**Assignment 1**

BIOL331 W2022

**Overview**

We have spent the first half of this course discussing different methods used in genomics research, as well as applications of these techniques (e.g., pangenomics and metagenomics). Below are some questions requiring you to think about these topics and how to interpret results. Some of the questions will require you to use the programs used in tutorials, namely, KBase, MEGA, SILVA, and MG-RAST.

Please answer the following questions. Where searching for results on the web or when running programs is involved, include a screen shot of your set up/results and put everything in a single PDF file.

Once completed, name the PDF file with YOUR name (Last, first) and student number. Upload your PDF to the OnQ site.

**Questions**

1. Go to the NCBI genome site (<https://www.ncbi.nlm.nih.gov/genome/>) and do a search for “bacteria”. Searching that term will deliver over 2000 pages of results. Go to a page number reflective of the first letter in your last name (eg. diCenzo, start at page 4) and choose an organism that has at least three genome assemblies. If the first page you chose doesn’t have a suitable set, then search the surrounding pages until you find an appropriate organism. Be sure to include a screenshot showing the relevant information as asked below. (3 points)
   1. Which organism (genus and species) did you pick?

* They just need to indicate the genus and species names for their organism. (1 point)
  1. How many genome sequences are available for your organism? And how many of them are “complete” genome assemblies?
* They should correctly indicate how many genome assemblies are available for their organism. There are at least one or two places where this is stated on NCBI, or they could just count them. (0.5 points)
* For the number that are complete, these are the assemblies that have a full black circle under ‘level’. They can either count them, or if they select just ‘complete’ in the options along the top of the table, the value should be given to them. It may also list the number elsewhere on the page. (0.5 points)
  1. What is the biggest and smallest genome size for your chosen organism?
     + They should sort the table of genome assemblies according to size and record the largest and smallest sizes. (1 point)

1. Go to KBase (<https://www.kbase.us>) and within the “Queen’s University BIOL 331 W2022” organization, access the module “BIOL331 W2022 Assignment One”. My lab recently sequenced the genomes of six multi-drug resistant bacteria that were isolated form patients at KGH. This module contains sequence and assembly data for one of these strains, named H17629. Use the information in this module to answer the following questions. (5 points)
   1. This module contains two sets of high-throughput sequencing data: “H17629\_read\_set\_1” and “H17629\_read\_set\_2”. Unfortunately, one of the read sets was too large to be analyzed with FastQC on KBase. Therefore, please access the FastQC reports through the links on onQ. Compare the results for the two read sets. Which read set has longer reads? Which read set has the higher average accuracy? Based on these results, which sequencing technologies do you think were used to generate read set 1 and read set 2?
      * Read set 2 has the longer reads (0.25 points)
      * Read set 1 has the higher average accuracy (0.25 points)
      * Read set 1 was generated using Illumina sequencing technology (0.25 points)
      * Read set 2 was generated with Nanopore sequencing technology. However, PacBio is an acceptable answer and so PacBio should also be given full marks. (0.25 points)
   2. Using the output provided in the box “Trim Reads with Trimmomatic”, what is the total number of reads from read set 1 that remained after trimming?
      * The correct answer is: 2,280,360 \* 2 + 51,167 + 29,181 = 4,641,068 (1 point)
      * I had an error on an onQ page that showed the equation (for a different example) with minus signs instead of plus signs. So even though an answer of “2,280,360 \* 2 - 51,167 - 29,181 = 4,641,068” is incorrect, it should nevertheless be given half a mark
      * Another way to do this calculation is: 2,371,554 \* 2 - 51,167 - 29,181 – 10846 \* 2 = 4,641,068
   3. Two genome assemblies were performed for H17629. Data summarizing the quality of the two assemblies are provided in the box titled “Assess Quality of Assemblies with QUAST”; H17629\_assembly\_1 is for an assembly created using the trimmed read set 1; H17629\_assembly\_2 is for an assembly created using both the trimmed read set 1 and the read set 2. Compare the results. For each assembly, provide the following information: i) number of contigs, ii) largest contig, and iii) length of the shortest contig required to cover 50% of the genome length. Does it make sense that H17629\_assembly\_2 has a longer total length than H17629\_assembly\_1 despite having fewer contigs? Why or why not?
      * Assembly 1: 100 contigs; longest contig is 498,634 bp; N50 is 200,632 (0.3 points)
      * Assembly 2: 8 contigs; longest contig is 4,801,038 bp; N50 is 4,801,038 (0.3 points)
      * Yes. Fewer contigs means it is a more complete genome and therefore has fewer gaps / less missing data. (0.4 points)
   4. Which assembly do you think is of better quality? Based on what you know about high-throughput sequencing technologies and genome assembly, why do you think that this assembly was better? (this is not asking for genome metrics, but conceptually, why would a genome assembled with just read set 1 or with both read sets 1 and 2 be better than the other?)
      * The better quality assembly is Assembly 2 (0.5 points)
      * Assembly 2 is a hybrid assembly that makes use of the best of both short-read and long-read technology. It uses the long-reads of Nanopore to generate a highly contiguous genome and the high accuracy of Illumina to minimize sequencing errors. In comparison, Assembly 1 uses only Illumina and thus has a more fragmented assembly due to the use of only short-reads. (0.5 points)
      * If the student says Assembly 1 is a better assembly and gives a logical reason for why they think this, they should be given half a mark.
   5. The strain H17629 has one chromosome and an unknown number of plasmids. Based on the assemblies, how many plasmids do you think H17629 contains? If it is not possible to definitely known how many plasmids there are, please state this and indicate the maximum number of plasmids that you think H17629 contains.
      * Based on the given information, it is not possible to be certain how many plasmids there are as we do not know if the assembly is complete (0.5 points)
      * The maximum number of plasmids is 7. (8 contigs – 1 chromosome = 7 contigs that could each be a plasmid) (0.5 points)
      * If the student says that there is 7 plasmids and either says or implies that this can be known definitively, they should get half a mark
2. The file “AtpD\_sequences.txt” contains the amino acid sequences of the AtpD protein of the six multidrug resistant bacteria, as well as the sequence from an outgroup species. AtpD is the beta subunit of the ATP synthase, and is commonly used in bacterial taxonomy. Using MEGA, align the sequences with the ClustalW algorithm and the PAM substitution matrix, and then save the alignment. Then use this alignment to create a Neighbour-Joining tree with 100 bootstrap replicates. Repeat the construction of a Neighbour-Joining tree (using the same alignment) another two times, giving you a total of three trees. (4 points)
   1. Provide screenshots of the “Bootstrap Tree” for all three phylogenies. Do all three trees have the same topology? Based on these trees, which strain is most closely related to H17629?
      * The topology should be the same for all three trees. (0.5 points)
      * H70375 should be the most closely related to H17629. (0.5 points)
      * It seems that if different settings were used when creating the Neighbour Joining tree, different topologies can be obtained. So mark this question based on the trees they provide. It is possible that their three trees may have different topologies although I doubt this. Note that the attached example has two trees that look different but are actually the same topology, so if they have these trees and say they do not have the same topology, that is wrong.
   2. How confident are you that strains H70375, H17629, and F48994 form a clade? How about F5446, S2568, and T64870?
      * They should be highly confident that the first group forms a clade (0.5 points) but skeptical that the second group forms a clade (0.5 points)
      * As with 4a, mark this based on their bootstrap values. Bootstrap values of about 80 or higher should give them good confidence. Lower bootstrap values should mean they are not confident.
   3. You should see that the bootstrap values differ across the three trees, even though all three trees were made from the same alignment and with 100 bootstrap replicates. How can we explain the observation that the bootstrap values differ across the three trees?
      * Each bootstrap replicate has a random subsample of the alignment columns, with each of the tree including results from 1000 bootstrap replicates. Thus, the 1000 random bootstraps of each tree would differ, giving different values. (1 point)
   4. Now use your alignment to create a Maximum Parsimony phylogeny by hand. Open your saved alignment and look only at the sequences for F48994, H17629, S2568, and T64870 (hint: it makes it easier if you delete the 3 sequences that you don’t need). Starting at the N-terminus of the alignment, find the first 3 sites that would be considered informative according to Maximum Parsimony. Use these 3 sites to create a Maximum Parsimony phylogeny of the strains F48994, H17629, S2568, and T64870. Show your work.
      * See attached photo. (1 point)
      * It is possible that the 3 sites may differ across students, but I think this should not occur. The sites they choose should have two nucleotides, with each present in two strains. If they do the method right but bick sites that do not meet these requirements, they should get 0.5 points.
3. Go to KBase (<https://www.kbase.us>) and within the “Queen’s University BIOL 331 W2022” organization, access the module “BIOL331 W2022 Assignment One”. My lab recently sequenced the genomes of six multi-drug resistant bacteria that were isolated form patients at KGH. These strains are named: F48994, H17629, F5446, H70375, S2568, and T64870 The genome assemblies for all six strains are provided in KBase. Use the information in this module to answer the following questions. (2 points)
   1. Average Nucleotide Identity (ANI) data is provided for all pairwise comparisons in the box “Compute ANI with FastANI”. Based on this data, group the six strains into species. How many species are represented by the 6 strains? For each species, which strains belong to that species? For example, if strains X and Y form one species, and strain Z does not group with any of the other strains, you would say there are two species: species 1 consisting of strains X and Y, and species 2 consisting of strain Z.
      * 3 species are present (0.25 points)
      * Species 1: S2568 and F5446 (0.25 points)
      * Species 2: F48994, H70375, and H17629 (0.25 points)
      * Species 3: T64870 (0.25 points)
   2. We are interested in these six strains in part because they are all resistant to the antibiotic ertapenem, which is classified as a carbapenem, which in turn is a sub-class of beta-lactams. Based on the results in the “Pangenome” box, which cluster (or clusters if there are multiple equally likely candidates) do you think is most likely responsible for ertapenem resistance, and why? Note: if the “Pangenome” box says “bad object”, you will need to hover your mouse over the “Pangenome” object in the “Data” box on the left side of the screen, and click on the three dots that appear. Then, click on the binoculars (which will say “Explore data” when you hover over them). This will open the data in a new page.
      * Cluster 2275 (subclass B1 metallo-beta-lactamase VIM-1) (0.5 points)
      * It is annotated as a beta-lactamase, which provides resistance to beta-lactams, and it is found in all 6 of the strains (0.5 points)
      * If a student says cluster 2275 and cluster 1738, they should get 0.75 marks. Cluster 1738 is no correct since it is a regulator protein.
4. The file “16S\_rRNA\_sequences.fasta” contain the 16S rRNA sequences for each of the 6 strains. Please use SILVA ACT(<https://www.arb-silva.de/aligner/>) to analyze the sequences. For construction of the phylogeny, ensure you select “De novo including neighbours”. (2 points)
   1. Using the “Alignment Results Table” from SILVA ACT, please taxonomically identify the 6 strains of bacteria.

A picture containing text

Description automatically generated

* 1. What species do each of the six strains belong to based on the phylogeny? Please include a picture of the phylogeny.
     + It is possible that different students get different phylogenies, so grade according to their phylogeny. The answers below are from when I ran the analysis.
     + 0.17 points per strain, to a maximum of 1 point
     + Only the species needs to be indicated – subspecies or strain does not need to be added
     + F48994 – *Enterobacter cloacae*
     + F5446 – *Escherichia coli*
     + H17629 – *Enterobacter cloacae* or *Enterobacter hormaechei* (unfortunately, this one wasn’t a clear answer)
     + H70375 – *Enterobacter cloacae* or *Enterobacter hormaechei* (unfortunately, this one wasn’t a clear answer)
     + T64870 – *Klebsiella pneumoniae*
     + S2568 – *Escherichia coli*

1. Go to MG-RAST (<https://www.mg-rast.org/mgmain.html?mgpage=analysis>) and select sequencing type “shotgun). Select three datasets from the “shotgun” list at random, and then run the analysis using the RefSeq database and the COG database. (4 points)
   1. Please provide a figure summarizing the taxonomic distribution at the class level and a figure summarizing the functional distribution at the ‘level2’ level.
      * These will be stacked bar graphs. There should be three stacked bars in each figure. One of them will be based on the RefSeq data (for the taxonomic distribution), and the other based on the COG data (for the functional distribution). (1 point)
   2. Is there more variability at the functional or taxonomic level across samples?

* This refers to differences between the samples. So, are the three stacked bar graphs showing the functional data more similar to each other than the three bar graphs showing the taxonomic data, or vice versa? (1 point)
* This will depend on the samples that they chose to use
* The question is not asking which graphs have more colours within a bar.
  1. In each sample, what is the most abundant class? What percentage of the total population is accounted for by that class?
* This will depend on the datasets that they chose. They should indicate the phylum that is most abundant in each of the three bars for the RefSeq graphs (i.e., which colour accounts for the biggest part of each bar?)
* 0.5 points for indicating the correct most abundant class, and 0.5 points for correctly indicating the abundance (you do not need to run the analysis to confirm the abundance, as long as it looks reasonable from the graph)
  1. Based on the rarefaction curves for the RefSeq data, which sample has the greatest “species richness”? Which samples, if any, are likely to represent the full “species richness” of the environment and how do you know?
* The sample with the greatest species richness will be the sample with the highest curve. (0.5 points)
* A sample would only represent the full species richness if the curves have more-or-less plateaued. If all curves are still going up, none represent the full species richness. (0.5 points)