IntroNGS Documentation

Release 0.1

Isheng J Tsai

Oct 25, 2022

Contents

Hi! I am Jason Tsai , an associate research fellow at Biodiversity Research Center, Academia Sinica, Taiwan.

This website contains the teaching slides that I taught every two years in NTNU and NTU, along with other things.

```
** or *GSB*
```

Contents:

CHAPTER 1

Course Introduction

This module aims to cover the backgrounds of next generation sequencing (NGS), and what you can do with it in research. This module will also provides hands-on exercises from real-world scenarios.

The lectures are jointly taught by Isheng Jason Tsai, Meiyeh Lu, John Wang of Biodiversity Research Center, Academia Sinica and JiaMing Chang of NCCU.

Assessment

There will be one paper review and presentation (7+3 min) for your exercise proposal. You will also have a written proposal and in-class final exam.

CHAPTER 2

Lecture notes

This page contains all the slides that I taught.

2.1 GSB 2022 R practicals

1. R intro exercise [updated v2022] Open
2. R exercise for diversity Open
3. DEseq exercise Open

2.2 Advanced Microbiology

1. Ecological Genomics [v2020] `Download`

2.3 TIGP Introduction to Genomics / (NGS)

1. Introductory lecture [v2020] `Download`
2. Introduction II Linux, R, and other-Tools [v2020] `Download`
3. Read Mappings [v2020] `Download`
4. Genome Assembly and case studies [v2020] `Download`
5. Comparative Genomics [v2020] `Download`
6. RNAseq and Genome annotation [v2020] `Download`
7. From Alignment to phylogenetic tree (Jiang Ming Chang) [v2020] `Download`
8. Population Genomics (John Wang) [v2020] `Download`

9. Amplicon / Metagenomics [v2020] `Download`
10. Study week (no class)
11. DNA/RNA preparation and different sequencing technologies (Meiyeh Lu) [v2020] `Download`
12. Midterm Exam (27t May)
13. Final presentation I (Students)
14. Final presentation II (Students)
15. R tutorial I [v2020] ; Attendance optional ; `Download`
16. Practical two: RNAseq mapping and DEseq2 (not updated yet) ; Attendance optional
17. no class

First assignment (dealine 25th March): Find a paper that has a combination of comparative, population, RNAseq or metagenomics in your field (at least 2). Write a protocol on how the bioinformatics part of the study was conducted (what tools, what version, input, output). As detailed as possible.

Second assignment (deadline 15th April):

1. Construct a BWT of the following sequence ANNABANANA . Show your working.
2. What is the output of last column?
3. Write out how you searched the string ANNA

Final presentation. Each of you will present a 10 minute talk about your "genomic projec proposal", followed by 5 minutes of questions. This can be resequencing, de novo assembly, RNAseq, amplicons, metagenomics, anything as long as it involves sequencing. This can not be your current work, so you need to think outside the box. The only required format is your first slide -> briefly introduce yourself and what you are currently working on. The second slide is the start of proposal with a title of your project. Any number of slides are fine as long as you can keep track of time. Order of students will be chosen randomly. Marks will be normalised by week, for example the marks of first group will be normalised based on the performance of presenters in week 14.

Midterm exam: Self explanatory. Two hours. All the guest speakers' lectures won't be tested.

Final marks submitted

2.4 Example Dataset (v2018 ; not updated yet)

1. `myoviridae_healthy.txt` (taken from R into with GGPLOT).
2. `worms.txt` (Example taken from R exercises and examples).
3. `Survey2.csv` (questionnaire survey).
4. `R examples in Lecture 2.`

2.5 TIGP B2

1. Comparative and Evolutionary Genomics [v2022] `Download`
2. Genome assemblies and case studies [v2022] `Download`

Assignment (take home question): First choose a group of species, or a species. Then please write a short review (~10 references) on how analyses of comparative/population genomics have been transformed by recent advances (algorithm and experimental approaches) in sequencing.

2.6 TIGP Microbial Diversity and Ecology

1. Fungal Diversity [v2019] `Download`
2. Genomics of Eukaryotic microorganisms `Download`

2.7 2021

1. NGS Introduction `Download`

Note: Email ijtsai at sinica.edu.tw if you have any problems/suggestions/want to use the slides

CHAPTER 3

Statistics

In 2015, the final students who registered this course were 12 (=5 (BIODIV/NTNU) + 7 (NTU)) and 29 further participating students/PIs.

CHAPTER 4

Genome skimming exercise (last updated 2022.04.14)

This page is part of the Ecology Master Class. We will take the sequences that we sampled and produced from MinION platform and see if we can retrieve the mitochondrial genome!

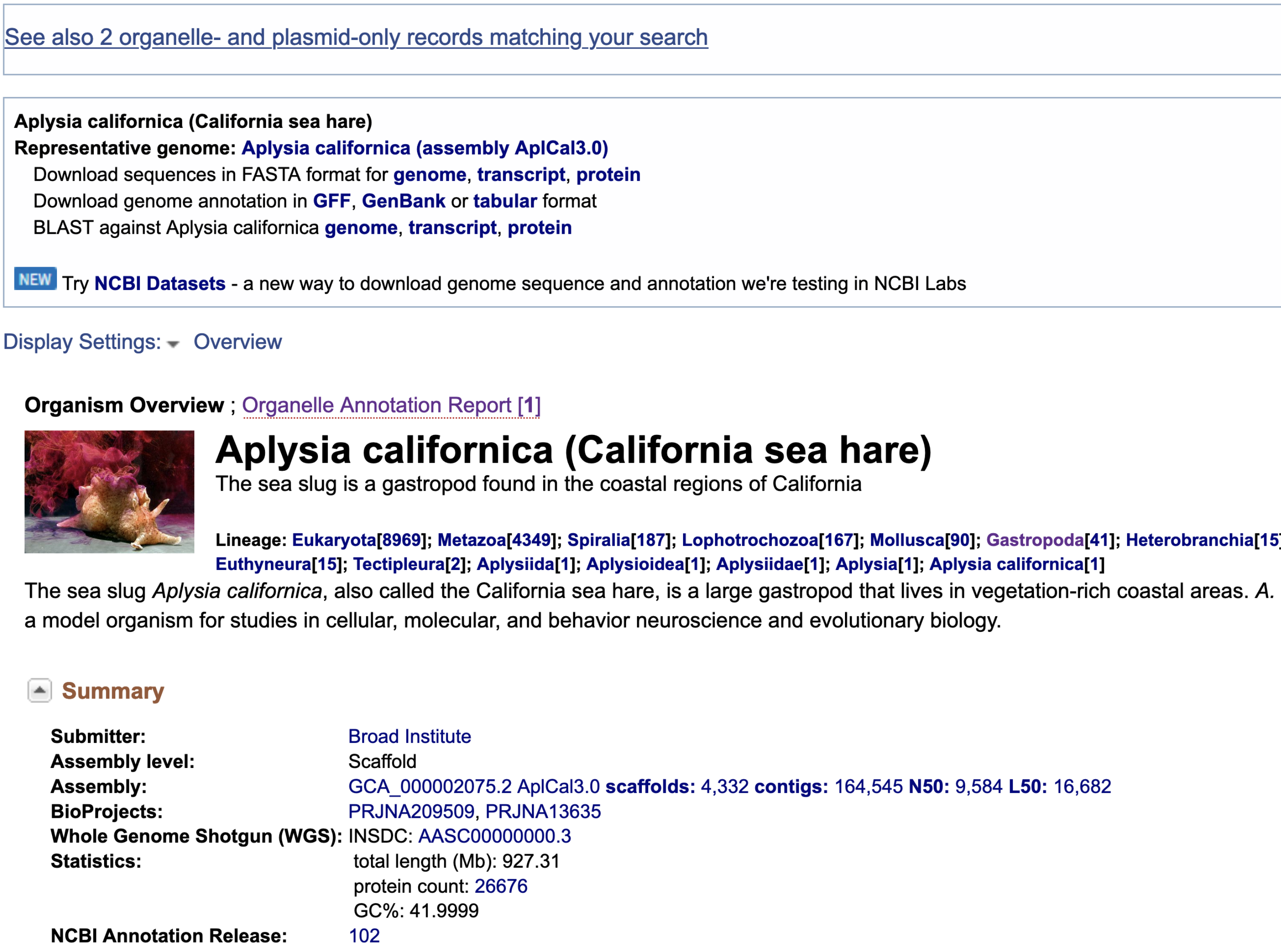
Relevant reading: 1. Genome skimming for next-generation biodiversity analysis

2. Utilisation of Oxford Nanopore sequencing to generate six complete gastropod mitochondrial genomes as part of a biodiversity curriculum

4.1 Step 1: Which species to choose/download?

Most of the complete mitochondrial genomes are available in the Organelle Genome Resources . So try to search for the complete mitogenome of most closely related species of your Sample.

Since we previously identified our sample as Aplysia xxxx, just type aplysia and click search. You should see the search result something like this.



See also 2 organelle- and plasmid-only records matching your search

Aplysia californica (California sea hare)
Representative genome: **Aplysia californica (assembly AplCal3.0)**
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Aplysia californica **genome**, **transcript**, **protein**

NEW Try **NCBI Datasets** - a new way to download genome sequence and annotation we're testing in NCBI Labs

Display Settings: ▾ Overview

Organism Overview ; Organelle Annotation Report [**1**]

Aplysia californica (California sea hare)

The sea slug is a gastropod found in the coastal regions of California

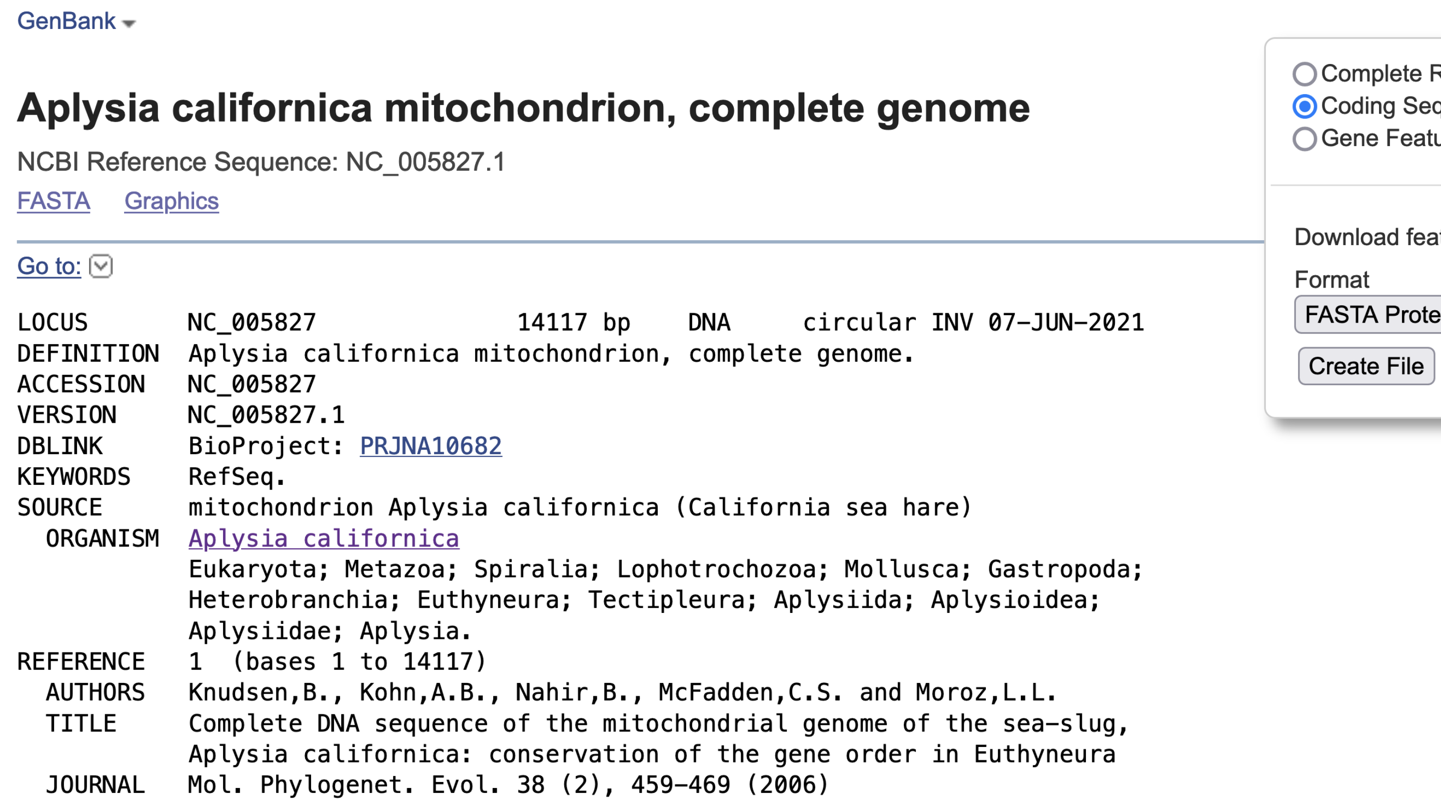
Lineage: Eukaryota[8969]; Metazoa[4349]; Spiralia[187]; Lophotrochozoa[167]; Mollusca[90]; Gastropoda[41]; Heterobranchia[15 Euthyneura[15]; Tectipleura[2]; Aplysiida[1]; Aplysioidea[1]; Aplysiidae[1]; Aplysia[1]; Aplysia californica[1]

The sea slug *Aplysia californica*, also called the California sea hare, is a large gastropod that lives in vegetation-rich coastal areas. *A.* a model organism for studies in cellular, molecular, and behavior neuroscience and evolutionary biology.

Summary

Submitter:	Broad Institute
Assembly level:	Scaffold
Assembly:	GCA_000002075.2 AplCal3.0 **scaffolds:** 4,332 **contigs:** 164,545 **N50:** 9,584 **L50:** 16,682
BioProjects:	PRJNA209509, PRJNA13635
Whole Genome Shotgun (WGS):	INSDC: AASC00000000.3
Statistics:	total length (Mb): 927.31 protein count: 26676 GC%: 41.9999
NCBI Annotation Release:	102

You should immediately found that representative organism in the group is the California sea hare Aplysia californica. And there's a Organelle Annotation Report 1 . Don't be afraid and click into it. You should then find the summary of the mitogenome and its accession number NC_005827.1 . Click!



GenBank ▾

Aplysia californica mitochondrion, complete genome

NCBI Reference Sequence: NC_005827.1

FASTA Graphics

Go to:

```
LOCUS       NC_005827              14117 bp    DNA     circular INV 07-JUN-2021
DEFINITION  Aplysia californica mitochondrion, complete genome.
ACCESSION   NC_005827
VERSION     NC_005827.1
DBLINK      BioProject: PRJNA10682
KEYWORDS    RefSeq.
SOURCE      mitochondrion Aplysia californica (California sea hare)
  ORGANISM  Aplysia californica
            Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Gastropoda;
            Heterobranchia; Euthyneura; Tectipleura; Aplysiida; Aplysioidea;
            Aplysiidae; Aplysia.
REFERENCE   1  (bases 1 to 14117)
  AUTHORS   Knudsen,B., Kohn,A.B., Nahir,B., McFadden,C.S. and Moroz,L.L.
  TITLE     Complete DNA sequence of the mitochondrial genome of the sea-slug,
            Aplysia californica: conservation of the gene order in Euthyneura
  JOURNAL   Mol. Phylogenet. Evol. 38 (2), 459-469 (2006)
```

○ Complete R
● Coding Seq
○ Gene Featu

Download fea

Format

FASTA Prote

Create File

Please download the coding sequences in the fasta protein format (See screenshot). This will act as the bait sequence to identify putative mitochondrial sequences from our sample.

4.2 Step 2: Upload the sequences to the server

Once the sequence file is available, you will need to copy the sequence to the server, where the raw sequences and programs reside. The server's address will be made available on the day of the class.

```
# scp: Secure Copy (from the SSH suite of computer applications
# for secure communication)
scp source_file_name destination_file_name

# Example 1
# From laptop/desktop to Server
# Need to open a terminal and go to the directory to where the sequence is
# usually @ ~/Downloads
# Need to replace groupx with your group number (e.g., group1, group2)
scp sequence.txt tigp2022@xxxxxxxxx:/home/tigp2022/group1/pep.fa

# Example 2
# copy from server to laptop/desktop
scp tigp2022@xxxxxxxxxx:/home/tigp2022/file_name ~/Desktop/filename

# Now please try to upload the protein fasta sequence to server
```

4.3 Login to the server and start understanding your sequence data

In the home directory, you will see a few fastq files that contains raw sequences of the samples that you have sequenced.

```
# home directory is /home/tigp2022/

# First do a pwd
# pwd = print working directory
# You should see that you are in /home/tigp2022/
pwd

# Try ls (abbreviation for list)
# You should see a list of fastq file and the folder Aoc which you just created
ls

# ls or any Linux commands can be added with different arguments
# What files have we got here?
ls -lrt

# now we want to move around the folders. We use cd (Change Directory) command
# change to the data directory
# Inspect using ls
# ../ means previous directory
cd data
pwd
ls -lrt
cd ../
pwd

# you can use cd ~/ to go back to your home directory (if you are lost)
cd ~/
```

Now that you know how to move around, you can try to inspect some files

```
# Go to the data folder again and find try to view a fastq file.
# Since they are gzipped. You need to use the command zless
zless Aoc.R1.fastq.gz

# use space to go page down, use arrows to go up and down.
# use q to quit viewing the file
```

Let’s start the analysis!

```
# cd to your groups's directory. This will the directory you will carry out your analyses
# cd means Change directory
# We will use group1 as an example
cd ~/group1
pwd

# you want to copy fastq file into the new working folder and renamed to data.fastq.gz
# ../ means previous directory
# . means current directory
cp ../data/Aoc.R1.fastq.gz .
cp ../data/Aoc.R2.fastq.gz .

# Since we have two fastq files which correspond to sequencing output of individual
```

(continues on next page)

(continued from previous page)

```
# sequencing runs. We will combine them using cat (short for for conCATnate) command
cat Aoc.R1.fastq.gz Aoc.R2.fastq.gz > data.fastq.gz

# data stats
# what does the output mean?
fastn2stats.py --fastn Aoc.R1.fastq.gz
fastn2stats.py --fastn Aoc.R2.fastq.gz
fastn2stats.py --fastn data.fastq.gz


# Search for closely related species
# See [Step 1]

# Copy the protein sequences from your desktop to your current working directory in the server using # And rename it to pep.fa
# Remember you can do it in one step!
# See [Step 2]
```

4.4 Search for putative mitogenome sequences

```
# Come back to original directory
# diamond makedb
diamond makedb --threads 8 --in pep.fa -d ref


# match reference
# what does the output say?
diamond blastx -b5 -c1 --threads 8 -d ref -q data.fastq.gz -o ref.matches.tsv


# get the ID out
awk '{print $1}' ref.matches.tsv | sort | uniq > ref.match.id


# get the reads out
fastq_subset.firstfield.pl ref.match.id data.fastq.gz data.fastq.gz.subseq.fq

# stats
fastn2stats.py --fastn data.fastq.gz.subseq.fq
```

4.5 Assembly

```
# flye
# flye if not working set --min-overlap 1000 or 1500
# if longer sequence than expected and failed to circlise then set --min-overlap 3000
flye --nano-raw data.fastq.gz.subseq.fq --out-dir out_nano --threads 8 --min-overlap 3000
```

4.6 Annotation using MITOS online

```
# 1. Go to the flye assembly folder and look around
cd out_nano
ls -lrt

# 2. try a few command. For example. How long is it?
# Any previous command you could use? or use the new seqstat command.
fastn2stats.py --fastn assembly.fasta
seqstat assembly.fasta

# 3. Print the sequence onto screen.
cat assembly.fasta
less assembly.fasta

# 4. Copy the sequence to your desktop/laptop using scp and try to blast to NCBI. What to you find?

# 5. Annotate using MITOS
http://mitos.bioinf.uni-leipzig.de/index.py
```

4.7 Alternative mitogenome annotation using MitoZ

```
# 5. Annotation using mitoZ; Result here:
# Copy the files from this to your working directory OR your desktop/laptop
# Have a browse
# You can copy the files to your desktop to take a look, too!
cd /home/tigp2022/mitoZ.result/Aoc/
```

4.8 Do you want to try other species?

4.9 Reference of the programs used

1. The flye assembler
2. DIAMOND which is a sequence aligner for protein and translated DNA searches, which is MUCH faster than BLAST
3. MITOS WebServer which annotates mitogenomes online
4. mitoZ which is a local tool for annotating mitogenomes (can be quite hard to install to run).

Note: Email ijtsai at sinica.edu.tw if you have any problems/suggestions about the genome skimming exercise

CHAPTER 5

References

This page contains all the goodies on the internet also relevant to this course.

5.1 Lecture slides

1. Workshop on all kinds of genomics[1] link
2. Introduction to differential gene expression analysis using RNA-seq link
3. Konrad Paszkiewicz. History of DNA and modern approaches to sequencing (2017) link
4. Introduction to bioinformatics using NGS link
5. Introduction to genome annotation link
6. EMBL predocs python course link

5.2 Good reviews / papers / videos

Genome assembly

1. Jang-il Sohn and Jin-Wu Nam (2016) The present and future of de novo whole-genome assembly

Metgenomics

1. Jovel *et al.*, Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics

RNAseq

1. Van Den Berge *et al* (2018)., RNA sequencing data: hitchhiker's guide to expression analysis

Population genomics

[1] This is led by a small group of faculty at various institutions around the world. I strongly recommend any students to study the materials in here.

1. Sònia Casillas and Antonio Barbadilla (March 2017) Molecular Population Genetics [#f2]_
2. The 100,000 Genomes Project - How We Get Results

Note: This module will be taught in English