# AMA.v1.4.0

An automatic pipeline for exploration of SRA datasets with sequences as a query

---

### Requirements

- Bioconda (Go to the official documentation).

It is critical to include the 'bioconda channel' in addition to the other channels as indicated in the official manual.

Please pay attention to the instructions in the installer while setting up the miniconda. Use `echo $PATH` to verify whether the 'conda installation' of Python is in your PATH variable.

---

### Installation

- Please download and extract the package.zip file into the desired location before starting the setup. Before downloading and processing fastq files with the pipeline, please make sure the system has enough disk space available.
- Use the `requirements.txt` provided in the package and create a virtual environment.
- Activate the environment and run the setup script.

```
$ cd AMA/
$ conda create --name env --file requirements.txt -y
$ conda activate env
$ perl setup.pl
```

This tool was tested on "Ubuntu 20.04.4 LTS" with "conda 4.11.0" using the installation procedure mentioned above.

Please refer the Troubleshooting section for additional resources.

---

### Example usage

- To get full usage info: `perl ama.pl --help`

- Minimal usage example: `perl ama.pl --input example/SraRunInfo.csv --sequences example/Arabidopsis_thaliana.TAIR10.ncrna.fa`

---

### Configuration file

Configuration file `conf.txt` requires the absolute path of the tools incorporated into the pipeline.

The user can choose between *blastn* or *bowtie2* by changing the 'execute flag' to either 0 or 1 in the configuration file while leaving the rest of the parameters to default values. By default, both the tools are enabled *ie.* `execute = 1`.

**Note**: If the user wishes to use a different installation than Bioconda, the user can manually install and specify the path of the tools in the configuration.

---

### Pipeline parameters

- **`--input`** (mandatory) The user can provide input in either of the following ways:
  - A single SRA run accession. eg: **`perl ama.pl --input SRR12548227 --sequences example/Arabidopsis_thaliana.TAIR10.ncrna.fa`**
  - A list of run accessions in a text file (1 run accession per line). eg: **`perl ama.pl --input example/list.txt --sequences example/Arabidopsis_thaliana.TAIR10.ncrna`**
  - The SRA runInfo exported directly from the NCBI-SRA web portal. Goto the SRA homepage and search for the desired keyword. Export the `SraRunInfo.csv` by clicking 'Send to' => File => RunInfo). eg: **`perl ama.pl --input example/SraRunInfo.csv --sequences example/Arabidopsis_thaliana.TAIR10.ncrna.`**
- **`--sequences`** (mandatory) The user should provide a fasta file containing the query sequences.
- **`--output`** (optional) The output directory to store the results. By default, the output will be stored into the **`results/`** directory of the package. eg: **`perl ama.pl --input example/SraRunInfo.csv --sequences example/Arabidopsis_thaliana.TAIR10.ncrna.fa --output /src/main/test/`**
- **`--mode`** (optional) Choose one of the three modes to run the pipeline.
  - The **`screen`** is the default mode which will only download a fraction of the data-set per SRA-run accession and analyse the file as per the given configuration.
  - The **`full`** mode will execute the pipeline by downloading the complete fastq file per SRA-run accession.

  - The **both** option searches for samples using a fraction of the data that meet the minimum alignment cutoff from either 'bowtie2' or 'blastn', and then automatically performs alignment by downloading the entire fastq file.

    eg: **perl ama.pl --input example/SraRunInfo.csv --sequences example/Arabidopsis**
    **--output /src/main/test/ --mode screen**

    There is a supporting **summary** mode, that will generate a unified alignment summary by examining the output files created by either screen-mode or full-mode. The summary mode should only be used when the user needs to recreate the summary stats from the pre-existing results. The user must enter **-mode summary** along with the previously used command parameters to re-generate the summary.

- **--config** (optional) Pipeline configuration. By default it will use the **conf.txt** generated by the setup script. eg: **perl ama.pl --input example/SraRunInfo.csv --sequences example/Arabidopsis_thaliana.TAIR10.ncrna.fa --output /src/main/test/ --mode screen --config conf.txt**

---

### Troubleshooting

- Errors related to Bioconda:

  Use **conda list** command to verify whether the packages mentioned in the `requirements.txt` are successfully installed into your environment.

  Note: The `requirements.txt` provided in this package was exported from `conda 4.11.0` installation running on `Ubuntu 20.04.4 LTS`.

- In case of any missing tool/ conflicting dependencies in the environment, the user can try using **conda search <tool name>** command to find the supported version of the tool and then manually install it by typing **conda install <tool name>** inside the environment. Please refer the official troubleshooting guide for further help.

  Note: On macOS and Linux, the supported tools and their dependencies aren't always the same. Even when all of the requirements are completely aligned, the set of available versions isn't necessarily the same. User may try setting up the environment using any of the supplementary `requirements-*.txt` provided in the `src/main/resources/` directory.

- Error using CPAN modules:

  Before beginning the setup, Mac OS users must ensure that they have write permission to the `/Users/\*/.cpan/` directory, and Linux users must ensure that their CPAN is properly configured.

---

**List of Perl modules and tools incorporated in the pipeline**

- Perl modules:
  - Config::Simple
  - Parallel::ForkManager
  - Log::Log4perl
  - Getopt::Long
  - Text::CSV
  - Text::Fuzzy
- Tools:
  - NCBI EDirect utilities
  - NCBI SRA Toolkit
  - FastQC
  - Trimmomatic
  - FASTX-Toolkit
  - NCBI Blast
  - Bowtie2
  - Samtools

---