

FungiSearch manual

(for taxonomic assignment of fungi)
Version 2, December 2024

Table des matières

1. Install programs (on Windows).....	2
2. Unzip the FungiSearch folder on your computer	3
3. Reference database	4
3.1. Reference database formatting	4
4. Update the EPPO list of fungi	6
5. Input the fastq.gz files R1 and R2 for the samples to analyze in the directory Data	6
6. Determine the normalization level to be applied to samples	7
6.1. Tables of OTUs or ASVs	7
6.2. Rarefaction curves	7
7. Taxonomy assignment.....	8
7.1. Parameters for the analysis.....	8
7.2. List of samples to analyze	8
7.3. Clustering method and assignment to fungal species	9
8. Results.....	9
9. Licence	10
10. References	10

1. Install programs (on Windows)

A Cygwin environment must be created. To this end, install Cygwin on your computer (<https://cygwin.com/install.html>). When you will be asked to select packages, click **Next** as the standard installation of Cygwin is sufficient for using FungiSearch pipeline. To use the pipeline properly, it is required to install different programs:

- o **perl for Windows** (<http://strawberryperl.com/>)

Note: To create a new database (optional), the Bio::SeqIO package must be installed¹. Its installation can take 1-2 hours.

- o **R** (<https://www.r-project.org/>) – R Core Team (2023)

Note: To create rarefaction curves (section 6), the package “vegan” must be installed in R.

- o **R Studio** (<https://rstudio.com/products/rstudio/download/>) - R Core Team (2023)

Note: the free version is sufficient for using FungiSearch pipeline

- o **USEARCH** (<https://www.drive5.com/usearch/>) - Edgar (2010)

Note: The 32-bit version (free of charge) is sufficient for using FungiSearch pipeline. The USEARCH file is an executable file. It must be renamed as usearch and added to one directory of your computer.

- o **Blast** (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST>) – Altschul et al. (1990)

In the list of files, select the one for Windows (for instance ncbi-blast-2.7.1+-win64.exe).

- o **Gzip for Windows** (<http://gnuwin32.sourceforge.net/packages/gzip.htm>).

Download complete package, except sources.

Cygwin's "bin", perl's "bin" as well as R, usearch, Gzip and blast+ must be placed in the environmental variable of your computer (Consult the tutorials available on internet to know how to add to Windows PATH environment variable).

To easily modify the different scripts, Notepad++ should be installed on your computer.

To verify that the various programs are accessible from Cygwin, use the following commands:

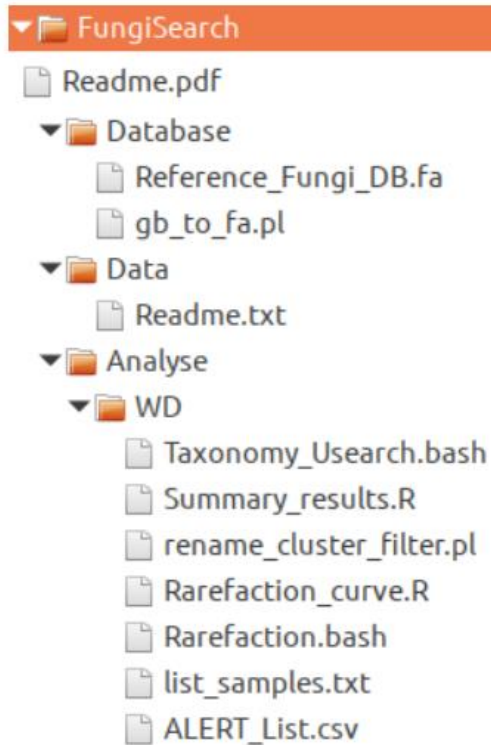
```
$ perl -v
$ rscript --version
$ usearch
$ blastn -version
$ gunzip --version
```

Note: For this study, the HTS analysis has been done on Windows 11, with USEARCH v.11, Blast -2.7.1, Strawberry Perl 5.24.2 and R version 4.3.1.

¹Command prompt (cmd.exe): \$ cpan install Bio::SeqIO (provided as figshare repository)

2. Unzip the FungiSearch folder on your computer

This folder is organised as illustrated here after:



It contains:

- Three specific directories
 - o Database
 - o Data
 - o Analysis (with its subdirectory WD)
- Five scripts needed for the HTS analysis
 - o Rarefaction.bash
 - o Rarefaction_curve.R
 - o Taxonomy_Usearch.bash
 - o rename_cluster_filter.pl
 - o Summary_results.R
- One script needed to create a reference database
 - o gb_to_fa.pl
- Two default databases
 - o Reference_Fungi_DB.fa: the reference database used for the assignment of fungi (built on 9 October 2022)*
 - o ALERT_List.csv: an EPPO list with the fungi on A1, A2 and alert lists download from the EPPO website on 30 December 2024*
- A list of samples
 - o list_samples.txt

*these databases can be updated (see sections 3 and 4)

Notes:

1) All commands must be written in the Cygwin environment (see Cygwin documentation to navigate under bash in directories). In the example below, the FungiSearch folder is placed on the disk C in the folder “Documents” of the computer owner:

```
$cd /cygdrive/c  
$cd Users/Owner/Documents/FungiSearch
```

- To access to the Database directory in the FungiSearch folder

```
$ cd Database
```

- To access to the Data directory from the Database directory

```
$ cd ../Data
```

- To access to the WD directory from the Data directory

```
$ cd ../Analyse/WD
```

2) The WD directory containing all the results of a HTS analysis will be erased when a new HTS analysis will be performed. It is therefore recommended to keep results of HTS analysis in another directory.

3. Reference database

The FungiSearch pipeline is provided with a reference nucleotide database built on 9 October 2022 from NCBI.

3.1. Reference database formatting

The reference database provided in the FungiSearch folder (Database directory) is transformed to be usable by the pipeline. To this end, access the Database directory of the FungiSearch folder using Cygwin command (see previous section to know how to navigate in Cygwin) and use the following command line.

```
$ makeblastdb -in Reference_Fungi_DB.fa -parse_seqids -dbtype nucl
```

Reminder: the use of the program gb_to_fa.pl requires the installation of the seq::Bio package in R.

3.2. Updated reference database

To work on an updated version of the reference database, there are two options described in paragraphs 3.2.1 and 3.2.2.

3.2.1. New reference database

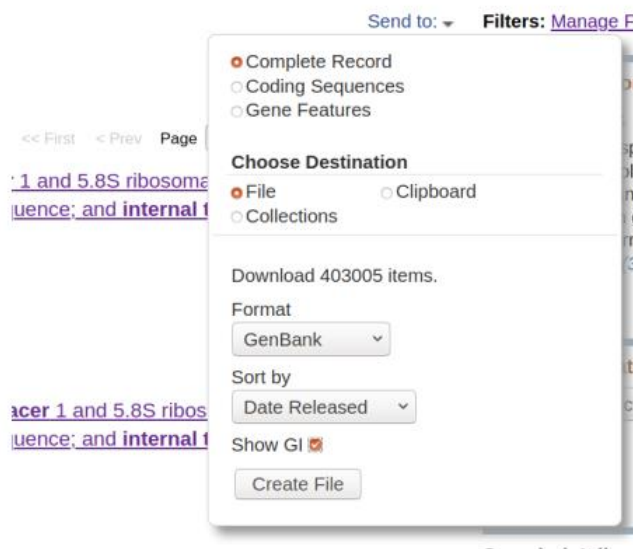
This option is the best option as it enables to work on a completely revised database (including the new names of fungi in case of taxonomical changes). However, the downloading process from NCBI can take several hours.

- Import the sequence.gb database from NCBI (<https://www.ncbi.nlm.nih.gov/nucleotide/>) in the Database directory using the ad hoc NCBI filters. In this study, we used the following filters:

```
(((((internal[All Fields] AND transcribed[All Fields] AND spacer[All Fields]) NOT uncultured[All Fields]) NOT sp.[All Fields]) NOT aff.[All Fields]) NOT cf.[All Fields]) NOT ("unidentified"[Organism] OR unidentified[All Fields])) NOT unverified[All Fields]) NOT "fungal endophyte"[Organism]) NOT "422523"[BioProject] AND (fungi[filter] AND is_nuccore[filter] AND ("300"[SLEN] : "10000"[SLEN]))
```

This list of filters must be copied and pasted in the Search box at the address www.ncbi.nlm.nih.gov/nucleotide/

- To send the created db to your computer, click on “Send to” button on the NCBI website, then select the correct parameters as indicated in the figure below (parameters=Complete record, File, GenBank format, sort by Date Release, Show GI), and finally click on “Create File”. The download process is long (several hours).



- Move the created reference database (name = sequence.gb) in the Database directory and make it usable by the FungiSearch pipeline using the following command lines

```
$ perl gb_to_fa.pl sequence.gb Reference_Fungi_DB.fa  
$ makeblastdb -in Reference_Fungi_DB.fa -parse_seqids -dbtype nucl
```

Reminder: the use of the program gb_to_fa.pl requires the installation of the seq::Bio package in R.

3.2.2. Addition of the missing period to the existing database

The missing part to the reference database (i.e. starting one day after the date of creation of the existing reference database to the date of analysis of the samples) can be uploaded from NCBI and concatenated to the existing reference database. This can be done by importing the sequence.gb database as detailed in the previous paragraph after indicating the missing period (for instance from 2022 10 10 to add the accessions introduced from 10 October 2022 to the reference database created on 9 October 2022) in the specific filter “Release date” of NCBI (see figure below).

Custom date range

2022 10 10 to YYYY MM DD

Apply Clear

The additional database (name=sequence.gb) is moved to the Database directory, transformed in a fasta file and then concatenated to the existing database (name=actual_DB.fa) using the following command lines.

```
$ perl gb_to_fa.pl sequence.gb additional_DB.fa
```

```
$ makeblast -in additional_DB.fa actual_DB.fa -out Reference_Fungi_DB.fa -parse_seqids -dbtype nucl
```

4. Update the EPPO list of fungi

The name of fungi (genus and species) from A1, A2 and alert lists of EPPO can be retrieved on the EPPO website (https://www.eppo.int/ACTIVITIES/quarantine_activities). The organisms are listed in a csv file named ALERT_List.csv. The file must be placed in the WD directory.

5. Input the fastq.gz files R1 and R2 for the samples to analyze in the directory Data

The FungiSearch pipeline is dedicated to the analysis of Illumina MiSeq paired-end reads. The naming convention for FASTQ files is in the format (for sample S1 in the example below):

```
Data\SampleName_S1_L001_R1_001.fastq.gz
```

```
Data\SampleName_S1_L001_R2_001.fastq.gz
```

6. Determine the normalization level to be applied to samples

6.1. Tables of OTUs or ASVs

The Rarefaction.bash (available in the WD directory) must be edited (preferably using Notepad++) prior to the analysis.

6.1.1. Samples to analyze

The same type of samples should be analyzed (i.e. do not mix mock communities and environmental samples in the same analysis). There are two options depending on the number of samples to analyze. The default option analyses all the samples from a HTS run. Another option can be selected when a limited number of samples are to be analyzed (e.g. 3 samples from an HTS run that do not necessarily follow each other). This second option is proposed as a comment in the program (i.e. # at the beginning of the command at line 19).

To activate this option, just remove the # at the beginning of the corresponding command line, add manually the name of the samples to be analyzed in the command line (e.g. S49, S50, S73) and put a # at the beginning of the command line of the default option (line 16).

6.1.2. Clustering method

Two options are proposed for the clustering of reads. The default option is UNOISE (creating ZOTUs - corresponding to ASVs). Another option using UPARSE (creating OTUs) can be selected. This second option is proposed as a comment in the program (i.e. # at the beginning of the command lines). To activate this option, just remove the # at the beginning of the corresponding command lines (lines 64 and 72) and put a # at the beginning of the command lines of the default option (lines 61 and 69).

6.1.3. Table

To launch the script, execute the following command from the WD directory in Cygwin

```
$ bash Rarefaction.bash
```

A table of ZOTUs (or OTUs) will be available in the Rarefaction folder (name=zotutab_raw.txt or otutab_raw.txt depending on the clustering method selected) created by the script in the WD directory (see Results section).

6.2. Rarefaction curves

In RStudio, open the script Rarefaction_curve.R (WD directory) and run it on the zotutab_raw.txt file (or otutab_raw.txt) created in the Rarefaction folder after adapting the path to the file (in line 6 of the R script). In the example below, the FungiSearch folder is placed on disk C in the folder "Documents" of the computer's owner.

```
>setwd ("C:/Users/Owner/Documents/FungiSearch/Analyse/WD/Rarefaction")
```

Notes:

- for the first use of the R program, it is needed to open the file `Rarefaction_curve.R` with an editing program (preferably Notepad++) and to activate the line 2 (installation of package `vegan`). After the installation of the package on your computer, it is important to deactivate line 2 (by adding a # at the beginning of the line) to avoid installing `vegan` package at each use of the program.
- the different components of the path are separated by / (not \).

A rarefaction curve (`Rarefaction_curve.pdf`) will be available in the `Rarefaction` folder.

7. Taxonomy assignment

The `Taxonomy_Usearch` program (available in the `WD` directory) must be adapted to the analysis to carry out. To this end, it is edited (preferably using Notepad++).

7.1. Parameters for the analysis

These parameters can be modified at the beginning of the program:

- Parameters for the creation of ZOTUs/OTUs
 - Number of bases removed at the 5' and 3' ends of the merged fastq file (default = 17)*
 - Normalization level (default = 10000)**
 - Cut off of reads (default = 0.05)***
- Parameters for Blast analysis
 - Percent identity (default = 99.5)
 - Number of hits (maximum number of aligned sequences to keep) (default = 1000)
 - Query-coverage (default = 85)

* number of bases removed to be adapted to the PCR primers used to build the HTS library

**normalized sequencing depth to be adapted according to the result of the analysis carried out in section 6

*** Cut off (expressed in % of reads) to be adapted to the proportion of false positive and false negative in experimental controls of HTS analysis (mock communities of fungi).

7.2. List of samples to analyze

There are two options depending on the number of samples to analyze. The default option analyses all the samples from a HTS run. Another option can be selected when a limited number of samples are to be analyzed (e.g. 3 samples from an HTS run that do not necessarily follow each

other). This second option is proposed as a comment in the program (i.e. # at the beginning of the command at line 30).

To activate this option, just remove the # at the beginning of the corresponding command line, add manually the name of the samples to be analyzed in the command line (e.g. S49, S50, S73) and put a # at the beginning of the command line of the default option (line 27).

7.3. Clustering method and assignment to fungal species

Two options are proposed for the clustering of reads. The default option is UNOISE (creating ZOTUs corresponding to ASVs). Another option using UPARSE (creating OTUs) can be selected. This second option is proposed as a comment in the program (i.e. # at the beginning of the command lines). To activate this option, just remove the # at the beginning of the corresponding command lines (lines 74, 76, 83, 91, 101 and 102) and put a # at the beginning of the command lines of the default option (lines 68, 70, 82, 90, 97 and 98).

To launch the script, execute the following command from the WD directory in Cygwin:

```
$ bash Taxonomy_Usearch.bash
```

Different files will be created in the WD directory (see Results section)

8. Results

(available in the WD directory)

- **Rarefaction:** this folder contains the tables of ZOTUs (or OTUs) and rarefaction curves
- **Taxonomy_USEARCH_date_hour.log:** this file contains all the “history” of the HTS run (total number of sequences, sequences that merged, number of singletons, number of filtered reads...for each sample).
- **Files:** this folder contains all the intermediate files created for each sample.
- **sample.fa:** this file (one for each sample) contains the sequences of the reads used for blast analysis (i.e. after application of a cut off of reads)
- **blast_results:** this folder contains the blast results for each sample (notably the corresponding accession number in the reference database, the size of the query sequence, the size of the subject sequence, and the E-value)
- **csv:** this folder contains one file summarizing in one table the read abundance (RA) of species identified for the different samples (=run_summary.csv) and one file for each sample summarizing in a table the list of species identified with their respective RA and HP values (=sample.csv). To display the results in percentages in the run_summary.csv file, it is needed to modify the format to %.

9. Licence

FungiSearch is free software: you can redistribute it and/or modify it under the terms of the CC-BY License as published by figshare.

FungiSearch is distributed without any warranty, without even the implied warranty of merchantability or fitness for a particular purpose. Bug reports and other feedbacks are welcome.

10. References

- FungiSearch development:

The pipeline has been developed in the framework of a European BIODIVERSA project (RESIPATH) at the Walloon Agricultural Research Centre (Belgium). The validation of the pipeline is described in:

Chandelier A, Hulin J, San Martin G, Debode F and Massart S. (2021). Comparison between real-time PCR and metabarcoding methods as tools for the detection of airborne inoculum of forest fungal pathogens. *Phytopathology* 111: 570-581.

An updated version of the pipeline has been used in:

Chandelier A, Schmitz S, Dubois B and San Martin G. Fungal pathogens detected in conifer seeds by high-throughput sequencing. *Phytopathology*. Manuscript submitted.

- Programs used:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460-2461.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>