Instructions for Doing Multi-Locus Sequencing Typing Using MLST-Blast

Johan Nylander

February 13, 2015

These are brief instructions for using the scripts download-mlst-data.pl and mlst-blast.pl for assigning strains to ST-types given fasta-formatted input data.

For the impatient

1. Get scripts download-mlst-data.pl and mlst-blast.pl from https://github.com/nylander/MLST-Blast
2. Prepare a special fasta-formatted input file (e.g., query.fas) with DNA data from one or several strains
3.
   ```
   download-mlst-data.pl
   ```
4.
   ```
   mlst-blast.pl -q query.fas -c mlst-data/BLASTDB/blastdb.fas -d blastdb -p mlst-data/PROFILES/PROFILES.txt
   ```

1 Approach taken

ST-types are assigned based on the unique combination of allele types found for a specific strain. Here, we use blastn to compare query sequences against a curated database (pubmlst.org) with known alleletypes, and then compare the result to a list of known ST types.

The first script provided, download-mlst-data.pl, will download subsets, or the entire database from pubmlst.org, and save the data locally. The script will also provide a fasta formatted file, blastdb.fas, to be used as blast database, together with a profiles file, PROFILES.txt, to be used as a lookup table when doing the ST-typing.

Normally, a set of genes (around 6–7) are needed to assign a strain to a unique ST-type. Here we allow for "missing genes", and report the number of candidate ST-types a strain might belong to, even though some gene might have failed in sequencing.

2 Requirements and installation

2.1 Perl

For running the scripts, Perl needs to be installed, and preferrably with the documentation system (perldoc). Perl is preinstalled on Macintosh and Linux operating systems. For MS Windows, this is a good start:
http://www.activestate.com/activeperl.

2.2 Blast+

For running the similarity searches, blastn and makeblastdb needs to be installed in the PATH. Both these programs can be downloaded at NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST).

2.3 Perl-scripts download-mlst-data.pl and mlst-blast.pl

The Perl scripts download-mlst-data.pl and mlst-blast.pl are necessary. They can be placed in the same folder as your input files, or installed in the PATH. Note that if the scripts are not in the PATH, they need to be started by typing 'perl scriptname' or './scriptname'. If the scripts lack execute permissions, change the mode by using the command 'chmod +x scriptname' (change scriptname to download-mlst-data.pl or mlst-blast.pl).

2.4 Internet access to pubmlst.org

The approach taken is highly dependent on the presence of a correctly formatted data base located at pubmlst.org, and specifically specified in the file http://pubmlst.org/data/dbases.xml. No data can be downloaded unless this file is accessible through the internet.

3 Input

Three files are needed as input for conducting the MLST search:

1. A FASTA-formatted sequence file with DNA sequences. The file should contain sequences from the loci to be used for strain typing. The file can contain sequences from one or several strains. In order to identify fasta entries as coming from the same strain, a strict format for the fasta header needs to be used, for example: ">Strain_xxx". Note that the strain name is followed by an underscore, and then som arbitrary identifier. No white spaces are allowed in the strain identifier. See example below (p. 5).
2. A sequence data base (blastdb), with fasta headers following this format: Species acronym followed by double dash, followed by gene name, followed by underscore and allele type number (e.g., ">ecoli|adk_1"). No white spaces are allowed in the fasta header. See example below (p. 6).
3. A file containing ST-profiles (PROFILES.txt) according to a specific format. See example below (p. 8).

The data-base file, and the ST-profile file can be downloaded and prepared using the script download-mlst-data.pl.

4 Output

From download-mlst-data.pl

- mlst-data/ – Folder with species data. See more details in the file README.txt in the same folder.
- mlst-data/BLASTDB/blastdb.fas – File with concatenated and formatted fasta sequences from all the species in the mlst-data folder.
- mlst-data/PROFILES/PROFILES.txt – File with ST-profile definitions for all the species in the mlst-data folder.

From mlst-blast.pl

- <query file>.mlst-blast.out – Main result from the ST-typing. See example below (p. 9).
- <query file>.blast.out.tab – Report from the blast search.
- <query file>.blast.out.raw – Details from the blast search. Optional file.

5 Usage

First time usage

1. Prepare the query file with all fasta headers conforming to the strict input format: strain name followed by underscore, then some identifier (e.g., ">StrainA_nnn").

2. Download the whole, or parts of, the data base at pubmlst.org, using the script download-mlst-data.pl. This will create a new folder in the current working directory named "mlst-data". The folder will also contain a README.txt file with detailed information of the folder content. Importantly, the mlst-data folder also contains a fasta file (mlst-data/BLASTDB/blastdb.fas) ready formatted to be used as a blast data base, and a file (mlst-data/PROFILES/PROFILES.txt) with ST-profile definitions.

```
download-mlst-data.pl
```

3. Format the blast database.

```
mlst-blast.pl -c mlst-data/BLASTDB/blastdb.fas
```

4. Do a similarity search with blastn against the downloaded data base using your query file, and assign strains to ST-types based on the definitions you provide in a "profiles"-file.

```
mlst-blast.pl -q query.fas
```

5. Done! Look in the current working directory for newly created text files.

Note that steps 3. and 4. above can be done in one step using the general set of arguments:

```
mlst-blast.pl -q query.fas -c mlst-data/BLASTDB/blastdb.fas -d blastdb -p mlst-data/PROFILES/PROFILES.txt
```

Normal usage

If the blast database is created (and named blastdb), and placed in the working directory together with the query file and the PROFILES.txt, then the usage simplifies to:

```
mlst-blast.pl -q query.fas
```

The command above will also work if the blast database is called blastdb and placed in mlst-data/BLASTDB/, and the PROFILES.txt is placed in mlst-data/PROFILES/.

Another example, using query.fas on already created database mydb, and profiles myprofiles.txt (both mydb and myprofiles.txt in folder myfolder):

```
mlst-blast.pl -q query.fas -d myfolder/mydb -p myfolder/myprofiles.txt
```

Note: In the example above, all the blast database files mydb.nhr, mydb.nin, and mydb.nsq need to be in folder myfolder!

The script can also be used to create the blast database only (by calling makeblastdb, which needs to be installed on the computer). The following command will create the files that make up the blast database (.nhr, .nin, .nsq).

```
mlst-blast.pl -c mydb.fas
```

5.1 Downloading subsets of the database using download-mlst-data.pl

If only a subset is needed from the pubmlst.org database, a restricted download can be made by specifying the acronyms for the species wanted as an argument to the download script. In order to list all availiable data currently on pubmlst.org, one can invoke the script using the argument '-q'. This will create a REPORT.txt with a list of species, acronyms, and genes that are currently available for download.

```
download-mlst-data.pl -q
```

Lets assume that we are interested in only one species, *Aspergillus fumigatus*. The command below will download all data for this species, and will store the data in a folder A_fumigatus. Note that the acronym used at pubmlst.org for *A. fumigatus* is "afumigatus", which can be found in the file REPORT.txt generated by the command above.

```
download-mlst-data.pl -s afumigatus -d A_fumigatus
```

The following command will download the data for the two species belonging to the genus *Enterococcus*:

```
download-mlst-data.pl -s efaecalis,efaecium -d Enterococcus
```

5.2 Changing parameters for blastn using mlst-blast.pl

See output from

```
mlst-blast.pl --help
```

5.3 Caveats

The approach taken is highly dependent on the presence of a correctly formatted data base located at pubmlst.org, and specifically specified in the file http://pubmlst.org/data/dbases.xml. No data can be downloaded unless this file is accessible. Furthermore, the parsing of the blast results by the mlst-blast.pl script is highly dependent on the formatting of both the blastdb.fas, and PROFILES.txt files.

Example query fasta file (input for the scripts)

```
>Apa_adk
GTAATTTGCTTGGCGCTCCGGGCGCGGGGAAAGGGACTCAGGCTCAGTTCATCATGGAGAAATATGGTATTCCGCAAATC
TCCACTGGCGATATGCTGCGTGCTGCGGTCAAATCTGGCTCCGAGCTGGGTAAACAAGCAAAAGACATTATGGATGCTGG
CAAACTGGTCACCGACGAACTGGTGATCGCGCTGGTTAAAGAGCGCATTGCTCAGGAAGACTGCCGCAACGGTTTCCTGT
TGGACGGCTTCCCGCGTACCATTCCGCAGGCAGACGCAATGAAAGAAGCGGGCATCAATGTTGATTACGTTCTGGAATTC
GACGTACCGGACGAACTGATTGTTGACCGTATCGTAGGCCGCCGCGTTCACGCGCCGTCTGGTCGTGTTTATCACGTTAA
ATTCAATCCGCCGAAAGTAGAAGGCAAAGACGACGTTACCGGTGAAGAGCTGACTACCCGTAAAGACGATCAGGAAGAGA
CCGTACGTAAACGTCTGGTTGAATACCATCAGATGACTGCACCGCTGATCGGCTACTACTCCAAAGAAGCGGAAGCGGGT
ACACC
>Apa_fumC
CGGCAGATAAGCTGTGGGGCGCACAAAACTCAGCGCTCGCTGGAGCATTTCCGCATTTCGACGGAGAAAATGCCCACCTC
ACTGATTCATGCGCTGGCGCTAACCAAGCGCGCAGCGGCAAAAGTTAATGAAGATTTAGGCTTGTTGTCTGAAGAGAAAG
CGAGCGCCATTCGGCAGGCGGCGGATGAAGTACTGGCAGGACAGCATGACGACGAATTCCCGCTGGCTATCTGGCAGACC
GGCTCCGGCACGCAAAGTAACATGAATATGAACGAAGTGCTGGCTAACCGGGCCAGTGAATTACTCGGTGGCGTGCGCGG
GATGGAACGTAAAGTTCACCCTAACGACGACGTGAACAAAAGCCAAAGTTCCAACGATGTCTTTCCGACGGCGATGCACG
TTGCGGCACTACTGGCGCTGCGCAAGCAACTCATTCCACAACTTAAAACCCTGACCCAGACGCTGAGTGAAAAATCCCGC
GCATTTGCTGATATCGTCAAAATCGGTCGTACCCACTTGCAGGACGCCACGCCGTTAACGCTGGGGCAGGAGATTTCCGG
CTGGGTAGCGATGCTGGAGCATAATCTCAAACATATCGAATACAGCCTGCCTCACGTAGCGGAACTGGCTCTGGGCGGTA
CAGCGGTGGGTACTGGACTAAATACCCATCCGGAATATGCGCGTCGCGTAGCAGATGAACTGGCAGTCATTACCTGTGCA
CCGTTTGTTACCGCACCGAACAAATTTGAAGCGCTGGCGACCTGTG
>Apa_gyrB
GACGGCCCGGTCTGCACCACATGGTATTCGAGGTGGTAGATAACGCTATCGACGAAGCGCTCGCGAGGTCACTGTAAAGA
AATTATCGTCACCATTCACGCCGATAACTCTGTCTCTGTACAGGATGACGGGCGCGGCATTCCGACCGGTATTCACCCGG
AAGAGGGCGTATCGGCGGCGGAAGTGATCATGACCGTTCTGCACGCAGGCGGTAAATTTGACGATAACTCCTATAAAGTG
TCCGGCGGTCTGCACGGCGTTGGTGTTTCGGTAGTAAACGCCCTGTCGCAAAAACTGGAGCTGGTTATCCAGCGCGAGGG
TAAAATTCACCGTCAGATCTACGAACACGGTGTACCGCAGGCCCCGCTGGCGGTTACCGGCGAGACTGAAAAAACCGGCA
CCATGGTGCGTTTCTGGCCAAGCCTTGAAACCTTCACCAATGTGACCGAGTTCGAATATGAAATTCTGGCGAAACGTCTG
CGTGAGTTGTCGTTCCTCAACTCCGGCGTTTCCATTCGTCTGCGCGATAAGCGCGACGGCAAAGAAGACCACTTCCACTA
TGAAGGCGGCATCAAGGCGTTCGTTGAATATCTGAACAAGAACAAAACGCCGATCCACCCGAATATCTTCTACTTTTCCA
CCGAAAAAGACGGTATTGGCGTCGAAGTGGCGTTGCAGTGGAACGATGGCTTCCAGGAAAACATCTACTGCTTTACCAAC
AACATTCCGCAGCGTGACGGCGGTACTCACCTGGCAGGTTTCCG
>Apa_icd
GTGTAATGAAAGTAAAGTAGTTGTTCCGGCACAAGGCAAGAAGATCACCCTGCAAAACGGCAAACTCAACGTTCCTGAAA
ATCCGATTATCCCTTACATTGAAGGTGATGGAATCGGTGTAGATGTAACCCCAGCCATGCTGAAAGTGGTCGACGCTGCA
GTCGAGAAAGCCTATAAAGGCGAGCGTAAAATCTCCTGGATGGAAATTTACACCGGTGAAAAATCCACACAGGTTTATGG
TCAGGACGTCTGGTTGCCTGCTGAAACCCTTGATCTGATTCGTGAATATCGCGTTGCCATTAAAGGCCCGCTGACCACTC
CGGTTGGTGGCGGTATTCGCTCTCTGAACGTTGCCCTGCGCCAGGAACTGGATCTCTACATCTGCCTGCGTCCGGTACGT
TACTATCAGGGCACTCCAAGCCCGGTTAAACACCCTGAACTGACCGATATGGTTATCTTCCGTGAAAACTCGGAAGACAT
TTATGCGGGTATCGAATGGAAAGCAGACTCTGCCGACGCTGAGAAAGTGATTAAATTCCTGCGTGAAGAGATGGGCGTGA
AGAAAATTCGCTTCCCGGAACATTGCGGTATCGGTATTAAGCCGTGTTCGGAAGAAGGCACCAAACGTCTGGTTCGTGCA
GCGATCGAATACGCAATTGCTAACGATCGTGACTCTGTGACCCTGGTGCACAAAGGCAACATCATGAAGTTCACCGAAGG
CGCGTTTAAAGACTGGGGCTACCAACTGGCGCGTGAAGAATTTGGCGGTGAACTGATCGACGGCGGCCCGTGGCTGAAAG
TTAAAAACCCGAATACCGGCAAAGAGATCGTCATTAAAGACGTGATTGCTGATGCATTCCTGCAACAGATCCTT
>Apa_mdh
CTGCTGGCGGTATTGGCCAGGCGCTTGCACTACTGTTAAAAACCCAACTGCCTTCAGGTTCAGAACTCTCTCTGTATGAT
ATCGCTCCAGTGACTCCCGGTGTGGCTGTCGATCTGAGCCATATCCCTACTGCTGTGAAAATCAAAGGTTTTTCAGGTGA
AGATGCGACTCCGGCGCTGGAAGGTGCAGATGTCGTTCTTATCTCTGCAGGCGTAGCGCGTAAACCGGGTATGGATCGTT
CCGACCTGTTTAACGTTAACGCCGGTATCGTGAAAAACCTGGTACAGCAAGTTGCGAAAACCTGCCCGAAAGCGTGCATT
GGTATTATCACTAACCCGGTTAACACTACAGTTGCGATTGCTGCTGAAGTGCTGAAAAAAGCCGGTGTTTATGACAAAAA
CAAACTGTTCGGCGTTACCACGCTGGATATCATTCGTTCCAACACCTTTGTTGCGGAACTGAAAGGCAAACAGCCAGGCG
AAGTTGAAGTGCCGGTTATTGGTGGTCACTCTGGTGTTACCATTCTGCCGCTGCTGTCACAGGTTCCTGGCGTTAGTTTT
ACCGAGCAGGAAGTGGCTGATCTGACCAAACGTATCCAGAACGCGGGTACTGAGGTGGTTGAAGCGAAAGCCGGTGGCGG
GTCTGCAACCCTGTCTATGGGCCAGGCAGCTGCACGTTTTGGTCTGTCTCTGGTTCGCGCACTGCAGGGCGAACAAGGCG
TTGTCGAATGTGCCTACGTTGAAGGCGACGGTCAGTACGCTCGTTTCTTCTCTCAACCGCTGCTGCTGGGTAAAAACGGC
GTGGAAGAGCGTAAATCTATCGGTACTCTGAGCGCATTTGAACAGAACGCGCTGGAAGGTATGCTGGATACGCT
>Apa_purA
ACGGTGTTGTGCTGTCTCCGGCTGCGCTGATGAAAGAGATGAAAGAACTGGAAGACCGTGGCATCCCCGTTCGTGAGCGT
CTGCTGCTGTCCGAAGCATGTCCGCTGATCCTTGATTATCACGTTGCGCTGGATAACGCGCGTGAGAAAGCGCGTGGCGC
GAAAGCGATCGGCACCACCGGTCGTGGTATCGGGCCTGCTTATGAAGATAAAGTAGCACGTCGCGGTCTGCGTGTTGGCG
ACCTTTTCGACAAAGAAACCTTCGCTGAAAAACTGAAAGAAGTGATGGAATATCACAACTTCCAGTTGGTTAACTACTAC
AAAGCTGAAGCGGTTGATTACCAGAAAGTTCTGGATGATACGATGGCTGTTGCCGACATCCTGACTTCTATGGTTGTTGA
CGTTTCTGACCTGCTCGACCAGGCGCGTCAGCGTGGCGATTTCGTCATGTTTGAAGGTGCGCAGGGTACGCTGCTGGATA
TCGACCACGGTACATATCCGTACGTAACTTCTTCCAACACCACTGCTGGTGGCGTGGCGACCGGTTCCGGCCTGGGCCCG
CGTTATGTTGATTACGTTCTGGGTATCCTCAAAGCTTACTCCACTCGTGTGGGGGCAGGTCCGTTCCCGACCGAACTGTT
TGATGAAACTGGCGAGTTCCTCTGCAAGCAGGGTAACGAATTCGGCGCAACTACGGGTCGTCGTCGTCGTACCGGCTGGC
TGGACACCGTTGCCGTTCGTCGTGCGGTACAGTTGAACTCCCTGTCTGGCTTCTGCCTGACTAAACTGGACGTTCTGGAT
GGCCTGAAAGAGGTTAAACTCTGCGTGGCTTA
>Apa_recA
CAGCGTGAAGGTAAAACCTGTGCGTTTATCGATGCTGAACACGCGCTGGACCCAATCTACGCACGTAAACTGGGCGTCGA
TATCGACAACCTGCTGTGCTCCCAGCCGGACACTGGCGAGCAGGCACTGGAAATCTGTGACGCCCTGGCACGTTCTGGTG
CAGTAGACGTTATCGTCGTTGACTCCGTGGCGGCACTGACGCCGAAAGCGGAAATCGAAGGTGAAATCGGCGACTCTCAC
ATGGGCCTTGCGGCACGTATGATGAGCCAGGCGATGCGTAAGCTGGCGGGTAACCTGAAGCAGTCCAACACGCTGCTGAT
CTTCATCAACCAGATCCGTATGAAAATTGGTGTGATGTTCGGTAACCCGGAAACCACTACCGGTGGTAACGCGCTGAAAT
TCTACGCCTCTGTTCGTCTCGACATCCGTCGTATCGGCGCGGTGAAAGAGGGCGAAAACGTGGTGGGTAGCGAAACCCGC
GTGAAAGTGGTGAAGAACAAAATCGCTGCGCCGTTTAAACAGGCTGAATTCCAGATCCTCTACGGCGAAGGTATCAACTT
CTACGGCGAGCTGGTTGACCTGGGCGTGAAAGAGAAGCTTATCGAGAAAGCAGGCGCGTGGTACAG
```

Example blast database file (part.) (input for the scripts)

```
>ecoli|adk_1
GGGGAAAGGGACTCAGGCTCAGTTCATCATGGAGAAATATGGTATTCCGCAAATCTCCACTGGCGATATG
CTGCGTGCTGCGGTCAAATCTGGCTCCGAGCTGGGTAAACAAGCAAAAGACATTATGGATGCTGGCAAAC
TGGTCACCGACGAACTGGTGATCGCGCTGGTTAAAGAGCGCATTGCTCAGGAAGACTGCCGTAATGGTTT
CCTGTTGGACGGCTTCCCGCGTACCATTCCGCAGGCAGACGCGATGAAAGAAGCGGGCATCAATGTTGAT
TACGTTCTGGAATTCGACGTACCGGACGAACTGATTGTTGATCGTATCGTAGGCCGCCGCGTTCATGCGC
CGTCTGGTCGTGTTTATCACGTTAAATTCAATCCGCCGAAAGTAGAAGGCAAAGACGACGTTACCGGTGA
AGAACTGACTACCCGTAAAGACGATCAGGAAGAAACCGTACGTAAACGTCTGGTTGAATACCATCAGATG
ACTGCACCGCTGATCGGCTACTACTCCAAAGAAGCGGAAGCGGGTA
>ecoli|adk_2
GGGGAAAGGGACTCAGGCTCAGTTCATCATGGAGAAATATGGTATTCCGCAAATCTCCACTGGCGATATG
CTGCGTGCTGCGGTCAAATCTGGCTCCGAGCTGGGTAAACAAGCAAAAGACATTATGGATGCTGGCAAAC
TGGTTACCGACGAACTGGTGATCGCGCTGGTTAAAGGGCGCATTGCTCAGGAAGACTGCCGTAATGGTTT
CCTGTTGGACGGCTTCCCGCGTACCATTCCGCAGGCAGACGCGATGAAAGAAGCGGGCATCAATGTTGAT
TACGTTCTGGAATTCGACGTACCGGACGAACTGATCGTTGACCGTATCGTCGGTCGCCGCGTTCACGCGC
CGTCTGGTCGTGTTTATCACGTTAAATTCAATCCGCCGAAAGTAGAAGGTAAAGACGACGTTACCGGTGA
AGAACTGACTACCCGTAAAGACGATCAGGAAGAAACCGTACGTAAACGTCTGGTTGAATACCATCAGATG
ACAGCACCGCTGATCGGCTACTACTCCAAAGAAGCTGAAGCGGGTA
>ecoli|fumC_1
CGAGCGCCATTCGTCAGGCTGCGGATGAAGTGCTGGCGGGGCAGCATGATGATGAATTCCCGCTGGCAAT
CTGGCAGACCGGTTCTGGCACACAAAGTAACATGAACATGAACGAAGTGCTGGCCAATCGGGCGAGTGAA
CTGCTCGGCGGCGTGCGCGGGATGGAACGTAAAGTTCACCCTAACGATGACGTGAACAAAAGCCAAAGTT
CCAACGATGTCTTTCCGACGGCGATGCACGTTGCGGCGCTGCTGGCGCTGCGCAAGCAACTCATTCCGCA
GCTTAAAACCCTGACACAGACGCTGAGTGAAAAATCGCGTGCATTTGCCGATATCGTAAAAATCGGTCGA
ACCCACTTGCAGGACGCCACGCCGCTAACACTGGGGCAGGAGATTTCCGGCTGGGTAGCGATGCTCGAGC
ATAATCTCAAACATATCGAATACAGCCTGCCTCACGTAGCGGAACTGGC
>ecoli|fumC_2
CGAGCGCCATTCGTCAGGCGGCGGATGAAGTACTGGCAGGACAGCATAACGATGAATTCCCGCTGGCTAT
CTGGCAGACCGGCTCCGGCACGCAAAGTAATATGAACATGAACGAAGTGCTGGCTAACCGGGCCAGTGAA
TTACTTGGCGGCGTGCGCGGGATGGAGCGTAAAGTTCATCCTAACGACGACGTGAACAAAAGCCAAAGTT
CCAATGATGTCTTTCCGACGGCGATGCACGTTGCGGCGCTGCTGGCGCTGCGCAAGCAACTCATTCCGCA
GCTTAAATCCCTGACACAGACGCTGAGTGAAAAATCGCGCGCATTTGCCGATATCGTCAAAATCGGTCGA
ACCCACTTGCAGGATGCCACGCCGCTAACACTAGGGCAGGAGATTTCCGGCTGGGTAGCGATGCTGGAGC
ATAATCTCAAACATATCGAATACAGCCTGCCGCATGTAGCGGAACTGGC
>ecoli|gyrB_1
GGTCTGCACGGCGTTGGTGTTTCGGTAGTAAACGCCCTGTCGCAAAAACTGGAGCTGGTTATCCAGCGCG
AGGGTAAAATTCACCGTCAGATCTACGAACACGGTGTACCGCAGGCTCCGCTGGCGGTTACCGGCGAGAC
TGAAAAAACCGGCACCATGGTGCGTTTCTGGCCCAGCCTCGAAACCTTCACCAATGTGACCGAGTTCGAA
TATGAAATTCTGGCGAAACGTCTGCGTGAGTTGTCGTTCCTCAACTCCGGCGTTTCCATTCGTCTGCGCG
ACAAGCGTGACGGCAAAGAAGACCACTTCCACTATGAAGGCGGCATCAAGGCGTTCGTTGAATATCTGAA
CAAGAACAAAACGCCGATCCATCCGAATATCTTCTACTTCTCCACCGAAAAAGACGGTATTGGCGTCGAA
GTGGCGTTGCAGTGGAACGATGGCTTCCAGGAAAACATCT
>ecoli|gyrB_2
GGTCTGCACGGCGTTGGTGTTTCGGTAGTAAACGCCCTGTCGCAAAAACTGGAGCTGGTTATCCAGCGCG
AGGGTAAAATTCACCGTCAGATCTACGAACACGGTGTACCGCAGGCCCCGCTGGCGGTTACCGGCGAGAC
TGAAAAAACCGGCACCATGGTGCGTTTCTGGCCTAGCCTCGAAACCTTCACCAATGTGACCGAGTTCGAA
TATGAAATTCTGGCGAAACGTCTGCGTGAGTTGTCGTTCCTCAACTCCGGCGTTTCCATTCGTCTGCGCG
ACAAGCGTGACGGCAAAGAAGACCACTTCCACTATGAAGGCGGCATCAAAGCGTTCGTTGAATATCTGAA
CAAGAACAAAACGCCGATCCACCCGAATATCTTCTACTTCTCCACCGAAAAAGATGGTATCGGCGTTGAA
GTGGCGTTGCAGTGGAACGATGGCTTCCAGGAAAACATCT
>ecoli|icd_1
CGACGCTGCAGTCGAGAAAGCCTATAAAGGCGAGCGTAAAATCTCCTGGATGGAAATTTACACCGGTGAA
AAATCCACACAGGTTTATGGTCAGGATGTCTGGCTGCCTGCTGAAACCCTTGATCTGATTCGTGAATATC
GCGTTGCCATTAAAGGTCCGCTGACCACTCCTGTTGGTGGCGGTATTCGCTCTCTGAACGTTGCCCTGCG
CCAGGAACTGGATCTCTACATCTGCCTGCGTCCGGTACGTTACTATCAGGGCACTCCAAGCCCGGTTAAA
CACCCTGAACTGACCGATATGGTTATCTTCCGTGAAAACTCGGAAGACATTTATGCGGGTATCGAATGGA
AAGCAGACTCTGCCGACGCCGAGAAAGTGATTAAATTCCTGCGTGAAGAGATGGGGGTGAAGAAAATTCG
CTTCCCGGAACATTGTGGTATCGGTATTAAGCCGTGTTCGGAAGAAGGCACCAAACGTCTGGTTCGTGCA
GCGATCGAATACGCAATTGCTAACGATC
>ecoli|icd_2
CGACGCTGCAGTCGAGAAAGCCTATAAAGGCGAGCGTAAAATCTCCTGGATGGAAATTTACACCGGTGAA
AAATCCACACAGGTTTATGGTCAGGATGTCTGGCTGCCTGCTGAAACCCTTGATCTGATTCGTGAATATC
GCGTTGCCATTAAAGGTCCGCTGACCACTCCTGTTGGTGGCGGTATTCGTTCTCTGAACGTTGCCCTGCG
CCAGGAACTGGATCTCTACATCTGCCTGCGTCCGGTACGTTACTATCAGGGCGCCCCAAGCCCGGTTAAA
CACCCTGAACTGACCGATATGGTTATCTTCCGTGAAAACTCGGAAGACATTTATGCAGGTATCGAATGGA
AAGCTGATTCTGCCGACGCTGAGAAAGTGATTAAATTCCTGCGTGAAGAGATGGGGGTGAAGAAAATTCG
CTTCCCAGAACATTGTGGTATCGGTATTAAGCCGTGTTCGGAAGAAGGCACCAAACGTCTGGTTCGTGCA
GCGATCGAATACGCAATTGCTAACGATC
>ecoli|mdh_1
GGTGTAGCGCGTAAACCGGGTATGGATCGTTCCGACCTGTTTAACGTTAACGCCGGCATCGTGAAAAACC
TGGTACAGCAAGTTGCGAAAACCTGCCCGAAAGCGTGCATTGGTATTATCACTAACCCGGTTAACACCAC
AGTTGCAATTGCTGCTGAAGTGCTGAAAAAAGCCGGTGTTTATGACAAAAACAAACTGTTCGGCGTTACC
ACGCTGGATATCATTCGTTCCAACACCTTTGTTGCGGAACTGAAAGGCAAACAGCCAGGCGAAGTTGAAG
TGCCGGTTATTGGCGGTCACTCTGGTGTTACCATTCTGCCGCTGCTGTCACAGGTTCCTGGCGTTAGTTT
TACCGAGCAGGAAGTGGCTGATCTGACCAAACGTATCCAGAACGCGGGTACTGAGGTGGTTGAAGCGAAA
GCCGGTGGCGGGTCTGCAACCCTGTCTATGGG
>ecoli|mdh_2
GGTGTAGCGCGTAAACCGGGTATGGATCGTTCCGACCTGTTTAACGTTAACGCCGGCATCGTGAAAAACC
TGGTACAGCAAGTTGCGAAAACCTGCCCGAAAGCGTGCATTGGTATTATTACTAACCCGGTTAACACCAC
```

```
AGTTGCGATTGCTGCTGAAGTGCTGAAAAAAGCCGGTGTTTATGACAAAAACAAACTGTTCGGCGTTACC
ACGCTGGATATCATTCGTTCCAACACCTTTGTTGCGGAACTGAAAGGCAAACAGCCTGGCGAAGTTGAAG
TGCCGGTTATTGGCGGTCACTCTGGTGTTACCATTCTGCCGCTGCTGTCACAGGTTCTTGGCGTTAGTTT
TACCGAGCAGGAAGTGGCTGATCTGACCAAACGCATCCAGAACGCGGGTACTGAAGTGGTTGAAGCGAAA
GCCGGTGGCGGGTCTGCAACCCTGTCTATGGG
>ecoli|purA_1
ATAACGCGCGTGAGAAAGCGCGTGGCGCGAAAGCGATCGGCACCACCGGTCGTGGTATCGGGCCTGCTTA
TGAAGATAAAGTGGCACGTCGCGGTCTGCGTGTTGGCGACCTTTTCGACAAAGAAACCTTCGCTGAAAAA
CTGAAAGAAGTGATGGAATATCACAACTTCCAGTTGGTTAACTACTACAAAGCTGAAGCGGTTGATTACC
AGAAAGTTCTGGATGATACGATGGCTGTTGCCGACATCCTGACTTCTATGGTGGTTGACGTTTCTGACCT
GCTCGACCAGGCGCGTCAGCGTGGCGATTTCGTCATGTTTGAAGGTGCGCAGGGTACGCTGCTGGATATC
GACCACGGTACTTATCCGTACGTAACTTCTTCCAACACCACTGCTGGTGGCGTGGCGACCGGTTCCGGCC
TGGGCCCGCGTTATGTTGATTACGTTCTGGGTATCCTCAAAGCTTACTCCACTCGTGT
>ecoli|purA_2
ATAACGCGCGTGAGAAAGCGCGTGGCGCGAAAGCGATCGGCACCACCGGTCGTGGTATCGGGCCTGCTTA
TGAAGATAAAGTGGCACGTCGCGGTCTGCGTGTTGGCGACCTTTTCGACAAAGAAACCTTCGCTGAAAAA
CTGAAAGAAGTGATGGAATATCACAACTTCCAGTTGGTTAACTACTACAAAGCTGAAGCGGTTGATTACC
AGAAAGTTCTGGATGATACGATGGCTGTTGCCGACATCCTGACTTCTATGGTGGTTGACGTTTCTGACCT
GCTCGACCAGGCGCGTCAGCGTGGCGATTTCGTCATGTTCGAAGGTGCGCAGGGTACGCTGCTGGATATC
GACCACGGTACTTATCCGTACGTAACTTCTTCCAACACCACTGCTGGTGGCGTGGCGACCGGTTCCGGCC
TGGGCCCGCGTTATGTTGATTACGTTCTGGGTATCCTCAAAGCTTACTCCACTCGTGT
>ecoli|recA_1
CGCACGTAAACTGGGCGTCGATATCGACAACCTGCTGTGCTCCCAGCCGGACACCGGCGAGCAGGCACTG
GAAATCTGTGACGCCCTGGCGCGTTCTGGTGCAGTAGACGTTATCGTCGTTGACTCCGTGGCGGCACTGA
CGCCGAAAGCGGAAATCGAAGGCGAAATCGGCGACTCTCACATGGGCCTTGCGGCACGTATGATGAGCCA
GGCGATGCGTAAGCTGGCGGGTAACCTGAAGCAGTCCAACACGCTGCTGATCTTCATCAACCAGATCCGT
ATGAAAATTGGTGTGATGTTCGGTAACCCGGAAACCACTACCGGTGGTAACGCGCTGAAATTCTACGCCT
CTGTTCGTCTCGACATCCGTCGTATCGGCGCGGTGAAAGAGGGCGAAAACGTGGTGGGTAGCGAAACCCG
CGTGAAAGTGGTGAAGAACAAAATCGCTGCACCGTTTAAACAGGCTGAATTTCAGATCCTCTACGGCGAA
GGTATCAACTTCTACGGCGA
>ecoli|recA_2
CGCACGTAAACTGGGCGTCGATATCGACAACCTGCTGTGCTCCCAGCCGGACACCGGCGAGCAGGCACTG
GAAATCTGTGACGCCCTGGCGCGTTCTGGCGCAGTAGACGTTATCGTCGTTGACTCCGTGGCGGCACTGA
CGCCGAAAGCGGAAATCGAAGGCGAAATCGGCGACTCTCACATGGGCCTTGCGGCACGTATGATGAGCCA
GGCGATGCGTAAGCTGGCGGGTAACCTGAAGCAGTCCAACACGCTGCTGATCTTCATCAACCAGATCCGT
ATGAAAATTGGTGTGATGTTCGGTAACCCGGAAACCACTACCGGTGGTAACGCGCTGAAATTCTACGCCT
CTGTTCGTCTCGACATCCGTCGTATCGGCGCGGTGAAAGAGGGCGAAAACGTGGTGGGTAGCGAAACCCG
CGTGAAAGTGGTGAAGAACAAAATCGCTGCGCCGTTTAAACAGGCTGAATTCCAGATCCTCTACGGCGAA
GGTATCAACTTCTACGGCGA
```

...

Example PROFILES.txt file (part.) (input for the scripts)

```
ecoli|ST_1 ecoli|adk_4 ecoli|fumC_2 ecoli|gyrB_2 ecoli|icd_4 ecoli|mdh_4 ecoli|purA_4 ecoli|recA_4
ecoli|ST_2 ecoli|adk_5 ecoli|fumC_3 ecoli|gyrB_2 ecoli|icd_6 ecoli|mdh_5 ecoli|purA_5 ecoli|recA_4
ecoli|ST_3 ecoli|adk_6 ecoli|fumC_4 ecoli|gyrB_3 ecoli|icd_7 ecoli|mdh_7 ecoli|purA_7 ecoli|recA_6
ecoli|ST_4 ecoli|adk_6 ecoli|fumC_5 ecoli|gyrB_4 ecoli|icd_8 ecoli|mdh_8 ecoli|purA_8 ecoli|recA_2
ecoli|ST_5 ecoli|adk_7 ecoli|fumC_6 ecoli|gyrB_5 ecoli|icd_9 ecoli|mdh_9 ecoli|purA_8 ecoli|recA_2
ecoli|ST_6 ecoli|adk_8 ecoli|fumC_7 ecoli|gyrB_1 ecoli|icd_1 ecoli|mdh_10 ecoli|purA_8 ecoli|recA_6
ecoli|ST_7 ecoli|adk_9 ecoli|fumC_8 ecoli|gyrB_5 ecoli|icd_1 ecoli|mdh_11 ecoli|purA_8 ecoli|recA_7
ecoli|ST_8 ecoli|adk_10 ecoli|fumC_9 ecoli|gyrB_5 ecoli|icd_10 ecoli|mdh_12 ecoli|purA_9 ecoli|recA_2
ecoli|ST_9 ecoli|adk_6 ecoli|fumC_4 ecoli|gyrB_3 ecoli|icd_7 ecoli|mdh_7 ecoli|purA_7 ecoli|recA_8
ecoli|ST_10 ecoli|adk_10 ecoli|fumC_11 ecoli|gyrB_4 ecoli|icd_8 ecoli|mdh_8 ecoli|purA_8 ecoli|recA_2
ecoli|ST_11 ecoli|adk_12 ecoli|fumC_12 ecoli|gyrB_8 ecoli|icd_12 ecoli|mdh_15 ecoli|purA_2 ecoli|recA_2
ecoli|ST_12 ecoli|adk_13 ecoli|fumC_13 ecoli|gyrB_9 ecoli|icd_13 ecoli|mdh_16 ecoli|purA_10 ecoli|recA_9
ecoli|ST_13 ecoli|adk_6 ecoli|fumC_6 ecoli|gyrB_5 ecoli|icd_9 ecoli|mdh_9 ecoli|purA_8 ecoli|recA_2
ecoli|ST_14 ecoli|adk_14 ecoli|fumC_14 ecoli|gyrB_10 ecoli|icd_14 ecoli|mdh_17 ecoli|purA_7 ecoli|recA_10
ecoli|ST_15 ecoli|adk_15 ecoli|fumC_15 ecoli|gyrB_11 ecoli|icd_15 ecoli|mdh_18 ecoli|purA_11 ecoli|recA_11

...
```

Example <query file>.mlst-blast.out file (output from the mlst-blast.pl script)

```
# Query,Species,ST-type,adk,fumC,gyrB,icd,mdh,purA,recA,
Apa,ecoli,131,53,40,47,13,36,28,29,
```