# CBS-miRSeq v.1.0 User Manual

Color and base-space comprehensive profiling for microRNA sequencing

Date: 24/01/2017

Unit of Immunology and Functional Genomics
Centro Cardiologico Monzino IRCCS, Milan, Italy
Rupesh K. Kesharwani, Email: bioinforupesh2009.au@gmail.com

# CONTENTS

## 1. Overview of CBS-miRSeq

MiRNAs play important regulatory role in many biological processes in the cell including proliferation and differentiation, apoptosis, brain development, heart development [1] and are involved in a variety of diseases, such as tumor [2], cardiovascular disease [3], chronic lymphocytic leukemia, and viral infections [4, 5]. The development of Next Generation Sequencing (NGS) platforms allowed researchers to characterize small RNA profiles in several tissues with higher speed, accuracy and resolution than other techniques. Hence, a number of online and local tools were developed to analyze small RNA-Seq data, but inaccurate processing, lacking of optimal parameterization and comprehensiveness, outdated reference genome and annotations, uploading and input format issues limit their use.

Our goal was to extend the findings of the previous studies and developments; therefore, we proposed a fully customized bioinformatics pipeline (Color and Base-Space microRNA-Seq - CBS-miRSeq) (**Fig. 1)** for the seamless processing of miRNA-Seq data for both color-space (csfasta) and base-space (fastq) short reads generated by SOLiD and Illumina sequencers, respectively. The pipeline is based on *bash*, *perl* and *R* scripts, which accomplish reads pre-processing, quality assessment, filtering, adapter trimming, mapping, identification of miRNA variants (isomiRs), discovery of novel miRNAs, differential analysis, miRNA:mRNA target prediction, functional analysis (Ontology and Pathways analysis) and summarization of the results.

The workflow helps researchers to use different modules flexibly, for datasets spanning from a few to hundreds of samples. Results and outputs in html, pdf, and csv formats enable easy visual inspection and understanding for further investigations**.**

The pipeline can be run sequentially on a single machine or parallel in a cluster/server/cloud systems.

**Features of CBS-miRSeq**

- ✓ Fully customized and flexible (modularity); it is useful to control the outputs from every analysis step or to start analysis at different points of the pipeline.
- ✓ CBS-miRSeq supports different formats of raw data input, such as color-space and base-space.

- ✓ CBS-miRSeq supports almost every reference species and annotations; optionally the pipeline allows user to download the newest release versions of reference genome and annotations from Ensembl and miRBase databases.
- ✓ Option to identify differentially expressed (DE) miRNAs by two statistical methods and/or obtain results at the intersection of both methods.
- ✓ Exploration of the distribution of Ensembl biotypes (tRNA, rRNA, snoRNA, miRNAs, protein coding genes) and visualization through bar and pie charts.
- ✓ Identification of isomiRs (RNA editing events) and complete visualization.
- ✓ Implementation of two miRNA gene target prediction algorithms to obtain the most consistent results at the intersection.
- ✓ Prediction of novel miRNAs and summarization of unique features.
- ✓ Functional annotation analysis for target genes of DE miRNAs as well as of novel miRNAs predicted by the pipeline.
- ✓ Generation of mapping and pre-processed statistical reports of short-reads (quality control, read length, adapter trimming, mapping statistics).
- ✓ Complete visual analysis of results and production of interactive html, pdf/postscript, and csv tables for further investigation.

The CBS-miRSeq is a powerful bioinformatics pipeline that researcher can use to perform comprehensive profiling for miRNA-Sequencing data and gain biological insight.

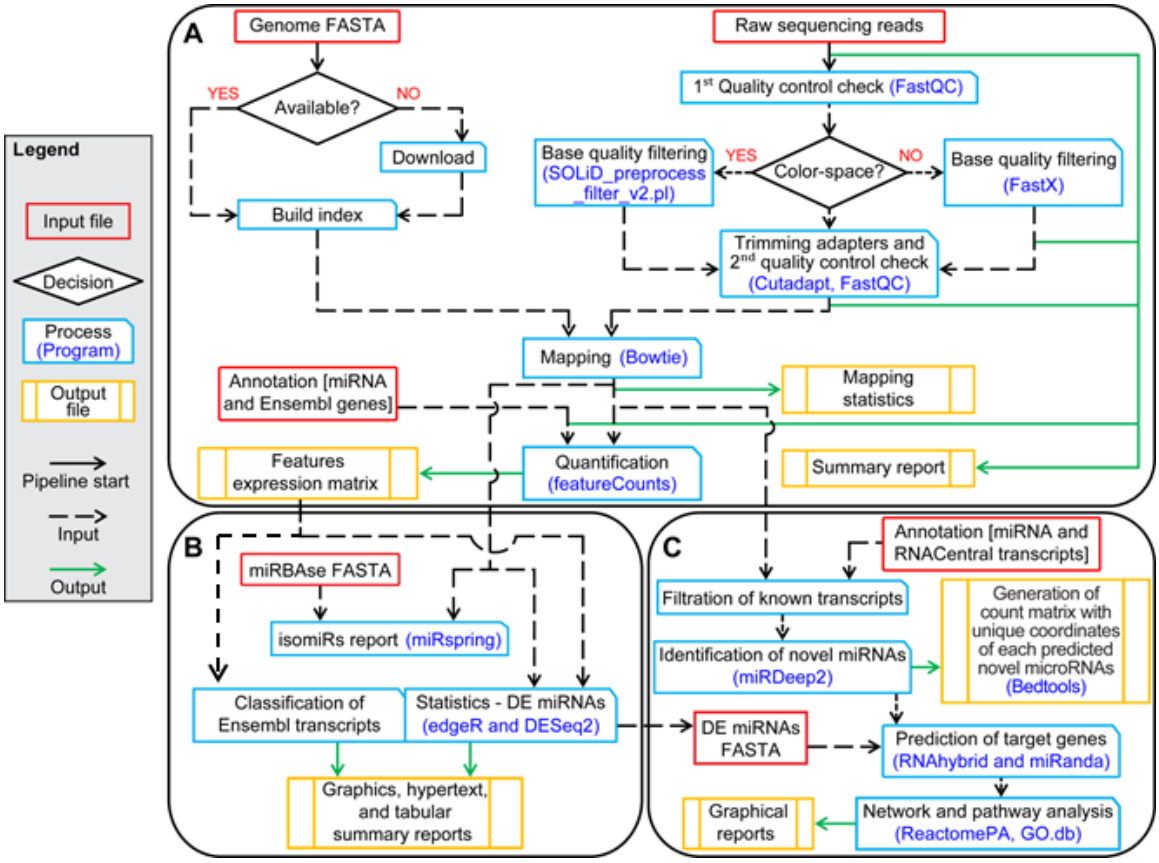



**Fig. 1: Overview of the CBS-miRSeq pipeline.** Each box presents one module of the pipeline, with the name of the tools integrated and required in a module highlighted in blue. (**A**) Module 1 performs preprocessing, quality control qualification, mapping, and quantification of raw reads. (**B**) Module 2 carries out differential expression analysis and isomiR detection. (**C**) Module 3 accomplishes identification of novel miRNA candidates, target gene prediction, and functional enrichment analysis of predicted targets.

## 2. Quick Start Virtual Machine

Along with full local installation of CBS-miRSeq, we also provide a Virtual Machine (VM) image to allow users to quickly test the workflow or run a small number of samples without need of an environment for parallelization. The VM image comes with pre-installed software and packages required by the CBS-miRSeq. The VM version can be run on a Windows, Mac, or Linux machine with at least 4GB of RAM and 200GB of free disk space.

To use CBS-miRSeq from an image, the user needs to follow these steps:

1. Download and install the free tool VirtualBox, which will allow user to run the VM image: https://www.virtualbox.org/
2. Download the CBS-miRSeq VM image (Ubuntu_16.10_CBS-miRSeq.1.0.ova; size 5.44 GB):
   https://drive.google.com/file/d/0ByG63sGTZ4JTSEVOSlhOVlE1UGs/view?usp=sharing
3. Import the image using Virtual box. For more help to import an image file please refer to https://youtu.be/Vu64isQS56Y
4. Once the ova file is imported, the user can use it pressing the start button.
5. Please login with the account **cbsmirseq**, using as password **cbsmirseq**.
6. Download the CBS-miRSeq pipeline (size: 231 KB) and copy it in the Desktop for easy access.
   https://drive.google.com/file/d/0ByG63sGTZ4JTWl8tdkdSMThZN0E/view?usp=sharing
7. Download the sample dataset (size: 204 MB) to test the workability of the pipeline.
   https://drive.google.com/file/d/0ByG63sGTZ4JTUHkxQlJmanV3VW8/view?usp=sharing

*Note:* One may use this Virtual machine to run real miRNA-Seq datasets, but this requires updating annotation and reference genome as required by species under study. In addition, user may need to increase the available memory (RAM and disk space).

*Note:* The pipeline has been successfully tested under Ubuntu v16.10, CentOS v7.1.1503 (Core) and CentOS v5.7 (Final).

## 3. Local Installation

The CBS-miRSeq needs **No installation**, however pipeline requires prerequisite dependencies (software/tools) that need to be previously install. Before running any module of the pipeline, user need to add all executable tool with absolute path to: **~/.profile or ~/.bashrc.**

We provide some utility scripts (CBS-miRSeq-SystemPackagesInstall.v1.0.sh, Install.CBS-miRSeq.dependencies.v1.0.sh, and CBS-miRSeq.Required.Packages.R) that would help user to install dependencies (please refer to READ_ME for more detail). Optionally, user may install it manually and set executable into the $PATH.

*For example:*

```
export PATH=$PATH:/path/to/installed_directory
```

*Or usually*

```
export PATH=$PATH:~/bin
```

*User can look at the available paths by typing:*

```
echo $PATH
```

**Important note**: *Analyses are launched from within the directory of CBS-miRSeq (working directory).*

## 4. Requirements

Following below, the instructions to correctly install pre-requisite software packages/tools, along with links to download them.

## Required bioinformatics tools

| Tool/Software | Function | Source for downloading |
|---|---|---|
| **bowtie ≤ v 1.0.0** | Mapping (short read alignment) | http://sourceforge.net/projects/bowtie-bio/files/bowtie/1.0.0/bowtie-1.0.0-linux-x86_64.zip/download |
| **Note:** miRDeep2 also required bowtie aligner but you don't need to install it again while installing mirdeep2 dependencies. | | |
| **miRDeep2 ≥ v 2.0.0.5 and its dependencies** | Novel miRNA discoveries | https://www.mdc-berlin.de/43969303/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/mirdeep2_0_0_7.zip |
| **R and Bioconductor ≥ v 3** | DE statistical and functional analysis | http://www.r-project.org/<br>http://www.bioconductor.org/install/ |
| **Note:** Installing R will also install the Bioconductor itself. Required packages will install automatically via *CBS-miRSeq.Required.Packages.R* whatever and whenever module needed during analyses (but make sure you are connected to the internet). However, required packages can be install manually.<br>"RColorBrewer","BiocInstaller","edgeR","DESeq2","gplots","VennDiagram","plotrix","ReportingTools","hwriter","lattice","S4Vectors","clusterProfiler", "reactome.db","ReactomePA","GO.db","biomaRt","DOSE","networkD3","igraph","magrittr","stringi","KEGGprofile","pathview","plyr","gridExtra","grid","grDevices","XML","rJava","crayon", "HTSFilter" | | |
| | | |
| **Cutadapt ≥ 1.3** | Trimming 3' adapter | https://cutadapt.googlecode.com/files/cutadapt-1.3.tar.gz |
| **Bfast ≥ 0.6.5a** | Only for solid2fastq conversion purpose and resultant used for first Quality Control (QC) | http://sourceforge.net/projects/bfast/ |
| **FastQC ≥ v0.10.1** | QC and preprocessing of short reads | http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.10.1.zip |

| | | |
|---|---|---|
| **FastX ≥ v 0.0.13** | Quality trimming and preprocessing of short reads | http://hannonlab.cshl.edu/fastx_toolkit/fastx_toolkit_0.0.13_binaries_Linux_2.6_amd64.tar.bz2 |
| **featureCounts ≥ v 1.4.6** | Quantification of digital gene expression | http://sourceforge.net/projects/subread/files/subread-1.4.6-p1/subread-1.4.6-p1-Linux-x86_64.tar.gz/download |
| **RNAhybrid ≥ v 2.1.1** | Prediction of miRNA target gene | http://bibiserv.techfak.uni-bielefeld.de/spool/download/bibiserv_1435230880_978/RNAhybrid-2.1.1-src.tar.gz |
| **miRanda ≥ v 3.3a** | Prediction of miRNA target gene | http://cbio.mskcc.org/microrna_data/miRanda-aug2010.tar.gz |
| **samtools ≥ v 0.1.1** | Utilities for post-alignment | http://sourceforge.net/projects/samtools/files/latest/download |
| **bedtols ≥ v 2.22** | Utilities for genomic features (co-ordinates) | https://github.com/arq5x/bedtools2/releases/download/v2.22.1/bedtools-2.22.1.tar.gz |
| **SAMStat ≥ v 1.5.1** | Utilities for displaying summary statistics of mapped-unmapped reads | http://sourceforge.net/projects/samstat/files/samstat-1.5.1.tar.gz/download |

*Note: Xcode and Xquartz are required for Mac users.*

## 5. Input Reference Genome and Annotation files

The CBS-miRSeq analysis requires a local file of the following (genome and other reference) annotations.

1) **Reference genome:** The CBS-miRSeq currently retrieves the genome automatically for 4 major species (Human, Mouse, Rat, and Zebrafish). Module 1a requires a version release and chromosomal assembly as input to fetch the reference genome fasta directly from the Ensembl genome browser. Alternatively, for other species user may download it manually.

2) **Annotation files:** The CBS-miRSeq also requires annotation files prior to analysis. We provide a utility module (*CBS-miRSeq.Annotations.Retrieval.v1.0.sh*) that helps to fetch the required annotations. Optionally, user can download it manually.

***a*) miRBase annotation:** To quantify digital gene expression of known miRNAs and conduct differential analysis, the pipeline requires miRBase references:

1. miRNA GFF3 file;
2. mature miRNA sequences in FASTA and a closely related mature miRNA fasta;
3. precursor miRNA sequences in FASTA;

***b*) Ensembl annotation:** To quantify other Ensembl biotypes, the pipeline requires a genomic features (.gtf) file.

***c*) RNACentral features:** In order to predict true novel miRNAs, the pipeline needs a ncRNAs features (.bed) file from the database RNAcentral.

*Note:* Please make sure that you have downloaded a right release of the miRBase, Ensembl annotation including genome, and RNAcentral features. Please visit miRBase site (http://www.mirbase.org/ftp.shtml) for more information about release and relation with other databases.

**For example: Relation of version release of the miRBase, Ensembl and RNAcentral databases.**

mmu == miRbase v21 == GenBank Assembly:GCA_000001635.2(mmu10) == Ensembl release v69(38.69) == RNAcentral 1.0/2.0

has == miRbase v21 == GenBank Assembly:GCA_000001405.15(hg38) == Ensembl release v78(38) == RNAcentral 1.0/2.0

dre == miRbase v21 == GenBank Assembly:GCA_000002035.2(Zv9) == Ensembl release v79(9) == RNAcentral 1.0/2.0

rno == miRbase v21 == GenBank Assembly:GCA_000001895.3(Rnor_5.0) == Ensembl release v79(5.0) == RNAcentral 1.0/2.0

*WARNING: The CBS-miRSeq pipeline TAKES NO GUARANTEE FOR THE RIGHT RELATION BETWEEN ONE DATABASE TO ANOTHER.*

3) **3' UTR:** To predict target gene for DE miRNAs or any miRNA of interest, user has to download their target 3' UTR. Details are illustrated in the following **Fig. 2**.

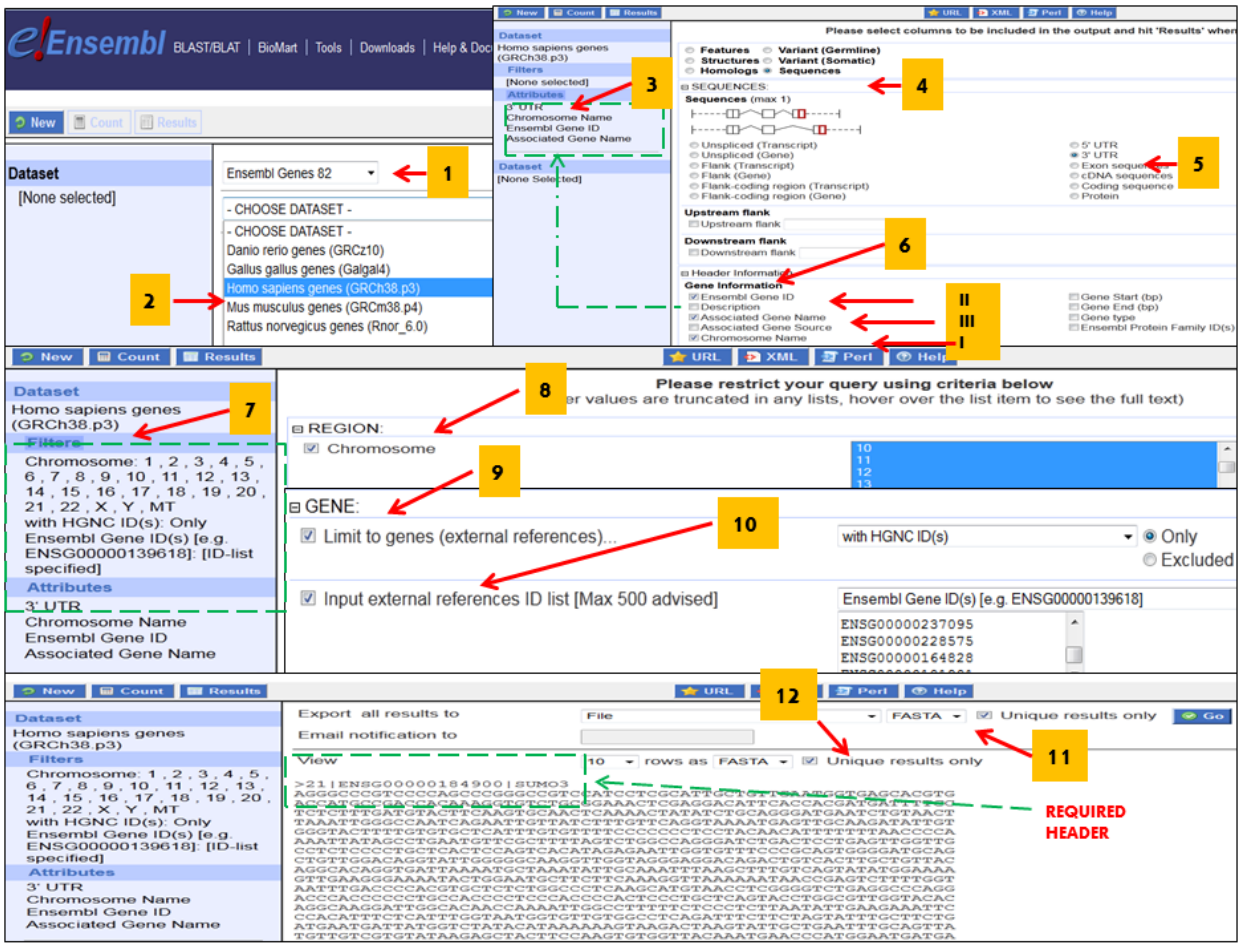



**Fig. 2: The snapshot represents step to retrieve 3'UTR Fasta, required by the CBS-miRSeq module3.**

**Sources for the retrieval of above required files**

| References | Source |
|---|---|
| Species reference genome | http://www.ensembl.org/info/data/ftp/index.html |
| miRBase GFF3 file | http://www.mirbase.org/ftp.shtml (e.g. hsa.gff3) |
| miRBase precursor miRNA fasta sequences | ftp://mirbase.org/pub/mirbase/CURRENT/hairpin.fa.gz |
| miRBase mature miRNA fasta sequences | ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz |
| Ensembl transcripts (GTF file) | http://www.ensembl.org/info/data/ftp/index.html |
| RNACentral features (BED file) | ftp://ftp.ebi.ac.uk/pub/databases/RNAcentral/releases/3.0/genome_coordinates/ |
| 3' UTR | http://www.ensembl.org/biomart/martview |

**Manually downloaded files can be prepared as:**

*a*) *Reference genome*: The downloaded genome must be in split fasta files and zipped.

1. Change the directory

```
cd /…/..path_where_ref_genome_has_been_downloaded
```

2. Unzip and concatenate into single genome fasta using:

```
gunzip *.gz
cat *.fa> name_of_genome_file.fa
```

*b*) *miRBase files:* Fasta files (precursor and mature fasta sequences) contain all species; here we advised to run following commands in order to prepare input for the pipeline.

1. Change the directory

```
cd /…/..path_where_miRBase_has_been_downloaded
```

2. Unzip

```
gunzip *.gz
```

3. Change the directory

cd CBS-miRSeq.v1.0/Utilities

```
perl extract_miRNAs.pl /../hairpin.fa sps > /../sps.mirbase_precursor.fa
perl extract_miRNAs.pl /../mature.fa sps > /../sps.mirbase_mature.fa
perl extract_miRNAs.pl /../mature.fa rel > /../rel.mirbase_mature.fa
```

***Note1:*** *Where sps is a three-letter code of the species under study (human=has; Mouse=mmu; Zebrafish=dre and Rat=nro).*

***Note2:*** *script extract_miRNAs.pl is part of miRDeep2.*

*c*) *Ensembl annotation:*

1. Change the directory

```
cd /…/..path_where_Ensemble_gtf_has_been_downloaded
```

2. Unzip

```
gunzip *.gz
```

*d*) *RNAcentral annotation:*

1. Change directory

```
cd /…/..path_where_RNAcentral_bed_has_been_downloaded
```

2. Unzip

```
gunzip *.gz
```

*e*) *3' UTR preparation:* the header should be like:

```
>2|ENSDARG00000022303|hig1
AGAGTTTTGTTGGTTTCCTGAGATATCCTTCGTACCACCTCATATTAAGGTGTCCTTAAAGTGTATTTATAGC
GAACTTTTTGTTTACTTCTTATTTCTTTACACCGTTTCCTATTTAAACCTAATAAGCCTCGGATCAGGAGGTT
TGTGTGTTCATGGCACGTTTGAGGAATGACAAAATAGTTCCTATTTTGAATCTGCCAACTATTATTTATTTCT
>2|ENSDARG00000022303
AGAGTTTTGTTGGTTTCCTGAGATATCCTTCGTACCACCTCATATTAAGGTGTCCTTAAAGTGTATTTATAGC
GAACTTTTTGTTTACTTCTTATTTCTTTACACCGTTTCCTATTTAAACCTAATAAGCCTCGGATCAGGAGGTT
TGTGTGTTCATGGCACGTTTGAGGAATGACAAAATAGTTCCTATTTTGAATCTGCCAACTATTATTTATTTCT
```

***e.g.*** **In general**

```
>chr|ENSEMBLEGene ID|Associated Gene Name
```

OR

```
> Chromosome Name| ENSEMBLEGene ID
```

## Customized/Genome UTR can be fetch from Ensembl biomart as shown in **Fig. 2**:

Go to http://www.ensembl.org/biomart/martview

1) Choose "Ensembl Genes"

2) Choose "Species"

3) Click on "Attributes"

4) Select "Sequences"

5) Click "3'UTR"

6) In Header Information, first Uncheck "Ensembl gene ID and Ensembl Transcript ID", then check the boxes sequentially (i) Chromosome Name, (ii) Ensembl gene ID (iii) Associated gene name.

7) Click on "Filters", Region and then select chromosomes

8) Under "Gene" tab (## Optional step)

9) Paste your target gene ids under "Input external references ID list".

i.e. Ensemble gene ids or Associate Gene symbol(s). (## Optional step)

10) Click on "Results"

11) Make sure it is set as FASTA

12) Check "Unique results only" and press "Go" and export as fasta text.

##When downloading is done, use the utility script *CBS-miRSeq.Prepare.UTR.v1.0.sh* to prepare the format required by the pipeline OR hit the following commands into your terminal:

**1. Change the directory**

cd /…/..path_where_UTR_fasta_has_been_downloaded

gunzip mart_export.txt.gz

**2. Paste as it is**

```
cat mart_export.txt | grep -v "Sequence unavailable" | awk 'BEGIN {RS = ">" ;
FS = "\n" ; ORS = ""} {if ($2) print ">"$0}' | sed /^$/d | perl -ane 'print
"$F[0]\n";' | sed '/^Sequence/d' > sps.UTR_fa_clean.fa
```

**3.** `perl -ane 'print "$F[0]\n";' sps.UTR_fa_clean.fa > sps.UTR_fa_clean.fa`

**4.** If you have Ensembl + Gene symbol together

```
## >1|ENSDARG00987654321|Genesymbol
sed -i -e 's/|ENSDARG[[:digit:]]\{11\}$/&&/g' sps.UTR_fa_clean.fa
```

***Note:*** *Where sps is a three-letter code of the speciesunder study (human=has; Mouse=mmu; Zebrafish=dre and Rat=nro).*

## 6. Guidelines to run the CBS-miRSeq

CBS-miRSeq allows running the pipeline in several ways. Customized features of each module may be launched one by one individually, which enables user to control the results in each analysis step. Depending on the necessity, user can deploy each sub-module. To focus on information in detail, command line arguments will let user know about inputs and parameters regarding execution of the modules.

### 6.1 Running Module 1

==========

***Aims: Downloading the reference genome, Building the index for mapping, Reads QC, Trimming of the 3' adapter, Mapping and Quantification of the features (miRNAs and Ensembl biotypes).***

==========

This Module is a wrapper for Module 1a and Module 1b. Navigate and open the configuration file (Input_Info/Module1_Input.txt) to fill the required input to run the pipeline.

```
$: cd..
$: bash ./CBS-miRSeq.module1.sh Input_Info/Module1_Input.txt
```

*Module 1a analysis description:* Built the bowtie index of the reference genome based on format of the input reads (color-space/base-space).

*Note: No restrictions regarding the reference organism, user may provide any reference genome. However, this module allow researcher to fetch a few model organism's genome such as Human, Mouse, Rat and Zebrafish (for more detail, please refer to READ_ME file into the CBS-miRSeq package).*

*Module 1b analysis description:* QC of the Reads, Trim the 3' adapter, Mapping, Quantification of the features (miRNAs and Ensembl biotypes) and summarization of datasets.

Configuration file of Module1_Input.txt

| Inputs | Files/Directories/Type | Description |
|---|---|---|
| **DIR_REF_GENOME** | /path/to…/dir | # Directory or path where reference genome index has to build and/or genome fasta already existed. |
| **SPS** | A String | # Name of analysis species (3 letter code; i.e. hsa) |
| **INDEX_COLOR** | A String | # Yes (In case of color-space reads from SOLiD) or No (in case of fastq reads from Illumina) |
| **ENS_RELEASE** | A numeric value | #Ensembl genome release (two digit integer i.e. 82) |
| **ENS_AssemblyGRCh** | A numeric value | # Ensembl chromosomal assembly (such as 38 or 36.67 based on assembly) |
| **DIR_INPUT_READS** | /path/to…/reads | # A directory where all reads are (csfasta and qual or fastq) |
| **Note:** Please make sure all input reads (csfasta and qual or fastq) are located in the same directory | | |
| **NumberOfSamples** | A numeric value | # Number of total sample |
| **MIR_ANNOTATION** | /path/to…/sps.gff3 | # miRBase gff3 file of species |
| **ENS_ ANNOTATION** | /path/to…/sps.gtf | # Ensembl gtf file |
| **MY_ADAPTER** | A sequence (“330201030313112312” / "ATCTCGTATGCCGTCTTCTGCTTG") | # An adapter sequence (can be color-space or base-space sequence) |
| **OUTPUT_DIR** | /path/to…./output | # A directory or path where results are produced |

## 6.2 Running Module 2

==========

***Aims: Differential expression analysis (DEA), Identification of Ensembl biotype, and Detection of iso-miRs.***

*Note: Please make sure that groups (along with their replicates) of the expression matrix are set in the right order of interest for proper comparison, i.e. Control x Treatment.*

*WARNING!! Output directory cannot be different than Module 1.*

==========

This script is a wrapper for Module 2a and Module 2b. Navigate and open the configuration file (Input_Info/Module2_Input.txt) to fill the required input to run the pipeline.

```
$: bash ./CBS-miRSeq.module2.sh Input_Info/Module2_Input.txt
```

*Module 2a analysis description:* Conducts differential expression between two experimental conditions and detection of Ensembl biotypes in each sample.

*Module2a analysis description:* Prediction of isomiR in each sample.

Configuration file of Module2_Input.txt

| Inputs | Files/Directories/Type | Description/Type |
|---|---|---|
| **OUTPUT_DIR** | /path/to…./output | # A directory or path where results are produced<br>Note: Directory should be the same as for Module 1 |
| **MIR_ANNOTATION** | /path/to…/sps.gff3 | # miRBase gff3 file of species |
| **ALL_SPS_MATURE_FASTA** | /path/to…./All.sps.mature_v21.fa | # FASTA of miRBase of all sps |
| **MIR_MATURE_FASTA** | /path/to…/sps.mature.fa | # miRBase sps mature fasta file |
| **SPS** | A String | # Name of species under study (3 letter code; i.e. hsa) |
| **conditionA** | A String | # Name of the Condition/group A (for example: Control) |
| **startCol_grpA** | A numeric value | # First Group start column of condition A from miRNA expression matrix obtained from Module 1.<br>**Tip:** always control condition start from "2" |
| **conditionB** | A String | # Name of the Condition/group B (for example: Treatment) |
| **startCol_grpB** | A numeric value | # Number of Start column of condition B from miRNA expression matrix obtained from Module 1.<br>**Tip:** The value depends on your second condition and its sample size; e.g. "5" for 3×3 samples.<br>**Note:** DEA uses condition 1 as reference.<br>Thus, DEA will be represented as log2 (group2 [B]/group1 [A]). |
| **LowCountsfeaturesFilterByCPM** | "yes" or " no" | # Pre-filter: Filter low/unexpressed tags before analysis (yes or no; [default Nothing]) |
| **CutoffLowCountsfeaturesFilter** | A numeric value like "1" | # Filtration cutoff either used for FilterByCPMCounts or for raw counts filtering if FilterByCPMCounts is set to no |
| **RemoveInconsistentFeatures** | "yes" or " no" | # Remove outliers from the matrix before performing DEA [yes or no; default Nothing/Optional] |
| **ThresholdToRemoveInconsistentFeatures** | A numeric value "1.5" | # Tags less than this threshold will be kept (applied for both edgeR and |

| | | |
|---|---|---|
| | | DESeq2) |
| **performIndependentFilter** | "yes" or " no" [default no] | # Independent filtering if yes, applied to both analysis methods |
| **plotType** | "pdf" or " eps" | # Plotting type should be pdf or postscript (ps) file [default Nothing] |
| **pval_Cutoff** | A numeric value (0.05) | # P-value cutoff for plotting [default Nothing] |
| **padj_Cutoff** | A numeric value (0.1) | # FDR cutoff for plotting [default Nothing] |
| **log2FC_Cutoff** | A numeric value (1) | # Log2Fold changes cutoff for plotting [default Nothing] |

### 6.3 Running Module 3

==========

***Aims: Discovery of Novel miRNA candidates, Prediction of target genes of DE miRNAs, Gene enrichment, Pathway analysis of known and novel miRNAs.***

*WARNING!! Output directory cannot be different than Module 1 and 2.*

==========

This script is a wrapper for Module 3a and Module 3b. Navigate and open the configuration file (Input_Info/Module3_Input.txt) to fill the required input to run the pipeline.

```
$: bash ./CBS-miRSeq.module3.sh Input_Info/Module3_Input.txt
```

*Module 3a analysis description:* Prediction of novel miRNAs in each sample.

*Module 3b analysis description:* Prediction of target genes, Gene enrichment, Pathway analysis of known and novel miRNAs.

Configuration file of Module3_Input.txt

| Inputs | Files/Directories/Type | Description/Type |
|---|---|---|
| **OUTPUT_DIR** | /path/to…./output | # A directory or path where results are written<br>Note: Directory should be the same as for Module 1 |
| **MIR_ANNOTATION** | /path/to…/sps.gff3 | # miRBase gff3 file of species |
| **RNAC_ ANNOTATION** | /path/to…/sps.bed | # RNAcentral bed file |
| **MIR_MATURE_FASTA** | /path/to…/sps.mature.fa | # miRBase sps mature fasta file |
| **REL_MATURE_FASTA** | /path/to…/rel.mature.fa | # miRBase related sps mature fasta file |
| **PRE_FASTA** | /path/to…/sps.hairpin.fa | # miRBase sps precursor fasta |

| | | |
|---|---|---|
| | | file |
| **SPS** | A String | # Name of species under study (3 letter code; i.e. hsa) |
| **SPECIES_NAME** | A String | # Name of species under study (full name; i.e. Human, Zebrafish, Mouse, Rat) |
| **INPUT_QUERY_FASTA** | /path/to…/DE/DE.mature.fa | # A File where DE_mature.fa is located, probably in "..../DE/DE_mature.fa" or any fasta that user wish to predict targets of |
| **TARGET_UTR_FASTA** | /path/to…/sps.3UTR.fa | # A 3' UTR fasta; can be fetch from Ensembl Biomart |
| **HYBRIDIZATION_THRESHOLD** | A negative value | # A negative value is required for the prediction of miRNA-mRNA hybrid (i.e. -10) |
| **entrezID** | "NO" or "YES" | # Based on header of UTR fasta: (YES = in case of entrezID; NO = in case of Gene symbol) |
| **ID_Type** | ID should be one of: SYMBOL/REFSEQ/ENSEMBL (## case sensitive) | ## ID of UTR fasta from the header |
| **targetHub** | "small" (recommended) or "big" or "FullNetwork" ## case sensitive | ## gene:miRNA network hub to plot |
| **pathways** | "reactome" or "kegg" | # Pathways analysis of target genes of miRNAs |
| **internet** | "yes" or "no" (no==internet in case not accessible) ## case sensitive | # To download recent path from KEGG pathways database |
| **plotType** | "pdf" or "eps" ## case sensitive | # Plot the results in pdf or ps format |
| **Ontology** | "DO" (DO = Disease Ontology; only for human samples) or "ANY" (i.e. BP, MF, CC) | # Gene Ontology analysis |
| **OrgDb** | "org.Hs.eg.db" for Human<br>"org.Mm.eg.db" for Mouse<br>"org.Dr.eg.db" for Zebrafish<br>"org.Rn.eg.db" for Rat | # Annotation package (Genome wide annotation) for organism under study |
| **pvalueCutoff** | A numeric value (0.05) | # The minimum P-value for enriched GO and Pathways |
| **qvalueCutoff** | A numeric value (0.1) | # The minimum adjusted P-value for enriched GO and Pathways |

# 7. Results

The CBS-miRSeq pipeline generates tables and plots of results at each step for visual analysis and further investigations.

## 7.1 **Output Structure**

The output directory of CBS-miRSeq results is structured in the following manner:

### Output Structure of Module 1

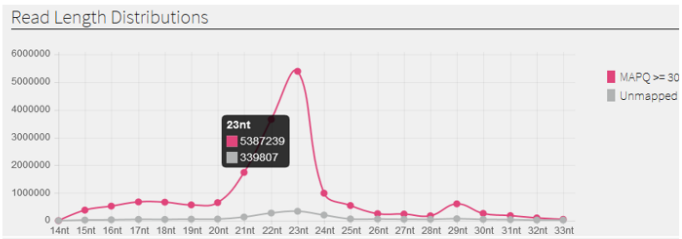
```
Output_dir/Index/
sps.genome_v38.84.fa
sps.genome_v38.84.fa.fai
*.ebwt

before_qc/
      sample_fastqc.zip

after_qc/
      sample.cln_fastqc.zip #(in case base-space reads)
      sample _T_F3.csfasta_fastqc.zip #(in case color-space reads)

clean_reads/
      sample.cln.fastq.zip #(in case base-space reads)
      sample_T_F3.csfasta #(in case color-space reads)
      sample_T_F3_QV.csfasta " "

discarded_reads/
      -     #(NULL; in case base-space reads)
      sample_U_F3.csfasta #(in case color-space reads)
      sample_U_F3_QV.csfasta " "

trim_reads/
      sample.cln.fastq.zip #(in case base-space reads)
      sample_T_F3.csfasta.fastq #(in case color-space reads)

reads_mapped/
      sample.sam
      sample.sorted.bam
      sample.sorted.bam.bai

visualization_summary/
      summary.stats.xls
      sample.samstat.html
log_files/
      all.map.log
      all.trim.log
      fastq_reads.log #(in case base-space reads)
```

# Visual Output from Module 1

**A**

| Sample | Total counts | Filtered reads | Reads after QC | Trimmed reads | Too short reads(<15bp) | Too long reads (>35bp) | Reads sent to mapping | Mapped reads |
|---|---|---|---|---|---|---|---|---|
| SRR433866_WT1.fastq | 38346007 | 14532776(37.90%) | 23813231(62.10%) | 23423768(98.4%) | 4098726(17.2%) | 0(0.0%) | 19714505 | 17934450(90.97%) |
| SRR433867_WT2.fastq | 35975387 | 9876413(27.45%) | 26098974(72.55%) | 24966875(95.7%) | 5101590(19.5%) | 0(0.0%) | 20997384 | 19253708(91.70%) |
| SRR433868_WT3.fastq | 33921150 | 7735525(22.80%) | 26185625(77.20%) | 25773309(98.4%) | 6758926(25.8%) | 0(0.0%) | 19426699 | 18251535(93.95%) |
| SRR433869_ATM1.fastq | 35328338 | 7528182(21.31%) | 27800156(78.69%) | 26622900(95.8%) | 2641170(9.5%) | 0(0.0%) | 25158986 | 22019121(87.52%) |
| SRR433870_ATM2.fastq | 32632356 | 5084572(15.58%) | 27547784(84.42%) | 23972147(87.0%) | 2969985(10.8%) | 0(0.0%) | 24577799 | 21551968(87.69%) |
| SRR433871_ATM3.fastq | 31967552 | 4916669(15.38%) | 27050883(84.62%) | 26701029(98.7%) | 6194940(22.9%) | 0(0.0%) | 20855943 | 19622921(94.09%) |

**B**




| | Number | Percentage |
|---|---|---|
| MAPQ >= 30 | 17934450.0 | 91.0 |
| MAPQ < 30 | 0.0 | 0.0 |
| MAPQ < 20 | 0.0 | 0.0 |
| MAPQ < 10 | 0.0 | 0.0 |
| MAPQ < 3 | 0.0 | 0.0 |
| Unmapped | 1780055.0 | 9.0 |
| Total | 19714504.0 | 100.0 |

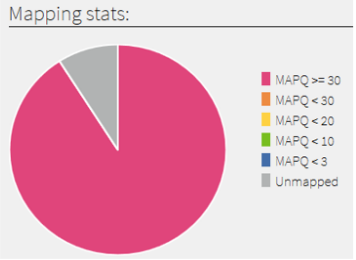



**Fig. 3: A summary report of each miRNA-Seq library generated by Module 1 of CBS-miRSeq. (A)** Summary statistics of miRNA sequencing data for each library. **(B)** Illustration of mapping statistics and read length distribution in a sample: Representation of the most mapped reads within typical length of a mature miRNA (21-23 nt).

## Output Structure of Module 2

```
isomiRs/
        sample.*.html
        sample.*.txt

quantification/
        sps.ensembl.report.txt
        sps.miRBase.report.txt
DE/
        sps_Deseq2_results.csv
        sps_DESeq2_analysis_plots.pdf (OR *.eps; a postscript plots)
        sps_edgeR_results.csv
        sps _edgeR_analysis_plots.pdf (OR *.eps; a postscript plots)
        hsa_merged.edgeR.DESeq2.csv
        sps_miR.Intersect.merged.statPvalue0.05.csv
        sps_comparison_plots.pdf
        Ensembl_biotypes_pie_bar.pdf
        RelativeExpressionCounts.pdf (word clouds of mean relative expression)
        DE_mature.fa
        DE_miR.txt
        sps_CBS-miRSeq.pipeline.DESeq2vsEdgeR.Rdata
         HTMLRports/
                                CBS-miRseq.Analysis.with.DESeq2.html
                                CBS-miRSeq.Analysis.with.edgeR.html
```

## Visual Output from Module 2

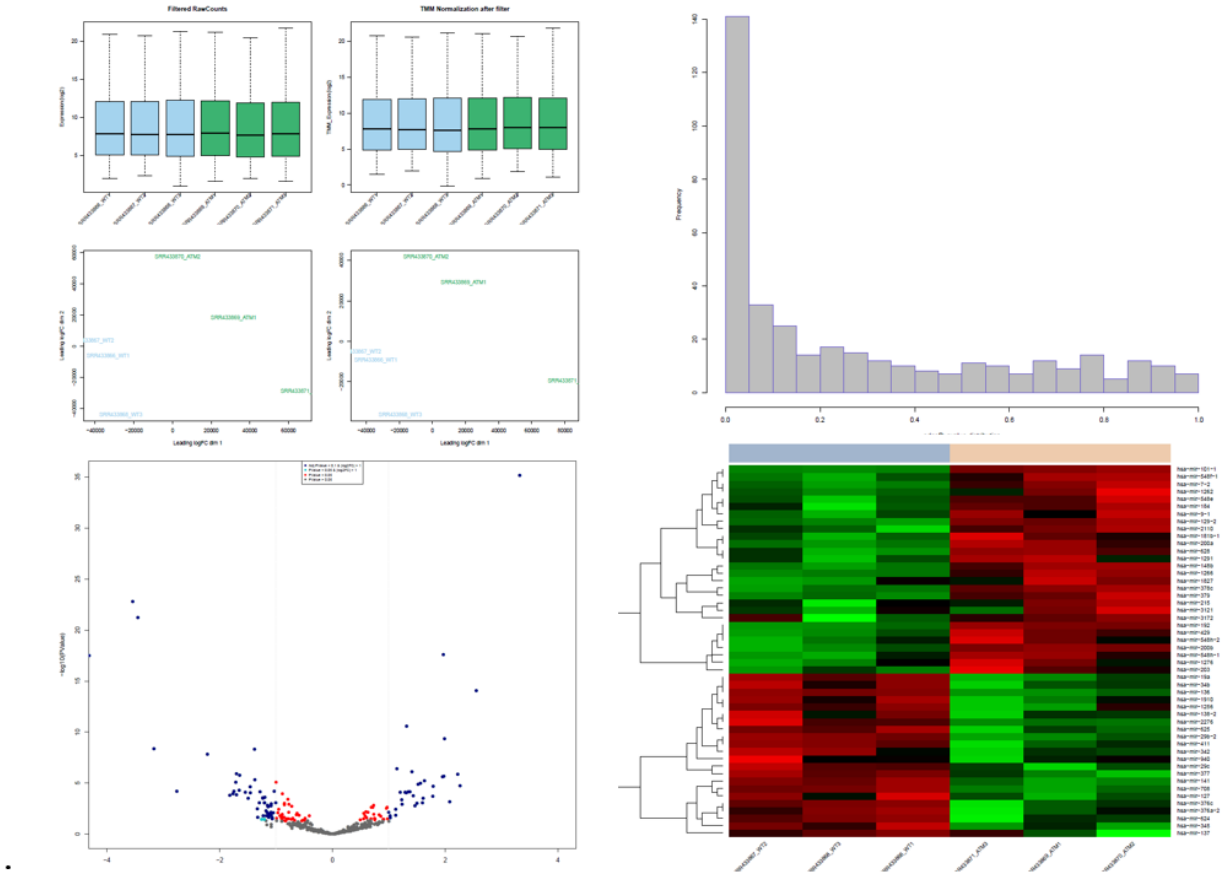



**Fig. 4: An example of output produced by Module 2 of CBS-miRSeq. (A)** Exploratory analysis (Box and PCA plots) of the samples. **(B)** P-value distribution of differential expression analysis. **(C)** Volcano plots of differentially expressed miRNAs based on statistical cutoff. **(D)** Heatmap of most differentially expressed miRNAs based on adjusted P-values.

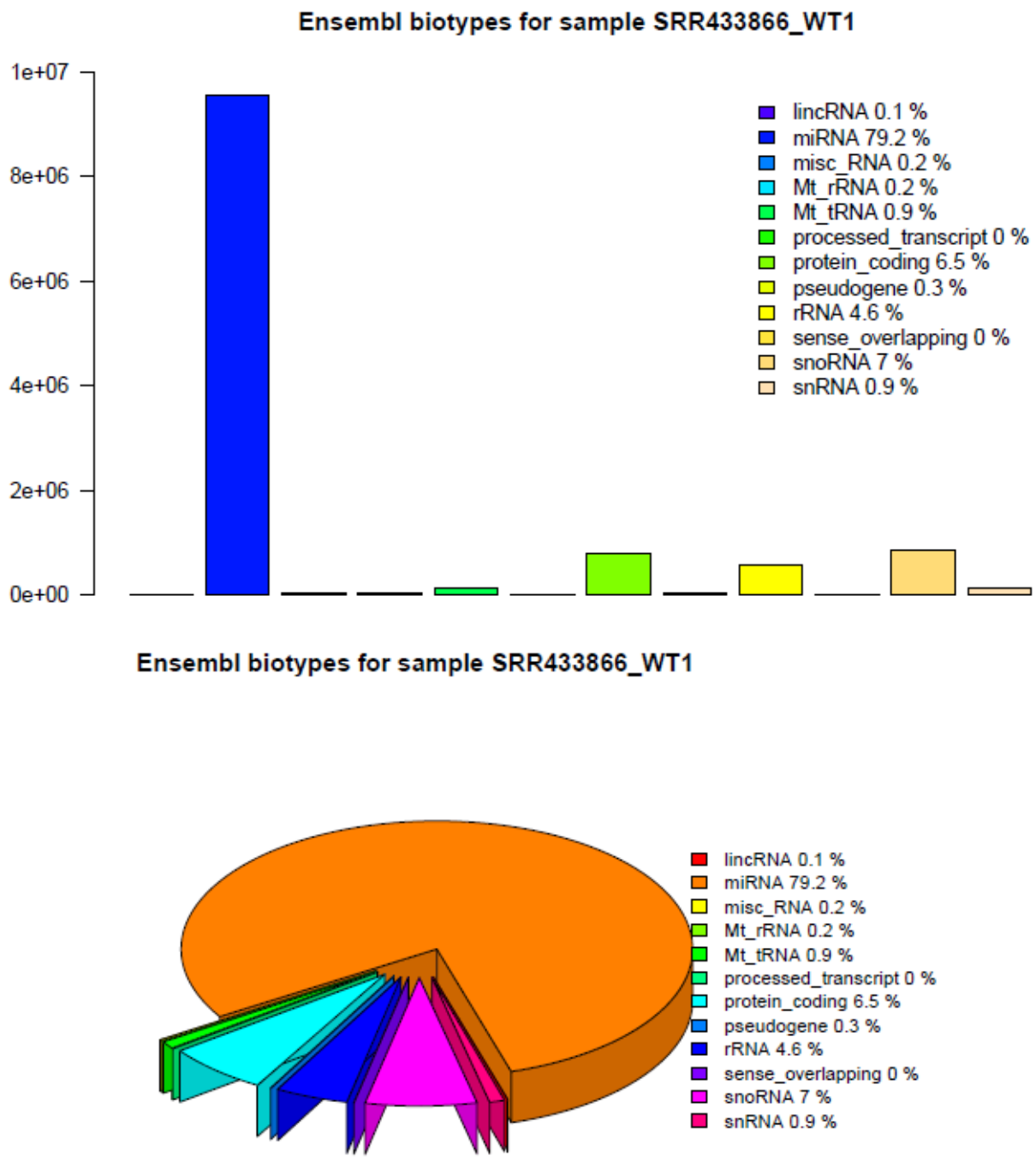
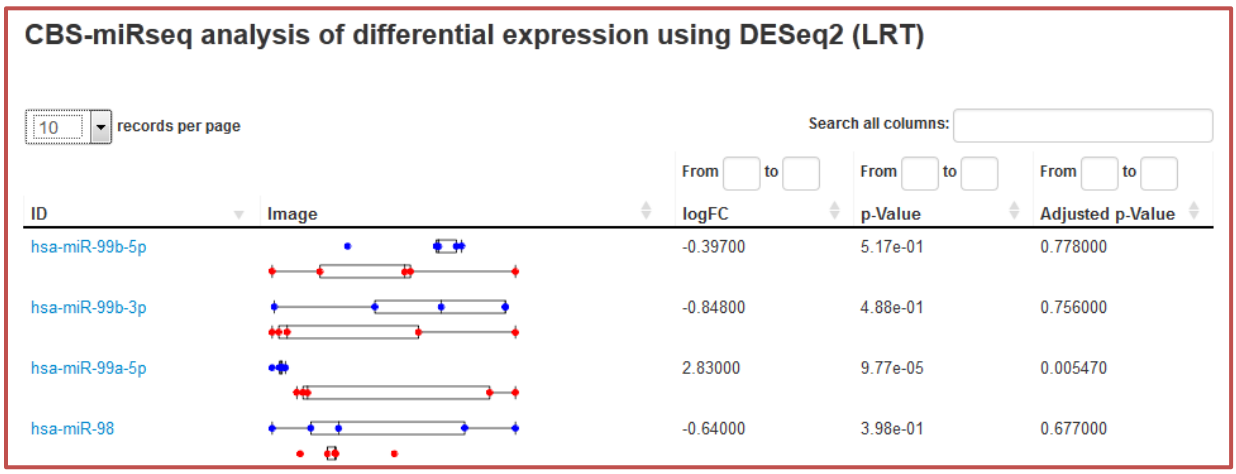
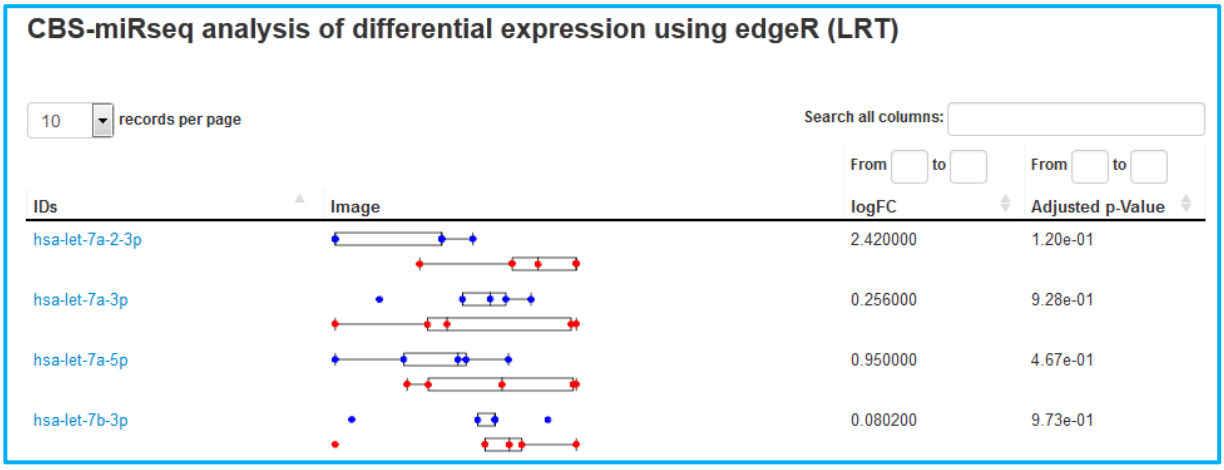



**Fig. 5: Pie and bar plots representing the distribution of detected transcripts in a sample.**

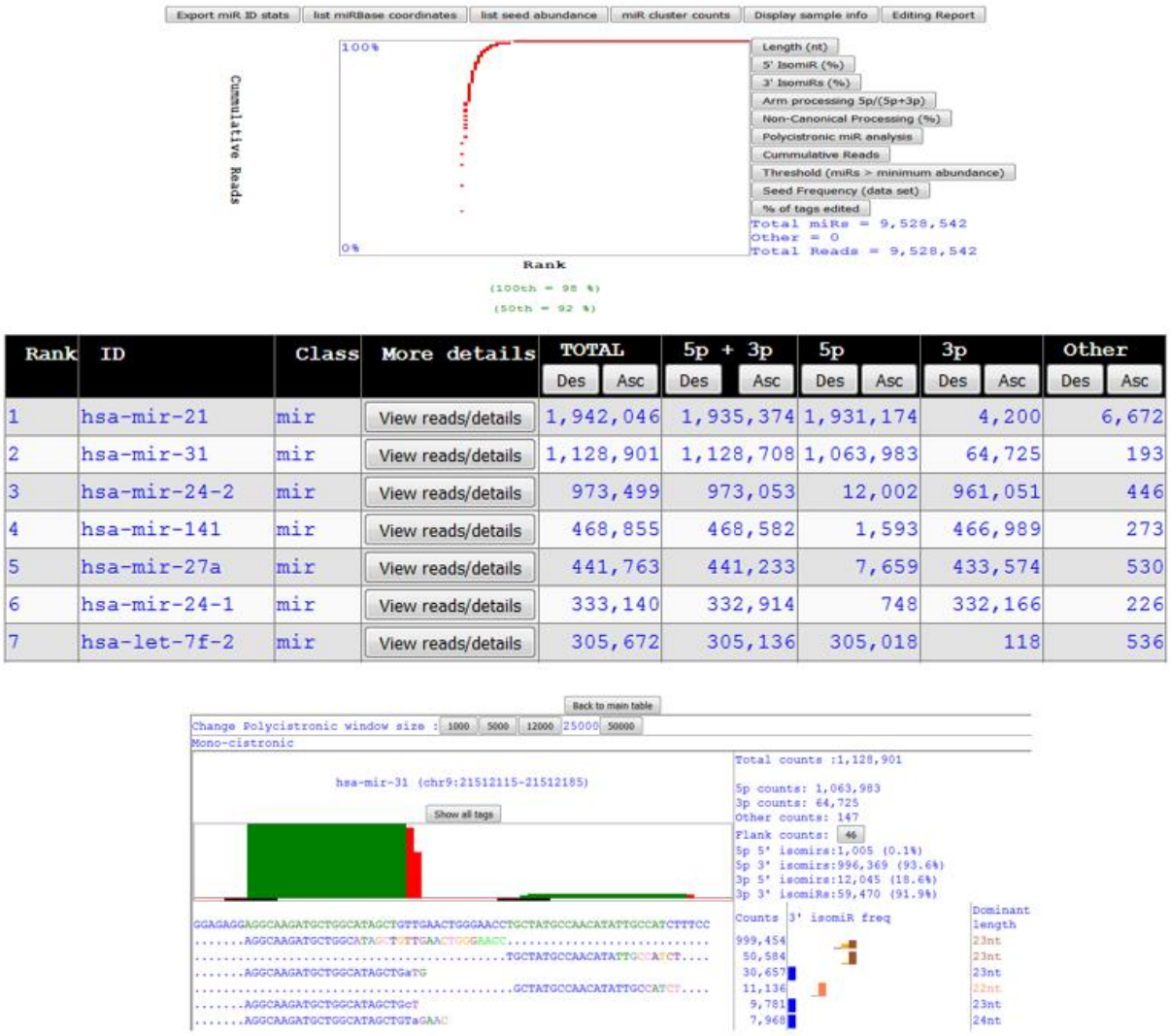



**A**




**B**

**Fig. 6: A representative HTML report generated by the CBS-miRSeq Module 2. (A)** an interactive HTML report generated by DESeq2 analysis. It contains miRNA IDs, log2FC, P-values and FDR values along with boxplot of expression distributions and miRBase hyperlink to each miRNAs. **(B)** The output of the analysis performed with edgeR contains for each miRNA a hyperlink to miRBase, boxplot of the expression in two conditions, logFC, and FDR.




| Rank | ID | Class | More details | TOTAL | | 5p + 3p | | 5p | | 3p | | Other | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Des | Asc | Des | Asc | Des | Asc | Des | Asc | Des | Asc |
| 1 | hsa-mir-21 | mir | View reads/details | 1,942,046 | | 1,935,374 | | 1,931,174 | | 4,200 | | 6,672 | |
| 2 | hsa-mir-31 | mir | View reads/details | 1,128,901 | | 1,128,708 | | 1,063,983 | | 64,725 | | 193 | |
| 3 | hsa-mir-24-2 | mir | View reads/details | 973,499 | | 973,053 | | 12,002 | | 961,051 | | 446 | |
| 4 | hsa-mir-141 | mir | View reads/details | 468,855 | | 468,582 | | 1,593 | | 466,989 | | 273 | |
| 5 | hsa-mir-27a | mir | View reads/details | 441,763 | | 441,233 | | 7,659 | | 433,574 | | 530 | |
| 6 | hsa-mir-24-1 | mir | View reads/details | 333,140 | | 332,914 | | 748 | | 332,166 | | 226 | |
| 7 | hsa-let-7f-2 | mir | View reads/details | 305,672 | | 305,136 | | 305,018 | | 118 | | 536 | |




**Fig. 7: An example of isomiRs detected in a sample predicted by Module 2 of the CBS-miRSeq pipeline.**

# Output Structure of Module 3

```
Novel_miR/
        novel.miRNA.counts.matrix.txt
        sample.*.sorted.csv.bed
        Sample/
            results.*.html
            results.*.csv
            (other files generated by miRDeep2)
            pdfs*/
            (Secondary structure of novel miRNAs )
miR_Target/

          sps_PathwaysEnrichment_Plots.pdf
          sps_GOenrich_Plots_DO001.pdf (OR *.eps; Post script files)
          sps_miR.mRNA_3d.network.html
          sps _Ontology_GOenriched.txt
          sps_Pathways_GOenriched.txt
          miranda_miR_Target_sps.tab
          miranda_miR_Target_str_sps.txt
          RNAhybrid_miR_Target_sps.tab
          RNAhybrid_miR_Target_str_sps.txt
          intersect.genes.uniq.pair.txt (shared targets of both prediction algorithms)
```

# Visual Output from Module 3

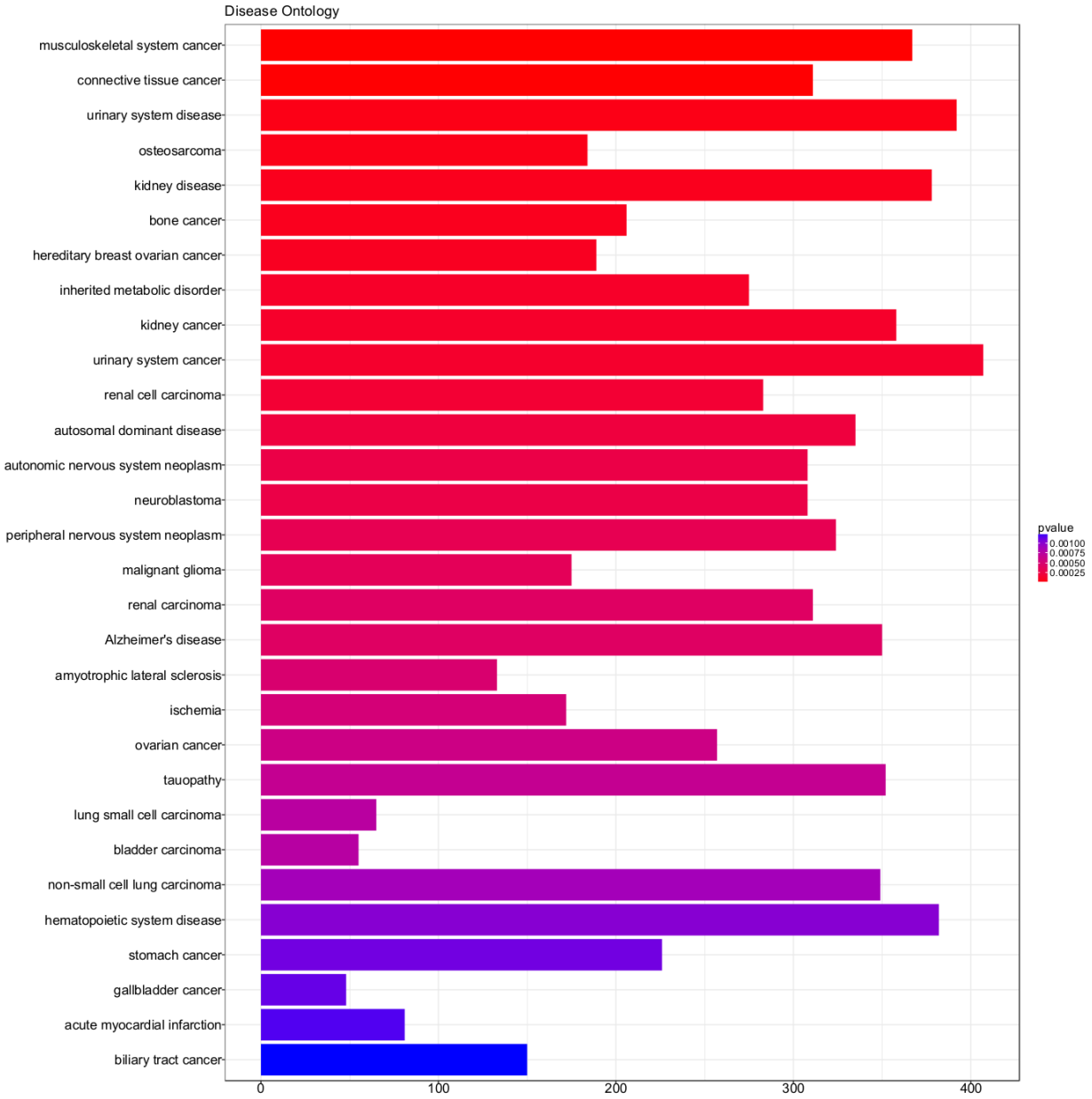



**Fig. 8: An example of Disease ontology (DO) enrichment analysis of predicted DE targets, generated by the CBS-miRSeq Module 3.**

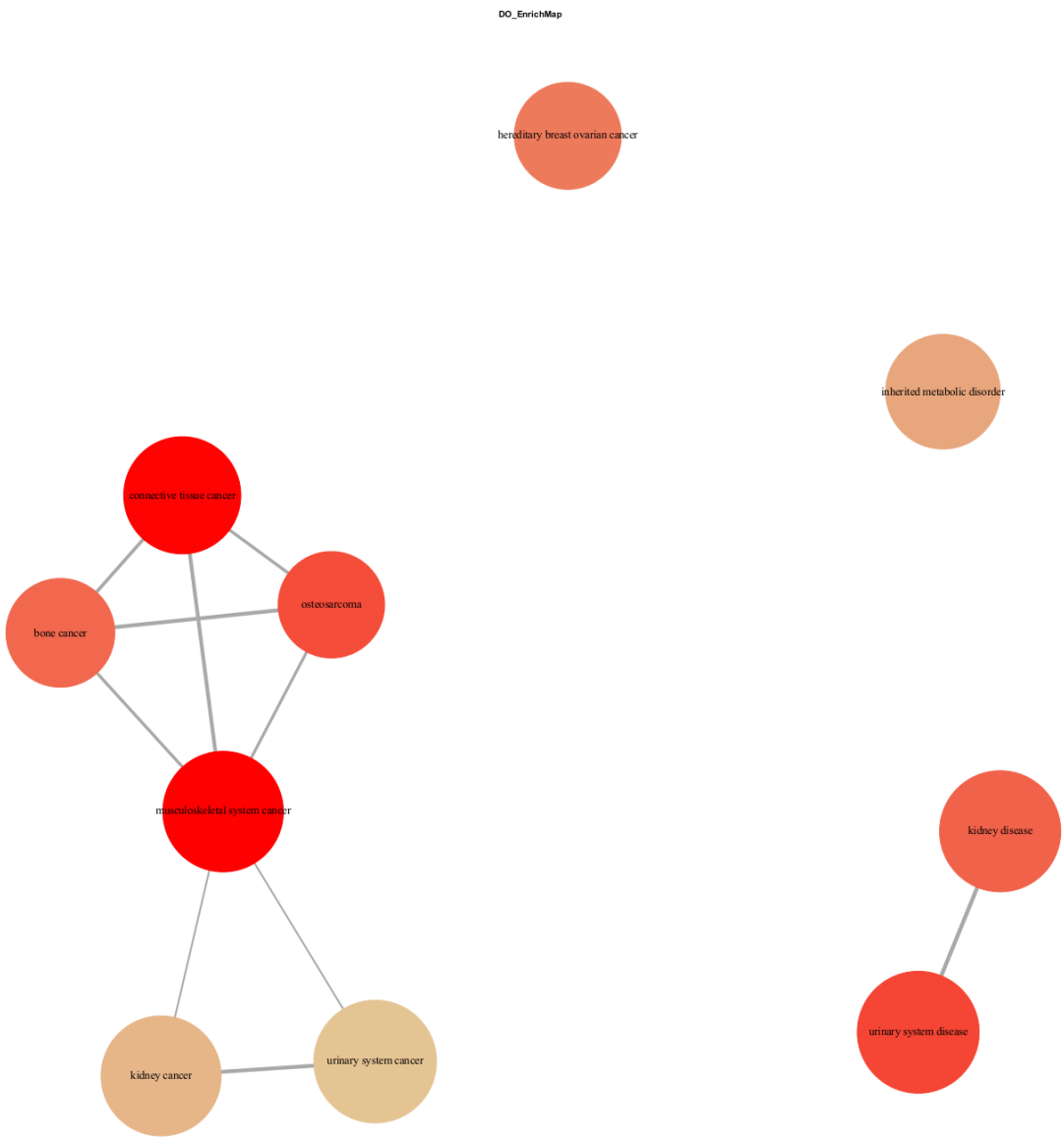



**Fig. 9: An example of enrichment map of DO terms, generated by Module 3 of the pipeline.**

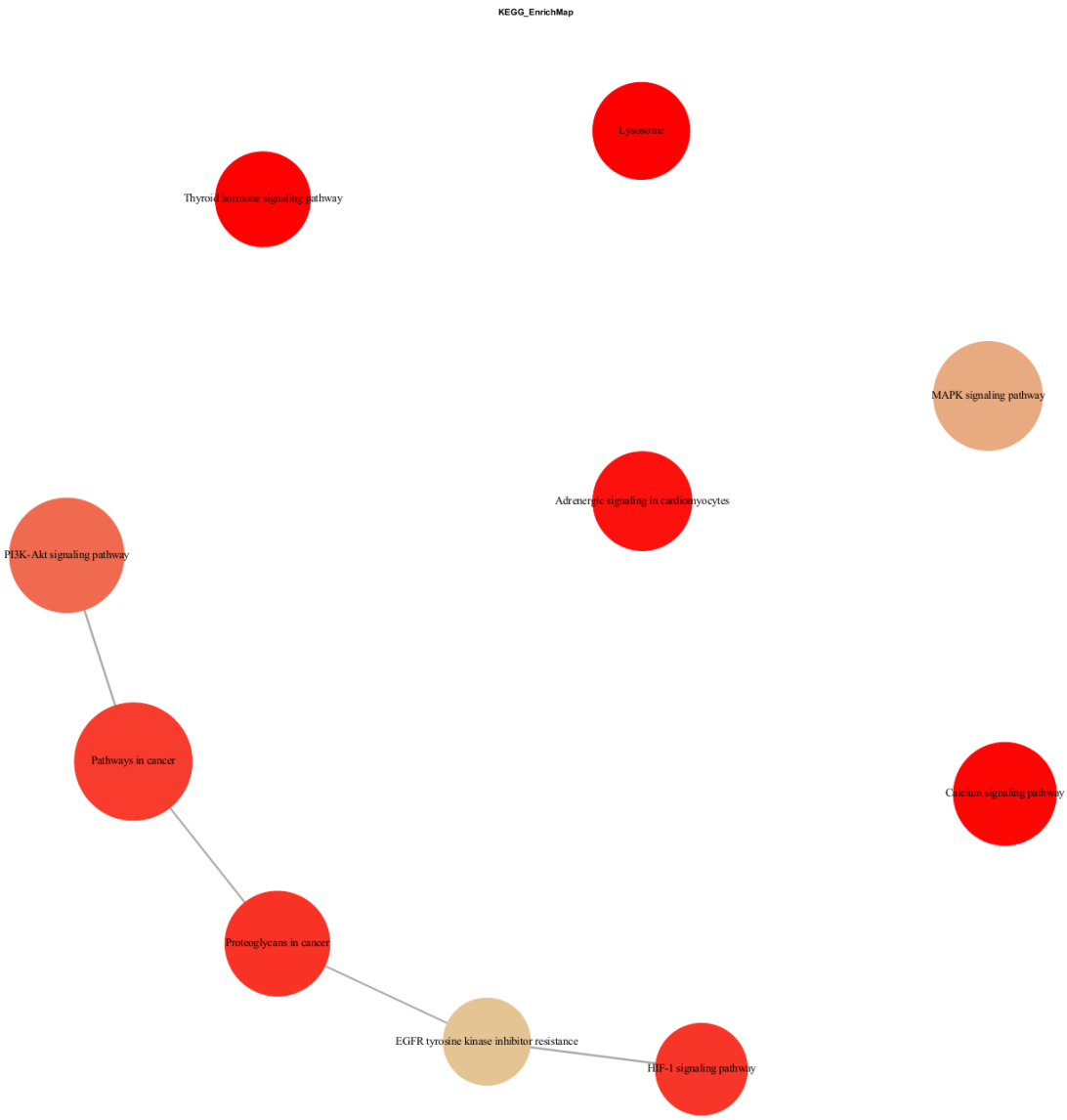



**Fig. 10: An example of KEGG pathways related to one category of enriched genes, produced by the Module 3b of the CBS-miRSeq pipeline.**

**Survey of miRDeep2 performance for score cut-offs -10 to 10**

| | novel miRNAs | | |
|---|---|---|---|
| **miRDeep2 score** | **predicted by miRDeep2** | **estimated false positives** | **estimated true positives** |
| 10 | 29 | 9 ± 3 | 20 ± 3 (70 ± 9%) |
| 9 | 32 | 9 ± 3 | 23 ± 3 (71 ± 9%) |
| 8 | 35 | 10 ± 3 | 25 ± 3 (72 ± 9%) |
| 7 | 39 | 10 ± 3 | 29 ± 3 (73 ± 9%) |
| 6 | 47 | 11 ± 4 | 36 ± 4 (76 ± 8%) |
| 5 | 53 | 12 ± 4 | 41 ± 4 (77 ± 7%) |
| 4 | 74 | 18 ± 4 | 56 ± 4 (76 ± 6%) |
| 3 | 114 | 46 ± 7 | 68 ± 7 (60 ± 6%) |
| 2 | 146 | 65 ± 8 | 81 ± 8 (56 ± 5%) |
| 1 | 188 | 83 ± 9 | 105 ± 9 (56 ± 5%) |
| 0 | 353 | 193 ± 14 | 160 ± 14 (45 ± 4%) |

**A**

**novel miRNAs predicted by miRDeep2**

| provisional id | miRDeep2 score | estimated probability that the miRNA candidate is a true positive | rfam alert | total read count | mature read count | loop read count | star read count | significant randfold p-value | miRBase miRNA | example miRBase miRNA with the same seed | UCSC browser | NCBI blastn | consensus mature sequence | consensus star sequence | consensus precursor sequence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr21_171151 | 2.8e+5 | 0.70 ± 0.09 | | 560247 | 559611 | 555 | 81 | | | | blat | blast | gagaacuuugaaggcccgaagu | ggugcagaucuuggguuguggu | ggugcagaucuuggguuguagugcaaauauucaaacgagaacuuugaaggcccgaagu |
| chr21_171433 | 5.4e+4 | 0.70 ± 0.09 | | 106119 | 104670 | 0 | 1449 | | | | blat | blast | acggcccuggcggagcgc | gccccgccggggucggccc | gccccgccggggucggccccagggccuggcggagcgc |
| chr21_171626 | 5.4e+4 | 0.70 ± 0.09 | | 106119 | 104670 | 0 | 1449 | | | | blat | blast | acggcccuggcggagcgc | gccccgccggggucggccc | gccccgccggggucggccccagggccuggcggagcgc |
| chr21_171093 | 5.4e+4 | 0.70 ± 0.09 | | 106119 | 104670 | 0 | 1449 | | | | blat | blast | acggcccuggcggagcgc | gccccgccggggucggccc | gccccgccggggucggccccagggccuggcggagcgc |
| chr21_171261 | 1.2e+4 | 0.70 ± 0.09 | | 24319 | 12196 | 12018 | 105 | | | | blat | blast | cagcccgagccgccuggga | gagggacggccgggggca | cagcccgagccgccuggguaacccgcagcuaggaauaauggaauaggaccgcgguucuauuuugguuggguuuucggaacugaggccuugauuaagaggaacggccgggggca |
| chr13_58455 | 9.1e+2 | 0.70 ± 0.09 | tRNA | 1784 | 1764 | 0 | 20 | | | mmu-miR-6904-3p | blat | blast | ucugauucucggaagcuaagcaggguuggccuggguaguacuuggauggga | gauaguacuuggauggga | ucugauucucggaagcuaagcaggguuggccuggguaguacuuggauggga |
| chr21_171709 | 5.3e+2 | 0.70 ± 0.09 | | 1045 | 577 | 346 | 122 | | | mmu-miR-6936-3p | blat | blast | gggcgcgcgccgcggcuggacga | ccccgccccacccgcgc | gggcgcgcgccgcggcuggacgaggcgccgccgccccccccacgcccgggggcaccccccucgcggccccucccccgccccacccgcgc |

**B**

**Fig. 11: An example of novel miRNA candidates predicted by miRDeep2, integrated into Module 3 of the CBS-miRSeq pipeline. (A)** The table contains the number of predicted miRNAs ordered by the mirDeep2 score (red color line), *i.e.* the probability of true positive novel miRNAs. **(B)** The table highlights novel miRNAs predicted along with score, expression counts, hyperlink to UCSC and NCBI, and sequences of the mature miRNAs (3p, 5p) and precursors.

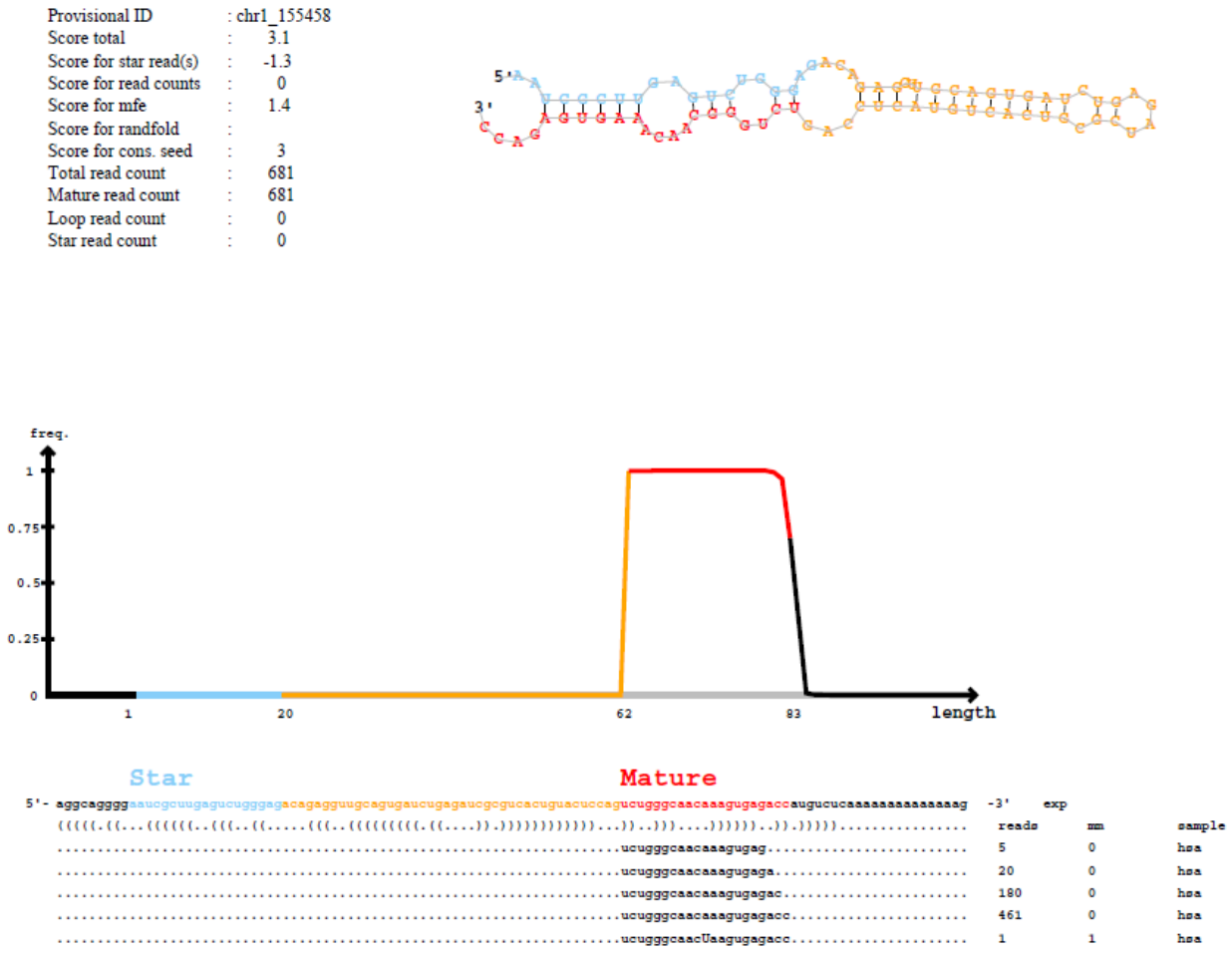



**Fig. 12: An example of the secondary structure and sequence read depth of a representative novel miRNA predicted by miRDeep2.**

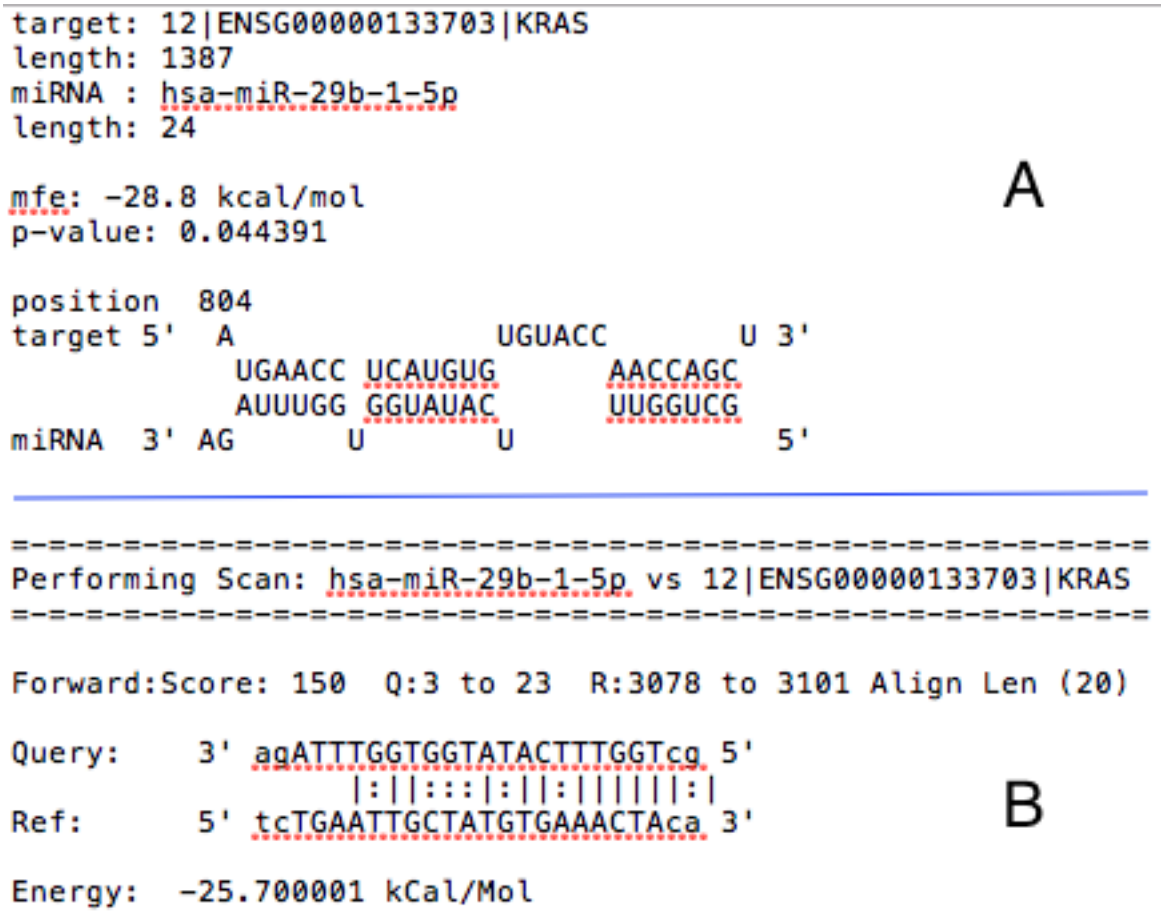



**Fig. 13: An example of a predicted target in a test dataset. (A)** A miRNA:mRNA hybridization structure predicted by RNAhybrid along with the lowest thermodynamics energy and P-value. **(B)** A miRNA:mRNA hybrid structure predicted by miRanda along with its lowest energy threshold. KRAS is a well-known target of tumor suppressor miRNAs such as miR-96 and miR-29 family. The results at intersection of both algorism are further considered for gene enrichment analyses.

## 8. Status/Error diagnostics

Each module of the CBS-miRSeq generates a log file within current script directory of CBS-miRSeq. The user can look at the corresponding log.txt for details about errors and bugs.

## 9. Availability

The pipeline is freely downloadable (http://www.labmedinfo.org/resources/software/CBS-miRSeq) with no restrictions. The tool is released under the terms of the GNU General Public License (GPL) version 3.0.

## 10. Requested Citation

Kesharwani RK et al. (2017) **CBS-miRSeq: a comprehensive tool for accurate and extensive analyses of microRNA-sequencing data.** *BMC Genomics.*

## 11. Contact information

For any query or feedback, feel free to contact Dr. Rupesh K. Kesharwani (bioinforupesh2009.au@gmail.com).

## 12. Abbreviation

sps = Name of the species/organism (usually 3 letter code used such as : has, dre, mmuetc)

rel = Name related species/organism (usually 3 letter code used such as : has, dre, mmuetc)

DE = Differentially expressed

DEA = differential expression analysis

dir = Directory

## 13. Acknowledgments

This work has been supported by institutional funds of Centro Cardiologico Monzino, Milano, Italy (to Dr. Gualtiero I. Colombo, project supervisor). Authors thank the Department of Electrical, Computer and Biomedical Engineering, Università degli Studi di Pavia, Italy for providing the virtual machine in order to test workability of pipeline.

## 14. References


1. Rougvie AE: Control of developmental timing in animals. *Nature Reviews Genetics* 2001, 2(9):690-701.
2. Esquela-Kerscher A, Slack FJ: Oncomirs - microRNAs with a role in cancer. *Nature reviews Cancer* 2006, 6(4):259-269.
3. Urbich C, Kuehbacher A, Dimmeler S: Role of microRNAs in vascular diseases, inflammation, and angiogenesis. *Cardiovasc Res* 2008, 79(4):581-588.
4. Mraz M, Pospisilova S: MicroRNAs in chronic lymphocytic leukemia: from causality to associations and back. *Expert Review of Hematology* 2012, 5(6):579-581.
5. Soifer HS, Rossi JJ, Saetrom P: MicroRNAs in disease and potential therapeutic applications. *Molecular Therapy* 2007, 15(12):2070-2079.


**Declarations:** Integrated miRSpring scripts come with CBS-miRSeq are not a part of the pipeline; however, it has been adopted to perform analyses only. Please cite all tools and software while using the pipeline.