LCAParse Documentation

Release 0.4

Richard Leggett

Aug 13, 2020

Contents

CHAPTER 1

Download and installation

1.1 Installing LCAParse

LCAParse can be downloaded from GitHub. You can either download the .zip file, or if you have git installed, you can type:

```
git clone https://github.com/richardmleggett/LCAParse.git
```

LCAParse is a Java application. To run it, you just need the LCAParse.jar file which can be found in the target directory. It can be executed by typing:

```
java -jar /path/to/LCAParse.jar -help
```

We also provide a script that executes the jar. This can be found in the bin directory. At the top of it can be found the line:

```
JARFILE=/Users/leggettr/Documents/github/LCAParse/target/LCAParse.jar
```

You should change this to point to the location of your LCAParse.jar file. You can then place the lcaparse script in a directory pointed to by your PATH variable, so that it is easily available without having to specify the full path.

Alternatively, add the bin directory to your path variable. On Linux, you would typically do this by adding the following command to your .bash_profile (or .profile on Ubuntu) or 'source' script:

```
export PATH=/path/to/LCAParse/bin:$PATH
```

Once you have done this (you may need to close and re-open your terminal window), you should be able to run LCAParse by typing:

```
lcaparse -help
```

1.2 Taxonomy files

LCAParse requires the nodes.dmp and names.dmp files from the NCBI Taxonomy. These are available as part of the taxdump download which can be obtained from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/.

For parsing accession IDs (if tax IDs are not in the Blast output), LCAParse also requires the nucl_wgs.accession2taxid file from the accession2taxid directory of the NCBI Taxonomy FTP site above.

CHAPTER 2

Running LCAParse

2.1 Preparing accession maps

Blast is capable of outputting the taxon ID of matches if a custom output format is specified. However for the default BlastTab and minimap2 outputs, it is necessary to map accession IDs to taxa using the NCBI accession2taxid data.

For speed and memory reasons, LCAParse uses a reformatted version of accession2taxid and you will need to create this file.

You can do this using the following command:

```
lcaparse -makemap -input /path/to/nucl_gb.accession2taxid -output /path/to/file_
↪prefix -taxonomy /path/to/taxonomy_files
```

where:

- `-input` specifies the name of a nucl_gb.accession2taxid file
- `-output` specifes a prefix to use for output filenames
- `-taxonomy` specifies the directory containing NCBI taxonomy files (files needed are nodes.dmp and names.dmp)

lcaparse will output six files:

- prefix_bacteria.txt - a mapping file between accession IDs and bacterial taxon IDs.
- prefix_viruses.txt - a mapping file between accession IDs and viral taxon IDs.
- prefix_archea.txt - a mapping file between accession IDs and archea taxon IDs.
- prefix_eukaryota.txt - a mapping file between accession IDs and eukaryote taxon IDs.
- preifx_other.txt - a mapping file between accession IDs and other taxon IDs.
- prefix_unclassified.txt - a mapping file between accession IDs and unclassified taxons.

You can merge these files if you need to. For example, if you want a mapping file for bacteria and viruses:

```
cat map_bacteria.txt map_viruses.txt > map_bacvir.txt
```

2.2 Running LCAParse

To run, type a command of the form:

```
lcaparse -input filename.txt -output /path/to/output_prefix -taxonomy /path/to/
↪taxonomy/dir -mapfile /path/to/mapfile.txt -format blasttab
```

where:

- `-input` specifies the name of an input Blast or minimap2 file
- `-output` specifies the prefix for output LCA results. Two files will be generated - a output_summary.txt and an output_perread.txt.
- `-taxonomy` specifies the directory containing NCBI taxonomy files (files needed are nodes.dmp and names.dmp)
- `-format` specifies the input file format - either 'nanook', 'blasttab' or 'PAF'.
- `-mapfile` specifies the name of an accession map file created as detailed above. This is needed for blasttab and PAF format files.

Other options:

- `-maxhits` specifies maximum number of hits to consider for given read (default 20)
- `-scorepercent` specifies minimum score threshold as percentage of top score for given read (default 90)
- `-limitspecies` limits taxonomy to species level (i.e. not strain)
- `-warnings` will show warnings for missing accession IDs and taxa

The summary output file consists of four tab separated columns:

- Read count
- Percentage of all reads
- Taxon ID
- Taxon path
- Taxon rank

The per read output file consists of three tab separate columns:

- Read ID
- Taxon ID
- Taxon name
- Taxon rank

2.3 Input formats

The 'blasttab' input file format is achieved using the Blast option:

```
-outfmt 6
```

The 'blasttaxon' format includes an additional taxa ID field and can be achieved using:

```
-outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send␣
↪evalue bitscore staxids'
```

The 'nanook' input file format also includes the subject title field:

```
-outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send␣
↪evalue bitscore stitle staxids'
```

LCAParse is a tool for carrying out taxonomy assignment using a Lowest Common Ancestor algorithm. It currently supports BlastTab and minimap2 PAF files.

CHAPTER 3

Further information

- To find out how to download and install LCAParse, see the *Download and installation page*.
- To find out how to run LCAParse, see the *Running LCAParse page*.
- Source code for LCAParse is on GitHub at https://github.com/richardmleggett/LCAParse.

CHAPTER 4

Follow us

Follow Richard Leggett on twitter @richardmleggett.

Comments or queries, please email richard.leggett@earlham.ac.uk.