BioE 131/231
Fall 2018

Lab Assignment #9

Overview and references

You will analyze the bacterial genome you assembled in the previous lab. You will summarize the quality of your assembly using a few different statistics, identify the genome's taxonomic origin, then obtain two genome annotations via different pipelines.

You are expected to keep a thorough record of everything you did in your notebook. Create a folder in your home directory for each lab, and keep all your files there. Try to create a directory hierarchy that makes sense, like the one we went over in Lab 1. Copy and paste any terminal commands you used into a Markdown section and explain what the input was, what the tool did, and what the output was. Plot any results in-line and explain them.

An iPython notebook containing your analysis is due at midnight on Wednesday of next week, spanning both this lab session and last week's. You can upload a link to your GitHub repo on bCourses.

Assembly statistics

https://github.com/sanger-pathogens/assembly-stats
https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics

rRNA Identification

https://en.wikipedia.org/wiki/16S_ribosomal_RNA
http://weizhong-lab.ucsd.edu/meta_rna/
https://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp

Bedtools

https://bedtools.readthedocs.io/en/latest/index.html

Annotation services

https://www.basys.ca/server3/basys/cgi/submit.pl
http://rast.nmpdr.org/

Generate assembly statistics

Locate your assembled genome. There should be a "contigs.fasta" and a "scaffolds.fasta" file in your SPAdes output directory. Contigs are contiguous sequences that could be assembled from your reads. Scaffolds are sets of contigs that have been stitched together in order, and are generally longer than contigs. Sometimes, the assembler can't tell what sequence connects two contigs in a scaffold, and inserts N's in the gap between them. Other times, the assembler has no additional information that could be used to determine the order and orientation of contigs in a scaffold. In this case, scaffolds == contigs.

Using the `assembly-stats` program, please calculate statistics on both your contigs and scaffolds file. Report the total length of all contigs (or scaffolds), the number of contigs (or scaffolds), and the N50 in your iPython notebook.

Why is N50 is useful statistic to calculate? Why not just list the mean or median contig length? In your assembly, are scaffolds longer than contigs, or are scaffolds approximately equal to contigs?

Extra Credit: The number of times a contig in your assembly was covered by the reads used to assemble it ("coverage") is listed at the end of the contig name in contigs.fasta. Extract the coverage from each FASTA header and plot a histogram of coverage for all contigs in your assembly.

Is coverage uniformly distributed? Does it look Gaussian? Is it bimodal or trimodal? What explains the presence of contigs with coverage an integer multiple of the mean coverage?

Identify the taxon from which your genome originated

We know that the genome originated from a taxon of bacteria. One component of bacterial ribosomes is the 16S ribosomal RNA subunit. This functional RNA is conserved throughout all bacteria, and is often used as a taxonomic marker gene. Much of the gene is highly conserved, as function ribosomes are required for protein synthesis, but some regions differ greater between bacterial taxa. These "hypervariable regions" can be used to determine the taxon from which a 16S rRNA gene originated.

To identify your genome, you must 1) scan over the entire genome to locate copies of the 16S ribosomal RNA gene, 2) extract the 16S rRNA genes from your assembly, and 3) compare these genes to a database of known 16S rRNA genes.

First, run the `rna_hmm3.py` program on your assembled contigs to locate rRNA genes. You must specify a path to the input assembly with `-i` and a path to the output General Feature Format (GFF) annotation file with `-o`. This program uses a Hidden Markov Model (HMM) that describes the structure of ribosomal RNA genes to find their coordinates inside a genome. After this program completes, delete all lines within other than those containing 16S_rRNA genes.

Next, use `bedtools getfasta` to extract nucleic acid sequences of the 16S rRNA genes from your assembly. You will need to specify the path to contigs.fasta with `-fi` and the path to the GFF file you obtained above, with `-bed`. The output will be in FASTA format.

Finally, open your web browser and head over to Ribosomal Database Project's SeqMatch tool. Copy and paste your 16S sequences one at a time into the window, or upload the resulting FASTA file from the previous step. This program will attempt to identify the 16S sequences as precisely as possible by comparing them to a database of high-quality, curated sequences, obtained from known bacteria.

You may not be able to obtain a "species"-level identification, but please write down your genus-level identification in your iPython notebook along with an explanation for how you came to this conclusion.

Genome annotation

Knowledge of a genome's taxonomic provenance can be used to infer quite a bit about its lifestyle. For example, if this genome belongs to the *Mycoplasma* genus, we could take a guess that its lacks a cell wall and has a parasitic relationship to an animal host. Given that we know the sequence of the genome, however, we needn't stop at an analysis based on taxonomic labels. We can look inside the genome and determine with high confidence whether it contains genes necessary to produce a cell wall, or virulence factors enabling the infection of an animal host. To do this, we need a program that can break the long genome sequence into genes, then identify their function by identifying orthologs with known function in other, more well-studied genomes. This is called genome annotation.

In this lab, you will upload your genome to two remote annotation services that will perform the annotation automatically: RAST and BASys. There are many more annotation services, but for the sake of time, focus on only two. You must create an account to upload your genome to RAST.

Research and write-up

Now that you have identified your genome and sent off your genome for annotation, take some time to research what is known about its genus and/or species. Search PubMed for recent publications and read through some abstracts. Once your RAST job has finished, you will be presented with a graphical interpretation of the pathways encoded by your genome, and information about its phylogenetic relatives.

Please summarize some of the information obtained in your annotation, placing it in the context of any papers you have found and read about related taxa. Use **no more than two pages** of text, citing **no more than two references**.

Here are some ideas for topics, but feel free to choose your own:

1. What environment do relatives of your bacterium live in? Is there evidence of adaptation to this environment in the genome?
2. Based on its genome, is your bacterium auxotrophic for any amino acids? Are its closest relatives also auxotrophic for these?
3. Horizontal gene transfer is common among bacteria. Is there any evidence for HGT in your genome?
4. CRISPR-Cas9 is so hot right now. Does your genome have a CRISPR system? Can you determine where the spacer sequences originated from?
5. Does your genome encode any known bacteriocins, antibiotics, or toxins?
6. Make an argument for why or why not your bacterium would be considered a human pathogen, using the genome and your research as evidence.
7. Does your genome encode any known antibiotic resistance genes? Do you expect it to be susceptible to penicillin, tetracycline, or chloramphenicol?