# How to create a custom reference database for BLAST in SEED2

Step 1 – dowload reference sequences as FASTA or create FASTA file from your edited sequences and put desired information into the titles of the sequences, e.g.:

<Species name>|<ACCESSION#>|<full taxonomy>

## Here is example of FASTA:

>Thelephora sp|UDB014120|k Fungi;p Basidiomycota;c Agaricomy cetes;o\_\_Thelephorales;f\_\_Thelephoraceae;g\_\_Thelephora;s\_Thelephor a sp

### GGAAGGATCATTACT...

>Dothiorella parva|KC898234|k Fungi;p Ascomycota;c Dothideomy cetes;o Botryosphaeriales;f Botryosphaeriaceae;g Dothiorella;s D othiorella parva - - X SEED v2.1.07a

Open FASTA file(s)

Open FASTQ file(s)

Add FASTQ file

File Paired-and reads MCBI Files | Settings | Console | Help

Parameter

Max-length

Min-length

Count

Value

0

AAGGATCATTACCGA

# Step 2 – open the file in SEED2







document & where you can read about the process that we used to generate these references

 Full length sequences and taxonomy references (188247 bacteria, 4626 archea, and 20246 eukarya sequences). This reference could be customized for alignments, but could also be used for classification. The uncompressed version is ~9.9 GB and the compressed version is 348 MB.

### https://unite.ut.ee/repository.php

#### QIIME release (download)

Main page

Help

Random page

Three sets of QIIME files are released, corresponding to the SHs resulting from clustering at the 97% and 99% threshold levels. The third set of files is the result of a dynamic use of clustering thresholds, such that some SHs are delimited at the 97% level, some at the 97.5% level, some at the 98% level, and so on; these choices were made manually by experts of those particular lineages of fungi. The syntax is the same throughout the three sets of files.

Each SH is given a stable name of the accession number type, here shown in the FASTA file of the dynamic set

#### >SH099456.05FU FJ357315 refs

CACAATATGAAGGCGGGCTGGCACTCCTTGAGAGGACCGGC...

SH099456 = accession number of the SH 05FU = global key release 5, organism group FUngi FJ357315 = GenBank/UNITE accession number of sequence chosen to refs = this is a manually designated RefS (reps = this is an automatically chosen RepS)

## Step 3 – save FASTA as local database

9

#### x Save as local database FASTA FILE(S): UNITE\_8.0\_ALL\_eukaryotes\_sh\_general\_release\_dynamic\_s\_all\_02.02.2019\_noUnidentifiedFunc Database folder: F:\LAB\FM DATABASE TEST\ Change folder File Titles Sequences Quality display sort Alignments Chimera-clean Extraction Cluster Database name: MY\_REFERENCE length hits tag: M1 2 3 4 6 7 8 9 101112131415161718192021 Save as FASTA 40861083 0 Save as FASTQ 4086667 0 408(635 0 1408624 0 Create database from sequences Save titles as TEXT 08FI620 0 140(619 0 66 0:618 Save as local database 1086.617 Thelephora\_splKP012968|SH1140868.08F617 0 10 Thelephoraceae spIMF405772|SH114086615 0 11 Thelephoraceae\_splMF405769|SH114087615 0 12 Thelephoraceae\_splUDB004461|SH11408614 0 Tomentella sp|UDB013066|SH1140872.0613 13 0 14 Thelephoraceae\_splUDB004558|SH1140(612 0 0% 15 Thelephoraceae\_splUDB008192|SH1140(612 16 Thelephora sp|UDB014120|SH1140875.0612 17 Tomentella\_sp|KT275603|SH1140877.08F611 0 18 Thelephoraceae\_splFR852149|SH114087610 0 ...build the database 19 Thelephoraceae\_splUDB012524|SH1140(610 20 Thelephoraceae\_splKP857986|SH114088610 0 21 Thelephoraceae\_splFN557576|SH114088609 23 22 Thelephoraceae\_splHQ204741ISH114086609 Save as local database 0 23 Thelephoraceae\_splUDB013016|SH1140(609 0 24 Thelephoraceae\_sp|UDB004893|SH1140(608 0 Database folder: F:\LAB\FM DATABASE TEST\ Change folder 25 Thelephoraceae splUDB004396(SH1140(608 0 26 Thelephoraceae\_spUF960840|SH114088;608 0 Database name: MY\_REFERENCE 27 Tomentella sp|UDB013200|SH1140888.0:607 0 28 Thelephoraceae splKC618522ISH114088607 0 29 Tomentella\_sp|KT275624|SH1140890.08F607 0 30 Thelephoraceae splHQ289866|SH114085606 0 x Seed Create database from sequences Visible region starts at 1 bp and ends at 45 bp Database MYREFERENCE has been created! File C:\seed data\working\db fata.fas has been saved -in C:\seed\_data\working\db\_fata.fas -dbtype nucl -ou Creating database: MYREFERENCE... OK This step will generate 3 binary 0%

files in database "folder":

| MYREFERENCE.nhr | 5.2.2020 11:36 | Soubor NHR | 26 329 kB |
|-----------------|----------------|------------|-----------|
| MYREFERENCE.nin | 5.2.2020 11:36 | Soubor NIN | 1 381 kB  |
| MYREFERENCE.nsq | 5.2.2020 11:36 | Soubor NSQ | 17 509 kB |

## **1** name your reference database

# Step 4 – use the reference database for BLAST identification

| FASTA FILE(S): ITS_OTUs [SeqCount: 5191 Ambiguous: 0 Min-len: 40 Max-len: 395 Max-qu     File   Titles   Sequences   Quality   display   sort   Alignments   Chimera-clean     No.   HEAD   length   hits tag: M1 2 3 4 5 6 7 8 9 1011121314     1   Otu1   170 0 1   IC C A A C C C T C A G G C     2   Otu2   191 0 1   T A T T C T C A A T T C T C     3   Otu3   192 0 1   A A T T C T C A A C C T A T     4   Otu4   153 0 1   C A A C C A T C A A G C C C     5   Otu5   227 0 1   A A T T C T C A A C C T G     6   Otu6   210 0 1   A T T C T C A A C C T G | al: 0 Min-qual: 13]<br>Extraction Clustering Identification Phylogeny<br>15161716152C2122222<br>NCBI BLAST<br>C T T C T B G G C<br>T A G G C T T G T<br>T A G G C T T G T<br>T A A C T G G T T T A A C A G C T G G T T T T G G G T A T T G<br>C A C T T G C T A T C A T T A G C A A G B T T C B T T G G<br>C A C C T T G C T A T C A T T A G C A A G B T T C B T T G G<br>C A C C T T G C T A T C A T T A G C A A G B T T C B T T G G<br>C A C C T T G C T A T C A T T A G C A A G B T T C B T T G G   |
|---|--|
| 7 0tu7   8 0tu8   9 0tu9   10 0tu10   11 0tu11   12 0tu12   13 0tu13   14 0tu14   15 0tu15   16 0tu16   17 0tu17   18 0tu18   19 0tu19   20 0tu20   21 0tu21   22 0tu22   23 0tu23   24 0tu24   25 0tu25   26 0tu26   27 0tu27   28 0tu28   29 0tu29   30 0tu30   | ks were closed READY FOR: blastn<br>TUS Sel Result Accession Description<br>BLAST settings<br>DATABASE SOURCE Select local database file(s)<br>remote (NCBI) number of threads select all<br>local 1 deselect al |
| Visible region starts at 1 bp and ends at 4   | F:\LAB\FM_DATABASE_TEST\   Specification   Image: Choose a BLAST program to run:   blastn   Parameters   Type: megablast   Tasks at the same time: 10   Results per sequence: 10   E-value threshold: 0.00001   Restart FAILED and UNDONE RUN (removes previous results)   |