Prokka

Annotation

Whole genome annotation maps and labels interesting parts of a genome sequence. We'll be annotating genes.

What ways are you aware of for finding genes in DNA sequence?

Prokka[1,2]

Prokka is a tool for quickly annotating genomes of bacteria, archaea, and viruses.

What are some broad differences between these organisms?

Prokka finds genes by comparing to a database of known genes. Take a look at the gene databases that are built in to Prokka.

```
prokka --listdb
```

Our Metagenome Assemblies

We'll use the genome assemblies made from reads from which red alder sequence was removed. In other words, we expect that these assemblies just have microbial DNA.

You probably made your own assembly so you can link to those, or you can link to the version below. Let's get them all together in a file.

```
mkdir ~/annotation
cd ~/annotation

ln -s /home/elavelle/canuMinit3/4956-3.contigs.fasta .
```

```
ln -s /home/elavelle/canuMinit4/3469-3.contigs.fasta .
ln -s /home/elavelle/CanuRun/AlderFiltered.contigs.fasta both_samples.fasta .
```

How many sequences do we have in each assembly?

What is the longest and shortest contig in each assembly?

What kinds of organisms do we have in our assembly based on your work with classifying reads with KrakenUniq?

Bacterial Genomes -- Some Context

Let's get a little context.

How long are bacterial genomes on average?

What range do bacterial genome sizes cover?

What is the average size of bacterial genes?

What is the average number of genes in a bacterial genome?

What percentage of bacteria genomes is typically covered by protein-coding genes? (hint: look up bacterial average gene density)

Bonus: What percentage of the human genome is covered by protein-coding genes?

Compare what you learned about bacterial genomes to what you know about our genome assemblies. Do you expect all of the contigs to contain full length genes?
Do you expect that we will find partial genes?
Do you expect that some contigs won't have any genes?
If we don't find a gene in a region does that mean that there isn't a gene there? Why or why not?

Running Prokka

Go into a screen and then activate the seqtools environment so that we can access prokka.

```
screen -S prokka
conda activate seqtools
```

Here are the parameters that we are going to use:
--metagenome
--addgenes
--addmrna
--kingdom Bacteria
--gcode 11
--genus Frankia
--prefix 4956-3
--outdir prokka4956-3

You can probably mostly guess what they mean, but it is a good idea to check out how the program defines them. You can just type "prokka" or "prokka --help" to get the usage note that tells you how to run the program and what each parameter does. Note that typing a program name followed by "--help" will get you the usage for most programs.

```
prokka
```

Now, let's get it running.

```
prokka --metagenome --addgenes --addmrna \
    --kingdom Bacteria --gcode 11 --genus Frankia \
    --prefix 4956-3 --outdir prokka4956-3 \
    4956-3.contigs.fasta
```

Now run it for the other two assemblies.

Annotation Output

Here is a list of the output files. We'll focus on the GFF file.

Table 2. Description of Prokka output files[2]

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

Let's take a look at the GFF file format.

GFF format[3]

1. **seqid** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seq ID must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.

2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **type** - type of feature. Must be a term or accession from the SOFA sequence ontology
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **phase** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attributes** - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent - see the GFF documentation for more details.

Let's look at the GFF file.

```
less prokka4956-3/4956-3.gff
```

Let's find out how many genes there are.

```
awk '$3~/gene/{print}' prokka4956-3/4956-3.gff
```

How many genes are there in the other two gff files (for the other sample's assembly and for the merged assembly?)

CDS stands for CoDing Sequence. It is the portion of the gene that gets turned into proteins. The CDS lines have functional information.

How many CDSs are there?

What does "hypothetical protein" mean? How many hypothetical proteins are there?

Find a gene function other than "hypothetical protein". Look up some information on it and share it with the group.

Can you find any overlapping genes? How can this happen?

References

1. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30(14):2068-9. https://academic.oup.com/bioinformatics/article/30/14/2068/2390517
2. https://github.com/tseemann/prokka
3. https://uswest.ensembl.org/info/website/upload/gff3.html