sam2lca

Release 1.0.0-beta

Maxime Borry

Apr 21, 2022

CONTENTS:

Homepage: github.com/maxibor/sam2lca

CHAPTER

ONE

SAM2LCA

Lowest Common Ancestor from a SAM/BAM/CRAM sequence alignment file.

1.1 TLDR

Analysis of sequencing reads aligned to a DNA database with NCBI accession numbers, using the NCBI taxonomy

```
sam2lca analyze myfile.bam
```

See all options

```
sam2lca --help
sam2lca update-db --help
sam2lca list-db --help
sam2lca analyze --help
```

For further infos, check out the sam2lca documentation and tutorial

1.2 Installation

1.2.1 With Conda (recommended)

```
conda install -c conda-forge -c bioconda -c maxibor sam2lca
```

1.2.2 With pip

```
pip install sam2lca
```

1.2.3 For development purposes, from the dev branch

```
git clone git@github.com:maxibor/sam2lca.git
git checkout dev
conda env create -f environment.yml
conda activate sam2lca
pip install -e .
```

or

```
pip install git+ssh://git@github.com/maxibor/sam2lca.git@dev
```

1.3 Documentation

The documentation of sam2lca, including tutorials, is available here: sam2lca.readthedocs.io

CHAPTER

TWO

PYTHON API

sam2lca.main.**list_available_db**(*dbdir*, *verbose=False*)

List available taxonomy databases

Parameters **db_dir** (*str*) – Path to sam2lca database directory

Returns List of available taxonomy databases list: List of available acc2tax databases

Return type list

sam2lca.main.**sam2lca**(*sam*, *output=None*, *dbdir='/home/docs/.sam2lca'*, *taxonomy='ncbi'*, *acc2tax='nucl'*, *process=2*, *identity=0.8*, *distance=None*, *length=30*, *conserved=False*, *bam_out=False*)

Performs LCA on SAM/BAM/CRAM alignment file

Parameters

- **sam** (*str*) – Path to SAM/BAM/CRAM alignment file
- **output** (*str*) – Path to sam2lca output file
- **dbdir** (*str*) – Path to database storing directory
- **taxonomy** (*str*) – Type of Taxonomy database
- **acc2tax** (*str*) – Type of acc2tax database
- **process** (*int*) – Number of process for parallelization
- **identity** (*float*) – Minimum alignment identity threshold
- **edit_distance** (*int*) – Maximum edit distance threshold
- **length** (*int*) – Minimum alignment length
- **bam_out** (*bool*) – Write BAM output file with XT tag for TAXID

sam2lca.main.**update_database**(*dbdir='/home/docs/.sam2lca'*, *taxonomy=None*, *taxo_names=None*, *taxo_nodes=None*, *taxo_merged=None*, *acc2tax='nucl'*, *acc2tax_json=None*)

Performs LCA on SAM/BAM/CRAM alignment file

Parameters

- **dbdir** (*str*) – Path to database storing directory
- **taxonomy** (*str*) – Name of Taxonomy database
- **names** (*str*) – names.dmp file for taxonomy database. None loads the NCBI taxonomy database
- **nodes** (*str*) – nodes.dmp file for taxonomy database. None loads the NCBI taxonomy database

- **merged** (*str*) – merged.dmp file for taxonomy database. None loads the NCBI taxonomy database
- **acc2tax** (*str*) – Type of acc2tax database
- **acc2tax_json** (*str*) – Path to acc2tax json file

CHAPTER

THREE

COMMAND LINE INTERFACE

To access the help menu:

```
$ sam2lca --help
```

The list of arguments of options is detailed below

3.1 sam2lca

sam2lca: Lowest Common Ancestor on SAM/BAM/CRAM alignment files
Author: Maxime Borry, Alexander Huebner
Contact: <maxime_borry[at]eva.mpg.de>
Homepage & Documentation: github.com/maxibor/sam2lca

```
sam2lca [OPTIONS] COMMAND [ARGS]...
```

Options

--version
Show the version and exit.

-d, --dbdir <dbdir>
Directory to store taxonomy databases

Default /home/docs/.sam2lca

3.1.1 analyze

Run the sam2lca analysis

SAM: path to SAM/BAM/CRAM alignment file

```
sam2lca analyze [OPTIONS] SAM
```

Options

-t, --taxonomy <taxonomy>
Taxonomy database to use

Default ncbi

-a, --acc2tax <acc2tax>
acc2tax database to use

Default nucl

-i, --identity <identity>
Minimum identity threshold NOTE: This argument is mutually exclusive with arguments: [distance].

Default 0.8

-d, --distance <distance>
Edit distance threshold NOTE: This argument is mutually exclusive with arguments: [identity].

-l, --length <length>
Minimum alignment length

Default 30

-c, --conserved
Ignore reads mapping in ultraconserved regions

-p, --process <process>
Number of process for parallelization

Default 2

-o, --output <output>
sam2lca output file. Default: [basename].sam2lca.*

-b, --bam_out
Write BAM output file with XT tag for TAXID

Arguments

SAM
Required argument

3.1.2 list-db

List available taxonomy and acc2tax databases

```
sam2lca list-db [OPTIONS]
```

3.1.3 update-db

Download/prepare acc2tax and taxonomy databases

```
sam2lca update-db [OPTIONS]
```

Options

-t, --taxonomy <taxonomy>
Name of taxonomy database to create (ncbi | gtdb)

Default ncbi

--taxo_names <taxo_names>
names.dmp file for Taxonomy database (optional). Only needed for custom taxonomy database (non ncbi or gtdb)

--taxo_nodes <taxo_nodes>
nodes.dmp file for Taxonomy database (optional). Only needed for custom taxonomy database (non ncbi or gtdb)

--taxo_merged <taxo_merged>
merged.dmp file for Taxonomy database (optional). Only needed for custom taxonomy database (non ncbi or gtdb)

-a, --acc2tax <acc2tax>
Type of acc2tax mapping database to build. NOTE: This argument is mutually exclusive with arguments: [acc2tax_json].

Options nucl|prot|plant_markers|gtdb_r207|test

--acc2tax_json <acc2tax_json>
JSON file for specifying extra acc2tax mappings NOTE: This argument is mutually exclusive with arguments: [acc2tax].

CHAPTER

FOUR

DATABASES

sam2lca uses two different type of databases:

- a **taxonomy** database to infer the Lowest Common Ancestor (LCA) and retrieve the names and lineage associated to a *taxonomic identifier* (TAXID)
- an **acc2tax** or *accession to TAXID* database, to match sequence accession to a taxonomic identifier

For each of these databases, sam2lca offers different possibilities.

4.1 Taxonomy databases

4.2 `ncbi`

If you're not sure what to use, stick with the default (`ncbi`)

4.3 `gtdb`

If you have bacteria and/or archea DNA sequencing data, you can altenatively choose to use the GTDB taxonomy, which is more phylogenetically consistent than the NCBI database. (see the GTDB article here: 10.1093/nar/gkab776).

To use the `GTDB` database with sam2lca, use:

```
--taxonomy gtdb --acc2tax gtdb_r207
```

This will work if you align your sequencing data against the `gtdb_genomes_reps` genomes.

As of 20/04/2022, only the latest GTDB release (r207) is available. For other (past or future) releases, please have a look at gtdb_to_taxdump and see **custom** section below, or open an issue on the sam2lca github repository.

4.4 `custom`

You can provide your own taxonomy database by providing the following files

- `names.dmp`
- `nodes.dmp`
- `merged.dmp`

For example:

```
sam2lca update-db --taxonomy my_custom_db_name --taxo_names names.dmp --taxo_nodes␣
↪node.dmp --taxo_merged merged.dmp
```

Make sure than the taxonomic IDs are matching the accession2taxid that you're using !

4.5 acc2tax - *accession to TAXID* databases

4.5.1 Nucleotide databases

- `nucl` for nucleotide/DNA sequences, made of:
  - *nucl_wgs* : nucleotide sequence records of type WGS or TSA
  - *nucl_gb* : nucleotide sequence records that are not WGS or TSA
- `plant_markers` for plant identication based on plant specific markers, made of:
  - *angiosperms353* : Angiosperms353 marker data extracted from treeoflife.kew.org with sequence headers reformatted as following:

    Original fasta header

    ```
    >5821 Gene_Name:dph5 Species:Cyperus_laevigatus Repository:INSDC Sequence_
    ↪ID:ERR3650073
    ```

    Reformatted fasta header

    ```
    >5821_Cyperus_laevigatus Gene_Name:REV7  Repository:INSDC Sequence_
    ↪ID:ERR3650073
    ```

    This reformating is necessary to ensure the uniqueness of sequence identifiers. The `fasta` file with reformatted headers (dumped from treeoflife.kew.org on October 21st, 2021) is available for download here: angiosperms353_markers.fa.gz
  - *ITS* : ITS plant markers data extracted from the planITS project. The ITS database is available ITS.fa.gz
  - *rbcL*: rbcL plant marker extraced from 10.3732/apps.1600110, using the version updated on 09.07.2021, shared by the authors here. Fasta headers were rewritten to ensure the uniqueness of sequence identifiers and the dabase is available rbcl.fa.gz.

    Original fasta header

    ```
    >123456 Grabowskia glauca
    ```

    _Reformatted fasta header

    ```
    >rbcL_0_Grabowskia_glauca
    ```

    * *18s SILVA*: 18S SSU markers extracted from the SILVA database. The fasta file is available directly from SILVA arb-silva.de/fileadmin/silva_databases/release_138_1/Exports/SILVA_138.1_SSURef_NR99_tax_silva.fasta.gz

4.5.2 Protein databases

- `prot` for protein sequences, made of:
  - *prot* : protein sequence records which have GI identifiers
  - *pdb* : protein sequence records from the Protein Data Bank

4.5.3 Test database

- `test` : local database to test sam2lca

4.5.4 Custom database

With sam2lca, you can provide a custom database to map accession numbers to TAXIDs.

To do so, sam2lca can accept a `JSON` file, with the `--acc2tax_json` flag in the `sam2lca update-db` subcommand in combination with `--acc2tax custom`.

For example:

```
sam2lca update-db --acc2tax_json acc2tax.json
```

This JSON file should be formatted as below:

```
{
    "mapfiles": {
        "[name_of_mapping]": [
            "path/url_to_compressed_accession2taxid.gz file"
        ]
    },
    "mapmd5": {
        "[name_of_mapping]": [
            "path/url_to_compressed_accession2taxid.gz md5sumfile"
        ]
    },
    "map_db": {
        "[name_of_mapping]": "Name of custom.db"
    }
}
```

An example `json` file can be found here: map_config.json

CHAPTER

FIVE

OUTPUT

sam2lca generates:

- a `JSON` file
- a `CSV` file
- (optionally), a `BAM` alignment file with the `XT` tag set to the NCBI Taxonomy IDs computed by the LCA.

5.1 JSON

A JSON file with NCBI Taxonomy IDs as keys.

- `name`: scientific name of the taxon
- `rank`: taxonomic rank of the taxon
- `count_taxon`: number of reads mapping to the taxon
- `count_descendant`: total number of reads belonging to the descendants of the taxon
- `lineage`: taxonomic lineage of the taxon

Example:

```
{
{
    "1": {
        "name": "root",
        "rank": "no rank",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {}
    },
    "2": {
        "name": "Bacteria",
        "rank": "superkingdom",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {
            "superkingdom": "Bacteria"
        }
    },
    "543": {
        "name": "Enterobacteriaceae",
        "rank": "family",
```

(continues on next page)

(continued from previous page)


```
        "count_taxon": 2152,
        "count_descendant": 2875,
        "lineage": {
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "561": {
        "name": "Escherichia",
        "rank": "genus",
        "count_taxon": 0,
        "count_descendant": 385,
        "lineage": {
            "genus": "Escherichia",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "562": {
        "name": "Escherichia coli",
        "rank": "species",
        "count_taxon": 0,
        "count_descendant": 385,
        "lineage": {
            "species": "Escherichia coli",
            "genus": "Escherichia",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "620": {
        "name": "Shigella",
        "rank": "genus",
        "count_taxon": 0,
        "count_descendant": 338,
        "lineage": {
            "genus": "Shigella",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "622": {
        "name": "Shigella dysenteriae",
        "rank": "species",
        "count_taxon": 0,
```


(continues on next page)

(continued from previous page)


```
        "count_descendant": 338,
        "lineage": {
            "species": "Shigella dysenteriae",
            "genus": "Shigella",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "1224": {
        "name": "Proteobacteria",
        "rank": "phylum",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "1236": {
        "name": "Gammaproteobacteria",
        "rank": "class",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "83333": {
        "name": "Escherichia coli K-12",
        "rank": "strain",
        "count_taxon": 0,
        "count_descendant": 385,
        "lineage": {
            "strain": "Escherichia coli K-12",
            "species": "Escherichia coli",
            "genus": "Escherichia",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "91347": {
        "name": "Enterobacterales",
        "rank": "order",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
```


(continues on next page)

(continued from previous page)


```
            "superkingdom": "Bacteria"
        }
    },
    "131567": {
        "name": "cellular organisms",
        "rank": "no rank",
        "count_taxon": 0,
        "count_descendant": 2875,
        "lineage": {}
    },
    "300267": {
        "name": "Shigella dysenteriae Sd197",
        "rank": "strain",
        "count_taxon": 338,
        "count_descendant": 338,
        "lineage": {
            "strain": "Shigella dysenteriae Sd197",
            "species": "Shigella dysenteriae",
            "genus": "Shigella",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    },
    "511145": {
        "name": "Escherichia coli str. K-12 substr. MG1655",
        "rank": "no rank",
        "count_taxon": 385,
        "count_descendant": 385,
        "lineage": {
            "strain": "Escherichia coli K-12",
            "species": "Escherichia coli",
            "genus": "Escherichia",
            "family": "Enterobacteriaceae",
            "order": "Enterobacterales",
            "class": "Gammaproteobacteria",
            "phylum": "Proteobacteria",
            "superkingdom": "Bacteria"
        }
    }
}
```


5.2 CSV

Rows: Taxons

Columns:

- `TAXID`: NCBI taxonomy ID
- `name`: Name of the taxon
- `rank`: Taxonomic rank
- `count_taxon`: number of reads mapping to the taxon

- `count_descendant`: number of reads belonging to the descendants of the taxon
- `lineage`: Taxonomic lineage of this taxon, each taxonomic level being separated by a – sign.

```
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| TAXID  | name                                    | rank         | count_taxon |␣
↪count_descendant | lineage
↪                                                                                ␣
↪                                                                                ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 1      | root                                    | no rank      | 0           |␣
↪2875             |
↪                                                                                ␣
↪                                                                                ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 131567 | cellular organisms                      | no rank      | 0           |␣
↪2875             |
↪                                                                                ␣
↪                                                                                ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 2      | Bacteria                                | superkingdom | 0           |␣
↪2875             | superkingdom: Bacteria
↪                                                                                ␣
↪                                                                                ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 1224   | Proteobacteria                          | phylum       | 0           |␣
↪2875             | phylum: Proteobacteria || superkingdom: Bacteria
↪                                                                                ␣
↪                                                                                ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 1236   | Gammaproteobacteria                     | class        | 0           |␣
↪2875             | class: Gammaproteobacteria || phylum: Proteobacteria ||␣
↪superkingdom: Bacteria
↪                                                                                ␣
↪                                                                                ␣
↪  |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-------------------------------------------------------------------
↪--------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 91347  | Enterobacterales                        | order        | 0           |␣
↪2875             | order: Enterobacterales || class: Gammaproteobacteria || phylum:␣
↪Proteobacteria || superkingdom: Bacteria
↪                                                                                ␣
↪                                                                                |
```

(continues on next page)

(continued from previous page)

```
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 543    | Enterobacteriaceae                      | family       | 2152        |␣
↪2875             | family: Enterobacteriaceae || order: Enterobacterales || class:␣
↪Gammaproteobacteria || phylum: Proteobacteria || superkingdom: Bacteria          ␣
↪                                                                                |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 561    | Escherichia                             | genus        | 0           |␣
↪385              | genus: Escherichia || family: Enterobacteriaceae || order:␣
↪Enterobacterales || class: Gammaproteobacteria || phylum: Proteobacteria ||␣
↪superkingdom: Bacteria                                                          ␣
↪       |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 562    | Escherichia coli                        | species      | 0           |␣
↪385              | species: Escherichia coli || genus: Escherichia || family:␣
↪Enterobacteriaceae || order: Enterobacterales || class: Gammaproteobacteria ||␣
↪phylum: Proteobacteria || superkingdom: Bacteria                                ␣
↪    |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 83333  | Escherichia coli K-12                   | strain       | 0           |␣
↪385              | strain: Escherichia coli K-12 || species: Escherichia coli ||␣
↪genus: Escherichia || family: Enterobacteriaceae || order: Enterobacterales ||␣
↪class: Gammaproteobacteria || phylum: Proteobacteria || superkingdom: Bacteria  ␣
↪ |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 511145 | Escherichia coli str. K-12 substr. MG1655 | no rank      | 385         |␣
↪385              | strain: Escherichia coli K-12 || species: Escherichia coli ||␣
↪genus: Escherichia || family: Enterobacteriaceae || order: Enterobacterales ||␣
↪class: Gammaproteobacteria || phylum: Proteobacteria || superkingdom: Bacteria  ␣
↪ |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
| 620    | Shigella                                | genus        | 0           |␣
↪338              | genus: Shigella || family: Enterobacteriaceae || order:␣
↪Enterobacterales || class: Gammaproteobacteria || phylum: Proteobacteria ||␣
↪superkingdom: Bacteria                                                          ␣
↪          |
+--------+-----------------------------------------+--------------+-------------+---
↪--------------+-----------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------------------+
```

(continues on next page)

(continued from previous page)

```
| 622    | Shigella dysenteriae                          | species      | 0           |␣
↪338              | species: Shigella dysenteriae || genus: Shigella || family:␣
↪Enterobacteriaceae || order: Enterobacterales || class: Gammaproteobacteria ||␣
↪phylum: Proteobacteria || superkingdom: Bacteria                                ␣
↪   |
+--------+---------------------------------------------+--------------+-------------+---
↪---------------+-------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------+
| 300267 | Shigella dysenteriae Sd197                  | strain       | 338         |␣
↪338              | strain: Shigella dysenteriae Sd197 || species: Shigella␣
↪dysenteriae || genus: Shigella || family: Enterobacteriaceae || order:␣
↪Enterobacterales || class: Gammaproteobacteria || phylum: Proteobacteria ||␣
↪superkingdom: Bacteria |
+--------+---------------------------------------------+--------------+-------------+---
↪---------------+-------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------+
```

5.3 BAM

Only generated when running `sam2lca analyze` with the `-b/--bam_out` flag

The input alignment file is written as a `bam` file, with the following extra tags:

- `XT` (of type `int/i`) set to the TAXID of the LCA assigned to the read
- `XN` (of type `string/Z`) set to the scientific name of the LCA assigned to the read
- `XR` (of type `string/Z`) set to the taxonomic rank of the LCA assigned to the read

```
escherichia_coli_180 355 NC_000913.3 38 1 68M = 148 186␣
↪GTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAA DFFAF?
↪DDHAFEBFHGEHIIIGFBFECBFGDBDF?G@HED?FHGHGE>=;@;@@=D@:5:.;;>:@CC AS:i:0 XS:i:0 XM:i:0␣
↪XO:i:0 XG:i:0 NM:i:0 MD:Z:68 YS:i:0 YT:Z:CP XT:i:543 XN:Z:Enterobacteriaceae␣
↪XR:Z:family
shigella_dysenteriae_504 147 NC_007607.1 181065 255 76M = 181033 -108␣
↪TGATGACAATTTATTGTCTTATCGTTGTTCTTATGGAACGCTTTTCTGATTGATTTCATATTGGCGAGAGAACAAG @CC>
↪CCCE@EGECHGGGEHEFCIGGGHDFIIIHIIIGJJIIJIJIJJJIJIGEHGEHGJIJJIHF@HGHHHHFDBF AS:i:0␣
↪XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:76 YS:i:0 YT:Z:CP XT:i:300267 XN:Z:Shigella␣
↪dysenteriae Sd197 XR:Z:strain
```

Reads belonging with these tags can be filtered with `samtools view` like this: `samtools view --tag [tag_name]:[value_to_filter] [YOURFILE.bam]`

For Example:

- Reads with the LCA's TAXID equal to `300267`: `samtools view --tag XT:300267 aligned.sorted.bam`
- Reads with the LCA's rank at `strain` level: `samtools view --tag XR:genus aligned.sorted.sam2lca.bam`
- Reads with the LCA's scientific name being `Shigella dysenteriae Sd197`: `samtools view --tag XN:"Shigella dysenteriae Sd197" aligned.sorted.sam2lca.bam`

CHAPTER

SIX

TUTORIAL

Using sam2lca to identify a plant taxon from `fastq` sequencing files.

In this tutorial, we'll use the Angiosperms353 plant markers database to identify a plant species present in our sequencing data. The Angiosperms353 database consists of up to 353 universal Angiosperms (flowering plants) gene markers that are derived from the 1000 plant transcriptomes project.

6.1 Installing all tools for this tutorial

For this tutorial, a dedicated conda-environment is available to ease the reproducibility.

Download the environment:

```
wget https://raw.githubusercontent.com/maxibor/sam2lca/master/docs/tutorial/
↪environment.yaml
```

Installing and activating environment:

```
conda env create -f environment.yaml
conda activate sam2lca_tutorial
```

6.2 Getting the reference database

First we need to download and decompress the reference `fasta` database, which in this case, has already been preformatted for sam2lca.

> For the sake of this tutorial, we use a reduced version of the angiosperms353 marker set. The full version is available for download, see sam2lca.readthedocs.io/en/latest/databases.html.

```
wget https://raw.githubusercontent.com/maxibor/sam2lca/master/docs/tutorial/data/
↪tutorial_db.fa.gz
gunzip tutorial_db.fa.gz
```

6.3 Indexing the database with Bowtie2

In this tutorial, we're going to work with the read aligner Bowtie2, but other aligners like BWA work also just fine.

Before being able to do any alignment, we need to index the Angiosperms353 database with Bowtie2:

```
bowtie2-build tutorial_db.fa angiosperms353
```

6.4 Preparing `fastq` sequencing files

> This step might be a bit long, especially if you have many references present in your database. You may want to speed it up by parallelizing it using the `--threads` option.

Prior to aligning the sequencing data, we need to download and process the sequencing data, e.g. to remove adapter sequences.

Downloading the paired-end DNA sequencing compressed `fastq` files

```
wget https://raw.githubusercontent.com/maxibor/sam2lca/master/docs/tutorial/data/
↪metagenome.1.fastq.gz
wget https://raw.githubusercontent.com/maxibor/sam2lca/master/docs/tutorial/data/
↪metagenome.2.fastq.gz
```

- Performing adapter-clipping and quality trimming with fastp

```
fastp -i metagenome.1.fastq.gz -I metagenome.2.fastq.gz -o metagenome_trimmed.R1.
↪fastq.gz -O metagenome_trimmed.R2.fastq.gz
```

6.5 Alignment with Bowtie2

After having prepared both the Angiosperms353 database and the FastQ sequencing files, we now align the sequencing data against the references using BowTie2.

> The important aspect here is to allow that multiple alignments can be reported for each read to ensure that all potential hits are reported. This is done by using the `-a` flag of Bowtie2 for reporting alignments, or `-k` for reporting up to *N* alignments.

Here, we will allow the reporting of up to 50 alignments per read.

```
bowtie2 -x angiosperms353 -k 50 -1 metagenome_trimmed.R1.fastq.gz -2 metagenome_
↪trimmed.R2.fastq.gz | samtools sort -O bam > metagenome.sorted.bam
samtools index metagenome.sorted.bam
```

6.6 Optional but (highly) recommended: bamAlignCleaner

Like many other tools working with `bam`/`cram` files, sam2lca relies on the file index to facilitate a rapid and parallel access to the aligned segments.

However, because we aligned our sequencing data against a database containing (most likely) a lot more reference sequences than what we can potentially align our reads to, the index of the file, based on the `bam`/`cram` header is huge and will slow down the I/O by many folds.

To circumvent this, we advise you to run bamAlignCleaner on your alignment files, and then reindexing them, before processing them further with sam2lca.

> For the sake of the tutorial, we use a smaller `fasta` reference sequence database. This step is therefore not really necessary in this tutorial.

```
bamAlignCleaner metagenome.sorted.bam | samtools sort > metagenome.cleaned.sorted.bam
samtools index metagenome.cleaned.sorted.bam
```

6.7 Running sam2lca

Once we have our alignment file, here in `bam` format, we can now run sam2lca to identify which plants shed some of its DNA in our sequencing file.

First, we need to set up the sam2lca acc2tax database for *plant markers*, with NCBI taxonomic identifiers. This step downloads and creates the taxonomy and accession to taxid conversion databases. These databases allow sam2lca to convert the accession ids of the reference genome sequences to their respective taxonomic ids and prepares this information to be accessible within sam2lca.

```
sam2lca update-db --taxonomy ncbi --acc2tax plant_markers
```

Let's check which sam2lca databases are now available:

```
sam2lca list-db
```

Finally, we run sam2lca with the *plant markers* database.

> To make sure that we don't accidentally run the LCA algorithm on DNA sequences that are unlikely to belong to the same clade, we will only run the LCA for all references aligned to each read that have a identity greater than 90%. Depending on the type of database, you might want to adjust the sequence identity threshold or try different ones.

```
sam2lca analyze --acc2tax plant_markers -b -i 0.9 metagenome.cleaned.sorted.bam
```

Let's look at the results that are summarized in the file `metagenome.cleaned.sorted.sam2lca.csv`

We see that the only species present in our data, is *Cannabis sativa*, which is indeed what was present in our sample ! (it was a simulated dataset)

```
+---------+--------------------+--------------+-------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| TAXID   | name               | rank         | count_taxon | count_descendant |␣
↪lineage
↪                                                                                    ␣
↪                                                                    |
```

(continues on next page)

(continued from previous page)

```
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 3398    | Magnoliopsida      | class        | 5          | 79               |␣
↪class: Magnoliopsida || clade: Embryophyta || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                 ␣
↪                                                                        |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 58024   | Spermatophyta      | clade        | 0          | 79               |␣
↪clade: Embryophyta || subphylum: Streptophytina || phylum: Streptophyta || kingdom:␣
↪Viridiplantae || superkingdom: Eukaryota                                          ␣
↪                                                                     |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 131567  | cellular organisms | no rank      | 0          | 79               |
↪                                                                                  ␣
↪                                                                                  ␣
↪                                                                  |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 2759    | Eukaryota          | superkingdom | 0          | 79               |␣
↪superkingdom: Eukaryota                                                           ␣
↪                                                                                  ␣
↪                                                                     |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 33090   | Viridiplantae      | kingdom      | 0          | 79               |␣
↪kingdom: Viridiplantae || superkingdom: Eukaryota                                 ␣
↪                                                                                  ␣
↪                                                                     |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 35493   | Streptophyta       | phylum       | 0          | 79               |␣
↪phylum: Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota         ␣
↪                                                                                  ␣
↪                                                                     |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 131221  | Streptophytina     | subphylum    | 0          | 79               |␣
↪subphylum: Streptophytina || phylum: Streptophyta || kingdom: Viridiplantae ||␣
↪superkingdom: Eukaryota                                                           ␣
↪                                                                           |
+---------+--------------------+--------------+------------+------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
```

(continues on next page)

(continued from previous page)

```
| 3193    | Embryophyta        | clade        | 0           | 79                |␣
↪clade: Embryophyta || subphylum: Streptophytina || phylum: Streptophyta || kingdom:␣
↪Viridiplantae || superkingdom: Eukaryota                                          ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 58023   | Tracheophyta       | clade        | 0           | 79                |␣
↪clade: Embryophyta || subphylum: Streptophytina || phylum: Streptophyta || kingdom:␣
↪Viridiplantae || superkingdom: Eukaryota                                          ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 78536   | Euphyllophyta      | clade        | 0           | 79                |␣
↪clade: Embryophyta || subphylum: Streptophytina || phylum: Streptophyta || kingdom:␣
↪Viridiplantae || superkingdom: Eukaryota                                          ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 1       | root               | no rank      | 0           | 79                |      ␣
↪                                                                                  ␣
↪                                                                                  ␣
↪                                                          |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 1437183 | Mesangiospermae    | clade        | 0           | 74                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                 ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 71240   | eudicotyledons     | clade        | 0           | 74                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                 ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 91827   | Gunneridae         | clade        | 0           | 74                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                 ␣
↪                                                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-----------------------------------------------------------------------------------
↪-----------------------------------------------------------------------------------
↪-------------------------------------------------------------+
| 1437201 | Pentapetalae       | clade        | 2           | 74                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                 ␣
↪                                                                  |
```

(continues on next page)

(continued from previous page)

```
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 71275   | rosids             | clade        | 1           | 72                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                  ␣
↪                                                                                   |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 91835   | fabids             | clade        | 3           | 71                |␣
↪clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina || phylum:␣
↪Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota                  ␣
↪                                                                                   |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 3744    | Rosales            | order        | 6           | 68                |␣
↪order: Rosales || clade: Embryophyta || class: Magnoliopsida || subphylum:␣
↪Streptophytina || phylum: Streptophyta || kingdom: Viridiplantae || superkingdom:␣
↪Eukaryota                                                                          |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 3481    | Cannabaceae        | family       | 21          | 62                |␣
↪family: Cannabaceae || order: Rosales || clade: Embryophyta || class: Magnoliopsida␣
↪|| subphylum: Streptophytina || phylum: Streptophyta || kingdom: Viridiplantae ||␣
↪superkingdom: Eukaryota                                                            |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 3482    | Cannabis           | genus        | 0           | 41                |␣
↪genus: Cannabis || family: Cannabaceae || order: Rosales || clade: Embryophyta ||␣
↪class: Magnoliopsida || subphylum: Streptophytina || phylum: Streptophyta ||␣
↪kingdom: Viridiplantae || superkingdom: Eukaryota                                  |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
| 3483    | Cannabis sativa    | species      | 41          | 41                |␣
↪species: Cannabis sativa || genus: Cannabis || family: Cannabaceae || order:␣
↪Rosales || clade: Embryophyta || class: Magnoliopsida || subphylum: Streptophytina␣
↪|| phylum: Streptophyta || kingdom: Viridiplantae || superkingdom: Eukaryota |
+---------+--------------------+--------------+-------------+-------------------+------
↪-------------------------------------------------------------------------------------
↪-------------------------------------------------------------------------------------
↪-----------------------------------------------------------------+
```

Note that out of all the reads in our sample, with the sam2lca parameters we used, 87.8% were classified (79 out of 90 trimmed reads). However, only 41 of them were assigned at the species level. This is due to the nature of these angiospersm353 markers: they are gene markers, hence relatively highly conserved among plants, which explains why the LCA is bringing many reads to a lower resolution taxonomic level.

Last but not least, we can also have a look at the `bam` file created by sam2lca: `metagenome.cleaned.sorted.sam2lca.bam`

```
$ samtools view metagenome.cleaned.sorted.sam2lca.bam | head -n 2
oneKP_189 329 5427_Ryparosa_wrayi 621 255 76M = 621 0␣
↪GGAGATTCAGAAGATGGCAAAATCAATTAAGGAACTGAAGAAGGAAAATTCATTCTTGAAGAGCAAGACTGAGAAA␣
↪D=BFCFCFHGIFCDGHBFH<@H?CFHIGHGIIGHDGEFGF@<=FHICHEIGHHHHFFFCECB>?;;ACDCCCDCDC AS:i:-
↪46 XM:i:9 XO:i:0XG:i:0 NM:i:9 MD:Z:0A2A3G28C1G10A11G5T0G7 YT:Z:UP XT:i:12908␣
↪XN:Z:unclassified sequences
oneKP_95 329 5921_Pappea_capensis 2336 255 76M = 2336 0␣
↪AGATTCAGTCTGATATTGACAAAAATAGTACCGACATCAACCGTCACAAGGTTCAAATAGAGACTGCCCAGAAAAT␣
↪DBDFHHFHIIIGIBHFHIIIIIIIGIIIIIIIIIIIIIIIIIAEHIIFGDHHHFFDBCECCCB@BCCCCCCCBCCBC AS:i:-
↪43 XM:i:8 XO:i:0XG:i:0 NM:i:8 MD:Z:28C2T2G14C14G1G0A2A5 YT:Z:UP XT:i:12908␣
↪XN:Z:unclassified sequences
oneKP_76 329 6270_Parartocarpus_venenosus 169 255 76M = 169 0␣
↪GACATAGACTATGGGAATGATGTGTTAACCTTGAAGCTTGGTGATTTAGGAACGTATGTGTTGAACAAACAAACTC ?D;
↪DBCBDAE:AEEAED?<C?<D?9?D?DDD?DADCC=.-;'>?77?).6;6>>A5-;;>>:;<5=1>;:2)(89> AS:i:-35␣
↪XM:i:9XO:i:0 XG:i:0 NM:i:9 MD:Z:26G2G0G10G5G5A5T0A1A13 YT:Z:UP XT:i:12908␣
↪XN:Z:unclassified sequences
oneKP_189 329 5427_Seidelia_firmula 609 255 76M = 609 0␣
↪GGAGATTCAGAAGATGGCAAAATCAATTAAGGAACTGAAGAAGGAAAATTCATTCTTGAAGAGCAAGACTGAGAAA␣
↪D=BFCFCFHGIFCDGHBFH<@H?CFHIGHGIIGHDGEFGF@<=FHICHEIGHHHHFFFCECB>?;;ACDCCCDCDC AS:i:-
↪46 XM:i:9 XO:i:0XG:i:0 NM:i:9 MD:Z:0A6G10G2G5C8C12A0T0G24 YT:Z:UP XT:i:12908␣
↪XN:Z:unclassified sequences
oneKP_95 329 5921_Placodiscus_turbinatus 980 255 76M = 980 0␣
↪AGATTCAGTCTGATATTGACAAAAATAGTACCGACATCAACCGTCACAAGGTTCAAATAGAGACTGCCCAGAAAAT␣
↪DBDFHHFHIIIGIBHFHIIIIIIIGIIIIIIIIIIIIIIIIIAEHIIFGDHHHFFDBCECCCB@BCCCCCCCBCCBC AS:i:-
↪43 XM:i:8XO:i:0 XG:i:0 NM:i:8 MD:Z:28C2T2G14A14A1G0A2A5 YT:Z:UP XT:i:12908␣
↪XN:Z:unclassified sequences
oneKP_144 73 6406_Cannabis_sativa 103 255 76M = 103 0␣
↪TCGAATACTGGACGGCTTTTCTTCGAACGAGCTATTGGAGCTTTTAGTATTGCCGAAATGCTGTTTACCTCTAGTC␣
↪FFFGHHHHJJJJJJJJJJJJJJHIJJIJJJJJIJJJJGIJJJJJJJJJJIIJJJJIIJIHFGGGH;DHCAEHFFDFB AS:i:0␣
↪XM:i:0 XO:i:0 XG:i:0NM:i:0 MD:Z:76 YT:Z:UP XT:i:3483 XN:Z:Cannabis sativa␣
↪XR:Z:species
```

Note the `XT`, `XN`, and `XR` tags at the end of each line. The value of these tags (the value after the `:`) is set by sam2lca to be respectively, the taxonomy IDs, the taxon name, and the taxon rank attributed to each sequencing read by the LCA algorithm.

To filter for these tags, we can make use of the filter expressions included in `samtools view`. For example, to obtain all reads assigned to the taxonomic level *class*, we can use `samtools view -e '[XR] == "class"' metagenome.cleaned.sorted.sam2lca.bam` or `samtools view --tag XR:class metagenome.cleaned.sorted.sam2lca.bam`.

Also note that for the first 5 reads (`oneKP_189`, `oneKP_95`, `oneKP_76`, `oneKP_189`), sam2lca classified them as "unclassified sequences" because they didn't pass the sequence identity threshold of 90%. Only the reads `oneKP_144` aligned to `6406_Cannabis_sativa` has an identity high enough to be classified by sam2lca.

CHAPTER

SEVEN

CONTRIBUTING

We welcome any contributions !

To further develop sam2lca, or add documentation, please read further:

7.1 Clone the sam2lca repository, and checkout the dev branch

```
git clone git@github.com:maxibor/sam2lca.git
git checkout dev
```

7.2 Install and activate the development environment

```
conda env create -f environment.yml
conda activate sam2lca
```

7.3 Install sam2lca with pip in editable mode

```
pip install -e .
```

7.4 Run the unit and integration tests

```
pytest -s -vv --script-launch-mode=subprocess
```

7.5 Build the documentation

```
cd docs
make html
```

The docs are built in the `docs/build/html` directory

7.5.1 Claim your sticker

Thanks for contributing to sam2lca ! If you want to spread the word about sam2lca, please get in touch with me to claim your sticker (maxime_borry[at]eva.mpg.de) !

CHAPTER

EIGHT

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

S

INDEX

Symbols

L

M

S

U