**Introduction to Next Generation Sequencing: Annotation**

**Introduction**

What is genome annotation? While it is valuable to have the nucleotide sequence of an organism’s genome, it is even more valuable to know what the different sections of that sequence code for. This is the basic idea of annotation. The goal is to annotate, at a minimum, all of the genes, their start and end positions, their introns, and their exons. If possible, annotation of promoter regions or major secondary structural elements, as well as common or important polymorphisms and mutations might also be useful.

How is annotation accomplished? The machinery of the cell helps us in our annotation efforts. If we remember the “central dogma”, DNA > RNA > Protein, there already exist in every cell DNA fragments including only the sequences of the genes. These are the transcribed messenger RNAs. The basic idea is to extract these messenger RNAs, convert them to DNA using reverse transcriptase, and sequence them (using the same NGS methods described in the first lab in this series). These DNA sequences are then aligned to the completed genome. Where they align, indicates where in the genome that gene is located.

In any given cell, different RNAs will be present because different proteins are necessary. Thus, in annotating a genome, it is important to extract RNA from roots, shoots, leaves, flowers, and fruits.

In the case of organellar genomes such as the chloroplast genome, where the gene content and even gene order are both well-known and generally well conserved among species, much of this process can be automated.

**Starting Files**

This workflow begins with the complete chloroplast genome we assembled last week, in FASTA format.

**Annotation**

There are a variety of programs that can do chloroplast genome annotations. We will be using GeSeq on the CHLOROBOX web platform. Navigate to the GeSeq website: https://chlorobox.mpimp-golm.mpg.de/geseq.html.

Upload the file, select a “linear” input sequence, and a “Plastid” sequence source. This will establish the defaults for a chloroplast annotation. As an option, scroll down and check the box next to “tRNAscan-SE” to add the functionality of that annotator.

As part of the default settings, the “Chloë” annotator will be checked and added to the process. I have found Chloë to be unreliable, and in some cases completely wrong. We might consider unchecking this box.

The annotator may run for some time, but is usually pretty quick for only one plastome. When complete, it will present a zipped folder to download, with all of the annotation files. Unzipping this folder, you will see that nine files have been generated (provided that the annotation conditions were as described in the above). Each sub-annotator will have its own output (files ending: \_blatN.txt, \_blatX.txt, \_Chloë v0.1.0.txt, and \_tRNAscan-SE v2.0.7). There will be an image of a graphical representation of the annotation (files ending: \_OGDRAW.jpg). Finally, there will be a series of summary files in different formats (files ending: \_GBSON.json, \_GenBank.gb, \_GFF3 + FASTA.gff3, and \_GFF3.gff3). With the exception of the image, all these files can be opened with a simple text editor.

We will take some time to go through these files.