Project: Finding AMRs

Purpose

To explore soil metagenomics using Galaxy tools for de novo genome assembly (Flye), assembled contig visualization (Bandage), finding antimicrobial resistance (AMR) genes (ABRicate) and binning contigs into larger groups/bins (MetaBAT2). You will also explore information about the taxonomic classifier GTDB-Tk. This will allow you to compare genomic assemblies and AMRs between soil and Zymo gut standard.

Learning Objectives

Use **Galaxy** tools to

1. Perform de novo genome assembly of long reads into 'contigs' using **Flye** tool
2. Visualize the contigs with **Bandage** tool
3. Find AMRs in contig assemblies using **ABRicate** tool
4. Bin contigs into larger MAGS using **MetaBAT2** tool
5. Learn about **GTDB-Tk** taxonomy classifier tool

Introduction

In this activity we will practice *de novo* metagenome assembly with the tool `Flye`, using a **soil** sample from the BioDIGS project sequenced with long-read Nanopore sequencing technology. We will have an opportunity to compare and contrast assembled contigs and antimicrobial resistance profiles of soil and gut and think about the differences and similarities in microbial diversity of the two environments.

In this activity we will also learn about MetaBAT2 and GTDB-Tk. The most up to date long read metagenomics workflow includes contig assembly (e.g. using Flye tool), followed by contig binning into larger metagenome-assembled genomes (MAGs) (with e.g. MetaBAT2 tool) and finally MAG classification (e.g. with GTDB-Tk). MetaBAT2 is an algorithm that bins (or groups) sequence fragments (contigs) into larger MAGs or draft genomes. Subsequently, MAGs can be taxonomically classified by GTDB-Tk.

You will not be executing GTDB-Tk in this activity to stay within a reasonable activity time frame. However, for your project work, you will have a chance to test GTDB-Tk on your genome assemblies and bins. A note to your future self working on a project - as a rule, after genome assembly, if your contigs are 500 kb or above, they will be considered large enough to be passed on to GTDB-Tk without the need for binning. Contigs of < 500 kb will be binned and passed on to GTDB-Tk for taxonomy classification as bins.

Overview of the approximate minimum times for a job to be completed on Galaxy using specified tools.

- Note, these times apply only to the specific input file we will be using in this activity, the nanopore-soil-subset that is 5.4 GB.

Flye	Bandage	ABRicate	MetaBAT2
5 hours	< 5 min	< 10 min	< 10 min

Activity 1 – Genome assembly with Flye

Estimated time: 30 min (activity time DOES NOT include the Flye run time on Galaxy)

The sample used in this activity is from the BioDIGS Project, sequenced using Oxford Nanopore Technologies' PromethION Instrument. A subset of this data was adapter-trimmed and filtered and is used in this activity.

Instructions

1. Run Flye in Galaxy on adapter-trimmed (with Porechop tool) and quality filtered (with fastp tool) nanopore soil subset nanopore-soil-subset-filtered to de novo assemble soil microbial genomes.
    a. Obtain .fastq file from a subset of Nanopore-sequencing soil study
        - https://usegalaxy.org/u/valerie-g/h/nanopore-soil-subset-filtered-1
    b. Name your new history **Finding soil AMRs**.
    c. Run **Flye** tool to assess sequence quality using the following **Tool Parameters**.
        - Under **Input Reads**: select your nano `pore-soil-subset-filtered` **.fastq** dataset.
        - Under **Mode**: select `--nano-raw` option, since the sequences were obtained using Nanopore sequencing technology.
        - Under **Perform metagenomic assembly**: select `Yes`.
        - Under **Generate a log file**: select `Yes`.
2. Explore Flye output **assembly info** file which is sorted by length (in base pairs, bp) of the contig (high to low).

Questions

1. How many contigs were assembled?

Note: Since each contig is represented by a separate row (or line) in the assembly info file, simply clicking on the assembly info file and recording the number of lines listed in the file will correspond to the number of contigs.

2. What is the longest contig size?

3. What percent of input was assembled into contigs?

Note: In the log file,

- Find the input number of bases going into flye assembly - this info corresponds to the “Total read length” value on top of the log file.
- Find the output number of bases after flye assembly - this info corresponds to the “Total length” value on the bottom of the log file.
- Calculate percent of input assembled into contigs using the 2 input and output values you obtained above.

4. Why do you think only a small fraction of reads was assembled into contigs?

5. Compare soil assembly to the Zymo gut standard assembly provided the following observations:

- For this activity you can consult back to your Prelab: Finding AMRs.
- The number of contigs assembled from Zymo gut standard D6331 subset was much smaller than for the filtered soil sample! Yet, the largest contig size for the Zymo gut standard was over 2 million bases (almost 10 times larger) and circular (while the largest contig for the soil sample was smaller and linear).

5A. Why do you think the number of contigs in the soil sample was so much higher than the number of contigs in the Zymo gut standard?

Note: It is NOT because of the difference in the size of the sequencing file. - Think about possible differences in the microbial diversity of the two samples.

5B. Why do you think it was possible to assemble a much larger and circular contig with the Zymo gut standard sample compared to the soil sample?

Activity 2 – Contig visualization with Bandage

Estimated time: 15 min

Instructions

1. Run **Bandage** Image tool in Galaxy to visualize contigs.
    - Run **Bandage Image** tool in Galaxy using your **Flye: graphical fragment assembly file** (in gfa1 format) as input, using default parameters.

Questions

1. Paste the resulting image below.

2. Describe contig profile based on the Bandage Image result.

Activity 3 – Finding AMRs

Estimated time: 30 min

Instructions

1. Run **ABRicate** tool in Galaxy using Flye consensus as input using the following **Tool Parameters**:
    - Under **Input Reads**: select your `Flye: consensus` output in FASTA format.
    - IMPORTANT: Under **Advanced Options**: select `NCBI Bacterial Antimicrobial Resistance Reference Gene Database` as your database option; the default 'resfinder' may not work well.
2. Explore **ABRicate report** file.
    - Note, Abricate output report has the following information.

Column	Description
FILE	The filename this hit came from
SEQUENCE	The sequence in the filename
START	Start coordinate in the sequence
END	End coordinate in the sequence
GENE	ABR gene name
COVERAGE	What proportion of the gene is in our sequence
COVERAGE_MAP	A visual of coverage map (gaps or no gaps)
GAPS	Was there any gaps in the alignment - possible pseudogene?
%COVERAGE	Proportion of gene covered

Column	Description
%IDENTITY	Proportion of exact nucleotide matches
DATABASE	The database this sequence comes from
ACCESSION	The genomic source of the sequence

Questions

1. How many AMR genes were detected? This is the number of rows in your file.

2. How many DIFFERENT AMR genes were detected and what are their GENE names?

3. What are the different AMR genes resistant to? What is their RESISTANCE?

4. How many DIFFERENT contigs had AMRs?

5. Research and write a small paragraph report on one of the AMR genes you found.
    - Use any search tools for your research, but we encourage you to use PubMed https://pubmed.ncbi.nlm.nih.gov/ where you can find many scientific articles on the topic if you search for e.g. your AMR gene name, or resistance name or using a sentence as input.
    - Talk about anything of interest, e.g., which microbes have the AMR of interest, what is the substance to which the gene shows resistance to, where could the resistance to this substance come from, what are possible health implications, etc.

5A. Report on one of the AMR genes you found.

5B. Why do you think the AMRs you found in soil differ from the AMRs you found in the gut (from your pre-lab)?

Activity 4 – Bin contigs with MetaBAT2

Estimated time: 20 min

Instructions

1. In Galaxy, find and click on **MetaBAT2** tool and explore tool parameters.
2. Run **MetaBAT2** tool in Galaxy using `Flye consensus` as input using the **following Tool Parameters**:
    - Under **Fasta file containing contigs**: select your Flye: consensus output in FASTA format.
    - Under **Output options**: from the **Extra outputs** dropdown menu select: `Process log file`.

Questions

1. Explore MetaBAT2 tool and parameters.

1A. What is the function of the MetaBAT2 tool based on the Galaxy tool description on top?

1B. Under Tool Parameters for MetaBAT2, find and record below the Minimum size of a contig for binning (a value given in basepairs, bp).

1C. Under Tool Parameters and Output options for MetaBAT2, find and record below the Minimum size of a bin as the output.

2. Explore MetaBAT2 tool output.

2A. Open MetaBAT2 Process log output file and record how many bins were formed from contigs.

2B. Open MetaBAT2 Process log output file and record how many bases in total were used to form bins.

2C. What percent of contig bases formed bins (given that 154,251,885 bases were in Flye output)?

2D. Based on percent of contigs that formed bins (from activity 4-2.1 above) did metaBAT2 do a good job of binning the contigs?

3. Examine MetaBAT2 Bin sequences output, which is a `DATASET COLLECTION`, where each collection is a separate bin.

3A. Click on the MetaBAT2 Bin sequences output. How many bins (or Galaxy 'folders') are there?

3B. Click on the MetaBAT2 Bin sequences output and then on bin 1. Without 'eyeballing' the fasta file, note how many sequences (contigs) were included in bin 1.

3C. Click on the MetaBAT2 Bin sequences output and then on bin 2. Without 'eyeballing' the fasta file, note how many sequences (contigs) were included in bin 2?

Activity 5 – Read about GTDB-Tk tool in Galaxy

Estimated time: 15 min

Instructions

1. In Galaxy, find and click on **GTDB-Tk Classify genomes.**

Questions

1. Read the GTDB-Tk Classify genomes tool's "What it does" part and summarize what GTDB-Tk does.

2. What is the ideal input sequence for GTDB-Tk classification: 1) raw sequences, 2) contigs or 3) large contigs(>500kb) and MAGs?

3. Visit https://gtdb.ecogenomic.org/stats/r220 to explore the database and database statistics.

3A. How many bacterial species are present in GTDB database Release 220?

3B. Scroll through the website. Although it has a lot of complex information, what is one thing you found interesting about GTDB-Tk content?

Grading Criteria

- Download as Microsoft Word (.docx) and upload on Canvas

Footnotes

Resources

- Google Doc

Contributions and Affiliations

- Valeriya Gaysinskaya, Johns Hopkins University
- Frederick Tan, Johns Hopkins University

Last Revised: June 2025