OTT manual

OneTwoTree Virtual-Machine Installation

OneTwoTree (OTT) offers an automatic procedure for alignment assembly and phylogeny reconstruction based on all sequence data available in NCBI GenBank for a given list of taxa names. The standalone version involves the installation of a large number of software tools, Python/Perl packages, and MySQL server. This section is intended for users that are interested to install and run OTT locally on a Linux-based computer. Otherwise, users are advised to use the on-line version of OTT, available at http://onetwotree.tau.ac.il.

The local installation is possible in two optional procedures:

1. VirtualBox, as described here.
2. Local installation of the pipeline and its dependencies. To follow this option please proceed to Appendix A.

For questions please contact: evolseq@post.tau.ac.il

VirtualBox installation

Follow these steps for installing and running OTT on VirtualBox:

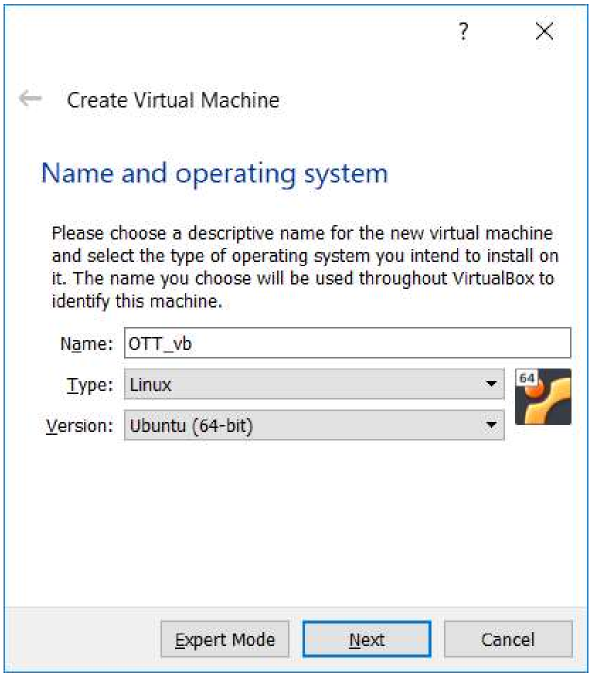
1. Download the file: OTT_VB.tar.gz from the OneTwoTree website: http://onetwotree.tau.ac.il/download/
2. Before installation:
   a. Your CPU must have 64bit capability and support either Intel or AMD virtualization technologies: VT-x or AMD-v.
   b. Go into your BIOS and enable VT-x/AMD-v.
   c. Disable Hyper-V platform in your Windows Feature list.
3. Install VirtualBox 5.2.6 on your local machine. Make sure to use 64bit host and have at least 50G available on your hard disk.

   Download from: https://download.virtualbox.org/virtualbox/

   Follow the instructions below for creating a new virtual machine:

   a. Click *New* in the VirtualBox manager window:

      In the “Create Virtual Machine” window, specify the name of the virtual machine (e.g., OTT_vb) and select Type: Linux and Version: Ubunto (64bit).



Create Virtual Machine

Name and operating system

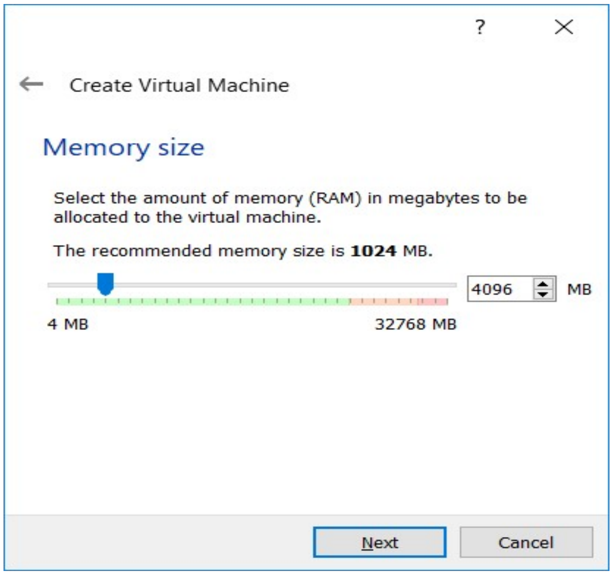
Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name: OTT_vb

Type: Linux

Version: Ubuntu (64-bit)

b. The next step will be to select the amount of RAM to be allocated in your machine. This depends on the available RAM of your host (choose 2048MB and above):



Create Virtual Machine

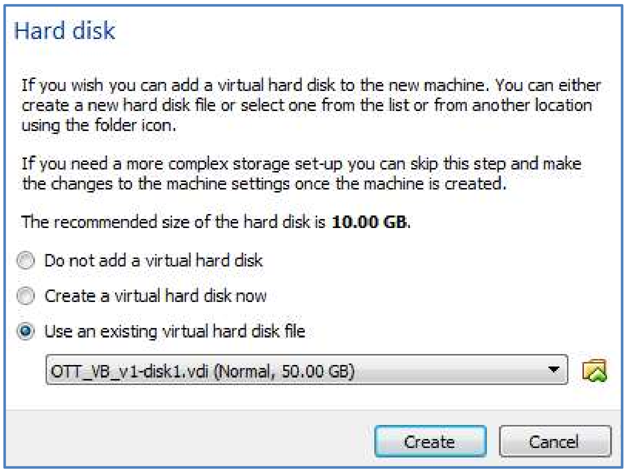
Memory size

Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.

The recommended memory size is **1024** MB.

4096 MB

c. Choose the option 'use an existing virtual hard disk' and select the downloaded VDI file ‘OTT_VB_v1-disk1.vdi’:



Hard disk

If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard disk is **10.00 GB**.

○ Do not add a virtual hard disk

○ Create a virtual hard disk now

● Use an existing virtual hard disk file

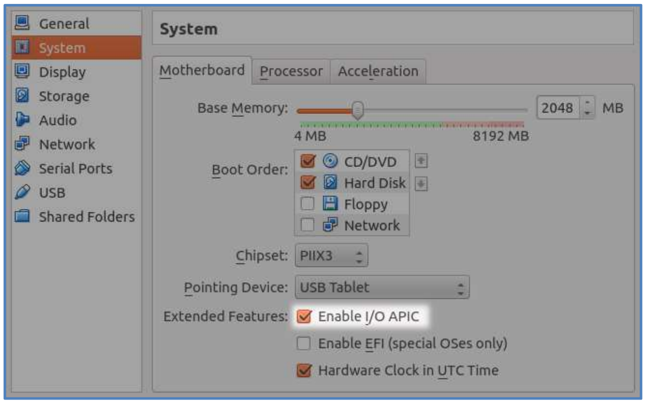
OTT_VB_v1-disk1.vdi (Normal, 50.00 GB)

Note: This .vdi file includes all installations that are needed for a default run. Several additional options are also already available: MrBayes, JmodelTest, JAVA, TRimal, automatic outgroup selection, treePL, and OpenMPI.

d. Enable more than 1 core to reduce running time (using 1 cpu the program will run but in some cases it will be very slow):

   i. To enable more than 1 CPU in your Virtual-Machine, modify the following Virtual-Machine settings:

Enable I/O APIC in the Motherboard settings tab



Base Memory: 2048 MB

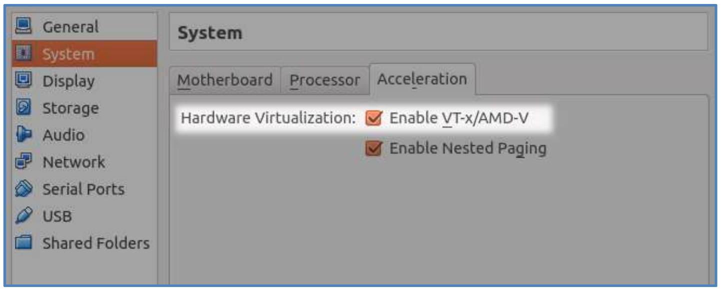
Boot Order: ☑ CD/DVD ☑ Hard Disk □ Floppy □ Network

Chipset: PIIX3

Pointing Device: USB Tablet

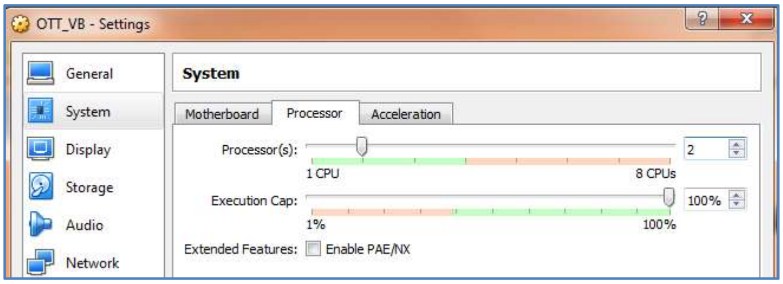
Extended Features: ☑ Enable I/O APIC
□ Enable EFI (special OSes only)
☑ Hardware Clock in UTC Time

Enable hardware virtualization (VT-x/AMD-V) in the Acceleration tab



Hardware Virtualization: ☑ Enable VT-x/AMD-V
☑ Enable Nested Paging

ii. Chose the number of CPUs (at least 2) for your Virtual-Machine. Go to System settings and set the number under the Processor tab:



Processor(s): 2

Execution Cap: 100%

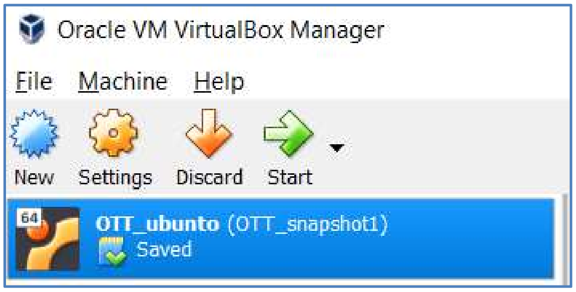
Extended Features: ☐ Enable PAE/NX

Note: If you run the Virtual-Machine on all available cores you may experience a better host performance when assigning an execution cap to the CPU.

e. You are ready to start the machine:

   i. Chose the machine you have just created.

   ii. Press the start arrow:



f. **Allow some time for initialization.** Once the linux window is open (see below) and you can start a new terminal by click on the terminal icon.

g. **User and password for this machine:**
   Username: ottadmin
   Password: OTTadmin (this is the root password in case you need to install software or if the Machine goes idle).

4. Once your machine is running you can run OTT:

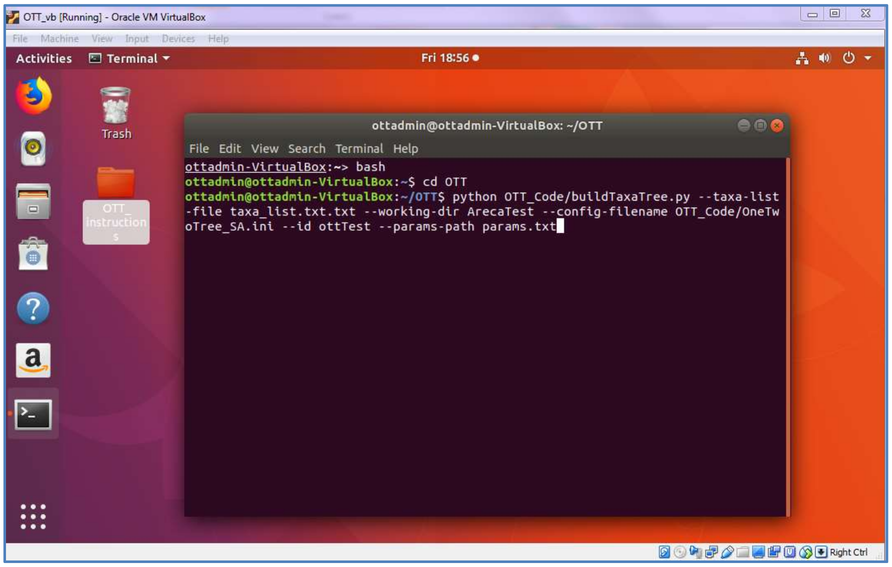
   a. Open a terminal and type 'bash'.

   b. Type ‘alias python=python3’ to make sure you run with the correct python version.

   c. cd (change directory) to /home/ottadmin/OTT/ (the working directory).

   d. Edit the file /home/ottadmin/OTT/taxa_list.txt to your requested taxa list and run the OTT command as in the example below:

Python OTT_Code/buildTaxaTree.py --taxa-list-file taxa_list.txt --working-dir OUTPUT/YOUR_DIR --config-filename OTT_Code/OneTwoTree_SA.ini --id ottRun --params-path params.txt


OTT_vb [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Fri 18:56
Trash
OTT_Instructions
ottadmin@ottadmin-VirtualBox: ~/OTT
File Edit View Search Terminal Help
Right Ctrl

```
ottadmin-VirtualBox:~> bash
ottadmin@ottadmin-VirtualBox:~$ cd OTT
ottadmin@ottadmin-VirtualBox:~/OTT$ python OTT_Code/buildTaxaTree.py --taxa-list
-file taxa_list.txt.txt --working-dir ArecaTest --config-filename OTT_Code/OneTw
oTree_SA.ini --id ottTest --params-path params.txt
```

e. An explanation of all arguments for the main script file can be found below.

5. This version is set for running OTT with default parameters and includes sequence databases for plants only. Several additional options are also already installed, which will enable running: MrBayes, JmodelTest, JAVA, TRimal, automatic outgroup selection, treePL, and OpenMPI.
   Further options may be enabled by installing additional software (see Appendix A for instructions). See Appendix C for instructions on how to replace the sequence database.

Arguments list: (can be modified both when using Virtual-Machine and the Standalone version)

Below is a description of each parameter:

- `--taxa-list-file` *<taxa_list.txt>*

  This is a plain text file that should contain a list of taxa names or TaxIds that should be included in your run (TaxIds are as specified in NCBI taxonomy). Each name should be separated by a new line. Parts of a name (e.g., a binomial species name) should be separated by a space or an underscore. See example list (also available at <OTT_working_dir>/taxa_list.txt):

```
Areca catechu
Areca concinna
Areca hutchinsoniana
264298
Areca_macrocalyx
Areca rheophytica
Areca triandra
Areca tunku
Narcissus
Areca vestiaria
```

- `working-dir` *`<Output_DirName>`*
  A path to a directory where all outputs will be created. The main result files will be found under <Output_DirName>/**SummaryDir:**

  Phylogeny file: Result_Tree_1519065090.tre

  MSA file: 1519065090-concat-aligned.fasta

- `--config-filename OneTwoTree_SA.ini`
  This is the path with all software paths. In case you install the optional software please update their paths accordingly (see Appendix A section 2 - Update OneTwoTree_SA.ini file).

- `--id` *`<any name, without spaces>`*
  A name that will be added to the results files for identification.

- `--params-file-path` *`<your paramsFile.txt>`*
  This is a plain text file that specifies the various run options. An example that specifies a run with all default options can be found at <OTT_working_dir>/params.txt. In case you wish to use different parameters:
  a) Make sure to install all needed software for the options you choose.
  b) Update the params.txt file according to the options specified in the ***ParamsOptions.xlsx*** file given with the OTT package (OTT_sa_v1.tar.gz) downloaded.

Appendix A - OneTwoTree Stand Alone Version 1.0

This section is intended for users that are interested to install and run OTT and all its dependencies locally on a Linux-based computer without a virtual machine.

1. Environment requirements:

Before running OTT, the following tool/packages must be installed. These are listed as required installations, which enable the default run of OTT, and optional installations for users that wish to change the default settings. The version listed under each installation is the one that was used to develop OTT.

Required Installations:

1. **Python** version 3.6.2 and above (type "python --v" to verify). The required modules are listed in Appendix B. Download path: https://anaconda.org/anaconda/python

   Python packages:

   To verify the correct package was installed type python in your bash terminal, import the package and print the version file, for example:

   ```
   >>> import Bio
   >>> print(Bio.__version__)
   ```

   1.1. Bio-python 1.69 and above.

   Download from: http://biopython.org/wiki/Download

   1.2. Numpy 1.3 and above.

   Download from: https://pypi.python.org/pypi/numpy

   1.3. Ete3 3.1.1

   Download from: https://pypi.python.org/pypi/ete3/

   1.4. Sql-connector 2.2.2b1

   Download from: https://pypi.python.org/pypi/mysql-connector-python-rf/2.2.2

   1.5. Pandas 0.20.3

   Download from: https://pandas.pydata.org/

2. **Perl** perl-5.16.3 (type "perl –v" to verify)

   Download from: http://www.perl.org/

   Additional Perl packages (obtain latest version through cpan or Sudo apt-get install**):**

   2.1. Bioperl

   2.2. Config

3. **MAFFT** mafft7149 (type “mafft” to verify).

   Download path: http://mafft.cbrc.jp/alignment/software/

4. **BLAST** ncbi-blast-2.2.25 and above (type “blastall” to verify)

   Download from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download

5. **OrthoMCL** orthomclSoftware-v2.0.3 and above (in VB we use v2.0.9). OrthoMCL requires the use of a **MySql server**. The definition of your MySql IP, username and password should be set in the ini file. For further instructions see OneTwoTree_SA.ini file (see section 2).

   The installation instructions for Orthomcl can be found at:

   https://github.com/apetkau/orthomcl-pipeline/blob/master/INSTALL.md

   - As explained in the link above Orthomcl dependencies include:
     - Blast (see section 1 item 4, above).
     - mcl - OTT use version mcl-14-137.

       Download from: https://www.micans.org/mcl/index.html

   Please note, the installation of OrthoMCL and the MySQL server may be complicated.
   To verify that it functions properly, verify that you can connect to the MySql server via python. The following code can be used for verification:

   ```
   $python
   Python 3.6.3
   [GCC 7.2.0] on linux
   Type "help", "copyright", "credits" or "license" for more information.
   >>> import mysql.connector
   >>> mysql.connector.connect(host='localhost',user='root',password='ottMysql')
   <mysql.connector.connection.MySQLConnection object at 0x7f893a344c18>
   >>>
   ```

6. **Raxml:** RAxML version 8.2.4, (type “raxmlHPC -version” to verify)

   Download from: https://github.com/stamatak/standard-RAxML

7. **R** version 3.0.1 and above [type “R”, to verify (possibly with a path to the bin directory)]

   Download from: https://cran.r-project.org/src/base/

   7.1. Make sure ‘ape’ package is installed:

   On root:

   ```
   >>R
   >install.packages(“ape”)
   ```

8. **CD-HIT** version 4.7 (type cd-hit to verify)

   Download from: http://weizhongli-lab.org/cd-hit/download.php

9. **Taxonome** version 1.5 (python package)

   Download path**:** https://pypi.python.org/pypi/Taxonome/1.5

Optional:

The following software are required only if you choose to change the default OTT parameters.

By default, you do not need to install them.

1. Clustering option:

1.1. BlastClust 2.2.25 (type “blastclust -“ to verify)

Download path: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download

2. Phylogeny inference software:

2.1 MrBayes v3.2.2 x64 (verify by typing “mb”) Download from: http://mrbayes.sourceforge.net/download.php

To enable model selection for MrBayes we require also:

2.1.1 JmodelTest version: jmodeltest-2.1.7

Download from: https://github.com/ddarriba/jmodeltest2/releases/tag/jModelTest-2.1.7-20141120

2.1.2 Java version used: java/java-1.7

Downloaded from: https://java.com/en/download/help/linux_install.xml

2.2 ExaML version 3.0.17 m (verify by typing "mpirun -np 1 examl --version").

Download from: https://github.com/stamatak/ExaML

2.2.1 Examl requires MPI. One option is OpenMPI version 1.6 (verify by typing "mpirun --version"). Download from: https://www.open-mpi.org/software/ompi/v1.6/

3. **Filter MSA software options:**

3.1 trimAl 1.2rev59 (verify by typing "trimal –version")

Download from: http://trimal.cgenomics.org/downloads

3.2 Guidance version: Guidance_v201_gcc620/guidance.v2.02

Download from: http://guidance.tau.ac.il/ver2/

Verify that the following file exists:

…/guidance.v2.02/www/Guidance/guidance.pl

3.3 GBLOCKS 0.91b (verify by typing "which Gblocks")

Download from http://molevol.cmima.csic.es/castresana/Gblocks.html

4. **Divergence time estimation and calibration:**

4.1 TreePL (*) version: 1.0 (verify by typing "TreePL)

Download from: https://github.com/blackrim/treePL

4.2 PLL-DPPDIV (*): version dppdiv-mpi-sse3

(type "which dppdiv-mpi-sse3" to verify)

Download from: https://github.com/ddarriba/pll-dppdiv

4.3 TreeAnnotator (*) version BEASTv1.8.0 (this tool is part of the BEAST package)

Download from: http://beast.community/install_on_unix

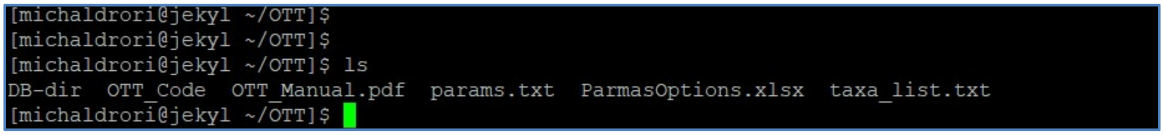
(*) These tools also use MPI. One option is OpenMPI version 1.6 (verify by typing "mpirun --version"). Download from: https://www.open-mpi.org/software/ompi/v1.6/

2. Download OneTwoTree

1. Download OTT_sa_v1.tar.gz to your working directory (from now on we will refer to this directory as <OTT_working_dir>) and extract the file:

   ```
   tar -xvzf OTT_sa_v1.tar.gz
   ```

   The following files and directories should be listed when typing the ls command:

   ```
   [michaldrori@jekyl ~/OTT]$
   [michaldrori@jekyl ~/OTT]$
   [michaldrori@jekyl ~/OTT]$ ls
   DB-dir  OTT_Code  OTT_Manual.pdf  params.txt  ParmasOptions.xlsx  taxa_list.txt
   [michaldrori@jekyl ~/OTT]$
   ```

2. If you intend on performing a name resolution process, you also need to download the naming databases ('The plant list' and 'Catalogue of life') from the website and place them under the <OTT_working_dir>/*DB-dir*/ directory, which is part of the extracted files:

   http://onetwotree.tau.ac.il/download.html/

3. **Update OneTwoTree_SA.ini file.** *OneTwoTree_SA.ini* file (located under <OTT_working_dir>/*OTT_Code/)* includes all configurations needed to run OTT-SAV. The following modifications should be made to this file before running OTT-SAV.

   a) Replace all "<Insert_Your_path>" with your local paths. Example:

   OTT_MAIN = <<Insert_Your_path>> /OTT_Code/

   Was changed to:

   OTT_MAIN = /home/ottadmin/OTT/OTT_Code/
   (/home/ottadmin/OTT is my working directory for OneTwoTree).

   b) The parameter GENBANK_GRP_LIST specifies the organism groups included in the sequence database. This parameter should be updated according to the selected groups (see section "Database creation" below). For example, the following should be specified if intending to use only sequences from plants and vertebrates:

      a. GENBANK_GRP_LIST = pln,vrt

   c) Update the following attributes of the MySQL server (this is the SQL server installed for OrthoMCL) at the beginning of the ini file (lines 3-5):

```
hostname = localhost # should provide an IP address in
                 case MySQL server is stored in a
                 remote host
username = <your username>
password = <your password>
```

d) Make sure that the stand alone flag is set to *on* (line #12):

```
SA_VERSION = on
```

4. Update the file *OneTwoTree_SA.ini* with your local paths for orthomcl and mcl directories under **[orthomcl],** marked in blue (the rest bolded paths should stay the same). Examples:

   mcl = <Insert_Your_mcl_path>/**mcl-14-137/src/shmcl/mcl**
   ortho_perl_path = <Insert_Your_orthomclSoftware_path>/**orthomclSoftware-v2.0.3/lib/perl**

3 Generation of a local sequence database

OTT is based on NCBI genbank sequences and thus it is necessary to create a local sequence database. General information on genbank can be found at https://www.ncbi.nlm.nih.gov/genbank/.

This step usually takes a while but needs to be performed only once.

To save disk space on your station we advise to download only the taxa group of interest:

mam (mamals), rod (rodents), pri (primates)
pln (plants), vrt (vertebrates), inv (invertebrates)

* OTT webserver (http://onetwotree.tau.ac.il/) includes by default all the groups mentioned above.

There are 2 options to create the database: download and extract NCBI files through the NCBI ftp site and then run the provided OTT script for creating the database. Alternatively, it is possible to use the provided script also for downloading the files from NCBI ftp. If you wish to download the files yourself (or in any situation where option 2 fails) please refer to Option 1 below, otherwise continue to Option 2:

* If you wish to add taxonomical groups or update your database you need to erase all directories and files under <OTT_working_dir>/DB-dir before you begin.

1. **Option 1.** Download NCBI files and then run GenBank_OTT.py to create OTT database:

Download and extract the following files from ftp://ftp.ncbi.nih.gov:

a) Use any web browser (or ftp client) to connect to NCBI ftp site at: ftp://ftp.ncbi.nih.gov/pub/taxonomy/

b) Download the file: taxdump.tar.gz

c) Extract the file to <OTT_working_dir>/DB-dir/ncbi_dmp/.

d) Use any web browser (or ftp client) to connect to NCBI ftp site at: ftp://ftp.ncbi.nih.gov/genbank/

e) Download the required sequence files. These files are marked with gbXXX*.seq.gz, where XXX stands for the group name. For example, for plants you need to download all files that begin with gbpln (e.g., gbpln4.seq.gz).

f) Extract the downloaded files to <OTT_working_dir>/DB-dir/gb_XXX/ (For plants: gb_pln).

g) Run the following command with the provided python script:

```
python <OTT_working_dir>/OTT_Code/ott_scripts/GenBank_OTT.py rod-pln
<DB-dir path> YES
```

Arguments for *GenBank_OTT.py*:

- The first parameter denotes the group names [mam (mammals), rod (rodents), pri (primates), pln (plants), vrt (vertebrates) and inv (invertebrates)]. To enable multiple groups, the group names should be concatenated with a dash (for example *rod-pln*). **The group names should be identical to those entered in *OneTwoTree_SA.ini* file** (line #23, see section 2, 3b).
- The second argument is the path to <OTT_working_dir>/DB-dir/
- The third argument is to denote whether to download NCBI files using OTT script or not:
  YES – OTT script will only create databases from files that were pre-downloaded (option 1 above).
  NO – OTT script will both download the files from NCBI ftp and create the databases (as explained under option 2, below).

2. **Option 2.** Run the following command with the provided python script. This will both download the files from NCBI ftp and create the database (the arguments are as detailed in option 1 above):

```
python <OTT_working_dir>/OTT_Code/ott_scripts/GenBank_OTT.py pln
<OTT_working_dir>/DB-dir/ NO
```

4 Outgroup Database creation - optional

In case an automatic outgroup detection is needed, an additional step is required that would enable a blast search. This step should be performed after the local sequence database is created. Run the following script which will create the BlastDB file under the DB-dir:

```
python
<OTT_working_dir>/OTT_Code/ott_scripts/Create_Blast_DB.py
rod-pln <OTT_working_dir>/DB-dir/
```

The first argument specifies the selected taxonomical groups (e.g., rod-pln) and the second argument the path to the database directory.

5 Run OTT standalone version:

The following command is used to run OTT (an explanation about the parameters can be found at the end of the Virtual-Machine section):

```
python <OTT_working_dir>/OTT_Code/buildTaxaTree.py --
taxa-list-file <your taxa_list.txt> --working-dir
<Output_DirName> --config-filename
<OTT_working_dir>/OTT_Code/OneTwoTree_SA.ini --id <any
name, without spaces> --params-file-path <your
paramsFile.txt>
```

Appendix B – python modules/packages used

Modules used by OTT, included by default in python version 3.6.2 and above:

argparse	fnmatch	operator	time
argparse	glob	os	unicodedata
codecs	hashlib	pickle	zipfile
collections	inspect	random	zlib
configparser	io	re	
copy	itertools	shutil	
csv	json	sqlite3	
ctypes	logging	string	
datetime	math	subprocess	
fileinput	mmap	sys	

Additional Python packages used by OTT. Packages listed with an asterisk are optional.

package Name	Version	package Name	Version
Taxonome*	1.5	Bio	1.69
ete3	3.1.1	pandas	0.20.3
mysql.connector	2.2.2b1	numpy	1.13.1

Appendix C – Replace NCBI sequence data

In case other taxonomical groups are needed (currently VB include only plants) or you would like to download the latest NCBI data, please follow the instructions below:

- Erase all directories and files under /home/ottadmin/OTT/DB-dir.
- Open a new terminal:
  - Type 'bash'
  - Type 'alias python=python3'
  - Type 'cd /home/ottadmin/OTT'
  - Type '*python OTT_Code/ott_scripts/GenBank_OTT.py* ***mam*** *DB-dir/ NO'*

    Make sure to replace the taxonomical group name to the groups you need, in the example above the groups mammals was chosen. If you need more than 1 group, concatenate by using '-' (for example*:* ***mam-rod).***

To enable an automatic outgroup detection (not mandatory), an additional step is required that would enable a blast search. This step should be performed after the local sequence database is created. Run the following script which will create the BlastDB file under the DB-dir:

- Type '*python OTT_Code/ott_scripts/Create_Blast_DB.py* ***mam*** *DB-dir/*

The first argument specifies the selected taxonomical groups (as explained above) and the second argument the path to the database directory.