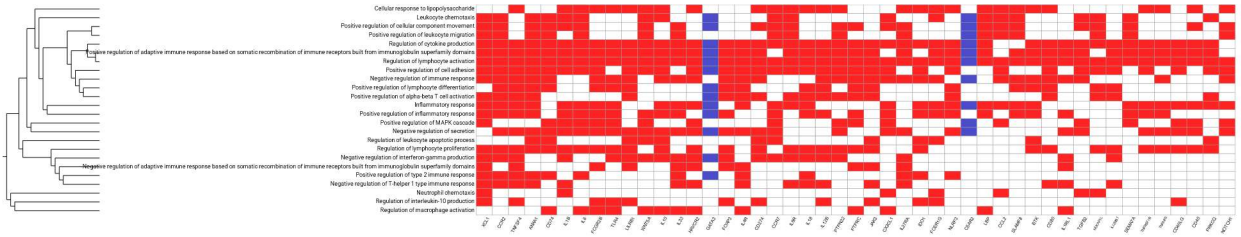


*Figure 30 Heatmap of prioritized features and processes under XBP1 arm*

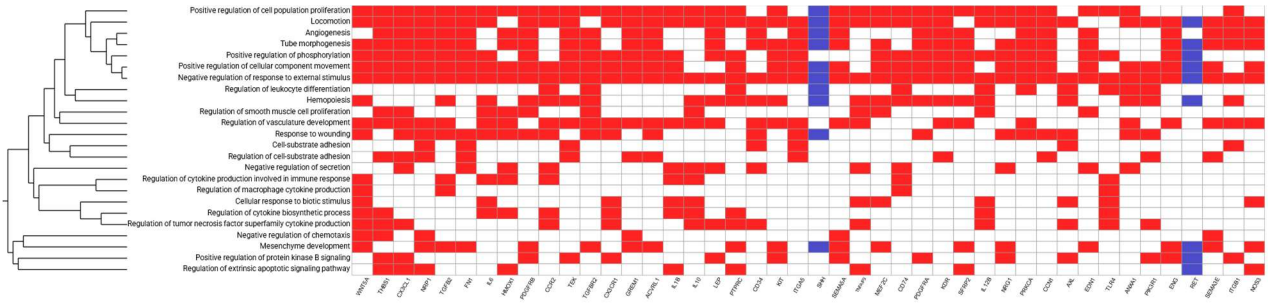


*Figure 31 Heatmap of prioritized features under RIDD arm*

The results yielded 200 lncRNA related to RIDD and XBP1 and 89 out of 90 noncoding molecules were common between the two signatures, suggesting that at least at the non-coding regulatory RNA level no major differences are found. As shown in both figures (Figure 30 and 31), the processes that seem to be deregulated in patients with IRE1 high and IRE1 low activity are mainly involved in immune signaling, inflammation and cell migration. This could explain why high IRE1 activity correlated with poorer survival outcome. It is an interesting hypothesis which requires further investigation. Most importantly, this kind of hypothesis was deducted after a 5min click-based analysis. This shows clearly how elimination of time and money-consuming tasks, like the from-scratch development of RNAseq analysis, can give rise to interesting questions and can advance our knowledge in the field.

## Conclusions

It has become apparent that with the explosion of OMIC technologies an analogous explosion in analyzing the data needs to follow, in order to utilize them at their full potential. The bottleneck from experimental results to biological insights relies on the analysis of the vast amount of data, due to the demand of high computational power and people with specific expertise. In an effort to overcome this issue, a local Galaxy instance was developed with emphasis on genomic and transcriptomic analyses.

Galaxy infrastructure was chosen for many reasons. Naturally one is its high popularity. Millions of people worldwide use Galaxy.org or have developed their own instances depending on their needs. Another important aspect is its layout simplicity. It is a very user-friendly platform, and no programming knowledge is required although this possibility to an experienced user is supported. Finally, many supporting venues that cover most usual problems encountered can be found online.

Tools that weren't publicly available were created using R or python languages and incorporated into DOCKERs. The idea of using this infrastructure as our back-end is reproducibility and scalability of the project. DOCKERs provide the possibility of "packaging" your application and using it on any server, eliminating the dependencies issues, making an ideal option for use on cloud services. Notably, the whole Galaxy instance can be incorporated in a docker, and this is one of the future steps we want to take.

Our Galaxy instance provides a large variety of tools for genomic and transcriptomic analyses. There is a section dedicated to variant calling annotation analysis, called VCF-analyzer, which performs the task of annotating, filtering and prioritizing variants produced by genomic studies. There is another section with tools useful for metagenomic analysis, called ANASTASIA. These tools are used for identification of the distinct organisms and/or pathways enriched in a biological sample. In addition, many tools are offered for various visualizations and text manipulation, tasks used routinely in every kind of analysis.

Special focus was given to the creation of a novel lncRNA workflow. The relative recent discovery of their implication in various phenotypes, including pathologies has set the need of establishing a standard protocol of analyzing these molecules. This workflow was created, in a module-based format. That means that the whole workflow can be divided into two sub workflows, and then these workflows consist of modules and the modules have one to three tools. This design offers

a way to build highly versatile frameworks depending on specific demands and can be scaled up extremely easily.

The lncRNA workflow is very comprehensive, covering analysis from BAM files to biological insights about the coding and non-coding networks and systemic processes that exist in a biological sample, meeting the growing need of a unified pipeline for non-coding biomarker discovery. Notably, our flexible design covers all potential users who might not have access to raw files but have downstream files.

In summary, the lncRNA pipeline, takes raw FASTQ files and performs quality control on those RAW files. Extensive reports are provided to the user so one can decide if they can proceed with downstream analysis. Alternatively, if someone is confident about their reads quality, they can directly import the BAM files and proceed from there. Different reference databases are used based on the type of RNA molecule in question. Regarding the lncRNAs extra steps of assessing the coding potential of the transcripts are taken in order to ensure the that the remaining RNAs are in fact non coding molecules. Once the different count matrices, representing miRNAs, mRNAs and lncRNAs are produced the pipeline performs differential expression analysis on each of them with non-parametric statistical tests to ensure robustness of results. After differential expression is finished, an intelligent pathway enrichment analysis is performed that takes into account not only the position of each feature on their interaction network but also the biological function that they reside in. In addition, tools that can test both established and novel interactions between RNA molecules are provided in order to unravel the full regulatory network of these molecules. All interactions can be visualized with tools of our instance.

Two case studies are presented in this thesis covering some of the possibilities provided by our instance. The first study explores RNA-seq data from a published study on triple negative breast cancer. The purpose of this meta-analysis was to give special emphasis on non-coding features. The results include features supported by literature to have implications on TNBC progression and also processes that are involved in non-coding signaling. For example, our pipeline confirmed the importance of mir221 in TNBC along with systemic processes like lipid metabolism and apoptosis. The second study explores the possibility of stratification of patients from TCGA based on a signature of interest and the consequent exploration of the different non-coding features expressed on the groups produced. Basically, we wanted to investigate which non-coding molecules are differentially expressed because of the differential IRE1 activity present in the different groups. First, stratification of patients and survival analysis was performed using a published IRE1 signature. After that, differential expression analysis was performed between the

groups produced with special focus on non-coding molecules. With this feature we could perform pathway enrichment analysis. Results indicated higher rates of immune signaling and inflammation in patients with higher IRE1 activity. These findings pose an interesting potential correlation between IRE1 activity, immune response and survival probability, worthy of further investigation.

Although we live in an era of continuous technological advancements there are still many unanswered questions regarding the flow of information in a cell. This knowledge is crucial because it will unlock many possibilities both for clinical and medical applications. OMICS technologies are promising but without the use of standardized tools and methods they will only contribute to the accumulation of stored purposeless data. In a world where the amount of data is exponentially increasing every day it is vital to have the means to use this data for our benefit. I believe that this instance will help with this effort, due to its variety and practicality of tools, potential for scalability and most importantly its user-friendly environment.

# REFERENCES


1. Richard P Horgan MRCOG MRCPI, Louise C Kenny MRCOG, 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics,18 July 2011, The Obstetrician & Gynaecologist, https://doi.org/10.1576/toag.13.3.189.27672Citations: 242, DOI:10.1576/toag.13.3.189.27672
2. Bo Wang, Vivek Kumar, Andrew Olson and Doreen Ware, Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing, Front. Genet., 26 April 2019, Sec. RNA, Volume 10 - 2019 | https://doi.org/10.3389/fgene.2019.00384 Casamassimi A., Federico A., Rienzo M., Esposito S., Ciccodicola A. (2017). Transcriptome profiling in human diseases: new advances and perspectives. Int. J. Mol. Sci. 18:E1652. 10.3390/ijms18081652
3. Taub, Floyd (1983). "Laboratory methods: Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs". DNA. 2 (4): 309–327. doi:10.1089/dna.1983.2.309.
4. Khan, M., Palakolanu, Sudhakar Reddy, Gupta, Ravi, 2022/05/22, Advancements in Developing Abiotic Stress-Resilient Plants: Basic Mechanisms to Trait Improvements. BOOK
5. Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. doi: 10.1038/nrg.2016.49
6. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484
7. Conesa, A., Madrigal, P., Tarazona, S. et al. A survey of best practices for RNA-seq data analysis. Genome Biol 17, 13 (2016). https://doi.org/10.1186/s13059-016-0881-8
8. Jiao, Y. P., Peluso, P., Shi, J. H., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature 546, 524–527. doi: 10.1038/nature22971
9. https://galaxyproject.org/galaxy-project/statistics/
10. 
11. "Galaxy Community Hub - Galaxy Community Hub". https://galaxyproject.org/galaxy-team/
12. Lazarus, R.; Taylor, J.; Qiu, W.; Nekrutenko, A. (2008). "Toward the commoditization of translational genomic research: Design and implementation features of the Galaxy genomic workbench". Summit on Translational Bioinformatics. 2008: 56–60.
13. https://www.docker.com/
14. The Variant Call Format and VCFtools, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011
15. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672
16. "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift", Cingolani, P., et. al., Frontiers in Genetics, 3, 2012.
17. COSMIC: the Catalogue of Somatic Mutations in Cancer (Tate et al., 2018)

18. Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam, ”MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics, Volume 31, Issue 10, 15 May 2015, Pages 1674–1676, https://doi.org/10.1093/bioinformatics/btv033
19. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014 Jul 15;30(14):2068-9. PMID:24642063
20. Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0
21. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.
22. Sean R. Eddy and the HMMER development team, “hmmer”, http://hmmer.org
23. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov 24;6:26. doi: 10.1186/1748-7188-6-26. PMID: 22115189; PMCID: PMC3319429.
24. Andrews, S. (2010), 'FASTQC. A quality control tool for high throughput sequence data' .
25. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
26. Shumate A, Wong B, Pertea G, Pertea M Improved transcriptome assembly using a hybrid of long and short reads with StringTie, PLOS Computational Biology 18, 6 (2022), doi.org/10.1371/journal.pcbi.1009730
27. S Anders, T P Pyl, W Huber: HTSeq — A Python framework to work with high-throughput sequencing data. bioRxiv 2014. doi: 10.1101/002824.
28. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England), 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
29. Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, Wei Li, , CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model, Nucleic Acids Research, Volume 41, Issue 6, 1 April 2013, Page e74, https://doi.org/10.1093/nar/gkt006
30. Yu-Jian Kang, De-Chang Yang, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei, Ge Gao , CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features, Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W12–W16, https://doi.org/10.1093/nar/gkx428
31. Schneider, H., Raiol, T., Brigido, M. et al. A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts. BMC Genomics 18, 804 (2017). https://doi.org/10.1186/s12864-017-4178-4
32. Guangyu Wang, Hongyan Yin, Boyang Li, Chunlei Yu, Fan Wang, Xingjian Xu, Jiabao Cao, Yiming Bao, Liguo Wang, Amir A Abbasi, Vladimir B Bajic, Lina Ma, Zhang Zhang, Characterization and identification of long non-coding RNAs based on feature relationship, Bioinformatics, Volume 35, Issue 17, 1 September 2019, Pages 2949–2956, https://doi.org/10.1093/bioinformatics/btz008

33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). *"limma powers differential expression analyses for RNA-sequencing and microarray studies."* Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.
34. Francesco Del Carratore, Andris Jankevics, Rob Eisinga, Tom Heskes, Fangxin Hong, Rainer Breitling, *RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets* , Bioinformatics, Volume 33, Issue 17, 01 September 2017, Pages 2774–2775, https://doi.org/10.1093/bioinformatics/btx292
35. Breitling, R., & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. Journal of bioinformatics and computational biology, 3(5), 1171–1189. https://doi.org/10.1142/s0219720005001442
36. Teng, X., Chen, X., Xue, H., Tang, Y., Zhang, P., Kang, Q., … He, S. (2019). NPInter v4.0: an integrated database of ncRNA interactions. Nucleic Acids Research, 1–6. https://doi.org/10.1093/nar/gkz969
37. Gong J, Shao D, Xu K, Lu Z, Lu ZJ, Yang YT, Zhang QC. RISE: a database of RNA interactome from sequencing experiments. Nucleic Acids Res. 2018 Jan 4;46(D1):D194-D201. doi: 10.1093/nar/gkx864. PMID: 29040625; PMCID: PMC5753368.
38. Tsukasa Fukunaga, Michiaki Hamada, Bioinformatics, Volume 33, Issue 17, 01 September 2017, Pages 2666–2674, https://doi.org/10.1093/bioinformatics/btx287
39. Shannon, P. et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), pp.2498–2504.
40. http://dx.doi.org/10.26240/heal.ntua.21609
41. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). " TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." Nucleic Acids Research. doi: 10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.
42. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010 Oct;1(4):274-8. doi: 10.4103/0974-7788.76794. PMID: 21455458; PMCID: PMC3059453.
43. Almanza A, Carlesso A, Chintha C, Creedican S, Doultsinos D, Leuzzi B, Luís A, McCarthy N, Montibeller L, More S, Papaioannou A, Püschel F, Sassano ML, Skoko J, Agostinis P, de Belleroche J, Eriksson LA, Fulda S, Gorman AM, Healy S, Kozlov A, Muñoz-Pinedo C, Rehm M, Chevet E, Samali A. Endoplasmic reticulum stress signalling - from basic mechanisms to clinical applications. FEBS J. 2019 Jan;286(2):241-278. doi: 10.1111/febs.14608. Epub 2018 Aug 4. PMID: 30027602; PMCID: PMC7379631.
44. Urra et al., Trends in Cancer. 2016, 2, 252
45. Logue et al., Nature Comm. 2018, **9**, 3267
46. Bashir et al., Life Sciences, 2021, 118740
47. Almanza et al., Nature Comm. 2022, **13**, 2493
48. Nassirpour, R., Mehta, P. P., Baxi, S. M., & Yin, M. J. (2013). miR-221 promotes tumorigenesis in human triple negative breast cancer cells. PloS one, 8(4), e62170. https://doi.org/10.1371/journal.pone.0062170 (Retraction published PLoS One. 2017 Apr 10;12 (4):e0175869)
49. Shen, Y., Zhang, B., Wei, X., Guan, X., & Zhang, W. (2022). CXCL8 is a prognostic biomarker and correlated with TNBC brain metastasis and immune infiltration.

International immunopharmacology, 103, 108454. https://doi.org/10.1016/j.intimp.2021.108454
50. Quang CT, Leboucher S, Passaro D, Fuhrmann L, Nourieh M, Vincent-Salomon A, Ghysdael J. The calcineurin/NFAT pathway is activated in diagnostic breast cancer cases and is essential to survival and metastasis of mammary cancer cells. Cell Death Dis. 2015 Feb 26;6(2):e1658. doi: 10.1038/cddis.2015.14. PMID: 25719243; PMCID: PMC4669815.
51. Lijun Cheng, Bryan P Schneider, Lang Li, A bioinformatics approach for precision medicine off-label drug drug selection among triple negative breast cancer patients, Journal of the American Medical Informatics Association, Volume 23, Issue 4, July 2016, Pages 741–749, https://doi.org/10.1093/jamia/ocw004
52. Maurel, M., Chevet, E., Tavernier, J., & Gerlo, S. (2014). Getting RIDD of RNA: IRE1 in cell fate regulation. Trends in biochemical sciences, 39(5), 245–254. https://doi.org/10.1016/j.tibs.2014.02.008
53. Lhomond, S., Avril, T., Dejeans, N., Voutetakis, K., Doultsinos, D., McMahon, M., Pineau, R., Obacz, J., Papadodima, O., Jouan, F., Bourien, H., Logotheti, M., Jégou, G., Pallares-Lupon, N., Schmit, K., Le Reste, P. J., Etcheverry, A., Mosser, J., Barroso, K., Vauléon, E., … Chevet, E. (2018). Dual IRE1 RNase functions dictate glioblastoma development. EMBO molecular medicine, 10(3), e7929. https://doi.org/10.15252/emmm.201707929
54. E.A. Milward, H. Hondermarck, in Encyclopedia of Cell Biology, 2016