PAREFIRST

User Guide

A tool for Degradome assisted miRNA prediction using high-throughput sequencing data

15/01/2020

Salma Alzahrani, Chris Applegate, Tamas Dalmay, Leighton Folkes, and Vincent Moulton

Contents

Introduction

PAREfirst analyses the mRNA fragment population through degradome analysis to extract clear signals of sRNA slicer activity. These signals of sRNA mediated cleavage are used to infer sRNA function. With this knowledge, the workflow is able to extract a subset of all sRNAs from a high-throughput sRNA sequencing experiment and focus on the functional sRNA subset for miRNA prediction.

For a full analysis, the workflow requires an input genome, a mRNA dataset (transcriptome), a sRNA dataset, and a degradome dataset. Alternatively, output from independent analyses can be provided in accepted input file formats, such as the PAREsnip2 format for functional analyses. In either scenario, users are encouraged to provide sRNA and degradome replicates to increase prediction confidence. An optional GFF annotation file may be provided, which will be used to annotate predicted sRNAs.

When running in GUI mode, users can input files using the File Wizard of the Database module and set the parameters of individual tools (such as miRCat2 and PAREsnip2).

The output of the workflow is a list of conserved functional miRNAs which are presented in a tabular format. The user can select a miRNA from the table to reveal all interactions that the miRNA was involved in, as well as any miRNA secondary structure predictions.

System requirements

The UEA sRNA Workbench has been tested on various platforms including:

- Mac OSX (Version 10.5 Leopard; 10.6 Snow Leopard; 10.7 Lion, 10.8 Mountain Lion, 10.9 Mavericks, 10.10 Yosemite, 10.13 High Sierra)
- Linux (Ubuntu Version 16.04)
- Windows 7 and 10

Currently the software is built and tested on the official Oracle builds of Java only. However, most of the software should behave in the same way under open builds but we cannot guarantee this.

Required:

Java 8 (At the time of writing Java 9 is not yet supported)

Recommended:

Intel i5 quad core (or similar) 16GB RAM

Input data and parameters

This section discusses the types of data that can be used to perform degradome assisted miRNA prediction with PAREfirst.

Input data

To perform analysis using PAREfirst for a specific organism, the user must input the following data:

- one or more sRNA library replicates,
- one or more degradome library replicates (or one or more PAREsnip2 output files for each provided sRNA library replicate),
- a genome file,
- and a reference sequence (either a transcriptome or gff3 file with corresponding genome)

When extracting the gene sequences from the gff3 and corresponding genome, the user has the option to include or exclude the UTRs.

The sRNA and degradome libraries must be in non-redundant FASTA format with the adapters trimmed. FASTQ to FASTA and adapter removal tools are provided within the UEA sRNA Workbench. Additionally, sequences containing any ambiguous bases will be discarded as they cannot be accurately aligned.

Launching PAREfirst from GUI

PAREfirst is a tool within the UEA sRNA Workbench. In order to run the sRNA Workbench in GUI mode, simply download the latest version from the UEA sRNA Workbench website and extract all the files from your downloaded zip archive to a new directory and then launch the Workbench.jar.

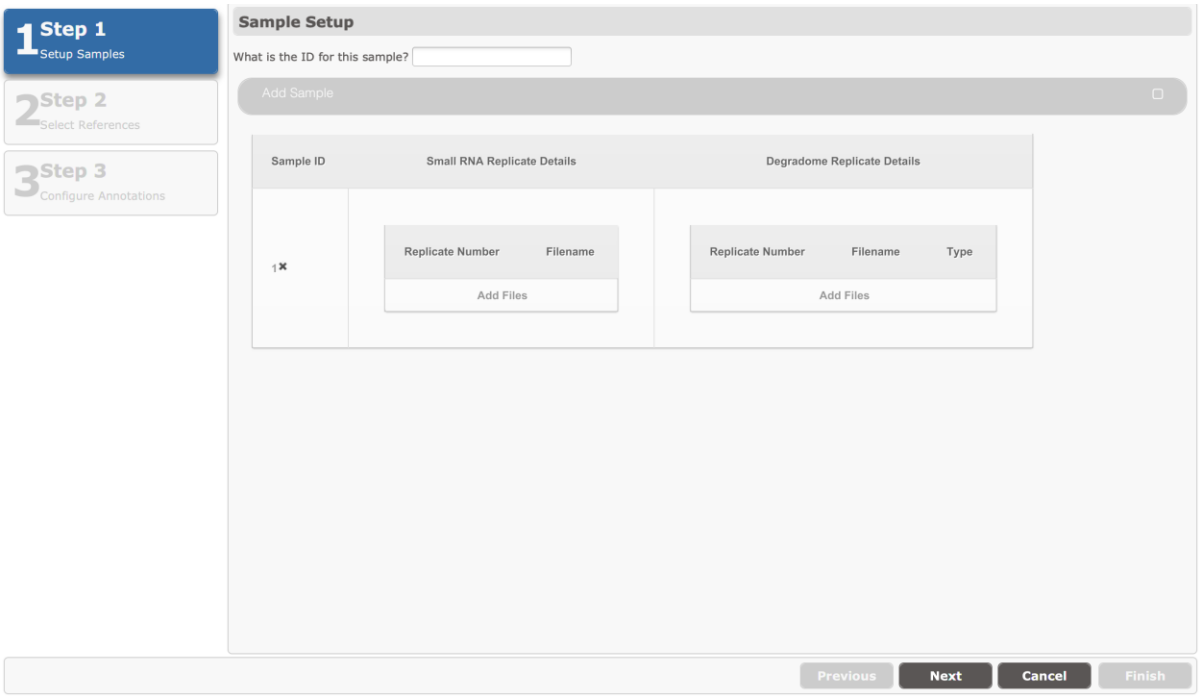
Next, click Open/Close menu -> Pre-configured Workflows -> Create PAREfirst Workflow

Note: If you would prefer to use the command line, please see section Launching PAREfirst from the command-line.

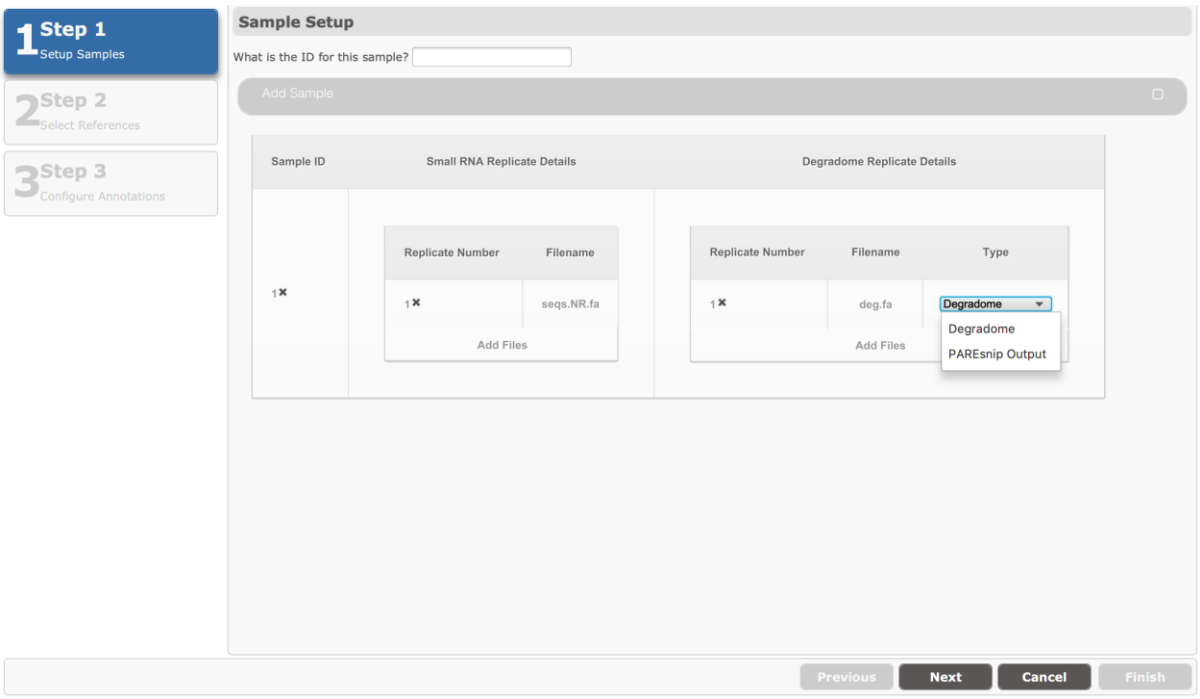
You will then be provided with the following screen

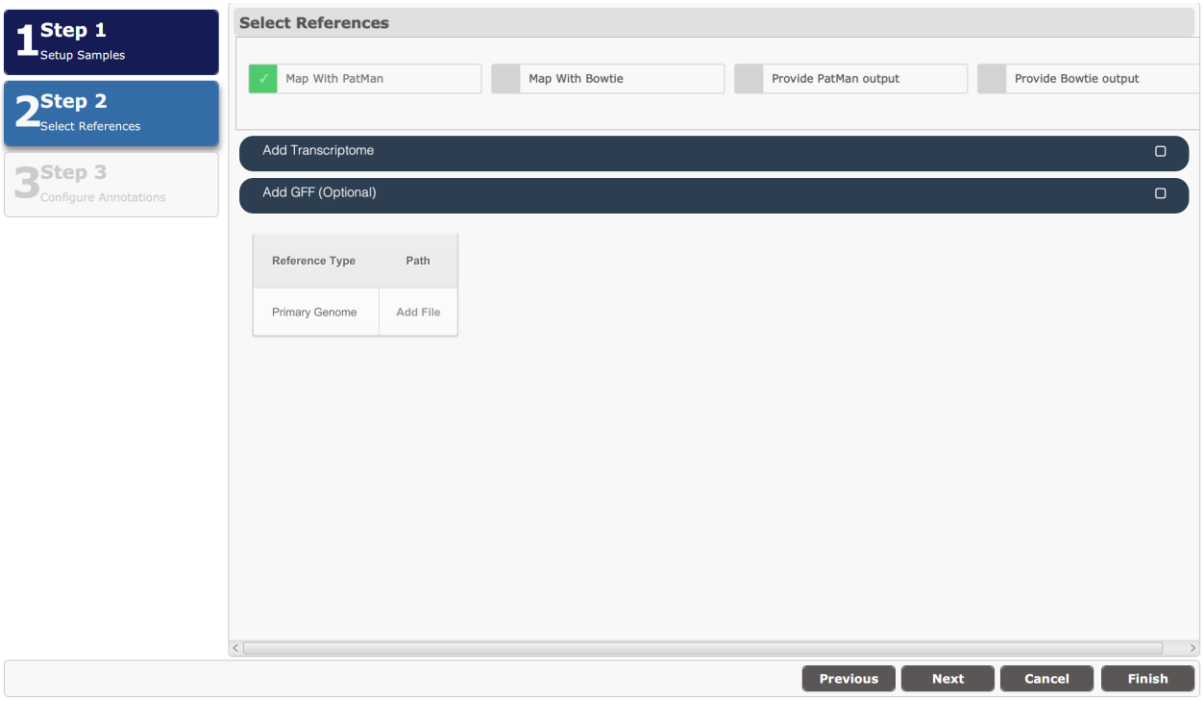
Loading data

Then, to set up the input files, double-click on the 'Database' node. First, add the sRNAome file and the target analysis file, where you can either provide Degradome file or the corresponding 'PAREsnip2 output' file for the inserted sRNAome. Specify which type of Degradome file you inserted using the drop-down menu, and then click 'Next' to continue.

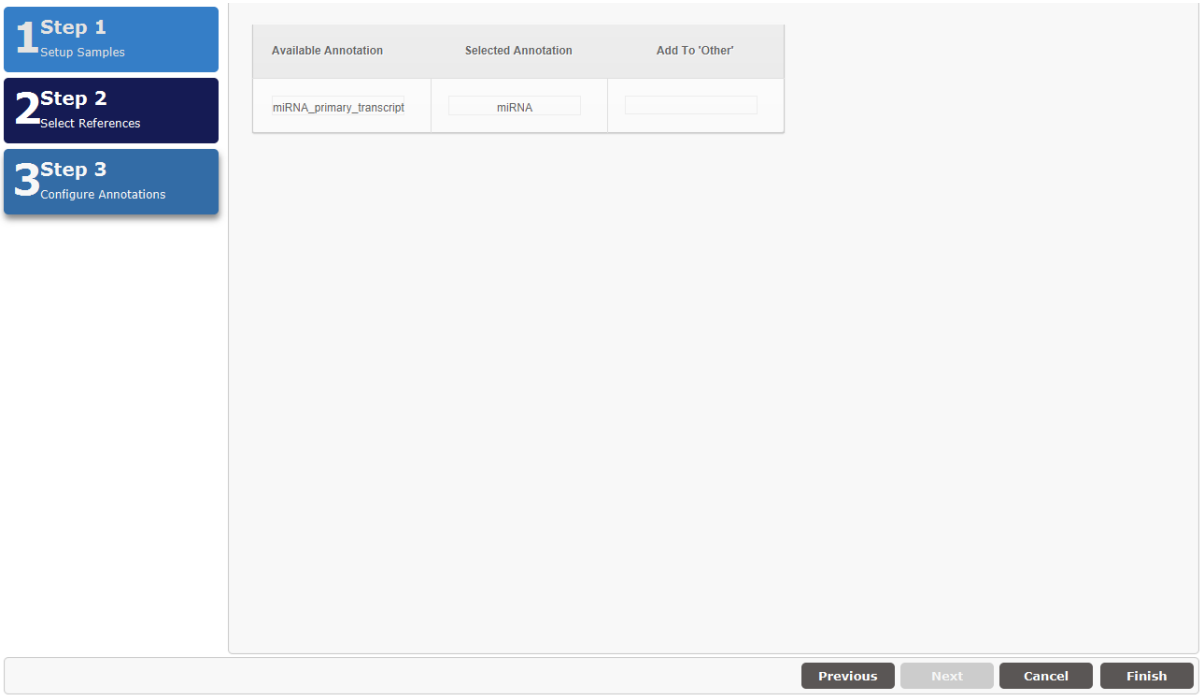


After that, you must provide the Genome file and the Transcriptome file, note that you can't add the Transcriptome file if you selected 'PAREsnip2 output' in the previous step. Once you finished setting-up the database, click 'Finish' and then go back to the home window by clicking on 'Home' in the menu. The database node will turn into blue indicating that is ready and the rest of the Workflow nodes will appear.

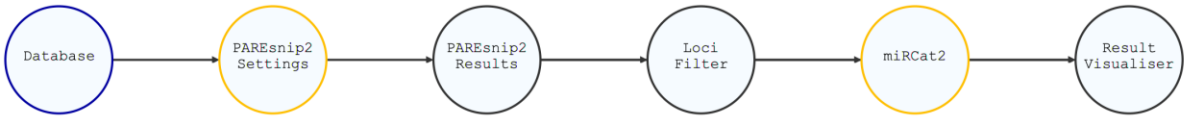




Optionally, you can add a GFF3 file for miRNA annotation. If you did so, click on ‘Next’ to configure the annotations, and drag ‘miRNA’ from ‘Available Annotation’ box to ‘Selected Annotation’ box. This step will add the miRNA annotation for miRCat2 miRNA prediction results.



The next step is to configure PAREsnip2 and miRCat2 where their nodes will be in yellow. Notice that the PAREsnip2 node will turn into yellow and you will be able to configure it only if you provided the Degradome file. You can configure these two tools as follow:



PAREsnip2 analysis

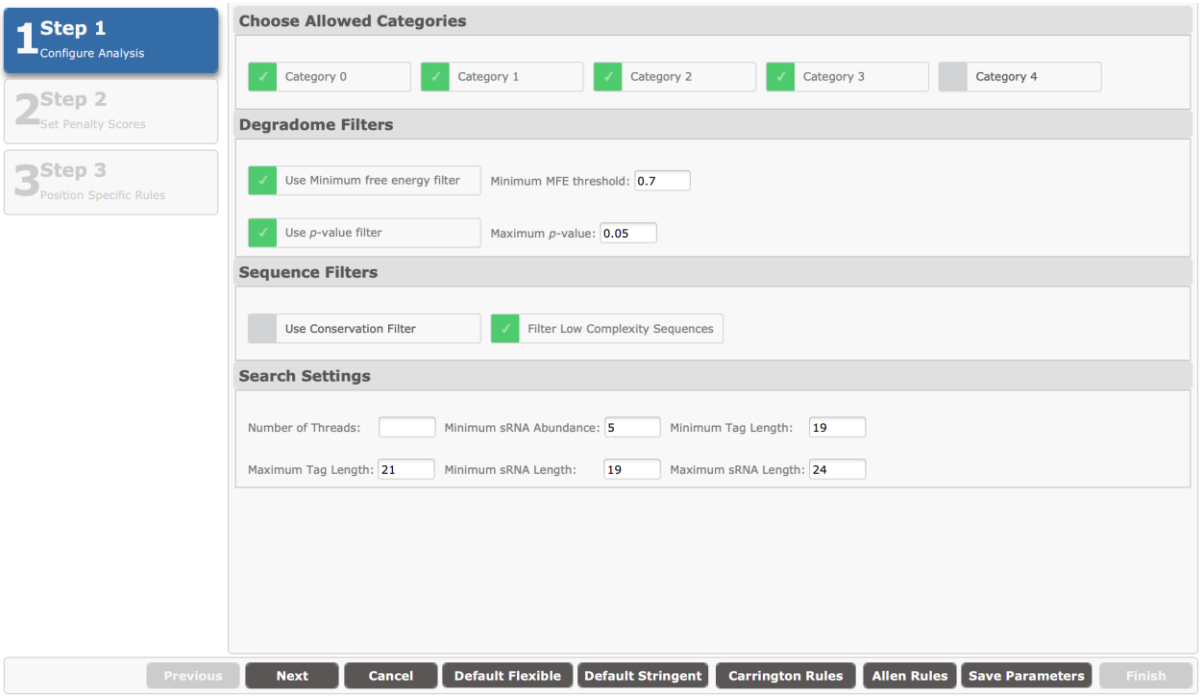
The PAREsnip2 module find targets of small RNAs using the degradome. The workflow version of the tool runs the original PAREsnip2 program but provides a GUI that is compatible with the workflow system. Users can set the PAREsnip2 parameters using the GUI (see the PAREsnip2 page for a detailed description of the input parameters, targeting rules and their default values). The results of the PAREsnip2 analysis are displayed in a table that can be searched, and the data are automatically exported to file in CSV format and can be found in the output directory.

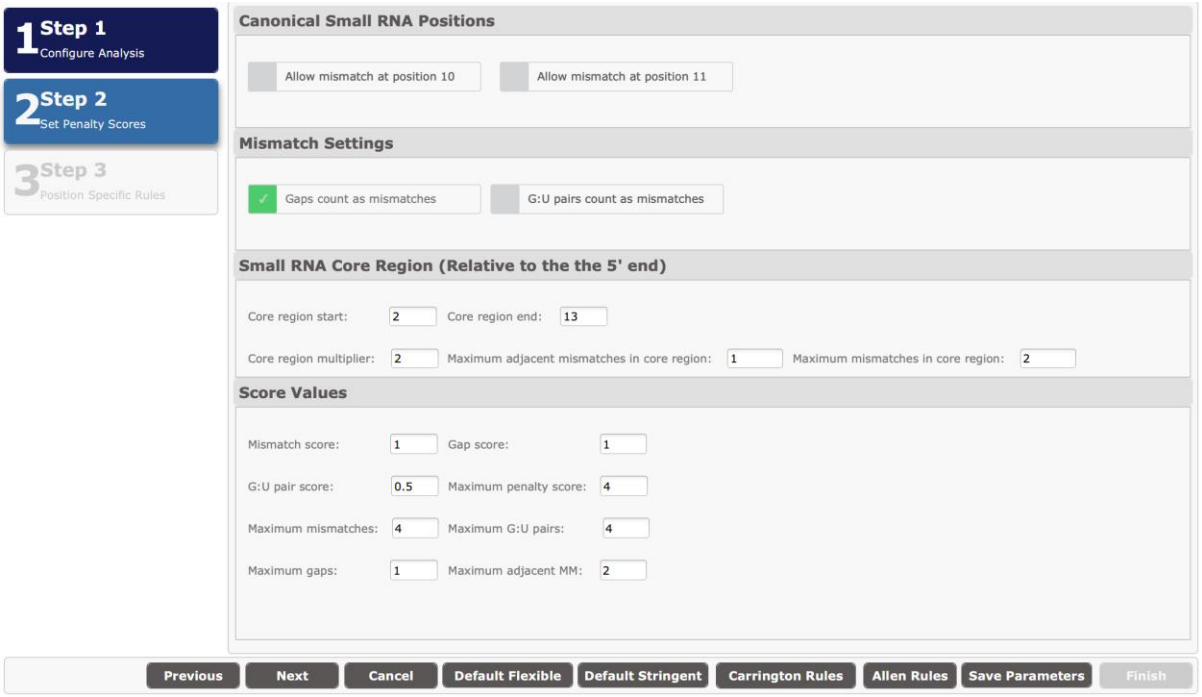
Note: Any sRNA in the duplex column of the output that is not in the sRNA DB is removed from the results.

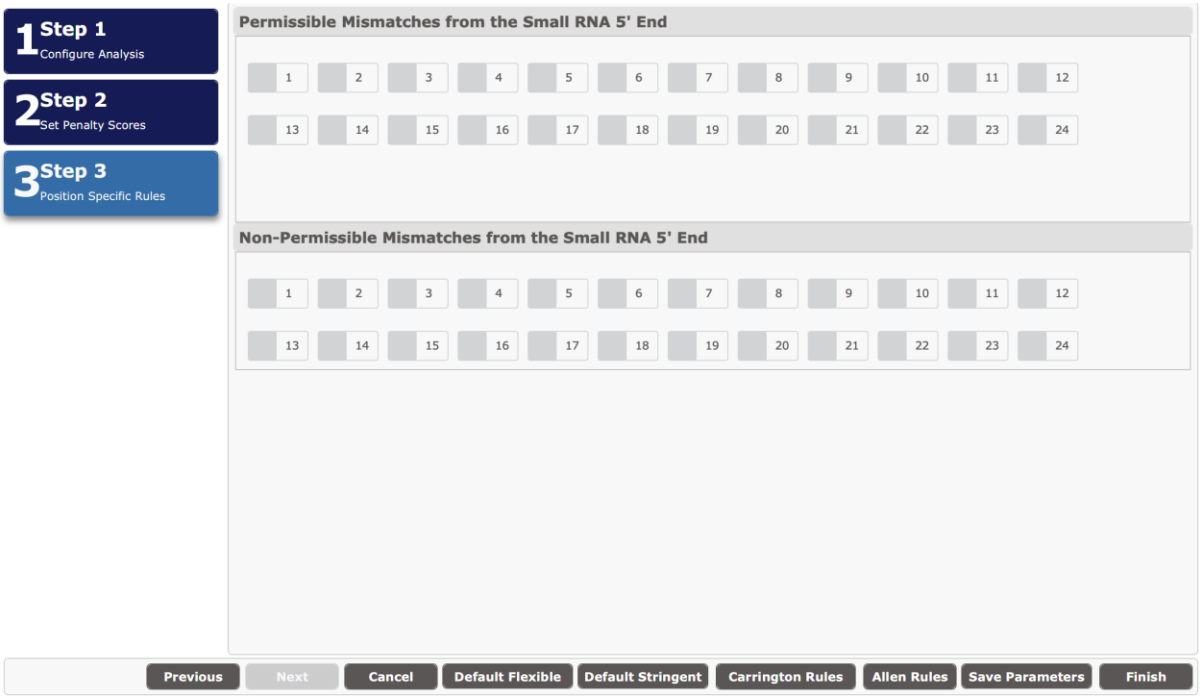
PAREsnip2 parameters

For PAREsnip2 configuration, double-click on the PAREsnip2 node where you can modify the configurations. Also, you may click on the ‘Default Flexible’ or ‘Default Stringent’ buttons to load the

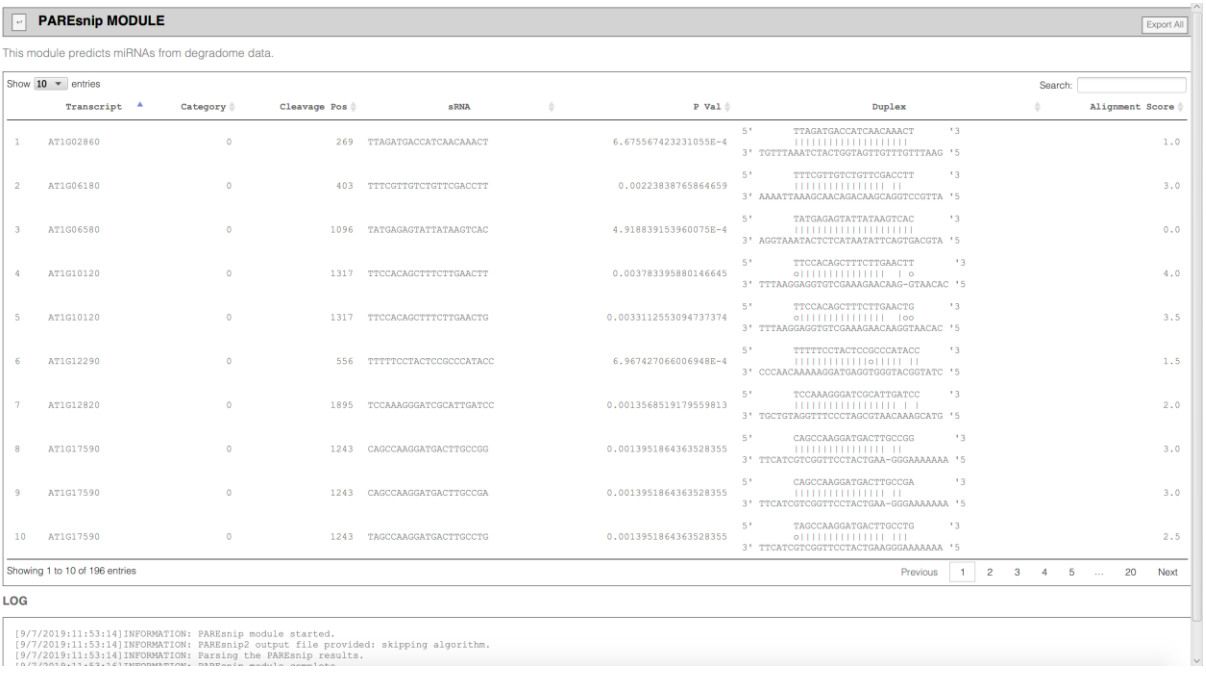
default configurations. Optionally, you can save the configuration by clicking ‘Save Parameters’. The output directory is fixed to the ‘PAREfirst data’ directory of the Workbench. After that, click ‘Next’ to set the targeting rules, where you may select to use the Fahlgren and Carrington rules or the Allen et al. rules by clicking on the ‘Carrington Rules’ or ‘Allen Rules’ button, respectively. Optionally, you can save the selected targeting rules by clicking ‘Save Rules’. Click on ‘Finish’ to go back to the Home window where the PAREsnip2 will turn into blue. A description of PAREsnip2 configuration can be found in the PAREsnip2 manual in the UEA sRNA Workbench website.







PAREsnip2 results



Loci filter

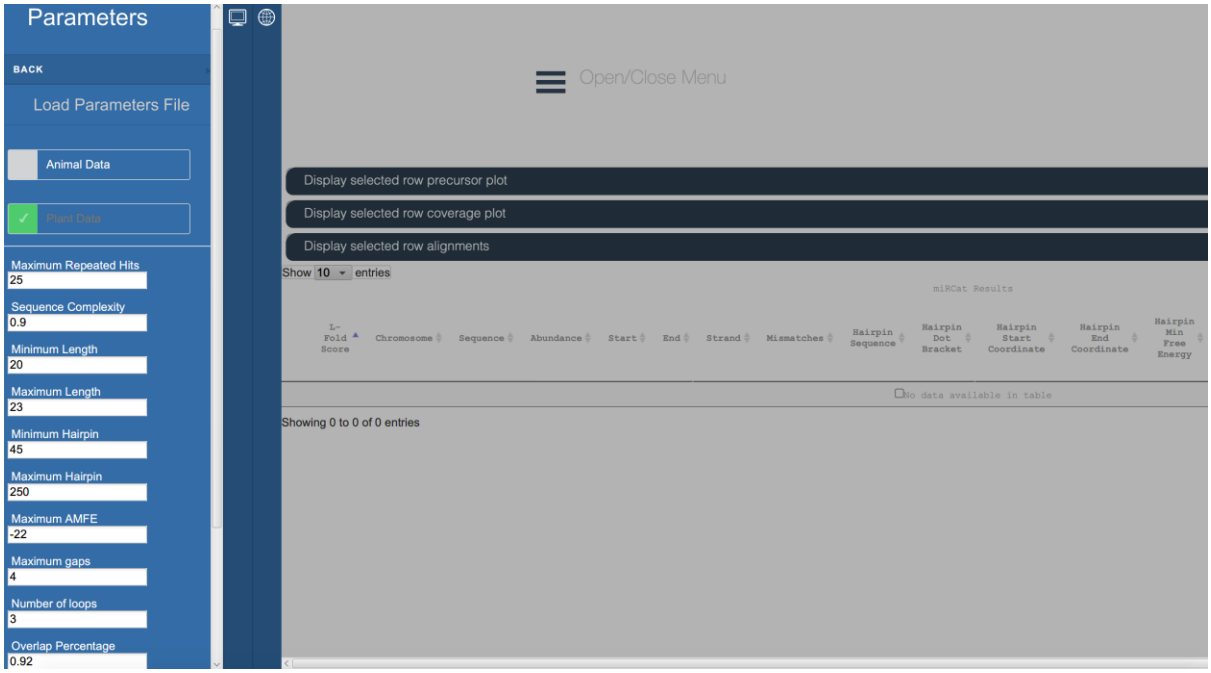
The Loci filter module will return the sRNAs that are involved in mRNA target interactions and the sRNAs that map to the loci of these functional sRNAs.

miRCa2t analysis

The miRCat2 module predicts miRNAs from high-throughput sRNA sequencing data. The workflow version of the tool runs the original miRCat2 program but provides a GUI that is compatible with the workflow system. Users can set the miRCat2 parameters using the GUI (see the miRCat2 page for a detailed description of the input parameters). The results of the miRCat2 analysis are displayed in a table that can be searched, and the data are automatically exported to file in CSV format that can be found in the output directory.

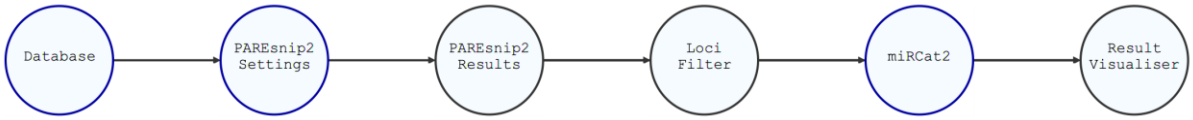
miRCat2 parameters

For miRCat2 configuration, double-click on its node and then open the miRCat2 setup menu and select ‘Settings’ then ‘Parameters’ to modify miRCat2 parameters. Optionally, you select ‘Animal data’ or ‘Plant data’ to load the default parameters for each, or you can load miRCat2 parameters file by clicking on ‘Load parameters’ button. The output directory for miRCat2 results will also be set to the ‘PAREfirst data’ directory of the Workbench. After that, click on ‘Continue Workflow’ then ‘Home’ buttons to go back to the home window where miRCat2 node will turn into blue. You can find detailed description of miRCat2 parameters in UEA sRNA Workbench website.

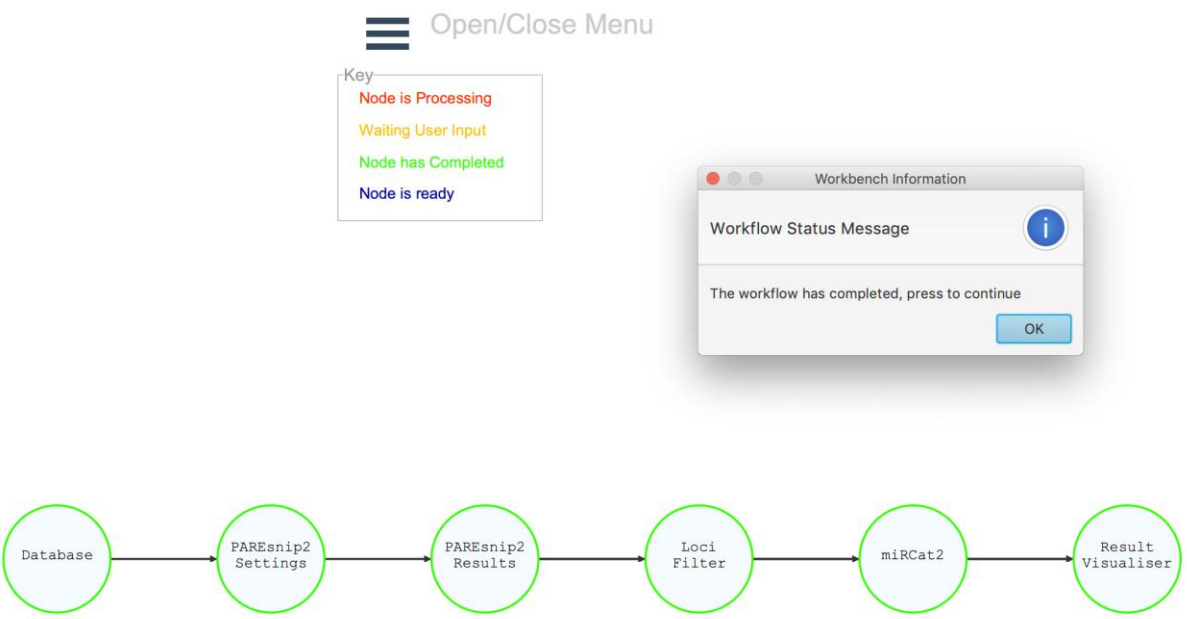


Permissive parameters search

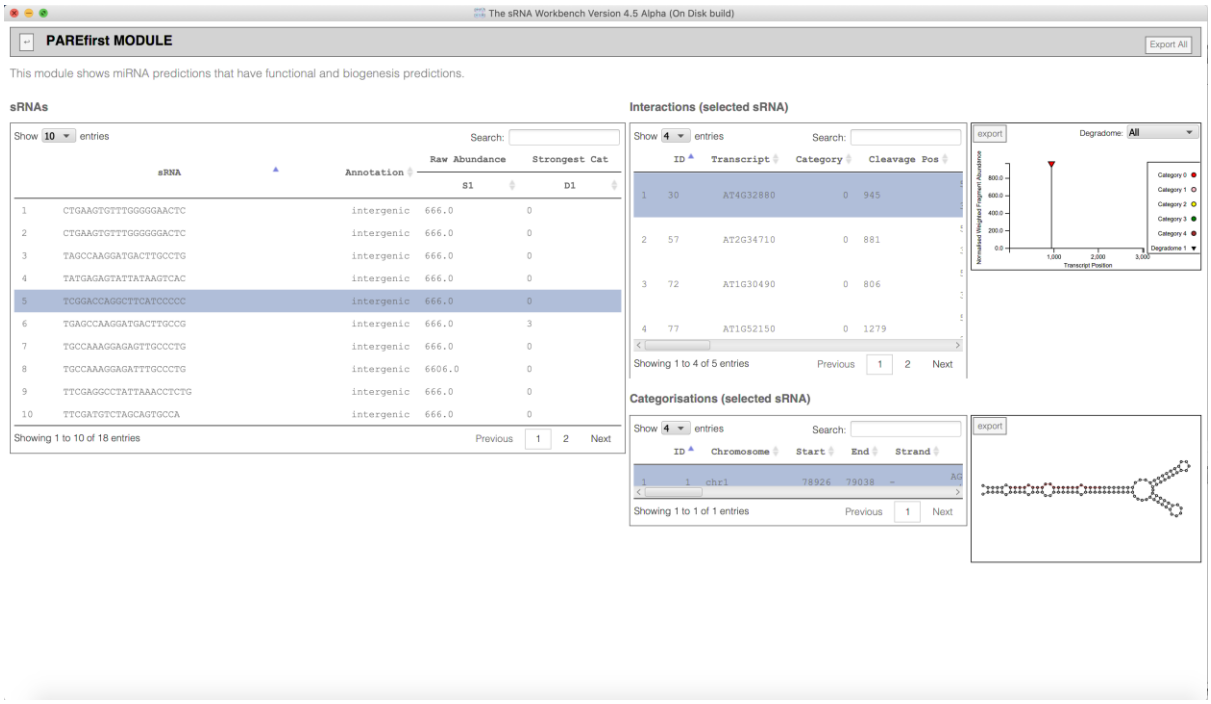
miRCat2 results



PAREfirst results



Finally, you can start the Workflow by clicking on ‘Begin Workflow’ in the Workflow menu. Once the analysis is completed, you can view the results by clicking on the ‘Results visualiser’ node where you can also export the t-plots, and hairpin precursors.. The resulted tables from PAREfirst, PAREsnip2 and miRCat2 can be found in this directory: Workbench_directory/User/PAREfirst_Data.



Launching PAREfirst from the command-line

To start the tool from the command line, navigate to the Workbench installation directory and issue the command:

```
java -jar Workbench.jar -tool parefirst -config config_file
```

The command line version of PAREfirst will only produce the three resulted tables. The configuration file (*config_file*) is in JSON format and it sets-up data for the PAREfirst workflow. A default config file can be found in the default parameters directory of the Workbench.

The configuration file sets-up data for PAREfirst workflow. The file is in JSON format (see example below).

```
{
    "srna_files":
    [
        {   "srna_filename": "/srna1_nr.fa"     },
        {   "srna_filename": "/srna2_nr.fa"     }
    ],
    "degradome_files":
    [
        {   "degradome_filename": "/degradome1.fa"  },
        {   "degradome_filename": "/degradome2.fa"  }
    ],
    "genome_filename": "/genome.fa",
    "transcriptome_filename": "/transcripts.fa",
    "annotation_filename": "/annotations.gff3",
    "mircat2_parameters": "/mircat2_params.cfg",
    "paresnip2_parameters": "/paresnip2_params.cfg",
    "paresnip2_rules": "/paresnip2_rules.cfg"
}
```