Advanced Topics in Applied Ecology
26 - 28 Jan 2022

dcv.uc
DEPARTMENT OF LIFE SCIENCES

The use of Molecular Markers

Tutorial – Organelle Genome Assembly and Annotation

LUIS CUNHA

Table of Contents

1. Context

1.1 Teaching Objectives

The purpose of this practical is to introduce you to a organelle genome assembly generated using illumine sequencing with paired-end reads. This process generates an paired reads (Read 1 and Read 2 files) that can be used to feed into the genome assembly pipeline. In this exercise there are several steps that you need to follow, namely:

- Select a species and get the genomic data (made available in google drive, R1 and R2);
- Find an appropriate seed sequence in Genbank;
- Create a folder for your work;
- Move the seed and the genomic data files to this folder
- Edit the config file and run the assembly;
- Evaluate the assembly and if OK proceed with the annotation (MITOS2 server);
- Develop Primers for COI and an extra gene (free for you to choose);
- Discuss and build your presentation.

Particular attention should be paid to the following points:

[1] The manual is designed to provide the individual steps required in the process. These steps may be used in a variety of ways to aid analysis.

1.2 Timetable

The session is timetabled for the 27th of January 2022

Any questions, drop me an email: luis.cunha@uc.pt

1.3 Main Software for the Genome Assembly

NOVOPlasty uses a file in FASTA format containing a seed (seed), which will be interactively **extended bidirectionally** during the organelle assembly. This seed sequence is not used to start the assembly, but to retrieve a sequence read of the target genome from the assembly of NGS data. Thus, unlike most assemblers, NOVOPlasty does not tries to assemble all reads, but extends the given seed until the genome circle is formed. Details about the entire process can be found in the article by tool description.

Dierckxsens N., Mardulyn P. and Smits G. (2016) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research doi: 10.1093/nar/gkw955

Software and guidance can be found at:

Dierckxsens, N., Mardulyn, P., & Smits, G. (2020). NOVOPlasty (Version 4.3.1) [Computer software]. GitHub. https://github.com/ndierckx/NOVOPlasty

You can find all the needed resources in the google shared folder exclusively created for this course:

https://bit.ly/3fZ13Oi

2. Installing the Needed Software:

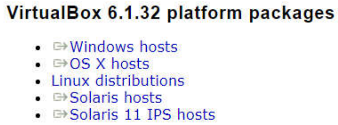
2.1 Installing Virtual Box (VB)

For this exercise you will need to have available on your machine a working copy of latest Virtual Box. VirtualBox is a powerful operative system virtualization product. It allows you to run a virtualized operated system within a “window” on your computer.

First, you need to pick one appropriate for your native operative system (VirtualBox runs on Windows, Linux, Macintosh) and installing it. You can get it here:

https://www.virtualbox.org/wiki/Downloads

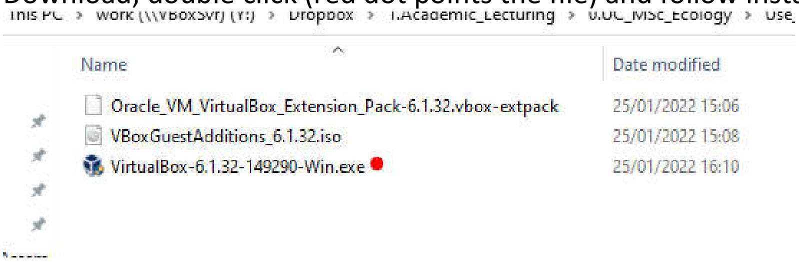
Choose your OS:



VirtualBox 6.1.32 platform packages

- ⇨Windows hosts
- ⇨OS X hosts
- Linux distributions
- ⇨Solaris hosts
- ⇨Solaris 11 IPS hosts

Download, double click (red dot points the file) and follow installation instructions:



This PC › work (\\VBoxSvr) (Y:) › Dropbox › 1.Academic_Lecturing › 0.UC_MSc_Ecology › Use_

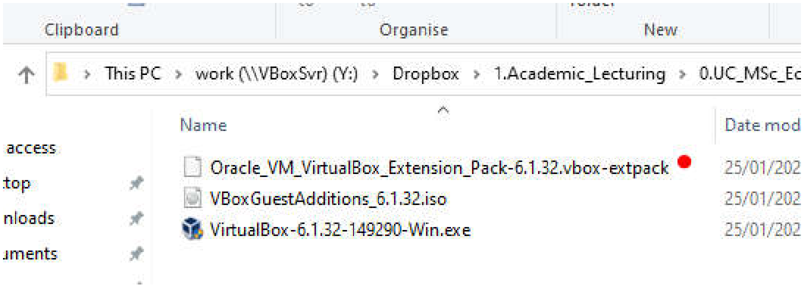
Name	Date modified
Oracle_VM_VirtualBox_Extension_Pack-6.1.32.vbox-extpack	25/01/2022 15:06
VBoxGuestAdditions_6.1.32.iso	25/01/2022 15:08
VirtualBox-6.1.32-149290-Win.exe ●	25/01/2022 16:10

2.2 Installing Extension Pack

You will also need to download the extension pack in the previous webpage:

LINK

and install it by double clicking on the file (red dot points the file):



This PC › work (\\VBoxSvr) (Y:) › Dropbox › 1.Academic_Lecturing › 0.UC_MSc_Ec

Name	Date mod
Oracle_VM_VirtualBox_Extension_Pack-6.1.32.vbox-extpack ●	25/01/202
VBoxGuestAdditions_6.1.32.iso	25/01/202
VirtualBox-6.1.32-149290-Win.exe	25/01/202

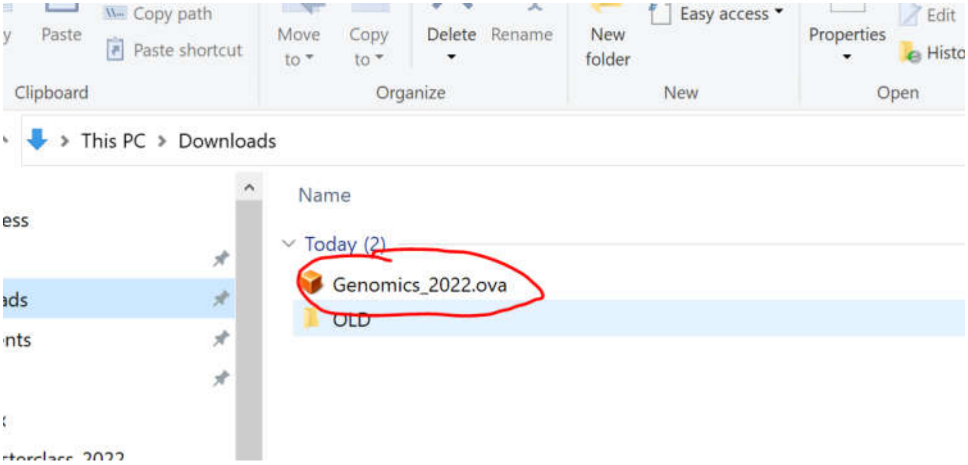
3. Loading the Virtual Box image and getting the Linux (Mint) Operative system to run:

3.1 Loading Image in Dropbox

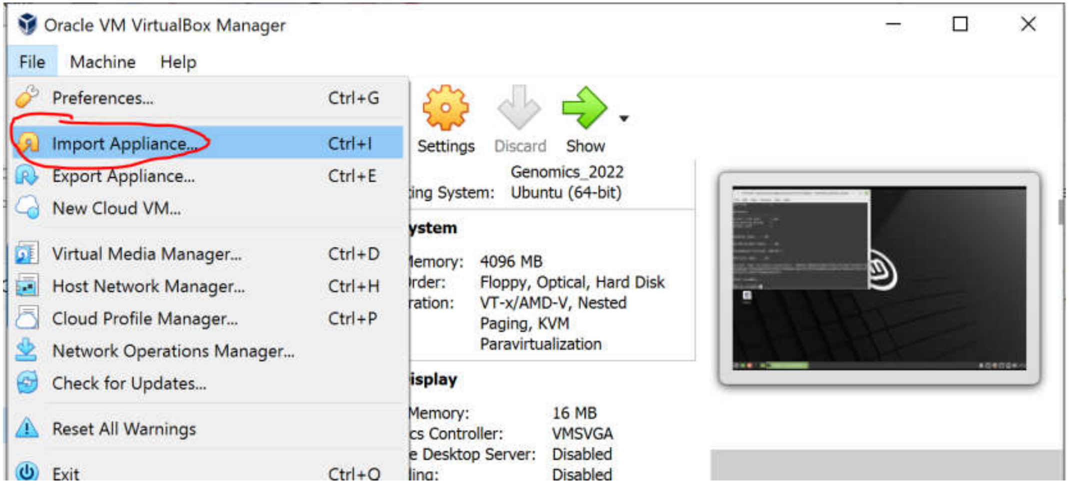
Download the image file from google shared folder or using the following link to obtain a copy: https://bit.ly/3s8Jl0v

3.2 Starting the operative system

You can double click, and it should open with the virtual box application.



If not, you can import using the VB interface. Follow the on screen instruction and leave the default options:

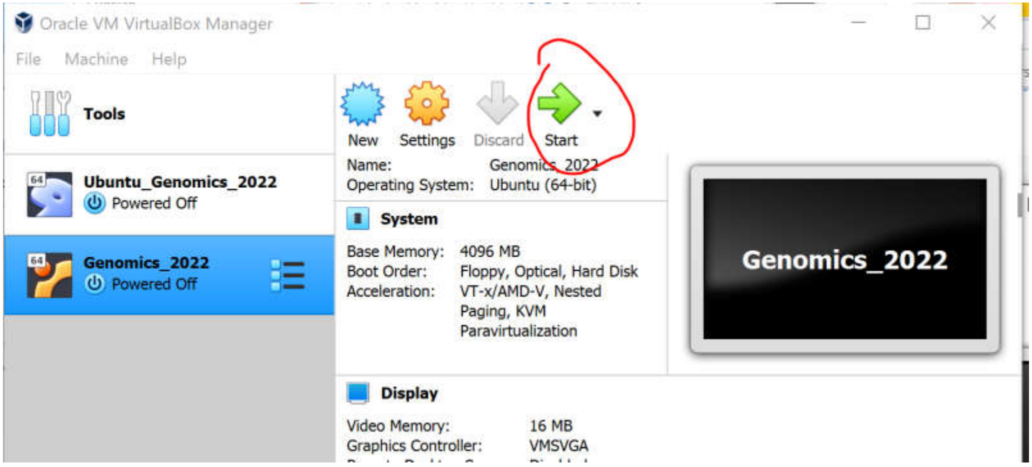

Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
New Cloud VM...
Virtual Media Manager... Ctrl+D
Host Network Manager... Ctrl+H
Cloud Profile Manager... Ctrl+P
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit Ctrl+Q
Settings
Discard
Show

Genomics_2022
ting System: Ubuntu (64-bit)

ystem
lemory: 4096 MB
rder: Floppy, Optical, Hard Disk
ration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

isplay
Memory: 16 MB
cs Controller: VMSVGA
e Desktop Server: Disabled
ling: Disabled

It should take a couple of minutes to load, but after importing you will see it in your virtual box interface, and you can start it:



Ubuntu_Genomics_2022 Powered Off

Genomics_2022 Powered Off

Name: Genomics_2022
Operating System: Ubuntu (64-bit)

System

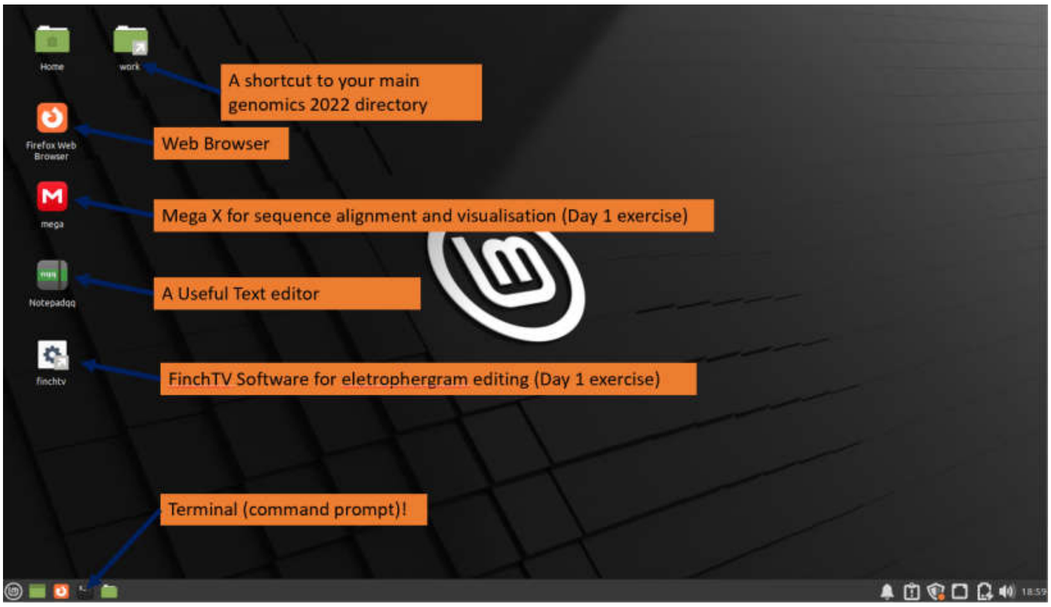
Base Memory: 4096 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VMSVGA

3.3 Understanding your operative system

You have now an operational Linux Mint operative system ready for some analysis. Check below the description of your desktop items. Also, you will notice that after opening your Mozilla web browser that some useful web pages bookmarks have been added.



A shortcut to your main genomics 2022 directory

Web Browser

Mega X for sequence alignment and visualisation (Day 1 exercise)

A Useful Text editor

FinchTV Software for eletrophergram editing (Day 1 exercise)

Terminal (command prompt)!

3.4 Additional software availability

The image that you will be provided already has all the needed software installed. However, if you prefer to get the software installed in your machines, you can use the following links to get appropriated software version to your operative system (OS).

- Alignment and sequence visualisation with Mega X: https://www.megasoftware.net/
- Sequence editing (eletropherograms) with FinchTV: https://digitalworldbiology.com/FinchTV
- Text editing with Notepad++: https://notepad-plus-plus.org/

4. Create a Folder and Gathering needed files:

For some Linux basic commands (Terminal/command prompt window), you can check the file: “0.Intro_Linux_Commands.pdf” available in google drive. For a more extended introduction to Linux, you can check the following link:
https://www.hostinger.com/tutorials/linux-commands

4.1 Open the terminal and create a folder for you to work exclusively in this task;

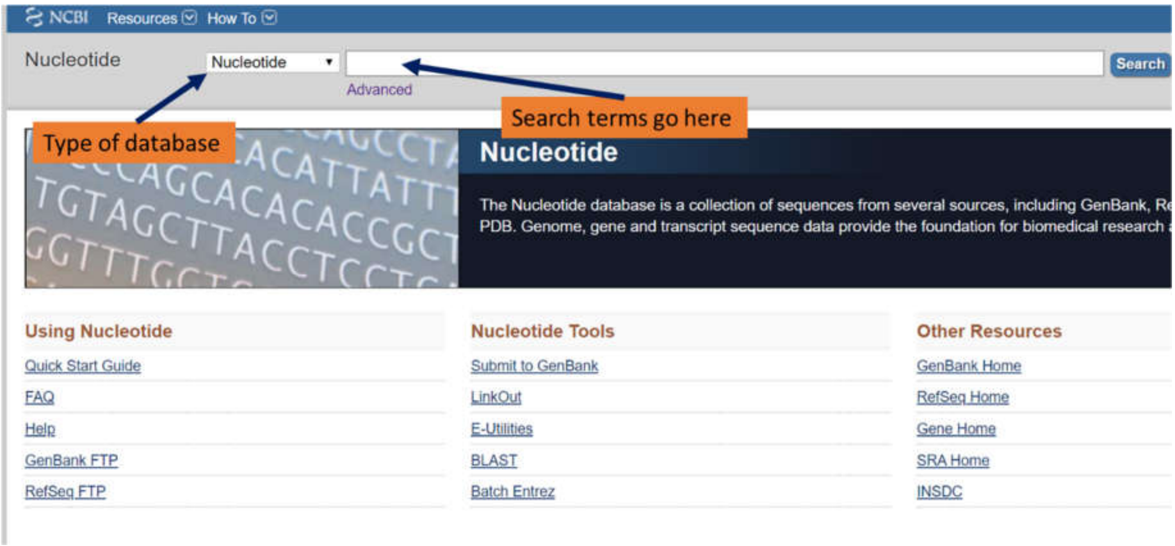
```
$ cd ~  # this takes you to your home directory
$ mkdir Name_of_working_folder  # create a directory
$ cd Name_of_working_folder # enter the directory
```

4.2 Open the browser and find and save a seed sequence for your species in Genbank nucleotide database;

There are different types of seed possible:

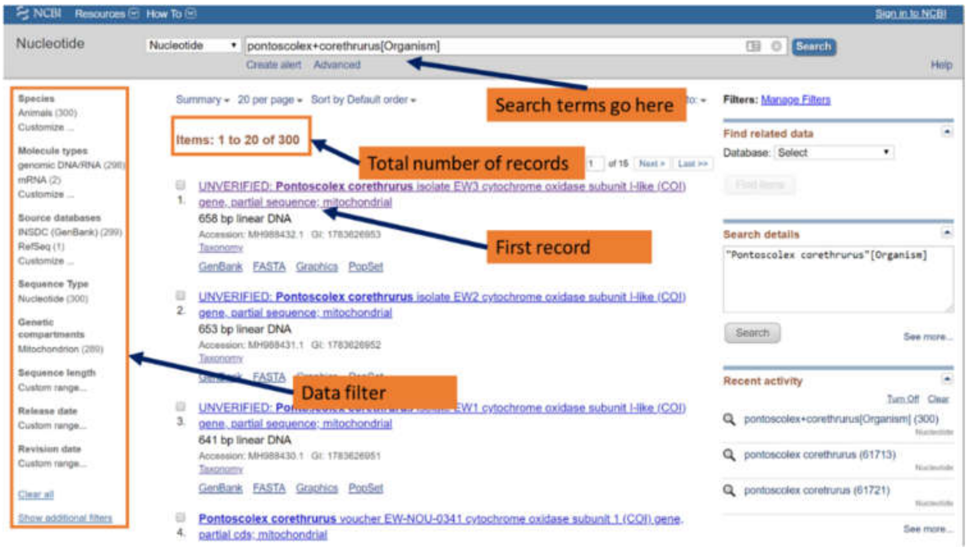
- A single read from the dataset that originates from the organelle genome.
- An organelle sequence derived from the same or a related species.
- A complete organelle sequence of a more distant species (recommended when there is no close related sequence available)

4.3 Open your browser and go to Pubmed - NCBI: https://www.ncbi.nlm.nih.gov/nuccore/



The Nucleotide database is a collection of sequences from several sources, including GenBank, Re
PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research a

Look for any sequence information for a specific species using term “Species name” with filter term “[Organism]”. This will provide all the sequence information available for this specific organism:



pontoscolex+corethrurus[Organism]

Search terms go here

Items: 1 to 20 of 300

Total number of records

First record

Data filter

1. UNVERIFIED: **Pontoscolex corethrurus** isolate EW3 cytochrome oxidase subunit I-like (COI) gene, partial sequence; mitochondrial
658 bp linear DNA
Accession: MH988432.1 GI: 1783626953
Taxonomy
GenBank FASTA Graphics PopSet

2. UNVERIFIED: **Pontoscolex corethrurus** isolate EW2 cytochrome oxidase subunit I-like (COI) gene, partial sequence; mitochondrial
653 bp linear DNA
Accession: MH988431.1 GI: 1783626952
Taxonomy
GenBank FASTA Graphics PopSet

3. UNVERIFIED: Po... EW1 cytochrome oxidase subunit I-like (COI) gene, partial sequence; mitochondrial
641 bp linear DNA
Accession: MH988430.1 GI: 1783626951
Taxonomy
GenBank FASTA Graphics PopSet

4. **Pontoscolex corethrurus** voucher EW-NOU-0341 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial

Search details

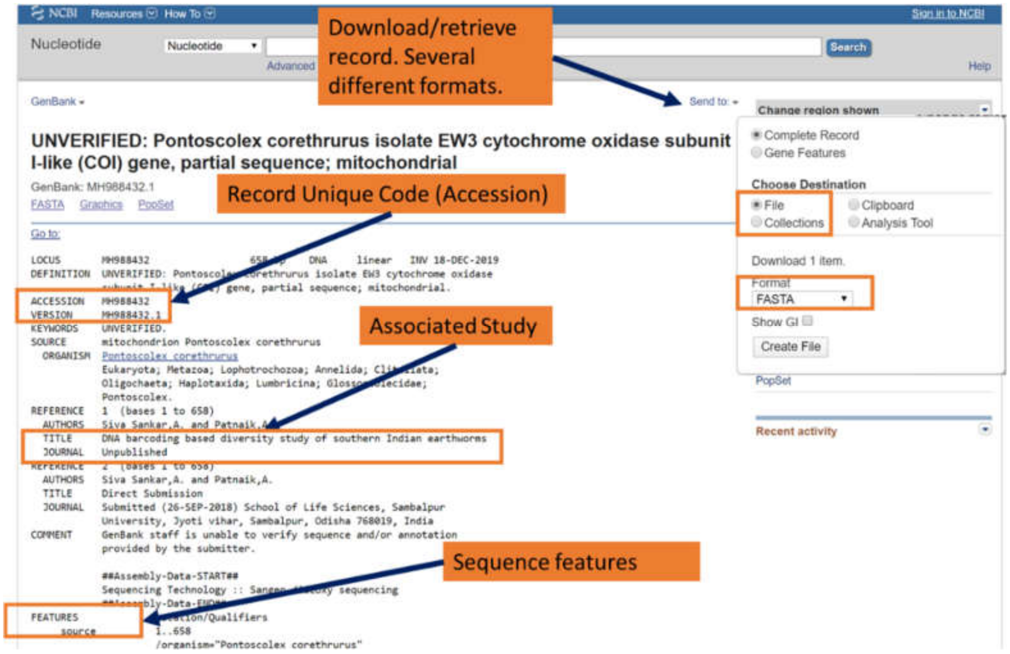
```
"Pontoscolex corethrurus"[Organism]
```

Recent activity

- pontoscolex+corethrurus[Organism] (300)
- pontoscolex corethrurus (61713)
- pontoscolex coretrurus (61721)

4.4 Take some time to explore the filtering options

(e.g. “Genetic compartments”). Pressing on the record, will take to a page with more information about it:



Download/retrieve record. Several different formats.

UNVERIFIED: Pontoscolex corethrurus isolate EW3 cytochrome oxidase subunit I-like (COI) gene, partial sequence; mitochondrial

GenBank: MH988432.1

Record Unique Code (Accession)

Associated Study

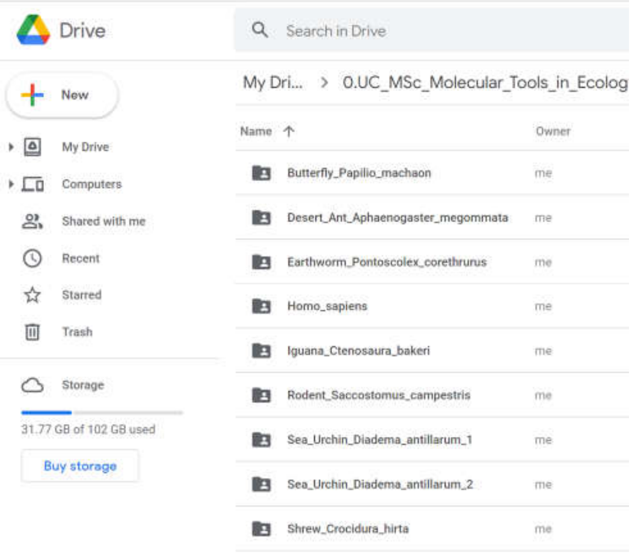
Sequence features

```
LOCUS       MH988432                 658 bp    DNA     linear   INV 18-DEC-2019
DEFINITION  UNVERIFIED: Pontoscolex corethrurus isolate EW3 cytochrome oxidase
            subunit I-like (COI) gene, partial sequence; mitochondrial.
ACCESSION   MH988432
VERSION     MH988432.1
KEYWORDS    UNVERIFIED.
SOURCE      mitochondrion Pontoscolex corethrurus
  ORGANISM  Pontoscolex corethrurus
            Eukaryota; Metazoa; Lophotrochozoa; Annelida; Clitellata;
            Oligochaeta; Haplotaxida; Lumbricina; Glossoscolecidae;
            Pontoscolex.
REFERENCE   1  (bases 1 to 658)
  AUTHORS   Siva Sankar,A. and Patnaik,A.
  TITLE     DNA barcoding based diversity study of southern Indian earthworms
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 658)
  AUTHORS   Siva Sankar,A. and Patnaik,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-SEP-2018) School of Life Sciences, Sambalpur
            University, Jyoti vihar, Sambalpur, Odisha 768019, India
COMMENT     GenBank staff is unable to verify sequence and/or annotation
            provided by the submitter.

            ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..658
                     /organism="Pontoscolex corethrurus"
```

4.5 Get read 1 and read 2 for your species

Go to our shared google drive folder and download and move the genomic R1 and R2 for your selected species to your working folder. This can be done using the command prompt ($ mv) or using the graphical interface (cut and paste). I always advise you to try to use the command prompt to improve your skills.



My Dri... > 0.UC_MSc_Molecular_Tools_in_Ecolog

Name ↑	Owner
Butterfly_Papilio_machaon	me
Desert_Ant_Aphaenogaster_megommata	me
Earthworm_Pontoscolex_corethrurus	me
Homo_sapiens	me
Iguana_Ctenosaura_bakeri	me
Rodent_Saccostomus_campestris	me
Sea_Urchin_Diadema_antillarum_1	me
Sea_Urchin_Diadema_antillarum_2	me
Shrew_Crocidura_hirta	me

5. Editing Config File and Running the Assembly:

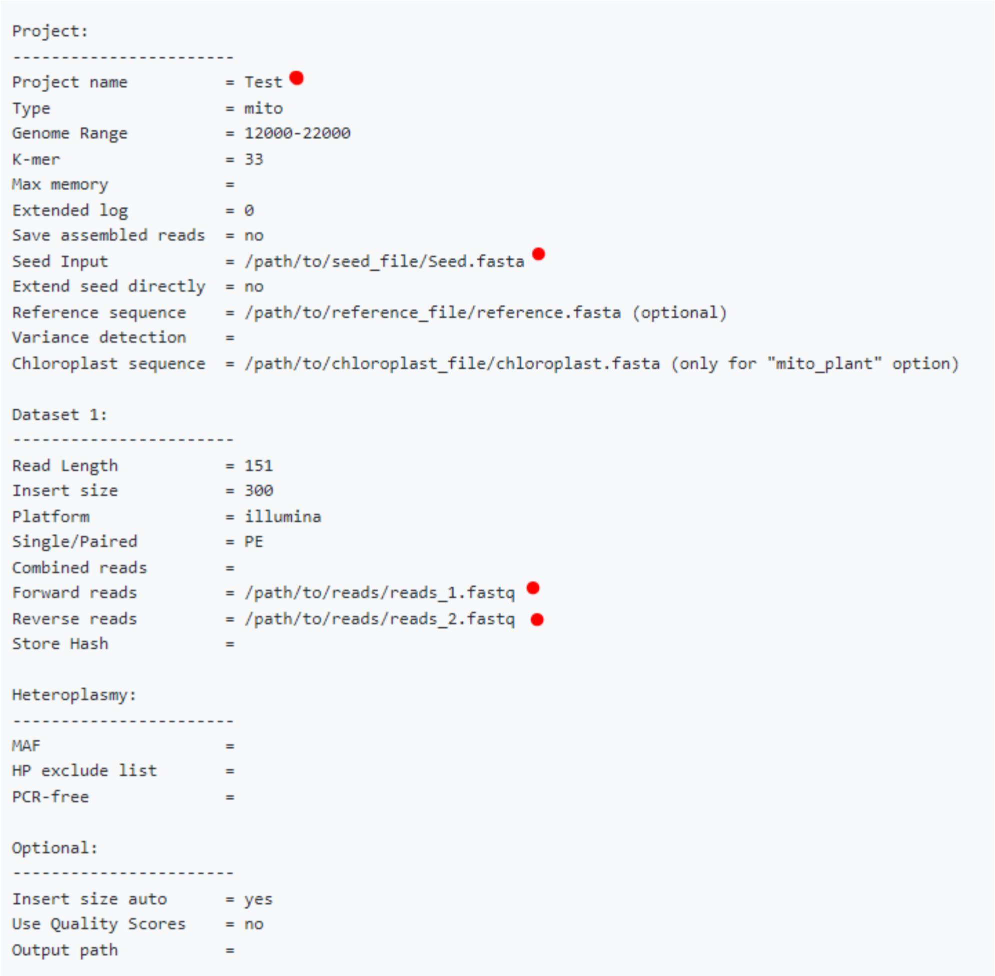
5.1 Get the config file template

Download the config file template (google drive) and move it to your working folder

5.2 Edit the config file and adjust the settings liking.

Every parameter of the configuration file is explained in annex 1. Discuss this with your lecturer if needed;

Example of configuration file:



```
Project:
-----------------------
Project name          = Test ●
Type                  = mito
Genome Range          = 12000-22000
K-mer                 = 33
Max memory            =
Extended log          = 0
Save assembled reads  = no
Seed Input            = /path/to/seed_file/Seed.fasta ●
Extend seed directly  = no
Reference sequence    = /path/to/reference_file/reference.fasta (optional)
Variance detection    =
Chloroplast sequence  = /path/to/chloroplast_file/chloroplast.fasta (only for "mito_plant" option)

Dataset 1:
-----------------------
Read Length           = 151
Insert size           = 300
Platform              = illumina
Single/Paired         = PE
Combined reads        =
Forward reads         = /path/to/reads/reads_1.fastq ●
Reverse reads         = /path/to/reads/reads_2.fastq ●
Store Hash            =

Heteroplasmy:
-----------------------
MAF                   =
HP exclude list       =
PCR-free              =

Optional:
-----------------------
Insert size auto      = yes
Use Quality Scores    = no
Output path           =
```

NOTE: You can always try different K-mer's. In the case of low coverage problems or seed errors, it's recommended to lower the K-mer (set between 21-39)!
If you are having fragmented assemblies, consider increasing the k-mer.

5.3 Run NOVOplasty

No further installation is necessary:
Novoplasty is a perl script, so you need to invoke perl to run it. Novoplasty script is found inside the directory path: “/home/genomics2022/work/apps/NOVOPlasty”. To run it, type:

$ **perl** /home/genomics2022/work/apps/NOVOPlasty/NOVOPlasty4.3.pl -c config.txt

5.4 Check output files:

After completing the mitochondrial assembly, NOVOPlasty will provide some files referring to the assembled mitogenome, in addition to a log file containing details about the parameters used, stages of the assembly process and metrics of the generated mitogenome. The main files generated are:

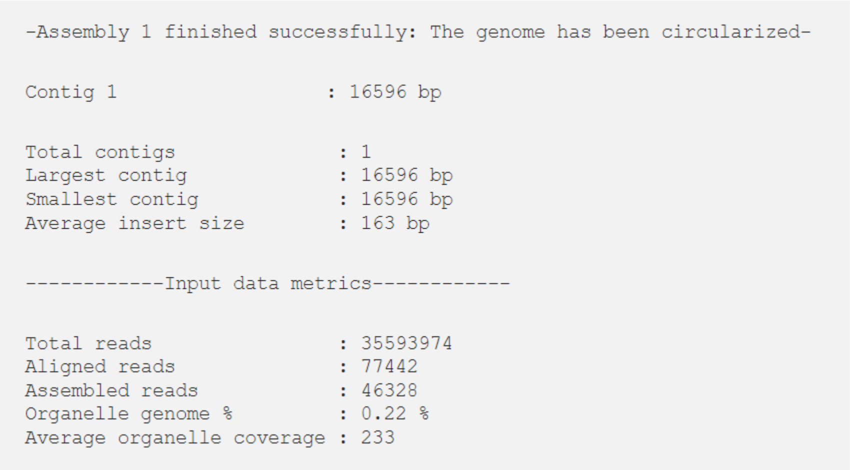
1. Contigs_project.txt

This file contains all the assembled contigs.

2a. Circularized_assembly_project.fasta

When NOVOPlasty is able to circularize one contig, without any additional contigs being produced, it will just output this circularized fasta file.

5.5 Check your genome metrics:



```
-Assembly 1 finished successfully: The genome has been circularized-

Contig 1                  : 16596 bp

Total contigs              : 1
Largest contig             : 16596 bp
Smallest contig            : 16596 bp
Average insert size        : 163 bp

------------Input data metrics------------

Total reads                : 35593974
Aligned reads              : 77442
Assembled reads            : 46328
Organelle genome %         : 0.22 %
Average organelle coverage : 233
```

As we can see, our mitogenome was able to circularize and provided us with a single contig containing 16596 bp, allowing us to complete the assembly of the example mitochondrial genome. More details about the library and the assembly process are provided in this log, demonstrating, for example, the 233x coverage of our organelle based on the dataset used. You should present this data in your presentation results.

6. Annotating the Assembly:

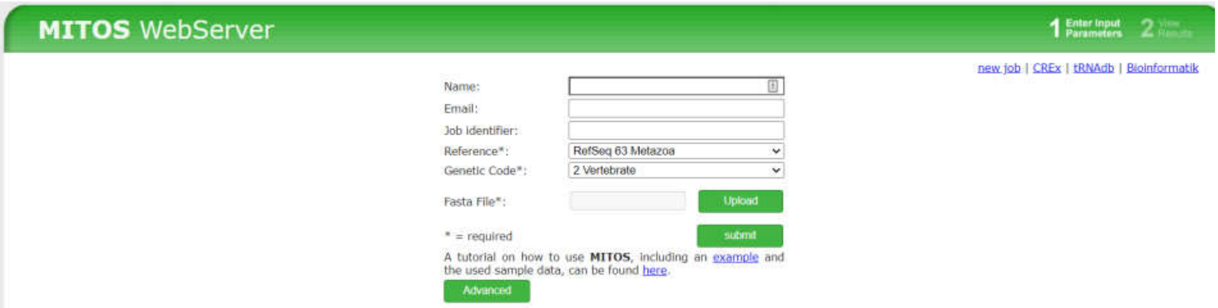
The MITOS pipeline is designed to compute a consistent de novo annotation of the mitogenomic sequences.

Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes Nucleic Acids Research 2019, 47(20):10543–10552 link

6.1 Accessing the MITOS2 web interface

Open your web browser. The MITOS pipeline is accessible online:
http://mitos2.bioinf.uni-leipzig.de/index.py

Fill the form according to the nature of your species (e.g., vertebrate or invertebrate, etc) give a name to your project (e.g., "Human_mito") and to the task (e.g., "Test_1"), ad your email, upload your file and submit:



Name:

Email:

Job identifier:

Reference*: RefSeq 63 Metazoa

Genetic Code*: 2 Vertebrate

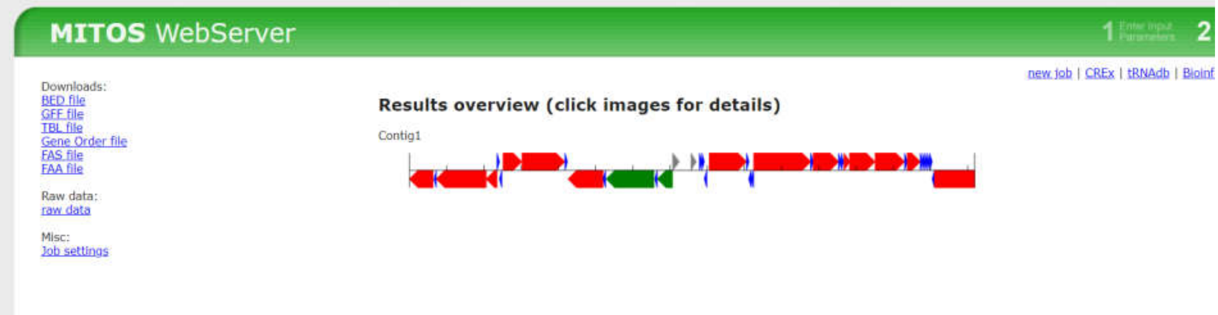
Fasta File*:

* = required

A tutorial on how to use **MITOS**, including an example and the used sample data, can be found here.

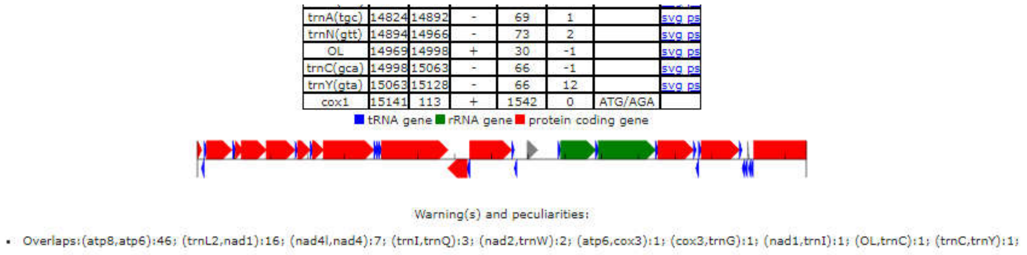
6.2 Check the annotation, take some time in understanding the output:

Once the annotation is ready, you will be redirected the results page. This page has a menu on the left that offering several downloads and the results in tabular and graphical form on the right.



Below the table you will see a visualization of the annotation. Genes located on the plus strand are drawn in the upper part. Genes annotated on the minus strand are shown in the lower region. A small vertical line is drawn every 1,000 nt. A click on the image will present a larger annotated version.

Note that the legend of the plot indicates which feature types have been searched by MITOS (tRNA, rRNA, protein).



trnA(tgc)	14824	14892	-	69	1		svg ps
trnN(gtt)	14894	14966	-	73	2		svg ps
OL	14969	14998	+	30	-1		svg ps
trnC(gca)	14998	15063	-	66	-1		svg ps
trnY(gta)	15063	15128	-	66	12		svg ps
cox1	15141	113	+	1542	0	ATG/AGA	

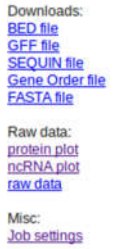
🟦 tRNA gene 🟩 rRNA gene 🟥 protein coding gene

Warning(s) and peculiarities:

- Overlaps:(atp8,atp6):46; (trnL2,nad1):16; (nad4l,nad4):7; (trnI,trnQ):3; (nad2,trnW):2; (atp6,cox3):1; (cox3,trnG):1; (nad1,trnI):1; (OL,trnC):1; (trnC,trnY):1;

6.3 The output:

Results are provided in commonly used file formats: BED, GFF, FASTA, and TBL format and a visual overview of the results is presented. Additionally, a file containing the gene order, i.e. the gene names in the order of their appearance on the genome, that can be used for genome rearrangement analyses (e.g. with CREx) is provided. Furthermore, the raw data, i.e. all files that have been generated by MITOS, is available as zip archive. More info at: http://mitos2.bioinf.uni-leipzig.de/help.py#output



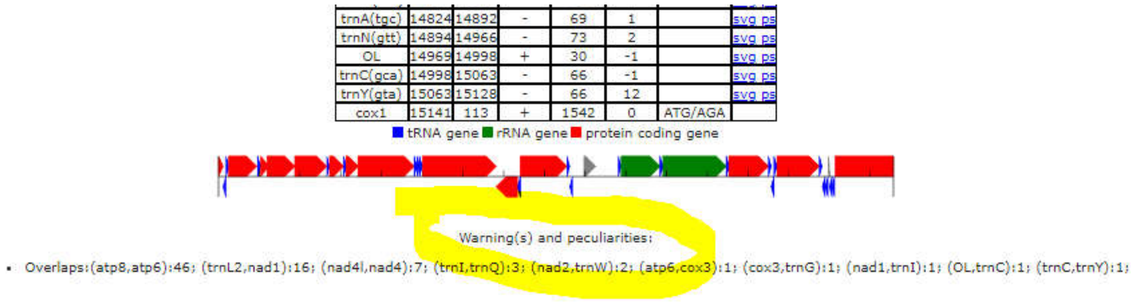
Downloads:
BED file
GFF file
SEQUIN file
Gene Order file
FASTA file

Raw data:
protein plot
ncRNA plot
raw data

Misc:
Job settings

6.4 Check gene numbers and any problems with the annotation:

The central part is the results table where you find all genes predicted by MITOS and their position in the provided sequence.



trnA(tgc)	14824	14892	-	69	1		svg ps
trnN(gtt)	14894	14966	-	73	2		svg ps
OL	14969	14998	+	30	-1		svg ps
trnC(gca)	14998	15063	-	66	-1		svg ps
trnY(gta)	15063	15128	-	66	12		svg ps
cox1	15141	113	+	1542	0	ATG/AGA	

🟦 tRNA gene 🟩 rRNA gene 🟥 protein coding gene

Warning(s) and peculiarities:

- Overlaps:(atp8,atp6):46; (trnL2,nad1):16; (nad4l,nad4):7; (trnI,trnQ):3; (nad2,trnW):2; (atp6,cox3):1; (cox3,trnG):1; (nad1,trnI):1; (OL,trnC):1; (trnC,trnY):1;

With this information you can create a table for your presentation.

Congratulations, you have reached the end of the tutorial.

For more advanced users you could:

1. Evaluate the sequence quality with SQUAT;
2. Run the heteroplasmy mode in Novoplasty;
3. Build a figure using OGDRAW.

I hope you have enjoyed, and would love to hear your feedback
Luis

Annex 1

Novoplasty Parameters

Project:

Project name   = Choose a name for your project, it will be used for the output files.
Type   = (chloro/mito/mito_plant) "chloro" for chloroplast assembly, "mito" for mitochondrial assembly and "mito_plant" for mitochondrial assembly in plants.
Genome Range   = (minimum genome size-maximum genome size) The expected genome size range of the genome.
Default value for mito: 12000-20000 / Default value for chloro: 120000-200000
If the expected size is know, you can lower the range, this can be useful when there is a repetitive
region, what could lead to a premature circularization of the genome.
K-mer   = (integer) This is the length of the overlap between matching reads (Default: 33).
If reads are shorter then 90 bp or you have low coverage data, this value should be decreased down to 23.
For reads longer then 101 bp, this value can be increased, but this is not necessary.
Max memory   = You can choose a max memory usage, suitable to automatically subsample the data or when you have limited
memory capacity. If you have sufficient memory, leave it blank, else write your available memory in GB
(if you have for example a 8 GB RAM laptop, put down 7 or 7.5 (don't add the unit in the config file))
Extended log   = Prints out a very extensive log, could be useful to send me when there is a problem (0/1).
Save assembled reads = All the reads used for the assembly will be stored in seperate files (yes/no)
Seed Input   = The path to the file that contains the seed sequence.
Extend seed directly = This gives the option to extend the seed directly, in stead of finding matching reads. Only use this when your seed
originates from the same sample and there are no possible mismatches (yes/no)
Reference (optional) = If a reference is available, you can give here the path to the fasta file.
The assembly will still be de novo, but references of the same genus can be used as a guide to resolve
duplicated regions in the plant mitochondria or the inverted repeat in the chloroplast.
References from different genus haven't beeen tested yet.
Variance detection = If you select yes, you should also have a reference sequence (previous line). It will create a vcf file
with all the variances compared to the give reference (yes/no)
Chloroplast sequence = The path to the file that contains the chloroplast sequence (Only for mito_plant mode).

You have to assemble the chloroplast before you assemble the mitochondria of plants!

Dataset 1:

Read Length = The read length of your reads.
Insert size = Total insert size of your paired end reads, it doesn't have to be accurate but should be close enough.
Platform = illumina/ion - The performance on Ion Torrent data is significantly lower
Single/Paired = For the moment only paired end reads are supported.
Combined reads = The path to the file that contains the combined reads (forward and reverse in 1 file)
Forward reads = The path to the file that contains the forward reads (not necessary when there is a merged file)
Reverse reads = The path to the file that contains the reverse reads (not necessary when there is a merged file)
Store Hash = If you want several runs on one dataset, you can store the hash locally to speed up the process (put "yes" to store the hashes locally)
To run local saved files, goto te wiki section of the github page

Heteroplasmy:

MAF = (0.007-0.49) Minor Allele Frequency: If you want to detect heteroplasmy, first assemble the genome without this option. Then give the resulting
sequence as a reference and as a seed input. And give the minimum minor allele frequency for this option
(0.01 will detect heteroplasmy of >1%)
HP exclude list = Option not yet available
PCR-free = (yes/no) If you have a PCR-free library write yes

Optional:

Insert size auto = (yes/no) This will finetune your insert size automatically (Default: yes)
Use Quality Scores = It will take in account the quality scores, only use this when reads have low quality, like with the
300 bp reads of Illumina (yes/no)
Output path = You can change the directory where all the output files will be stored.