miRge-build

Release 0.0.1

Oct 26, 2020

Contents

Enables building small-RNA libraries for the organism of choice to use in the miRge pipeline.

CHAPTER 1

Update: package migration to python3.8

Storage for library building tools. It is designed to allow a user to build specialty libraries for any species of interest to use with miRge. Refer to documentation on how to install and use miRge-build.

If you use miRge-build, please cite DOI:00.000000/ab.00.0.000 .

1.1 Links

- Documentation
- Source code
- Report an issue
- Project page on PyPI (Python package index)

CHAPTER 2

Table of contents

2.1 Installation

miRge-build is developed and tested on Linux environment.

2.1.1 Dependencies

- miRge-build installation requires `python 3.8 or newer`
- **Bowtie v1.2.3 - please pick one based on your OS.**
  - After downloading Bowtie, extract it (`unzip bowtie-1.2.3-linux-x86_64.zip`),
  - Change directory to bowtie `cd bowtie-1.2.3-linux-x86_64` and type `pwd` to get full path of the directory (pwd: present working directory).
  - **Add that path to the environment PATH: `export PATH=$PATH:"pwd <path> "`.**
    - Example: `export PATH=$PATH:"/home/user/software/bowtie-1.2.3-linux-x86_64"`
- Requires scipy for enabling novel miRNA analysis `python3.8 -m pip install --user scipy==1.4.1`
- Requires scikit-learn for enabling novel miRNA analysis `python3.8 -m pip install scikit-learn==0.23.1`
- Requires biopython for parsing all input FASTA files `python3.8 -m pip install biopython==1.77`

2.1.2 Quick installation

The easiest way to install miRge-build is to use `pip3` on the command line:

If you have root previlages, then install miRge-build as follows:

```
sudo python3.8 -m pip install miRge-build
```

if you have only user previlages:

```
python3.8 -m pip install --user miRge-build
```

This will download the software from PyPI (the Python packaging index), and install the miRge-build binary into `$HOME/.local/bin`. If an old version of miRge-build exists on your system, the `--upgrade` parameter is required in order to install a newer version. You can then run the program like this:

```
~/.local/bin/miRge-build --help
```

If you want to avoid typing the full path, add the directory `$HOME/.local/bin` to your `$PATH` environment variable.

2.1.3 Installation with conda

Yet to be implemented

2.1.4 Uninstalling

To uninstall type:

```
pip uninstall miRge-build
```

2.2 User guide

2.2.1 Parameters

To view command-line parameters type `miRge-build -h`:

```
usage: miRge-build [options]

miRge-build (Enables building small-RNA libraries for an organism of choice to use in␣
↪the miRge3.0 pipeline)
optional arguments:
  -h, --help  show this help message and exit
  --version   show program`s version number and exit

Options:
  -g,     --genome             genome file in fasta format (.fna, .fasta or .fa)␣
↪(Required)
  -mmf,   --mature-mir         mature miRNA file in fasta format (Required)
  -hmf,   --hpin-mir           hairpin miRNA file in fasta format (Required)
  -mtf,   --mature-trna        mature tRNA file in fasta format (Required)
  -ptf,   --pre-trna           precursor tRNA file in fasta format (Required)
  -snorf, --snorna             snoRNA file in fasta format (Required)
  -rrf,   --rrna               rRNA file in fasta format (Required)
  -ncof,  --ncrna-other        all other non-coding RNA in fasta format (Required)
  -mrf,   --mrna               mRNA file in fasta format (Required)
  -spnf,  --spike-in           user defined custom spike-in file in fasta format␣
↪(Optional)
```

(continues on next page)

(continued from previous page)

```
  -agff,   --ann-gff            miRNA annotation gff file (Required)
  -ngrs,   --gen-repeats        the genome repeats file with .gtf extension (Optional:␣
↪output however enables novel miRNA prediction in the miRge pipeline)
  -db,     --mir-DB             name of the database to be used (Options: miRBase,␣
↪miRGeneDB) (Required)
  -on,     --organism-name      name of the organism [Note: name should be one word␣
↪and use "_" as separator if necessary] (Required)
  -cpu,    --threads            the number of processors to use for trimming, qc, and␣
↪alignment (Default: 1)
  -pbwt,   --bowtie-path        path to system`s directory containing bowtie binary␣
↪(Required if bowtie is not in the environment path)
```

2.2.2 File format options

Having the right file format is important before making miRge libraries. When dealing with new species which are not available in the set of miRge3.0 libraries, it is important to prioritize what is essential. Novel miRNAs runs scipy cKDTree during library preparation and it consumes a lot of computational resources and time depending on the genome size (up to 10 hours). Making a general build without novel miRNA detection is more straight forward and faster to build libraries.

General format options

Example usage

Example command usage:

```
miRge-build -g genome.fasta -mmf nematode_mature_miRBase.fa -hmf hairpin_miR.fa -mtf␣
↪mature_trna.fasta -ptf pre_trna.fasta -snorf snorna.fasta -rrf rrna.fasta -ncof␣
↪ncrna_other.fasta -mrf mrna.fasta -agff nematode_miRBase.gff3 -db miRBase -on␣
↪roundworm -cpu 10  -ngrs WBcel235_genome_repeats.GTF
```

Output command line:

```
bowtie version: 1.2.3

Library indexing in progress...

Building the kdTree of roundworm_genome_repeats.GTF....

Building the kdTree of roundworm_genome_repeats.GTFtakes: 1.4s
Transforming roundworm_genome.fa takes: 0.9s

miRge-build is complete in 108.2122 second(s)
```

Output directory structure:

```
DB = '--mir-DB'; name of the database used (miRBase or miRGeneDB)

Organism
├── annotation.Libs
│   ├── organism_DB.gff3
│   ├── organism_genome_repeats.pckl (if `-ngrs` is opted)
```

(continues on next page)

(continued from previous page)

```
    ├── organism_miRNAs_in_repetitive_element_DB.csv (if `-ngrs` is opted)
    └── organism_merges_DB.csv
├── fasta.Libs
    ├── organism_genome.pckl (if `-ngrs` is opted)
    └── organism_merges_DB.fa
└── index.Libs
    ├── organism_genome*.ebwt
    ├── organism_hairpin_DB*.ebwt
    ├── organism_mirna_DB*.ebwt
    ├── organism_mature_trna*.ebwt
    ├── organism_pre_trna*.ebwt
    ├── organism_rrna*.ebwt
    ├── organism_snorna*.ebwt
    ├── organism_mrna*.ebwt
    ├── organism_ncrna_others*.ebwt
    ├── organism_mature_trna*.ebwt
    └── organism_spike-in*.ebwt (Optional)
```

Name of the database

miRge uses miRBase or miRGeneDB as the reference database. So, it is mandatory to use `-db` option to either `-db miRBase` or `-db miRGeneDB`. Reference miRNA database `-db` and annotation GFF `-agff` files can be found at miRGeneDB and miRBase.

Name of the organism

miRge-build creates and stores all the libraries under the folder which is named after the organism. It is recommended to use a simple name and avoid any special character (use "_" if the name needs to be seperated by a space). Example: `-on human`; `-on horse`; `-on golden_lemur`; `-on my_database` etc.

Fasta format

Parameters with `-g`, `-mmf`, `-hmf`, `-mtf`, `-ptf`, `-snorf`, `-mrf`, `-spnf` should be in FASTA format as shown below. `-spnf` or –spike-in is optional if the user is interested in adding an additional database with spike-in reads.

FASTA Format:

```
>Header or Identifier
NUCLEOTIDE SEQUENCE
Ex:
  >hsa-let-7a-5p
  TGAGGTAGTAGGTTGTATAGTT
```

NOTE:

```
The Header ID of hairpin miRNA FASTA should match the mature miRNA FASTA file. This␣
↪is required for accurate isomiR annotation.
miRge-build fetches 2bp upstream to 5p and 6bp downstream to 3p mature miRNA from the␣
↪hairpin miRNA based on the matching ID.
Exception: If the mature miRNA name contains XXXX-5p, XXXX-3p, XXXX-[5|3]p*,  XXXX_5p␣
↪or XXXX_3p where XXXX matches the hairpin miRNA ID.
Also, if this is not possible, miRge will not throw any errors, however, and it will␣
↪proceed with the user provided files.
```

(continues on next page)

(continued from previous page)

Novel miRNA options

Novel miRNA prediction requires the genome file (which is provided in the general format) and genome repeats file in GTF format, `-ngrs`. As mentioned previously, novel miRNA analysis consumes a lot of computational resources and time.

Custom annotation options

This is **optional**, that two files under the `annotation.Libs` subdirectory requires users input manually.

merges

This file structured as `organism_merges_database.csv` allows users to define a miRNA family for miRNAs with similar sequences. This method is described in detail in the original miRge manuscript (Baras et al Plos One, 2015). Below is the guide to format the file, where `hsa-miR-376b-5p/376c-5p` is the name of the miRNA family seperated by `/` followed by the family members such as `hsa-miR-376b-5p` and `hsa-miR-376c-5p` all separated by `,`. The next such miRNA family should begin in a new line. Here, four such examples are shown below.

```
hsa-miR-376b-5p/376c-5p,hsa-miR-376b-5p,hsa-miR-376c-5p
hsa-miR-518c-3p/518f-3p,hsa-miR-518c-3p,hsa-miR-518f-3p
hsa-miR-642a-3p/642b-3p,hsa-miR-642a-3p,hsa-miR-642b-3p
hsa-miR-3155a-3p/3155b,hsa-miR-3155a-3p,hsa-miR-3155b
hsa-miR-3689b-3p/3689c,hsa-miR-3689b-3p,hsa-miR-3689c
```

_miRNAs_in_repetitive_element_

This file structured as `organism_miRNAs_in_repetitive_element_database.csv` allows users to define miRNAs that overlap with repeat elements in the genome. This eliminates miRNA reads to be identified as novel miRNAs or identifying one as A-to-I editing, both of which might be misleading.

Below is the guide to format the file, where miRNA names which overlaps with repeat elements are separated by `,`. The `gene_id` and `transcript_id` of a repeat element should follow the miRNA name. See the example below:

```
hsa-miR-28-5p,gene_id "L2c"; transcript_id "L2c_dup8856";
hsa-miR-28-3p,gene_id "L2c"; transcript_id "L2c_dup8856";
hsa-miR-95-5p,gene_id "L2c"; transcript_id "L2c_dup382";
hsa-miR-95-3p,gene_id "L2b"; transcript_id "L2b_dup437";
hsa-miR-181c-5p,gene_id "MamRTE1"; transcript_id "MamRTE1_dup11";
```

Resources

- The genome repeats can be obtained from UCSC
- The database sequences for other small RNA can be obtained from UCSC or Ensembl
- Bowtie-v1.2.3 - please pick one based on your OS.

2.3 MIT License