Genome Annotation with prokka (prokaryote genome annotation)

Adam m France afrance3@uncc.edu

Using the contigs that were assembled from the previous lab(s) we are going to use prokka to annotate the genome in order to find the regions of the genome that code for genes and proteins. The annotations will be analyzed both qualitatively using the IGV genome browser and quantitatively using the grep command.

Prokka was installed on a local PC using homebrew.

Methods:

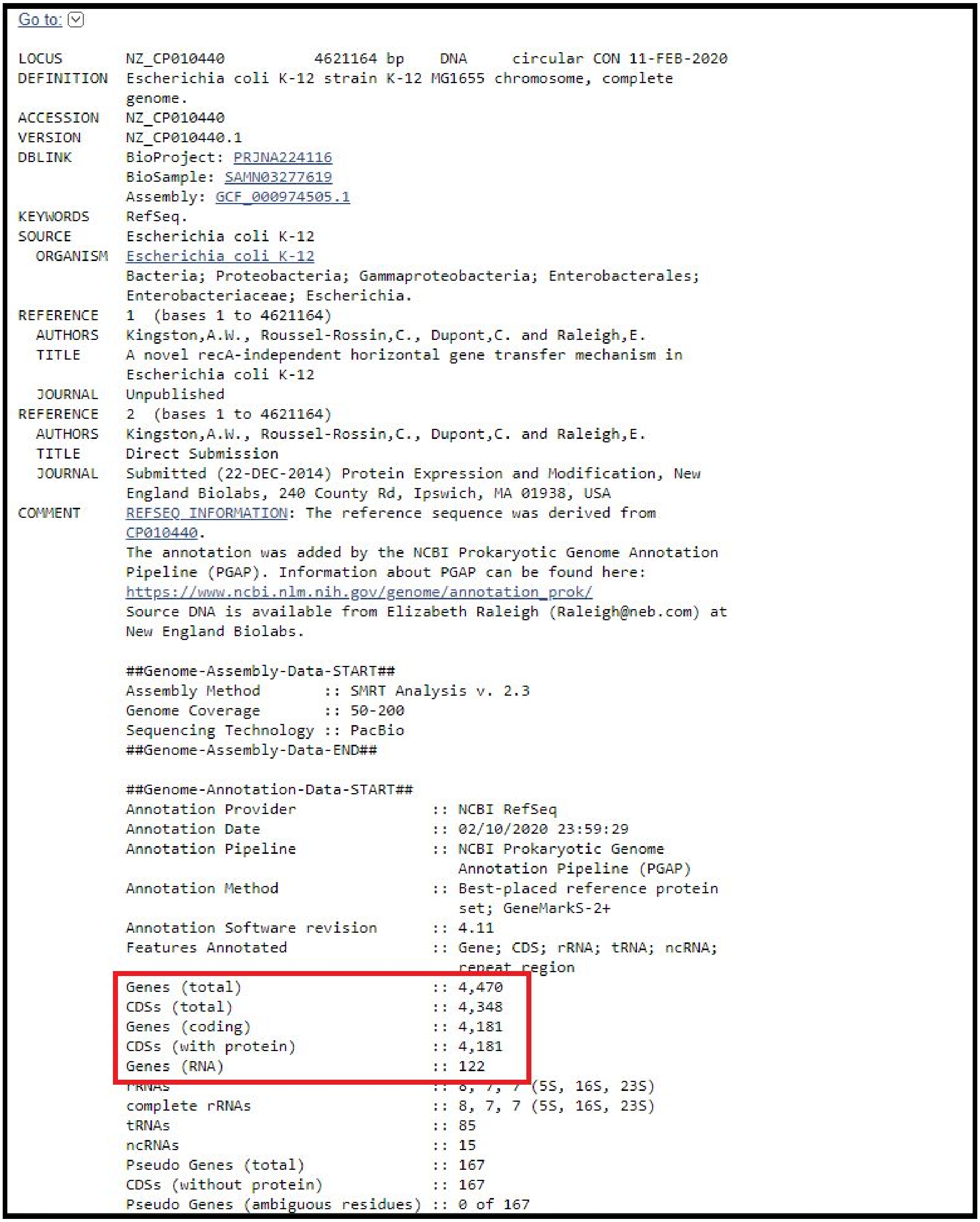
1. Ran the setupdb command from porkka.
2. Ran listdb command from prokka to confirm that e.coli would be in the listed genera.
3. Ran prokka on the contigs file produced from the assembly lab.
4. Used the grep linux command on the .faa file to count the number of ‘>’ , which would tell us how many genes prokka found.
5. USed the grep linux command on the .faa file to count the number of hypothetical proteins.
6. Visualized annotations in IGV browser on PC.

Results:

We see that the number of genes roughly the same (< 100 difference)

File	Genes	Hypothetical Proteins
Genbank page	4347	636
NCBI reference	4242	4
With pac bio (clr)	4313	479
With pac bio (ccs)	4258	482
Without pac bio	4305	475

Table 1.



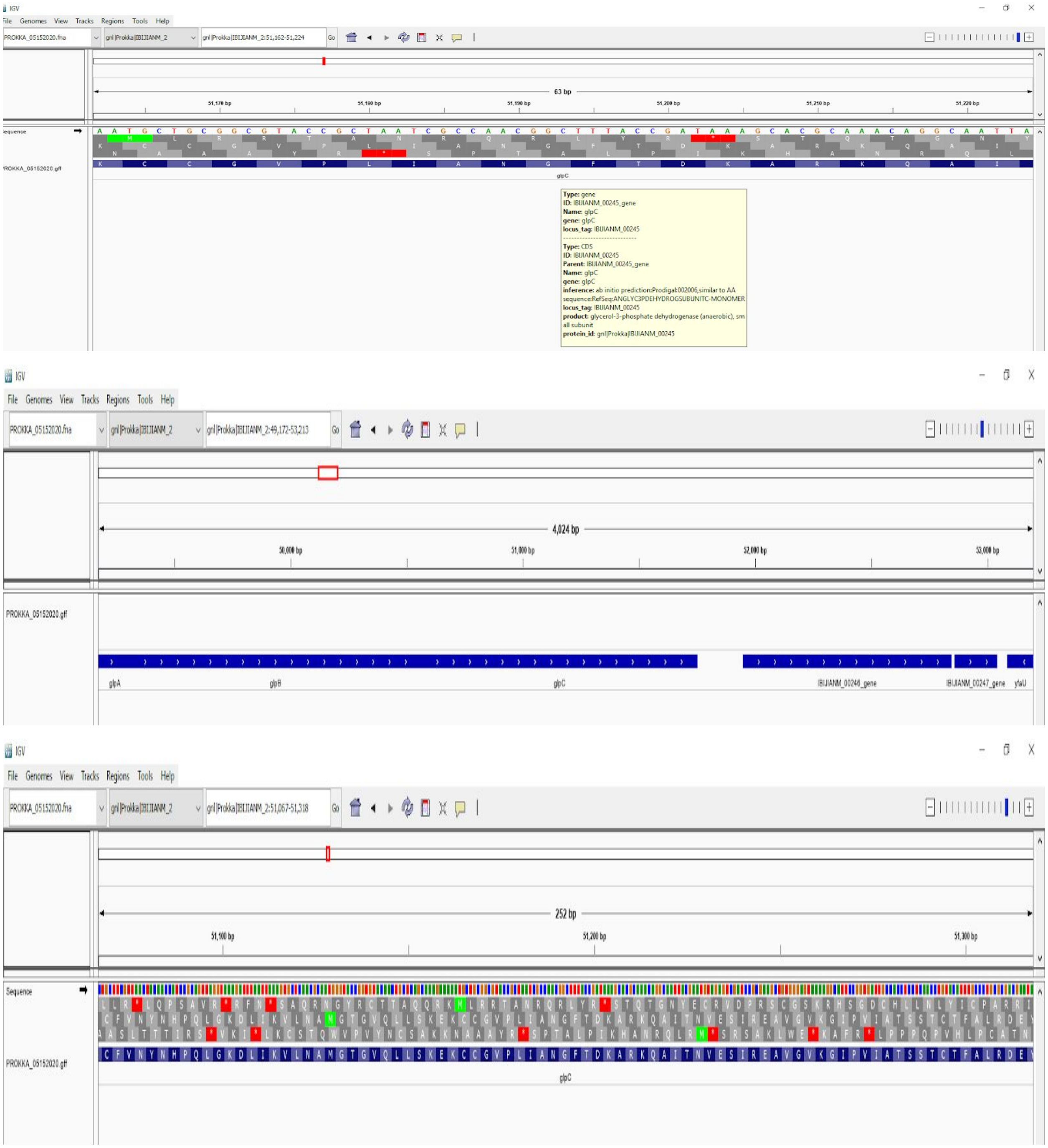
```
Go to:

LOCUS       NZ_CP010440          4621164 bp    DNA     circular CON 11-FEB-2020
DEFINITION  Escherichia coli K-12 strain K-12 MG1655 chromosome, complete
            genome.
ACCESSION   NZ_CP010440
VERSION     NZ_CP010440.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN03277619
            Assembly: GCF_000974505.1
KEYWORDS    RefSeq.
SOURCE      Escherichia coli K-12
  ORGANISM  Escherichia coli K-12
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4621164)
  AUTHORS   Kingston,A.W., Roussel-Rossin,C., Dupont,C. and Raleigh,E.
  TITLE     A novel recA-independent horizontal gene transfer mechanism in
            Escherichia coli K-12
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 4621164)
  AUTHORS   Kingston,A.W., Roussel-Rossin,C., Dupont,C. and Raleigh,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-DEC-2014) Protein Expression and Modification, New
            England Biolabs, 240 County Rd, Ipswich, MA 01938, USA
COMMENT     REFSEQ INFORMATION: The reference sequence was derived from
            CP010440.
            The annotation was added by the NCBI Prokaryotic Genome Annotation
            Pipeline (PGAP). Information about PGAP can be found here:
            https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
            Source DNA is available from Elizabeth Raleigh (Raleigh@neb.com) at
            New England Biolabs.

            ##Genome-Assembly-Data-START##
            Assembly Method       :: SMRT Analysis v. 2.3
            Genome Coverage       :: 50-200
            Sequencing Technology :: PacBio
            ##Genome-Assembly-Data-END##

            ##Genome-Annotation-Data-START##
            Annotation Provider               :: NCBI RefSeq
            Annotation Date                   :: 02/10/2020 23:59:29
            Annotation Pipeline               :: NCBI Prokaryotic Genome
                                                 Annotation Pipeline (PGAP)
            Annotation Method                 :: Best-placed reference protein
                                                 set; GeneMarkS-2+
            Annotation Software revision      :: 4.11
            Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                                 repeat_region
            Genes (total)                     :: 4,470
            CDSs (total)                      :: 4,348
            Genes (coding)                    :: 4,181
            CDSs (with protein)               :: 4,181
            Genes (RNA)                       :: 122
            rRNAs                             :: 8, 7, 7 (5S, 16S, 23S)
            complete rRNAs                    :: 8, 7, 7 (5S, 16S, 23S)
            tRNAs                             :: 85
            ncRNAs                            :: 15
            Pseudo Genes (total)              :: 167
            CDSs (without protein)            :: 167
            Pseudo Genes (ambiguous residues) :: 0 of 167
```

Figure 1. A e.coli k-12 assembly/annotation from genbank, in the box highlighted in red we see that the total number of predicted Genes is 4,470.



PROKKA_05152020.fna | gnl|Prokka|IBIJIANM_2 | gnl|Prokka|IBIJIANM_2:51,162-51,224

Type: gene
ID: IBIJIANM_00245_gene
Name: glpC
gene: glpC
locus_tag: IBIJIANM_00245

Type: CDS
ID: IBIJIANM_00245
Parent: IBIJIANM_00245_gene
Name: glpC
gene: glpC
inference: ab initio prediction:Prodigal:002006,similar to AA sequence:RefSeq:ANGLYC3PDEHYDROGSUBUNITC-MONOMER
locus_tag: IBIJIANM_00245
product: glycerol-3-phosphate dehydrogenase (anaerobic), sm all subunit
protein_id: gnl|Prokka|IBIJIANM_00245

PROKKA_05152020.fna | gnl|Prokka|IBIJIANM_2 | gnl|Prokka|IBIJIANM_2:49,172-53,213

PROKKA_05152020.fna | gnl|Prokka|IBIJIANM_2 | gnl|Prokka|IBIJIANM_2:51,067-51,318

Figure 2. The IGV browser loaded with the .gff file from the prokka annotation and the reference genome for e.coli k-12. The three screenshots are at different ‘magnifications’ of the genome, with the base pair range indicated near the top. At the very top picture we can see the individual codons, with the amino acid letters being indicated as you zoom out. The blue bars represent genes and their arrows the direction of transcription. Moving the mouse over one of these bars will bring up a description of that gene.

Discussion:

If we look in table 1. We see that the number of genes predicted from the refseq on gen bank (4,470) is roughly the same number we get from the files produced by our PROKKA runs. Each of the contig files searched had about 100 less genes than the 4,470 listed on genbank. The hypothetical proteins are interesting because by using grep on the NCBI file, the result was only 4, so there must be some difference in formatting between files for there to be such a discrepancy. Also from the genbank webpage we get roughly 200 more hypothetical proteins, but the method used to obtain that number was CTRL+F , and not grep, so again there may be an issue with formatting. It should also be noted that the two ncbi annotations are from different assemblies than what I used here, even though they are all of the same strain (e. Coli k12).

Link for ncbi assembly:
https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2/

Commands:

```
$ brew install prokka
$ brew install artemis

$ prokka --setupdb

$ prokka --listdb
[17:57:54] Looking for databases in: /home/linuxbrew/.linuxbrew/Cellar/prokka/1.14.6/db
[17:57:54] * Kingdoms: Archaea Bacteria Mitochondria Viruses
[17:57:54] * Genera: Enterococcus Escherichia Staphylococcus
[17:57:54] * HMMs: HAMAP
[17:57:54] * CMs: Archaea Bacteria Viruses

$ prokka contigs.fasta --outdir contigs1 --force --compliant --genus Escherichia --usegenus
$ prokka contigs_ccs.fasta --outdir ccs_cont --force --compliant --genus Escherichia --usegenus
$ prokka contigs_clr.fasta --outdir clr_cont --force --compliant --genus Escherichia --usegenus

$ grep '>' PROKKA_05152020.faa | wc                        4258   20781  222670

$ grep 'hypothetical protein' PROKKA_05152020.faa | wc     482    1469   17970

$ grep '>' PROKKA_05152020.faa | wc                        4313   21081  225489

$ grep 'hypothetical protein' PROKKA_05152020.faa | wc     479    1460   17859

$ grep 'hypothetical protein' GCF_000005845.2_ASM584v2_protein.faa | wc
4      40     339
```

```
$ grep '>' GCF_000005845.2_ASM584v2_protein.faa | wc
4242   46230  399675
```