BINF7001 ASSESSMENT 3 (2024)

De novo genome assembly and phylogenomics

Research Problem

You are a genomics specialist at the Queensland Genomics Institute. You are approached by a client with a set of paired-end Illumina reads generated from a bacterial genome (*see Dataset allocation on Blackboard*).

You are asked to assemble the genome *de novo*, identify the putative species (and strain, if possible) from which the genome is sequenced, and its evolutionary relationship with other closely related species based on 16S ribosomal RNA genes as a phylogenetic tree. You are also asked to compare this tree against that inferred from protein sequences coded by a housekeeping gene.

Specific Aims/Tasks

1. Obtain your assigned genome sequence reads from ENA. See dataset allocation on Blackboard. You can use **`wget`** to download FASTQ files directly.

2. Assemble the genome *de novo* and predict protein-coding genes *ab initio*.

3. Identify 16S ribosomal RNA sequence in your assembled genome. Compile the 16S rRNA set using the 7 sequences provided in Practical 9, incorporating the sequence from the assembled genome. Using the **EIGHT (8)** sequences, construct a multiple sequence alignment and infer a phylogenetic tree depicting the evolutionary relationship of these species. Identify the appropriate outgroup to root the tree.

4. Choose **ONE (1)** housekeeping gene from your allocated list. Identify the corresponding homologous protein set following the clustering steps in Practical 9; this should be a single-copy group. Using protein sequences from the **EIGHT (8)** species (including your target species; per above), construct a multiple sequence alignment and infer the protein tree. Identify the appropriate outgroup to root the tree; this outgroup should be the same as your 16S rRNA tree.

5. Present your key findings in a **video presentation** of **no more than five (5) minutes**.

6. Summarise your findings in a **technical report** of **no more than six (6) pages**.

7. Submit your **video with the attached technical report** on Blackboard by **2pm, Thursday, 10 October 2024**.

> *You will be allocated a dataset at the beginning of Module 2. During Practicals 8 and 9, you will learn all relevant techniques using an example dataset. You are strongly advised to analyse your dataset in parallel to the example dataset during these sessions.*

Assessment Guidelines

Assessment 3 constitutes **30 marks** of the course. Please submit your video presentation with the technical report (in a .pdf or .docx file) as an attachment on Blackboard by **2pm, 10 October 2024**. Name your file as follows: **Surname_Assessment3_BINF7001_2024.pdf** (or .docx).

Part A: Video Presentation (70% of Assessment 3)

The video presentation represents the **highlights** from your technical report. You may follow the general structure of the technical report (Part B below) in your video presentation.

You **DO NOT** need to present **ALL** results that are documented in your technical report. You are free to choose what results (that you think are important) to be included in the presentation.

However, the content (*40*) of your presentation **MUST** provide the following information:

- Methodology/techniques used in the research project (*10*)
- Three key findings from the research project (*15*)
- Recognition of limitations/problems or alternative approaches (*10*)
- Conclusions drawn appropriately from the findings (*5*)

The video presentation **MUST** be no longer than **FIVE (5) minutes**. You should aim to include **no more than SIX (6)** slides in your presentation. You may attach the slides (in additional to the Technical Report below) as a supporting document.

Part B: Technical Report (30% of Assessment 3)

The technical report provides a summary of your results and serves as supporting evidence to your video presentation. You may choose to present relevant information in the form of tables and bullet-points, and you may use no more than **THREE (3) figures/images** in the report. Each figure/image should contain a self-explanatory legend. No cited references are necessary.

The technical report **MUST NOT** be more than **SIX (6) pages** in length, and **MUST include** the following information:

1. **Genome data and assembly** (maximum two pages; *8*)
   1.1 Sequencing data (*3*)
   (a) Sequencing platform from which sequence reads were generated
   (b) Length of reads
   (c) Total number of reads
   (d) G+C content

1.2 Genome assembly (*5*)

(a) Brief description of your approach (name(s) of program(s), key parameters used)
(b) N50 scaffold length
(c) Maximum scaffold length
(d) Total scaffold length
(e) Total number of scaffolds
(f) Percentage of reads that are assembled into contigs
(g) Mean coverage for all contigs
(h) Genome size estimate based on *k*-mer coverage (optional)

2. *Ab initio* gene prediction (maximum one page; *4*)

(a) Brief description of your approach (name(s) of program(s), key parameters used)
(b) Total number of predicted genes
(c) Average gene length
(d) Length and function of the longest gene

3. Phylogenomic analysis (maximum two pages; *8*)

(a) Brief description of your approach (name(s) of program(s), key parameters used)
(b) Total number of homologous protein groups, and number of single-copy groups
(c) 16S rRNA gene tree (8 taxa, rooted using outgroup)
(d) protein tree of the chosen housekeeping gene (8 taxa, rooted using outgroup)
(e) one key difference/similarity between the two trees
(f) one plausible explanation as to why such a difference occurs

4. Recommendations (maximum one page; *10*)

(a) Comment on the limitations of:
  (i) the generated sequencing data from this work, and
  (ii) the approaches you adopted in your research.

(b) Make **two (2)** recommendations to your client for future work (open question). The recommendations may be constructive suggestions of a different approach, how your client can achieve a better genome assembly, or generate better-quality data.

Assessment 3 Marking Criteria

Part A: Video Presentation (70%)

High	Medium	Low	Mark
CONTENT (40%)			
Methodology/techniques used in the research project were clearly presented with sufficient detail.	Methodology/techniques used in the research project were presented with sufficient detail with some ambiguity and uncertainty.	Methodology/techniques used in the research project were not presented, or presented with minimal detail.	/10
Three key findings are presented clearly.	Two key findings are presented clearly, one key finding was presented with some ambiguity.	One key finding is presented clearly, two key findings were presented with some ambiguity, or less than three key findings were presented.	/15
Problems or alternative approaches were clearly recognised and identified.	Problems or alternative approaches were recognised and identified, to some extent.	Problems or alternative approaches were not recognised and not identified.	/10
Conclusions are plausible, and are drawn appropriately from the findings.	Conclusions may be plausible but without appropriate support from the findings.	Conclusions are not plausible, or with no support from the findings.	/5
PRESENTATION (30%)			
Overall findings were presented clearly with precision, and with high confidence.	Overall findings were presented with some precision and with some confidence.	Overall findings were presented with no precision and with little or no confidence.	/15
Information on the video/ slides is presented clearly.	Information on the video/ slides is presented with some ambiguity and uncertainty.	Information on the video/ slides is not presented clearly or is poorly presented.	/10
Presentation time (video length) does not exceed the allocated time.	Presentation time (video length) exceeded the allocated time by one minute.	Presentation time (video length) exceeded the allocated time by more than one minute.	/5

Part B: Technical Report (30%)

High	Medium	Low	Mark
Sequencing data and genome assembly (8%)			
Sequencing data support the video presentation, and are presented precisely and clearly.	Sequencing data support the video presentation, and are presented with some ambiguity.	Sequencing data do not support the video presentation, and/or are not presented (clearly).	/3
Methodology and results support the video presentation, and are presented precisely and clearly.	Methodology and results somewhat support the video presentation, and are presented with some ambiguity.	Methodology and results do not support the video presentation, and/or are not presented (clearly).	/5
***Ab initio* gene prediction (4%)**			
Methodology and results support the video presentation, and are presented precisely and clearly.	Methodology and results somewhat support the video presentation, and are presented with some ambiguity.	Methodology and results do not support the video presentation, and/or are not presented (clearly).	/4
Phylogenomic analysis (8%)			
Methodology and results (two trees) support the video presentation, and are presented precisely and clearly.	Methodology and results (two trees) somewhat support the video presentation, and are presented with some ambiguity.	Methodology and results (two trees) do not support the video presentation, and/or are not presented (clearly).	/8
Recommendations (10%)			
Problems or alternative approaches were clearly recognised and identified.	Problems or alternative approaches were recognised and identified, to some extent.	Problems or alternative approaches were not recognised and not identified.	/4
Two recommendations are plausible and clearly presented, and are drawn appropriately from the findings.	Two recommendations are somewhat plausible and/or presented with ambiguity, or without appropriate support from the findings.	Two recommendations are not plausible and/or not presented, or presented with no support from the findings.	/6

IMPORTANT: This assessment task evaluates students' abilities, skills and knowledge without the aid of Artificial Intelligence (AI). Students are advised that **the use of AI technologies to develop responses in strictly prohibited** and may constitute student misconduct under the Student Code of Conduct.