Lab 3- Part 2
Annotating a Bacterial Genome

In this lab we'll be annotating a bacterial genome. When annotating a genome there are a few things we want to learn.

1. What are the genes in this organism?
2. What do these genes do?
3. What genetic pathways are represented in this organism?
4. How does this organism interact with the world?

To answer these questions, we are going to use the program Prokka [1]. Prokka is essentially a pipeline and optimization program. It uses other programs that identify components of a genome. The table below shows what programs it calls.

Program	Component of the genome annotated
Prodigal [2]	Coding sequences (CDS)
RNAmmer [3]	Ribosomal Genes (rRNAs)
Aragorn [4]	Transfer RNAs
SignalP [5]	Signal leader peptides
Infernal [6]	Non-coding RNA

Once the program Prodigal identifies the coding sequences in the genome, Prokka hierarchically searches using the BLAST algorithm to identify these coding sequences. It starts with genes that have protein or RNA evidence in literature. These genes are in the database UniProt [7]. It then shifts down to using genes in RefSeq identified in completed bacterial genomes. It then goes down to less reliable genes such as those in Pfam [8]. At each step in this process, the algorithm for identification gets slower. Therefore, it saves a lot of time to pull out as many sequences as you can quickly, and by matching to genes for which there is a lot of evidence.

Section 0:
Prokka was installed wrong by UW-IT, so we are going to reinstall it. It's pretty fast.

1. Go to https://github.com/tseeman/prokka
    a. If you are installing on your own Mac, use homebrew option if you have it
2. There is a green button labeled clone or download. Using it, download the Zip file
3. Once it's downloaded and unzipped, move the whole folder to the Applications folder. (If you don't do this, you'll need to change the directories below)
4. In terminal type
   $ /Applications/prokka-master/bin/prokka --setupdb

5. This will set up the databases we need for annotation.

Section 1:
Let's try running it.

1. Download the Annotate.zip file on the google drive

2. Open Terminal
3. In Terminal run Prokka as below
   $ /Applications/prokka-master/bin/prokka –outdir Sau –prefix Sau scaffolds.fasta

   This will take a while to run. This will make a new folder, Sau (that's what --outdir is telling it to do), and create a series of files that start with Sau.
   These files are described below.

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

   While we're waiting. We're going to look at two ways we annotate once we have genes. One is the gene ontology (GO) and the other is KEGG. The GO database organized genes by cellular locations, biological pathways, and molecular functions. The KEGG database is similar.

4. We are going to use the KEGG database to try to determine what pathways are active in the *Staphylococcus aureus* genome that we looked at during assembly last week. To do that go to: http://www.kegg.jp/blastkoala/

5. Input the Sau.faa in the upload fasta section. You will need to give an email, and then when you get the email they send, you will need to submit the job. It will take about an hour to get it back. What it's doing is BLASTing each of the protein sequences in the .faa file against the entire KEGG database, then classifying those that find a match with the biological pathway they belong in.
6. What types of functions does this bacteria seem to have pathways that enable it to do?
7. Are there any pathways that are populated in the KEGG pathways that seem surprising? Why are they surprising? Why do you think they are showing up in the annotation of the bacterial genome?

Section 2:
We are going to explore one of the important characteristics of bacteria, antibiotic resistance. We are going to compare a series of bacterial genomes and look for antibiotic resistance genes in them.

1. In the Annotate folder, you will find more fasta files. These need to be annotated with PROKKA, and the .faa files will be used to find antibiotic resistance genes. You may work with a partner to annotate all the genomes.
2. Once you have a completed annotation, you can start to look for antibiotic resistance genes. We are going to use The Comprehensive Antibiotic Resistance Database (CARD) to identify these genes [9]. This database stores information on all the genes that have been identified that confer antibiotic resistance.
3. Go to https://card.mcmaster.ca/
4. Select Analyze
5. Select RGI
6. Choose a .faa file, select Protein Sequence as the data type and submit the run.
7. In a few minutes you will be able to see the results visually.
8. For each bacterial genome or assembly, what is the most common antibiotic resistance gene?
9. From the data, are there resistance genes that are commonly found in multiple bacterial species? Are there ones that you only see in one species of bacteria?
10. Do you think the *Staphylococcus aureus* that we used for assembly last week is methicillin-resistant? What evidence did you use to make that decision?

References:

1. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
2. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
3. Lagesen, K. *et al.* RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
4. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32,** 11–16 (2004).
5. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–6 (2011).

6. Kolbe, D. L. & Eddy, S. R. Fast filtering for RNA homology search. *Bioinformatics* **27,** 3102–3109 (2011).
7. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32,** D115-9 (2004).
8. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Research* **42,** (2014).
9. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* gkw1004 (2016). doi:10.1093/nar/gkw1004