PhyloMagnet Documentation

Release 0.0.1

Max Emil Schön

Oct 29, 2020

Contents

Contents:

CHAPTER 1

Installation Instructions for PhyloMagnet

1.1 Nextflow

To run PhyloMagnet, you will need to install Nextflow, a pipeline execution framework. This is however quite simple:

```
curl -s https://get.nextflow.io | bash
```

1.2 Singularity

We provide a singularity container with all necessary tools installed and configured. To use it, you first need to install *Singularity* (either version 2 or version 3) itself:

```
VERSION=2.6
wget https://github.com/singularityware/singularity/releases/download/$VERSION/
↪singularity-$VERSION.tar.gz
tar xvf singularity-$VERSION.tar.gz
cd singularity-$VERSION
./configure --prefix=/usr/local
make
sudo make install

# for installation of version 3 see the Singularity website
```

1.3 Singularity container

Then, download the container from singularity-hub or build it locally with the singularity recipe:

```
# singularity 2

singularity pull --name PhyloMagnet.simg shub://maxemil/PhyloMagnet:latest
# or
sudo singularity build PhyloMagnet.simg Singularity

# singularity 3

singularity pull --name PhyloMagnet.sif shub://maxemil/PhyloMagnet:latest
# or
sudo singularity build PhyloMagnet.sif Singularity
```

Note: We highly recommend using the provided Singularity container to install all needed Software. Installing everything directly on the machine can be achieved using the conda environment.yml file.

If you want to see the version of the tools installed in the container, simply use conda to list all installed packages:

```
singularity exec PhyloMagnet.sif conda list -n PhyloMagnet-<version>
```

1.4 PhyloMagnet

now you can either get PhyloMagnet from github (clone or download from github.com/maxemil/PhyloMagnet) or let Nextflow handle that as well:

```
nextflow run maxemil/PhyloMagnet --help

# or
git clone https://github.com/maxemil/PhyloMagnet

nextflow run PhyloMagnet/main.nf --help
```

CHAPTER 2

Command line options

To get a quick overview of all available command line options, run

```
nextflow run maxemil/PhyloMagnet --help
```

2.1 Query options

2.1.1 Fastq input (`--fastq`)

Provide one or several short read samples (with wildcards in double quotes) in fastq format to PhyloMagnet. The files can be in gzip compressed form. E.g. `"fastq/*.fastq.gz"`.

2.1.2 BioProject list (`--project_list`)

Provide a list of BioProject identifiers (e.g. `PRJNA324704`) in a single file, one ID per line.

2.1.3 BioProject run IDs (`--is_runs`)

Boolean value that modifies the `--project_list` option such that it expects run IDs instead of Project IDs (e.g. `SRR3656745`). Default is `false`

2.1.4 Database (`--database`)

Choose if you would like to download the sra data from ncbi's (`ncbi`) or ena's (`ena`) servers (depending on where you are, this might make a huge speed difference). Default is `ena`

2.2 Reference options

2.2.1 EggNOG reference Ids (`--reference_classes`)

A file with identifiers from the EggNOG database, e.g. `COG0051` or `ENOG410XPEN`. One ID per line.

2.2.2 Archived reference packages (`--reference_packages`)

A single path to one or several compressed references packages (see the utility script `make_reference_packages.sh` in the `utils` folder) from a previous PhyloMagnet run. Can contain wildcards if put in double quotes. e.g. `"my_rpkg/*.tgz"`

2.2.3 Local reference sequences (`--local_ref`)

A single path to one or several local fasta files containing orthologous groups of proteins. the referece sequences should be annotated with their taxonomy ID in the NCBI taxonomy (e.g. `562.NC_011750` as the sequence record's header, `562` being the taxID).

2.3 Output options

2.3.1 Output queries (`--queries_dir`)

Output directory for queries; assembled contigs, placement results and summary tables/figures get saved here. Defaults to `queries`.

2.3.2 Output references (`--reference_dir`)

Output directory for references; fasta files, alignments, model files and trees get saved here. Defaults to `references`.

2.4 Run parameters

2.4.1 Phylogenetic method (`--phylo_method`)

Phylogentic tool used for the reconstruction of the reference tree. Only used for references from EggNOG and local files, not packages. Accepted values: `iqtree`, `fasttree`, `raxml`.

2.4.2 Alignment method (`--align_method`)

Alignment tool used to compute reference alignments. Only used for references from EggNOG and local files, not packages. Accepted values: `maftt-*`, `prank`.

2.4.3 Taxonomic lineage (`--lineage`)

the lineage(s) to report occurences for. Can be either a list of labels provided by the user (e.g. `Rickettsiales`, `Holosporales`), a taxonmic rank (e.g. `family`), or both (e.g. `Rickettsia,family`)

2.4.4 No. of CPUs (`--cpus`)

No. of CPUs to use. For more fine-grained uage of resources per process change the file `nextflow.config`.

2.4.5 Threshold for plotting taxonomic labels (`--plot_threshold`)

Threshold for filtering low frequent taxon labels from summary plots. e.g. label would not get plotted when present in only 1 out of 4 trees for default value. Threshold is checked per sample. Default: 0.25, accepted values between o and 1

2.4.6 Threshold for including taxonomic labels (`--aLWR_threshold`)

Threshold of accumulated likelihood weight ratio (aLWR, see `gappa`'s documentation) to include labels in the summary table. Default 0.8, accepted values between o and 1

2.4.7 MEGAN VM options file (`--megan_vmoptions`)

File with options that are passed on to the Java virtual machine running MEGAN. Most importantly, state here the amount of memory that is available, e.g. *-Xmx16G* for 16GB. By default the file is expected to be in the execution directory. Example file *MEGAN.vmoptions* is included in the repository.

2.4.8 Location of MEGAN (`--megan_dir`)

The directory MEGAN's source files are located. When Manually installing MEGAN it could be something like */usr/local/megan*. Leave as default if the singularity image is used.

2.4.9 Location of Python3 (`--python3`)

Location of the python3 executable that has all needed packages available. Should usually be */usr/bin/env python3*, leave as default is using the singularity image.

CHAPTER 3

Example Usage

You can use PhyloMagnet in a number of different scenarios:

3.1 Using Bioproject IDs and eggNOG references

```
nextflow run maxemil/PhyloMagnet --reference_classes eggnog.txt \
                     --project_list bioprojects.txt \
                     --phylo_method fasttree \
                     --queries_dir queries_output \
                     --reference_dir ref_output \
                     --lineage Rickettsiales,family \
                     --megan_vmoptions MEGAN.vmoptions \
                     --cpus 20
```

3.2 Using SRA run IDs and local references

```
nextflow run maxemil/PhyloMagnet --local_ref customOG.fasta \
                     --project_list sra_runs.txt \
                     --is_runs true \
                     --phylo_method fasttree \
                     --queries_dir queries_output \
                     --reference_dir ref_output \
                     --lineage Rickettsiales,family \
                     --megan_vmoptions MEGAN.vmoptions \
                     --cpus 20
```

3.3 Using local FastQ file and local + eggNOG references

```
nextflow run maxemil/PhyloMagnet --reference_classes eggnog.txt \
                        --local_ref customOG.fasta \
                        --phylo_method fasttree \
                        --fastq local_metagenome.fastq.gz
                        --queries_dir queries_output \
                        --reference_dir ref_output \
                        --lineage Rickettsiales,family \
                        --megan_vmoptions MEGAN.vmoptions \
                        --cpus 20
```

CHAPTER 4

Troubleshooting

4.1 Out of Memory

A common source of error is the memory allocation to MEGAN6. If the MEGAN process crashes with a 'Out of Memory' Error, try to increase the parameter in the 'MEGAN.vmoptions' file, e.g. set it to 16GB:

```
-Xmx16G
```

The memory should be around the same as the size of the query (fastq) file.

4.2 data.jar not found

Another problem might be that the 'includeLocalRef' process exits saying it cannot find the 'data.jar'. Make sure you are using the location of MEGAN6, which is where PhyloMagnet copies the jar from.

CHAPTER 5

Contact

github: ‘@maxemil <ttps://github.com/maxemil>’ ‘maxemil.github.io <https://maxemil.github.io/>’

CHAPTER 6

Citation

If you use PhyloMagnet, please cite the article: Max E Schön, Laura Eme, Thijs J G Ettema, PhyloMagnet: fast and accurate screening of short-read meta-omics data using gene-centric phylogenetics, Bioinformatics, btz799, https://doi.org/10.1093/bioinformatics/btz799

Also cite all the software that PhyloMagnet uses internally, depending on your choice of tree and alignment software:

diamond: doi:10.1038/nmeth.3176 and https://github.com/bbuchfink/diamond

megan: doi:10.1371/journal.pcbi.1004957 and doi:10.1186/s40168-017-0233-2 and http://ab.inf.uni-tuebingen.de/software/megan6/

papara: doi:10.1093/bioinformatics/btr320 and https://cme.h-its.org/exelixis/web/software/papara/index.html

epa-ng: doi:10.1093/sysbio/syy054 and https://github.com/Pbdas/epa-ng

gappa: doi:10.1101/647958 and https://github.com/lczech/gappa

iqtree: doi:10.1093/molbev/msu300 and http://www.iqtree.org/

FastTree: doi:10.1371/journal.pone.0009490 and http://www.microbesonline.org/fasttree/

raxml-ng: doi:10.1093/bioinformatics/btz305 and https://github.com/amkozlov/raxml-ng

mafft: doi:10.1093/molbev/mst010 and https://mafft.cbrc.jp/alignment/software/

prank: doi:10.1007/978-1-62703-646-7_10 and http://wasabiapp.org/software/prank/

CHAPTER 7

Indices and tables

- genindex
- modindex
- search