Detecting variation

Today, we'll be comparing reads from a 2018 Trunk River metagenome to a MAG (metagenome-assembled genome) from the same metagenome to examine variation in the population. Since these steps require some computing power and time, I've pre-computed the full analysis and pulled out a subset as an example.

Step 0: Data acquisition

Make a new working directory in your group folder. Copy the starting data to the new folder and decompress it.

In [ ]:
```
cp /datahaus/variants/mini_example/student_data.zip .
```

In [ ]:
```
unzip student_data.zip
```

Step 1: read mapping

The first step is to map the metagenome reads to the reference "genome" - the contigs from the metagenomic assembly. Though we're only interested in one MAG for now, I've mapped reads to contigs from all the MAGs.

- *Why map all the reads to all the contigs? Why not just map to the MAG we're interested in?*

Knowing which reads map where, I've extracted a subset of reads that map to the MAG from bin28. We'll work with that for set for now.

First, build a bowtie index of the fasta file containing the bin contigs (`near_lem_2018_bin.28.fa`). `near_lem_bin28` will be the base name of the generated index files.

In [ ]:
```
bowtie2-build near_lem_2018_bin.28.fa near_lem_bin28
```

Now map the metagenome reads (`near_lem_2018_R1_bin28.fastq.gz` and `near_lem_2018_R2_bin28.fastq.gz`) to this reference using bowtie2. This may take ~15 minutes; `screen` might be useful.

In [ ]:
```
bowtie2 --local -p 4 -x near_lem_bin28 \
-1 near_lem_2018_R1_bin28.fastq.gz -2 near_lem_2018_R2_bin28.fastq.gz \
-S near_lem_bin28_vs_near_lem_bin28.sam --no-unal
```

Then convert the .sam file to a .bam file.

In [ ]:
```
samtools view -b near_lem_bin28_vs_near_lem_bin28.sam \
> near_lem_bin28_vs_near_lem_bin28.bam
```

Step 2: Gene Predictions

InStrain can give us a gene-level analysis of variation if we provide it a gene file. We'll use prodigal to predict genes. (If you tried prokka during the genome assembly & annotation workshop, prodigal is what prokka uses to predict genes).

In [ ]:
```
prodigal -a near_lem_bin28.faa -d near_lem_bin28.fna \
-f gff -i near_lem_2018_bin.28.fa -o near_lem_bin28.gff
```

Step 3: Find variable positions

Now, use inStrain's profile function to compare the reads to the reference genome (~15 mins for a single genome, longer for metagenomes).

In [ ]:
```
inStrain profile -o instrain_near_lem_bin28 -p 4 \
-g near_lem_bin28.fna \
near_lem_bin28_vs_near_lem_bin28.bam near_lem_2018_bin.28.fa
```

InStrain's output will go to the directory `instrain_near_lem_bin28`. The `output` folder contains several .tsv files, which are described in detail here: https://instrain.readthedocs.io/en/latest/example_output.html .

Note: all positions in these files are numbered starting from 0. Be careful when comparing to 1-based genome positions!

Briefly, *_SNVs.tsv contains a list of all single-nucleotide variants (SNVs). *gene_info.tsv, *scaffold_info.tsv, and *genome_info.tsv contain summaries of variation at the gene, scaffold, and genome (in our case, MAG) level. *linkage.tsv contains info about how frequently SNVs occur on the same read.

- *What type of variation would linkage information be useful for detecting?*

These files are all tab-separated text files and can be downloaded to your computer and opened in your favorite spreadsheet program.

Step 4: Exploration

Bin28 has been identified as purple nonsulfur bacterium from the genus *Rhodovulum*.

Some questions to ask:

- How many SNVs are present in the genome?
- Which genes in the Rhodovulum MAG have the most SNVs? The least? What do they do? (BLAST is your friend)
- Are there regions with multiple highly-linked SNVs?