System Requirements:

1. Linux operating systems* such as, Ubuntu**, openSUSE, or Red Hat.
2. Perl environment. Typically, all Linux OS have an integrated Perl environment.

*If you do not have access to a Linux operating system, you can emulate it within other systems, e.g. Windows, using Virtual Box. Please see Step 0.

**This software was built and tested with Ubuntu-20 and SLES-15.

Installation

(Optional) Step 0: Installation Guide for AnnotIEM on non-Linux Operating Systems:

This step is for the installation of AnnotIEM for non-Linux Operating Systems through emulation of this system. For this task we recommend utilizing Virtual Box, available for download at https://www.virtualbox.org/wiki/Downloads. Follow their instructions for download and installation.

After installation of the Virtual Box, a Perl environment, must be installed within the Virtual Box.

All other installations and scripts, must be then done within the Virtual Box that has the Perl environment.

Step 1: Installation of AnnotIEM

Once all system requirements are met.

1. Download the AnnotIEM folder from Github.

No further installation outside the download of the folder is needed. The databases used for base AnnotIEM are already provided in the downloadable folder. This includes by default the following databases:

NCBI 16S rRNA Refseq

RDP (11.4)

SILVA (138_1)

GTDB (27.0)

Step 2: Installation of the Basic Local Alignment Search Tool (BLAST)

If QIIME2 was installed within the Virtual Box, this step can be disregarded.

1. Download and unzip the NCBI rpm file.

   The latest version for installation can be found at https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.

   For example, the file should be named like “ncbi-blast-2.16.0+-1.x86_64.rpm”.

2. Install BLAST by running the following commands:

```
> sudo apt install rpm
> cd <Directory of Your System>,      e.g. home/AnnoIEM/ >
> rpm -Uvh ncbi-blast-2.16.0+-1.x86_64.rpm
> export PATH=$PATH:<Path of BLAST in Your System>/ncbi-blast-2.16.0+-1.x86_64/bin
```

More detailed installation instructions are provided at https://www.ncbi.nlm.nih.gov/books/NBK52640.

Step 3: Run AnnotIEM

Please see “Input File Parameters” section for the correct formatting of your query sequence data.

1. Import your query sequence data (in FASTA format) into the downloaded AnnotIEM folder (AnnotIEM).
2. Open the Linux Terminal
3. Set the directory of your system to the path of the AnnotIEM folder by using the following command

```
> cd <Path to the Folder in Your System>/AnnotIEM
```

4. Run the following command to start your AnnotIEM run

```
> perl AnnotIEM-master-V2.pl <Name of the Sequence File>.fasta
```

Additional Customization for Advanced Users:

AnnotIEM is optimized to run with the four databases. It is possible to run AnnotIEM with only one, multiple, or all databases provided. However, it is advised to use three to four databases to achieve the best results.

Four preformatted databases are provided with the base installation of AnnotIEM. In the current version, the user may download the four databases (RDP, SILVA, NCBI, and GTDB) for a single time on each computer. However, it is possible to download the most updated version of the databases from the mentioned links.

Step 1: Download of the Databases

Please use the following links to download the four databases.

1. SILVA
   Link: https://www.arb-silva.de/no_cache/download/archive

   At this link are multiple releases, e.g. [release_138_2].
   Click on the folder with the latest release, and then go within the “[Exports]” folder and download the library labeled as:

   SILVA_<version number>_SSUParc_tax_silva.fasta.gz.

2. RDP***

Link: https://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/sequences/by-database/rdp.fasta

After clicking the link, save the site as a FASTA file.
***There are many versions of the RDP database with different settings, the version provided at the link is the latest version without any clustering or filtering.

3. NCBI 16S rRNA
   Link: https://ftp.ncbi.nlm.nih.gov/blast/db/

   Download the file “16S_ribosomal_RNA.tar.gz”.
   This file is preformatted, for usage in AnnotIEM please run the following code:

```
> tar -zxvf 16S_ribosomal_RNA.tar.gz
```

4. GTDB
   Link: https://data.ace.uq.edu.au/public/gtdb/data/releases/

   At this link are multiple releases, e.g. release207.
   Click on the folder with the latest release, then then click the next folder with the same release number. Afterwards, go within the “genomic_files_all/” folder, and download the library labeled as:

   ssu_all_r<version number>.tar.gz.

Step 2: Format the databases

After unzipping the files, the RDP, SILVA, and GTDB databases are in the FASTA file format. They must be formatted for usage with BLAST using the following command:

```
> makeblastdb -in <Name of the Downloaded Database> -dbtype nucl -out
  <Name of the Formatted Database>
```

Once the formatting is completed make sure that all of the formatted databases are kept in “/AnnotIEM/Databases” folder.

Step 3: Edit the Code

To ensure AnnotIEM recognizes your new databases, open AnnotIEM-master-V2.pl file, and edit the database names in lines 39-42 to match those that were formatted in Step 2. The line should look as such:

```
blastn -db /DATABASES/formatted-name
```

Input File Requirements:

1. The input sequence file must be in FASTA format and must not contain any hyphens in the name.

2. The AnnotIEM code runs with strict parameters. To be considered as a hit, the sequence identity must be ≥ 95% and Query Coverage ≥ 95%.
3. It is important that the16S sequences in the input file are trimmed of primers and adapters.

Output Files:

After the AnnotIEM has finished running, the result files can be found in the folder "RESULT-FILE-<Name of the Sequence File>-MonthDay-Year-hhmmss". In the table below are all the output files of AnnotIEM generated with the description of their content. All the files generated are tab separated text files.

The final output file is <Name of the Sequence File>-Annotation-Final-Result.

File	**Description**
<Name of the Sequence File>-Annotation-Final-Result	Contains the recommended annotation and mentions the rank of the recommended taxa. If the annotation is not satisfactory or not found it is marked as "Problematic"
<Name of the Sequence File>-Annotation-with-Parameters	Contains a detailed annotation at both the species and genus level with all associated parameters
<Name of the Sequence File>-LOGFILE	Contains the log for each sequence
<Name of the Sequence File>-Parsed-Output-reformated-ncbi	For each sequence, provides the top 10 hits from the NCBI database
<Name of the Sequence File>-Parsed-Output-reformated-silva	For each sequence, provides the top 10 hits from the SILVA database
<Name of the Sequence File>-Parsed-Output-reformated-RDP	For each sequence, provides the top 10 hits from the RDP database
<Name of the Sequence File>-Parsed-Output-reformated-GTDB	For each sequence, provides the top 10 hits from the GTDB database

Simulated Demo data:

A simulated demo is also included within the code. This is a small part of the data used in this manuscript. The sequence file is named "COPSACV4_1.fasta", and since the databases used for this file cannot be re-distributed, the output of this run is used for demo.

The following code was run using the NCBI, SILVA, RDP and EzTaxon databases:

```
> perl AnnotIEM-master-V2.pl COPSACV4_1.fasta
```

Commented [JA1]: Should this number match the numbers above

The main output file is “CopsacV4_1-Annotation-Selected-Taxonomy-Marked”. All other interim files are also included in the folder. This input demo file contains 5000 sequences, and approximately 14 hours were required for the run.

Commented [JA2]: Shouldn’t this file name match what we put as output files above?