Downstream analysis

A pipeline of the Vonaesch Lab for the curnagl cluster

Introduction

- This bioinformatic pipeline slides attempts to describe the steps to performs various downstream analysis using assembled genomes from both short and long reads sequencing of bacterial isolates.
- Analysis
  - Gene annotations
  - Comparative genomic
  - Strain diversity (fastANI)

Annotation with Prokka

Prokka is a software tool to annotate bacterial, archaeal and viral genomes genomes quickly and produce standards-compliant output files.

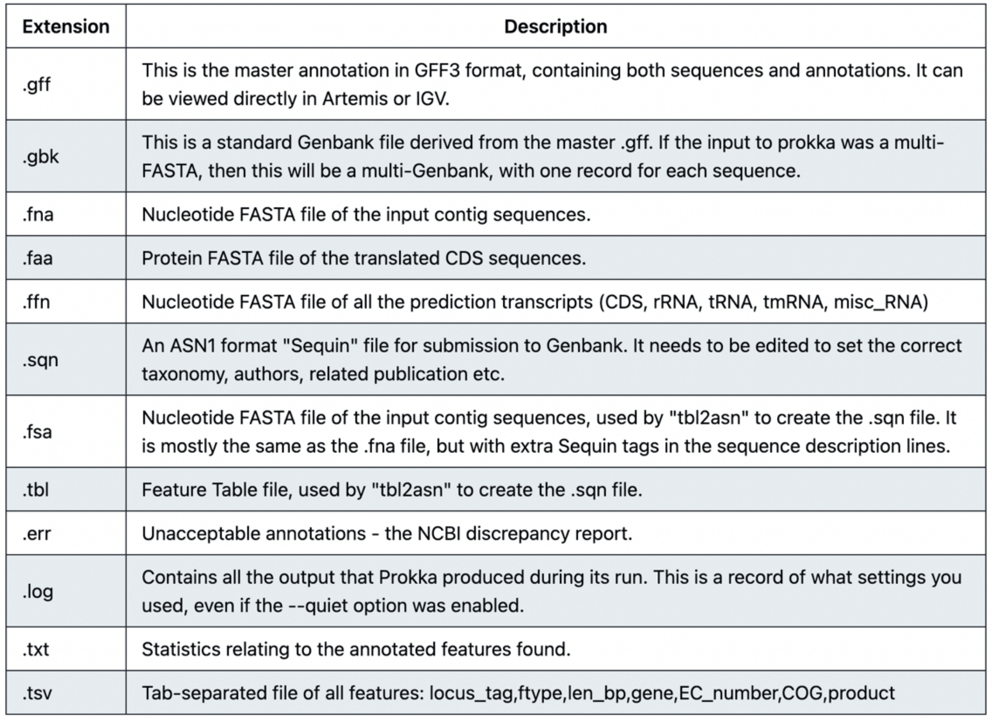
- To load Prokka correctly, we need to set up the databases, if necessary, follow the instruction on:

  install_prokka.sh

- Annotate your genomes using *Prokka*:

  📜 07_prokka.sh

  - Prokka produce a set of files per genomes:



Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag,ftype,len_bp,gene,EC_number,COG,product

Extract 16S rRNA gene

- If necessary, the 16S rRNA genes (as well as the 5S and 23S) can be extracted from our assemblies, for taxonomic assignation or comparative genomic (with ASVs of the community analysis).
- identify and extract the 16S rRNA gene, use *Barrnap*:
  - barrnap.sh
- Barrnap first identify the genes and produce a feature file .gff. Then, it extract the gene and save it in a fasta file that can be downloaded and blast for identification.

Annotation with DRAM

- DRAM is an annotation software that calls genes in our genomes, and annotate the translated proteins using different database, in our case, kofam (KEGG), pfam, dbCAN (CAZymes) and peptidases.
- This script uses the conda environment dram. It is activated at the start of the script.
- Annotate your genomes with *DRAM* using:

  📜 09_dram.sh
- Download the following files for further analysis:

  dram_annotations/annotations.tsv

  genome_summaries/product.html

Strain diversity with fastANI

- FastANI is a fast alignment-free computation of whole genome Average Nucleotide Identity (ANI). ANI is defined as mean nucleotide identity of orthologous gene pairs shared between two microbial genomes. You can run pariwise comparison of your genome using:

```
fastANI –q query_genone.fasta –r ref_genome.fasta –o name_outpute_file
```

- Or run many to many genomes pairwise comparison, using a text file with the list of the genomes to compare, using:

```
fastANI --ql list_of_genome --rl list_of_genome –o name_outpute_file
```

- fastANI is reporting the ANI for pair genomes in the output file:
  - Column 1: query genome
  - Column 2: reference genome
  - Column 3: ANI
  - Column 4: number of fragment that aligned successfully on the total number of fragment in the query
  - Column 5: total number of fragments in the query genomes