SAGESA AMR Bioinformatics Africa Course: 25 – 27 May 2022

Genome annotation hands-on exercise
Instructor: Stan Kwenda

This hands-on exercise will be done using the Linux command-line interface (CLI) and a popular bacterial genome annotation tool called Prokka. Prokka is already set-up in your environment and the assemblies are also already available. We will only focus on annotating protein coding genes.

1. List the options available to use when performing annotation with Prokka.
   prokka --help

2. How many databases are pre-installed?

   prokka --listdb

   *Prokka comes with a set of databases for the most common Bacterial genera

3. Perform annotations using prokka
   First change directory to the Genome Assemblies directory:
   cd ~/Genome Assemblies

   Now execute the following (It should take about 4 min to execute each annotation):

   a. prokka --outdir ./Ecoli-E37364 --force --prefix EcE37364 --compliant --genus Escherichia --species coli --strain EcE37364 --kingdom Bacteria --usegenus --cpus 7 --norrna --notrna ./Assemblies/Ecoli/Ecoli-E_37364.fasta > ./Ecoli-E37364/prk_raw.log 2>&1 &

   *At the end you should have 12 output files that prokka generates. Repeat above step for isolates Ecoli-A38843 and Styphi-B_38329. You can use the commands below:

   b. prokka --outdir ./Ecoli-A38843 --force --prefix EcA38843 --compliant --genus Escherichia --species coli --strain EcA38843 --kingdom Bacteria --usegenus --cpus 7 --norrna --notrna ./Assemblies/Ecoli/Ecoli-A_38843.fasta > ./Ecoli-A38843/prk_raw.log 2>&1 &

   c. prokka --outdir ./Styphi-B_38329-02 --force --prefix StB38329 --genus Salmonella --species typhi --strain StB38329 --kingdom Bacteria --usegenus --cpus 7 --rawproduct --norrna --notrna ./Assemblies/Styphi/Styphi-B_38329.fasta > ./Styphi-B_38329-02 /prk_raw.log 2>&1

   You can quickly explore the generated output files using any of the following commands, e.g.

   head -n 20 filename (e.g. head -n20 ./Ecoli-E37364/EcE37364.txt)
   tail -n20 filename (e.g. tail -n20 ./Ecoli-E37364/EcE37364.tsv)
   more filename (e.g. more ./Ecoli-E37364/EcE37364.tsv) # hit q to exit

*It's usually a good idea to quickly glance through the log file and ensure that everything ran as expected. You can do this using either:

```
more ./Ecoli-E37364/EcE37364.log        # hit q to exit
less -S ./Ecoli-E37364/EcE37364.log     # hit q to exit
```

4. How many genes (i.e. protein coding genes) were predicted?
   *This can be achieved by looking at the summary/statistics file which summarizes number of features that have been annotated. You can use the commands below:

   ```
   grep "CDS" ./Ecoli-E37364/EcE37364.txt
   cat ./Ecoli-E37364/EcE37364.txt          # This will printout the whole file to screen
   ```

5. How many contigs are present in this assembly?
   Using the summary file you can quickly get the number of contigs from the assembly file used for predicting genes.

   ```
   grep "contigs" ./Ecoli-E37364/EcE37364.txt
   cat ./Ecoli-E37364/EcE37364.txt          # This will printout the whole file to screen
   ```

   If --compliant option was used this number will represent only contigs greater than 200 bp in length. Have a look at the prokka options to understand why this is done or necessary.

6. How many hypothetical proteins were predicted?

   ```
   grep -c -i "hypothetical protein" ./Ecoli-E37364/EcE37364.tsv
   ```

   - **Hypothetical proteins** are predicted proteins with unknown functions (i.e. not yet recognized) predicted to be expressed from an open reading frame (ORF) but lacking any experimental evidence of translation.
   - **Conserved hypothetical proteins** are proteins that are conserved among organisms from several phylogenetic lineages but without functional validation

7. Are there any conserved hypothetical proteins?

   ```
   grep -c -i "conserved hypothetical" ./Ecoli-E37364/EcE37364.tsv
   ```

8. How many genes were annotated as putative?

   ```
   grep -c -i "putative" ./Ecoli-E37364/EcE37364.tsv
   ```

To list the putative genes:

grep -c -i "putative" ./Ecoli-E37364/EcE37364.tsv | less

- **Putative genes** can share sequence similarities to already characterized genes and thus can be inferred to share a similar function, yet the exact function of putative genes remains unknown.
- **Probable protein** is a protein exhibiting extensive sequence similarity to a characterized protein or conserved region
- **Unknown protein** are structures that have been experimentally shown to exist but are not characterized in protein chemical terms or cannot be linked to a known gene e.g. uncharacterized protein families.
- **Proteins of unknown functions** are experimentally documented but no known functional or structural domain is observed

*If --rawproduct is not set, Prokka will clean gene product names (protein names) comprising terms like possible, probable, predicted etc and name them as putative.

9. How many protein coding genes were annotated?

   egrep "[[:space:]]CDS[[:space:]]" ./Ecoli-E37364/EcE37364.tsv | egrep -c -i -v "hypothetical|putative"

Group discussion (Optional)

1. Why are hypothetical sequences pervasive across bacterial species?
2. What proportions of genes in different bacterial genomes remain unannotated?
3. What factors affect annotation completeness?

Group task (Optional)

Now that we have successfully annotated our genome(s) using Prokka, let's try to do some functional annotations, i.e. gene ontology enrichment analysis. However, note that this enrichment analysis works best when one has gene expression data such as is generated using RNA-seq. For this optional tutorial we will use an online tool, ShinyGO, which can be accessed at the following url: http://bioinformatics.sdstate.edu/go/

1. First generate a list of genes from the one of the output files from Prokka (filename with .tsv extension).
   cat EcE37364.tsv | cut -f4 | sort -u | sed '/^$/d' | head -n50 | tr '\n' ',' > shinygo_list1.txt

cat shinygo_list1.txt # copy and paste the genes into the input tab on shinyGO

2. Click the gene tab, and check the “Detailed Description” box. If any of the genes in your list are involved in previously annotated pathways and/or biological processes, these will be provided under the Description column.
3. Since we provided E . coli genes, it’s good to first check if the results provided are for your particular species, if not then select your specific species in the dropdown and resubmit your genes
4. You can explore results provided in the other tabs, e.g. Enrichment, Chart etc., however, remember that results from these tabs will be most useful if our gene list was from a gene expression experiment.

---------- END ----------