B2 Gene Prediction & Annotation: Preliminary Results

By: Celine Al-Noubani, Vishank Raghavan, Kyungbeom Kim, Sizhe Fang, Xuejiao (Jessica) Yuan

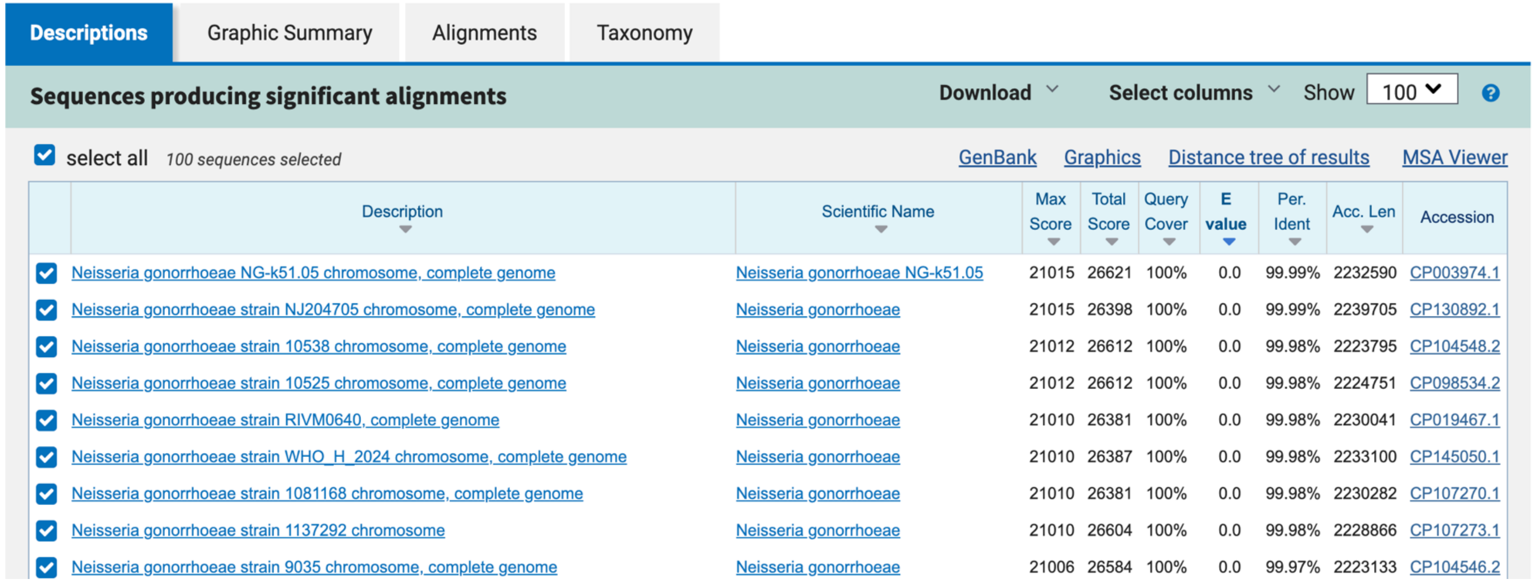
Georgia Tech

Breakdown & Student Roles

- Gene Prediction
    - GeneMarkS-2 (Jessica)
    - Glimmer (Kyungbeom Kim)
    - Prodigal (Vishank)
    - Augustus (Sizhe)
    - Barrnap (Vishank)
    - GeMoMa (Celine)
    - BLAST (Sizhe)
- Decision on Merging (Sizhe)
- Gene Annotation
    - InterPro (Celine)
    - EggNog (Kyungbeom)
- Standardization (Vishank)

Reference Genome

- Genome assembly consists of contigs from an unknown bacterial species.
- Randomly selected 5 sequences from the contigs.
- Conducted BLASTN searches against the NCBI database.
- Identified as belonging to ***Neisseria gonorrhoeae***.



Sequences producing significant alignments

☑ select all *100 sequences selected*

GenBank Graphics Distance tree of results MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Neisseria gonorrhoeae NG-k51.05 chromosome, complete genome	Neisseria gonorrhoeae NG-k51.05	21015	26621	100%	0.0	99.99%	2232590	CP003974.1
☑	Neisseria gonorrhoeae strain NJ204705 chromosome, complete genome	Neisseria gonorrhoeae	21015	26398	100%	0.0	99.99%	2239705	CP130892.1
☑	Neisseria gonorrhoeae strain 10538 chromosome, complete genome	Neisseria gonorrhoeae	21012	26612	100%	0.0	99.98%	2223795	CP104548.2
☑	Neisseria gonorrhoeae strain 10525 chromosome, complete genome	Neisseria gonorrhoeae	21012	26612	100%	0.0	99.98%	2224751	CP098534.2
☑	Neisseria gonorrhoeae strain RIVM0640, complete genome	Neisseria gonorrhoeae	21010	26381	100%	0.0	99.98%	2230041	CP019467.1
☑	Neisseria gonorrhoeae strain WHO_H_2024 chromosome, complete genome	Neisseria gonorrhoeae	21010	26387	100%	0.0	99.98%	2233100	CP145050.1
☑	Neisseria gonorrhoeae strain 1081168 chromosome, complete genome	Neisseria gonorrhoeae	21010	26381	100%	0.0	99.98%	2230282	CP107270.1
☑	Neisseria gonorrhoeae strain 1137292 chromosome	Neisseria gonorrhoeae	21010	26604	100%	0.0	99.98%	2228866	CP107273.1
☑	Neisseria gonorrhoeae strain 9035 chromosome, complete genome	Neisseria gonorrhoeae	21006	26584	100%	0.0	99.97%	2223133	CP104546.2

	RefSeq	GenBank
Provider	NCBI RefSeq	WTSI
Name	GCF_900087635.2-RS_2025_02_13	Annotation submitted by WTSI
Date	Feb 13, 2025	Sep 15, 2016
Genes	2,431	2,351
Protein-coding	2,114	2,351

Gene Prediction Tools: Ab Initio

We explored several options:

GeneMarkS-2

- Suited for small bacterial genomes

GLIMMER

- Uses Markov models to find open reading frames

Augustus

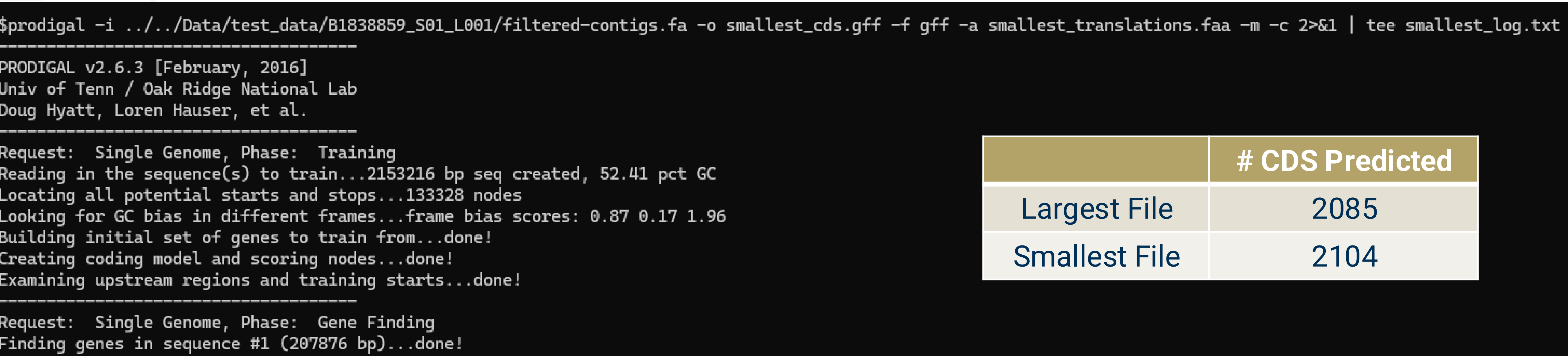
- Applies Hidden Markov Models, mainly for eukaryotes but adaptable to prokaryotes

We are focusing on:

Prodigal

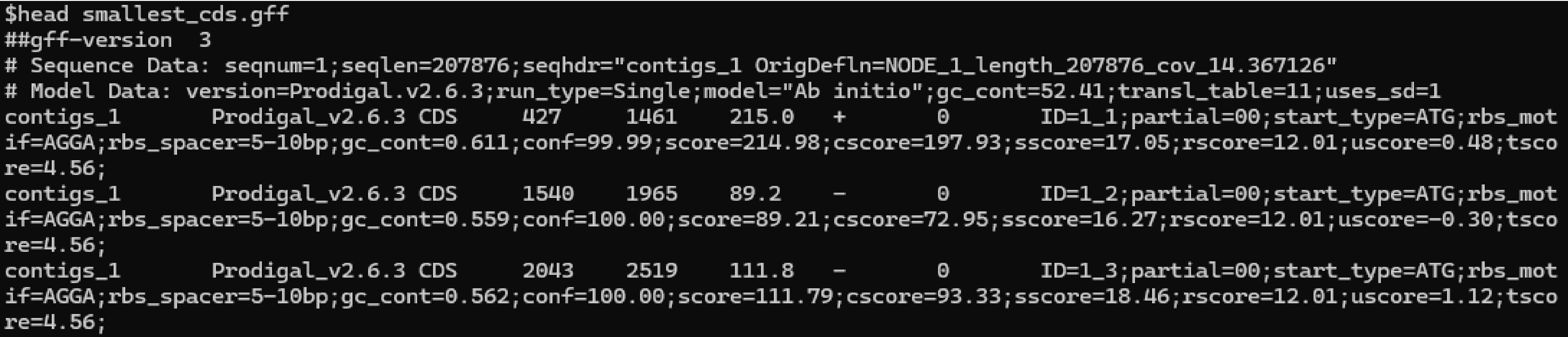
- Optimization for microbial genomes
- Ability to work in both standard and metagenomic modes
- Speed and accuracy, especially with fragmented assemblies

Prodigal: Command & Output



```
$prodigal -i ../../Data/test_data/B1838859_S01_L001/filtered-contigs.fa -o smallest_cds.gff -f gff -a smallest_translations.faa -m -c 2>&1 | tee smallest_log.txt
-------------------------------------
PRODIGAL v2.6.3 [February, 2016]
Univ of Tenn / Oak Ridge National Lab
Doug Hyatt, Loren Hauser, et al.
-------------------------------------
Request:  Single Genome, Phase:  Training
Reading in the sequence(s) to train...2153216 bp seq created, 52.41 pct GC
Locating all potential starts and stops...133328 nodes
Looking for GC bias in different frames...frame bias scores: 0.87 0.17 1.96
Building initial set of genes to train from...done!
Creating coding model and scoring nodes...done!
Examining upstream regions and training starts...done!
-------------------------------------
Request:  Single Genome, Phase:  Gene Finding
Finding genes in sequence #1 (207876 bp)...done!
```

	# CDS Predicted
Largest File	2085
Smallest File	2104



```
$head smallest_cds.gff
##gff-version  3
# Sequence Data: seqnum=1;seqlen=207876;seqhdr="contigs_1 OrigDefln=NODE_1_length_207876_cov_14.367126"
# Model Data: version=Prodigal.v2.6.3;run_type=Single;model="Ab initio";gc_cont=52.41;transl_table=11;uses_sd=1
contigs_1       Prodigal_v2.6.3 CDS     427     1461    215.0   +       0       ID=1_1;partial=00;start_type=ATG;rbs_mot
if=AGGA;rbs_spacer=5-10bp;gc_cont=0.611;conf=99.99;score=214.98;cscore=197.93;sscore=17.05;rscore=12.01;uscore=0.48;tsco
re=4.56;
contigs_1       Prodigal_v2.6.3 CDS     1540    1965    89.2    -       0       ID=1_2;partial=00;start_type=ATG;rbs_mot
if=AGGA;rbs_spacer=5-10bp;gc_cont=0.559;conf=100.00;score=89.21;cscore=72.95;sscore=16.27;rscore=12.01;uscore=-0.30;tsco
re=4.56;
contigs_1       Prodigal_v2.6.3 CDS     2043    2519    111.8   -       0       ID=1_3;partial=00;start_type=ATG;rbs_mot
if=AGGA;rbs_spacer=5-10bp;gc_cont=0.562;conf=100.00;score=111.79;cscore=93.33;sscore=18.46;rscore=12.01;uscore=1.12;tsco
re=4.56;
```

Gene Prediction Tools: Homology-based

BLAST

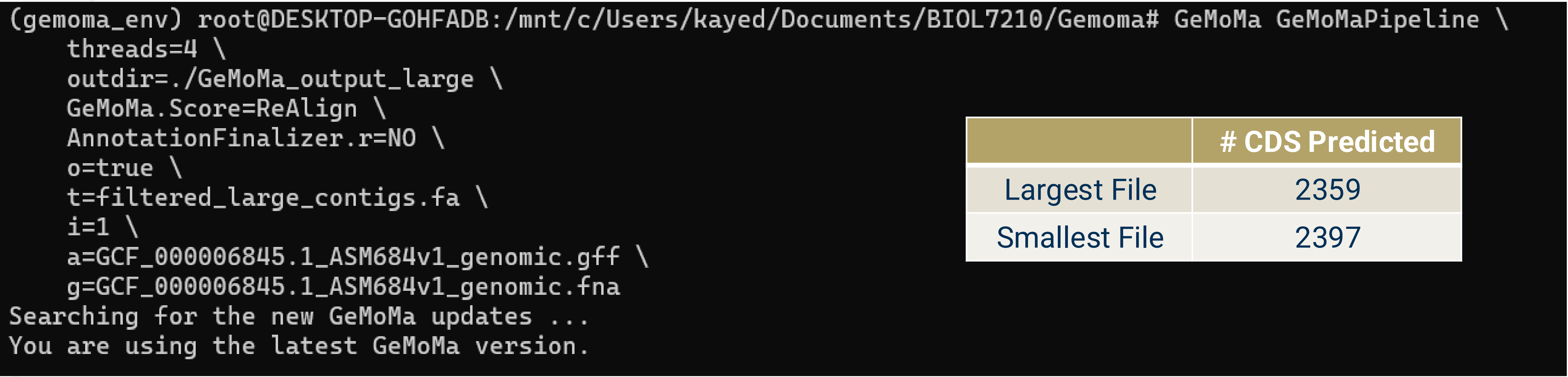
- Uses tblastn to compare reference proteins to our genome

We are focusing on →

GeMoMa

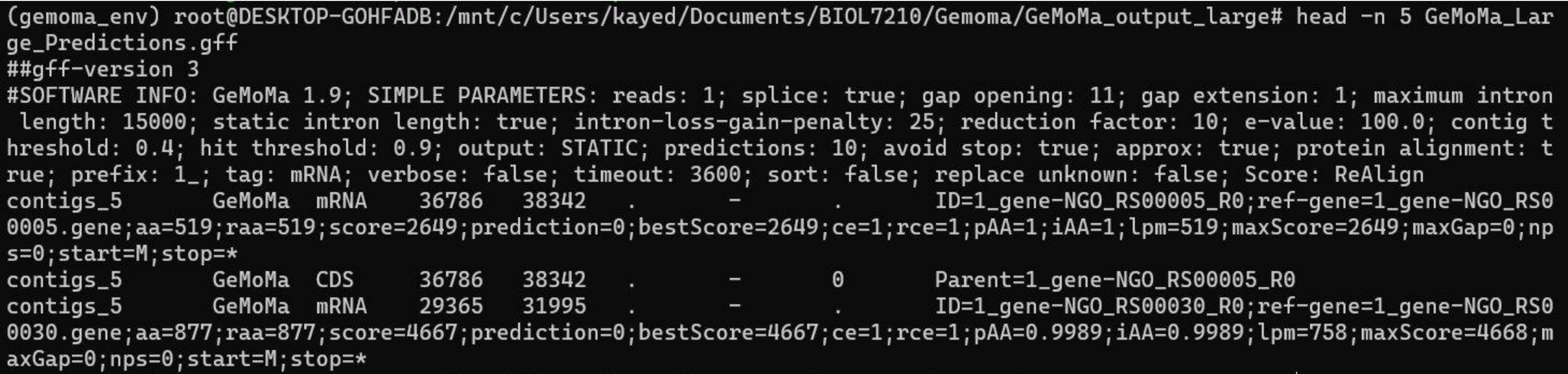
- Identifies coding sequences by comparing proteins from a reference genome
- Looks at conserved regions for better accuracy
- Allows refinement of gene boundaries

GeMoMa: Command & Output



```
(gemoma_env) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/Gemoma# GeMoMa GeMoMaPipeline \
    threads=4 \
    outdir=./GeMoMa_output_large \
    GeMoMa.Score=ReAlign \
    AnnotationFinalizer.r=NO \
    o=true \
    t=filtered_large_contigs.fa \
    i=1 \
    a=GCF_000006845.1_ASM684v1_genomic.gff \
    g=GCF_000006845.1_ASM684v1_genomic.fna
Searching for the new GeMoMa updates ...
You are using the latest GeMoMa version.
```

	# CDS Predicted
Largest File	2359
Smallest File	2397



```
(gemoma_env) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/Gemoma/GeMoMa_output_large# head -n 5 GeMoMa_Lar
ge_Predictions.gff
##gff-version 3
#SOFTWARE INFO: GeMoMa 1.9; SIMPLE PARAMETERS: reads: 1; splice: true; gap opening: 11; gap extension: 1; maximum intron
 length: 15000; static intron length: true; intron-loss-gain-penalty: 25; reduction factor: 10; e-value: 100.0; contig t
hreshold: 0.4; hit threshold: 0.9; output: STATIC; predictions: 10; avoid stop: true; approx: true; protein alignment: t
rue; prefix: 1_; tag: mRNA; verbose: false; timeout: 3600; sort: false; replace unknown: false; Score: ReAlign
contigs_5       GeMoMa  mRNA    36786   38342   .       -       .       ID=1_gene-NGO_RS00005_R0;ref-gene=1_gene-NGO_RS0
0005.gene;aa=519;raa=519;score=2649;prediction=0;bestScore=2649;ce=1;rce=1;pAA=1;iAA=1;lpm=519;maxScore=2649;maxGap=0;np
s=0;start=M;stop=*
contigs_5       GeMoMa  CDS     36786   38342   .       -       0       Parent=1_gene-NGO_RS00005_R0
contigs_5       GeMoMa  mRNA    29365   31995   .       -       .       ID=1_gene-NGO_RS00030_R0;ref-gene=1_gene-NGO_RS0
0030.gene;aa=877;raa=877;score=4667;prediction=0;bestScore=4667;ce=1;rce=1;pAA=0.9989;iAA=0.9989;lpm=758;maxScore=4668;m
axGap=0;nps=0;start=M;stop=*
```

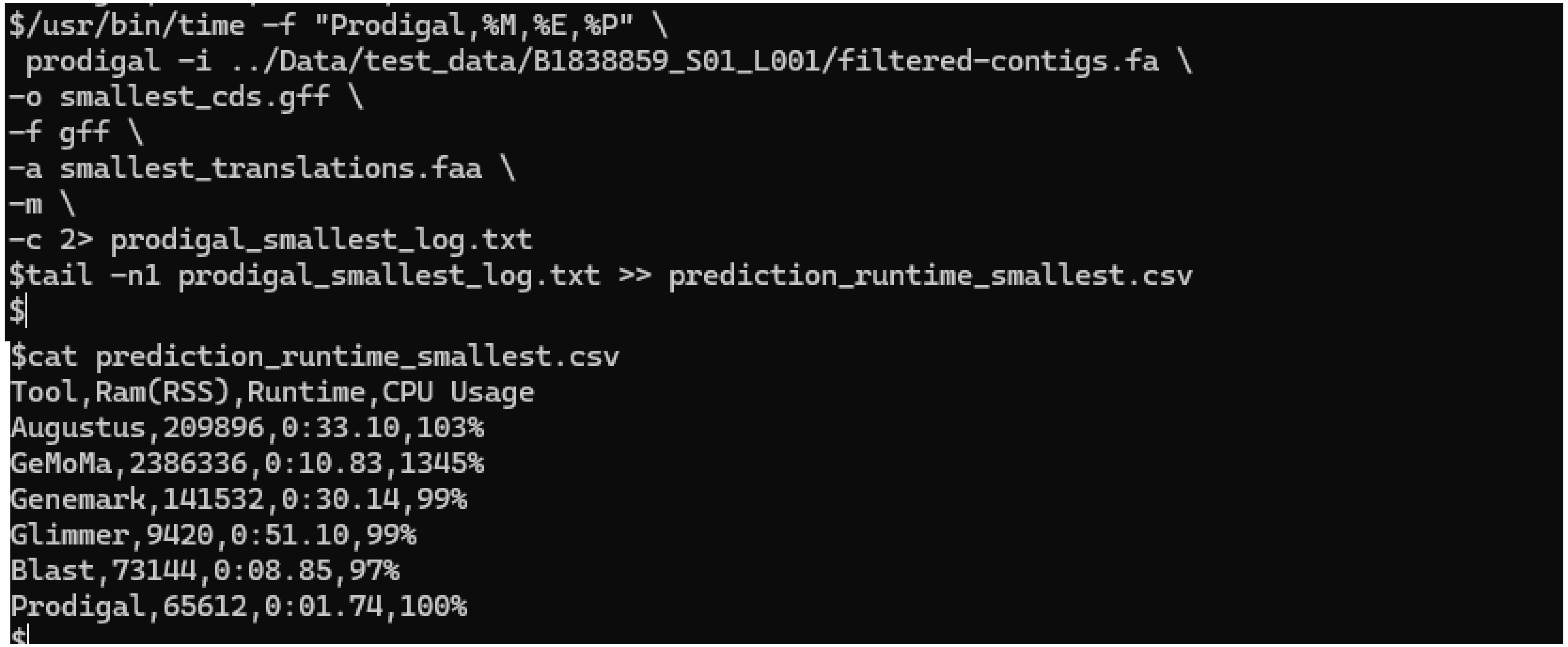
Runtime System Specifications

OS: Ubuntu 22.04.3 LTS (5.15.167.4-microsoft-standard-WSL2)

CPU: Ryzen 9 7900X (12 Cores, 24 Threads)

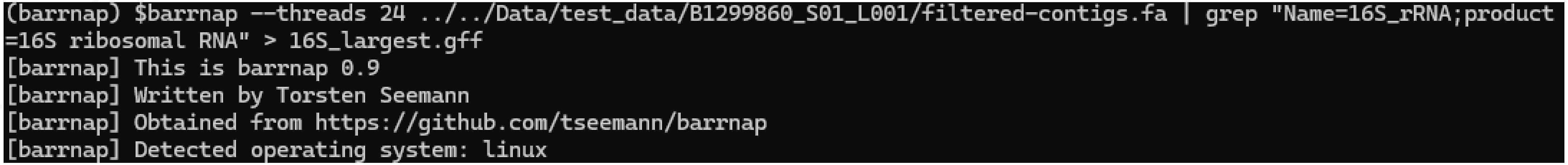
RAM: 64 GB DDR5 CL30 6000 MT/s

GPU: Nvidia RTX 4070 Super (12GB GDDR6X VRAM)

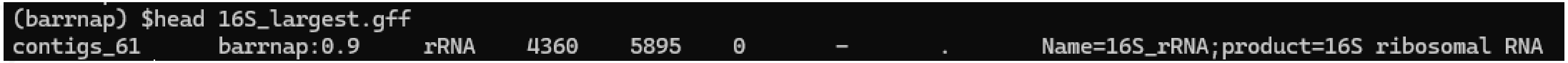


```
$/usr/bin/time -f "Prodigal,%M,%E,%P" \
 prodigal -i ../Data/test_data/B1838859_S01_L001/filtered-contigs.fa \
-o smallest_cds.gff \
-f gff \
-a smallest_translations.faa \
-m \
-c 2> prodigal_smallest_log.txt
$tail -n1 prodigal_smallest_log.txt >> prediction_runtime_smallest.csv
$|
$cat prediction_runtime_smallest.csv
Tool,Ram(RSS),Runtime,CPU Usage
Augustus,209896,0:33.10,103%
GeMoMa,2386336,0:10.83,1345%
Genemark,141532,0:30.14,99%
Glimmer,9420,0:51.10,99%
Blast,73144,0:08.85,97%
Prodigal,65612,0:01.74,100%
$|
```

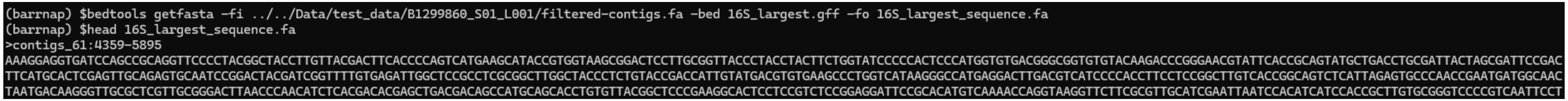
Barrnap 16S rRNA



```
(barrnap) $barrnap --threads 24 ../../Data/test_data/B1299860_S01_L001/filtered-contigs.fa | grep "Name=16S_rRNA;product
=16S ribosomal RNA" > 16S_largest.gff
[barrnap] This is barrnap 0.9
[barrnap] Written by Torsten Seemann
[barrnap] Obtained from https://github.com/tseemann/barrnap
[barrnap] Detected operating system: linux
```



```
(barrnap) $head 16S_largest.gff
contigs_61      barrnap:0.9     rRNA    4360    5895    0       -       .       Name=16S_rRNA;product=16S ribosomal RNA
```



```
(barrnap) $bedtools getfasta -fi ../../Data/test_data/B1299860_S01_L001/filtered-contigs.fa -bed 16S_largest.gff -fo 16S_largest_sequence.fa
(barrnap) $head 16S_largest_sequence.fa
>contigs_61:4359-5895
AAAGGAGGTGATCCAGCCGCAGGTTCCCCTACGGCTACCTTGTTACGACTTCACCCCAGTCATGAAGCATACCGTGGTAAGCGGACTCCTTGCGGTTACCCTACCTACTTCTGGTATCCCCCACTCCCATGGTGTGACGGGCGGTGTGTACAAGACCCGGGAACGTATTCACCGCAGTATGCTGACCTGCGATTACTAGCGATTCCGAC
TTCATGCACTCGAGTTGCAGAGTGCAATCCGGACTACGATCGGTTTTGTGAGATTGGCTCCGCCTCGCGGCTTGGCTACCCTCTGTACCGACCATTGTATGACGTGTGAAGCCCTGGTCATAAGGGCCATGAGGACTTGACGTCATCCCCACCTTCCTCCGGCTTGTCACCGGCAGTCTCATTAGAGTGCCCAACCGAATGATGGCAAC
TAATGACAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTTACGGCTCCCGAAGGCACTCCTCCGTCTCCGGAGGATTCCGCACATGTCAAAACCAGGTAAGGTTCTTCGCGTTGCATCGAATTAATCCACATCATCCACCGCTTGTGCGGGTCCCCGTCAATTCCT
```

	Barrnap	
	Largest	Smallest
Elapsed Time (seconds)	0.36	0.36
RAM (RSS)(kB)	276852	284944
CPU Usage (%)	248%	241%

CDS Prediction Result Comparison

Prediction Metrics (Ab Initio)

Large

Tool	Total CDS	Mean CDS Length
GeneMark	2317	658
Glimmer	2574	754
Prodigal	**2085**	852
Augustus	1872	900
Ground Truth (Latest Ref. Genome)	**2114**	~940

Small

Tool	Total CDS	Mean CDS Length
GeneMark	2318	619
Glimmer	2550	911
Prodigal	**2104**	851
Augustus	1891	898
Ground Truth (Latest Ref. Genome)	**2114**	~940

Prediction Metrics (Homology)

Large

Tool	Total CDS	Mean CDS Length
BLAST	2598	872
GeMoMa	**2359**	735
Ground Truth (Latest Ref. Genome)	**2114**	~940

Small

Tool	Total CDS	Mean CDS Length
BLAST	2702	865
GeMoMa	**2397**	746
Ground Truth (Latest Ref. Genome)	**2114**	~940

Performance Comparison(ab Initio)

	GeneMark		GLIMMER		Augustus		Prodigal	
	largest	smallest	largest	smallest	largest	smallest	largest	smallest
Elapsed Time (seconds)	30.43	30.14	56.20	51.10	30.41	33.10	1.70	1.74
RAM (RSS)(kB)	141372	141532	9096	9420	137620	209896	62678	65612
CPU Usage (%)	99%	99%	99%	99%	104%	103%	99%	100%

Performance Comparison(Homology-based)

	GeMoMa		Blast	
	largest	smallest	largest	smallest
Elapsed Time (seconds)	10.79	10.83	8.55	8.85
RAM (RSS)(kB)	2373596	2386336	73144	73144
CPU Usage (%)	1354%	1345%	99%	97%

Optimal Prediction Tools

Ab initio: Prodigal – closest to reference CDS count, close CDS length, very fast runtime, and low resource usage

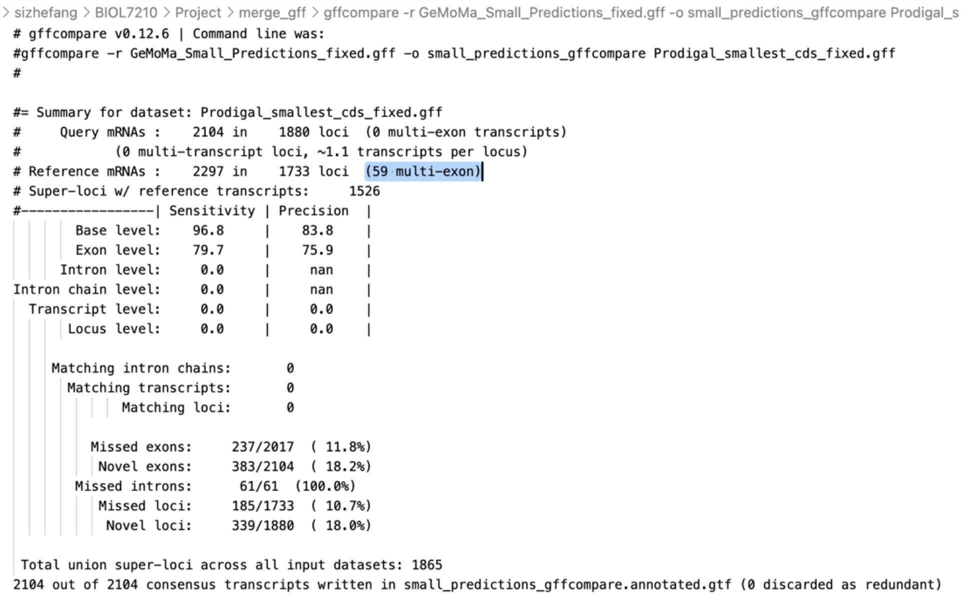
Homology: GeMoMa – closest homology tool to reference CDS count, reasonable CDS length, and efficient runtime via parallelization

16S rRNA: Barrnap – Only tool used, but is fast via parallelization and yields accurate results

Decision on Merging

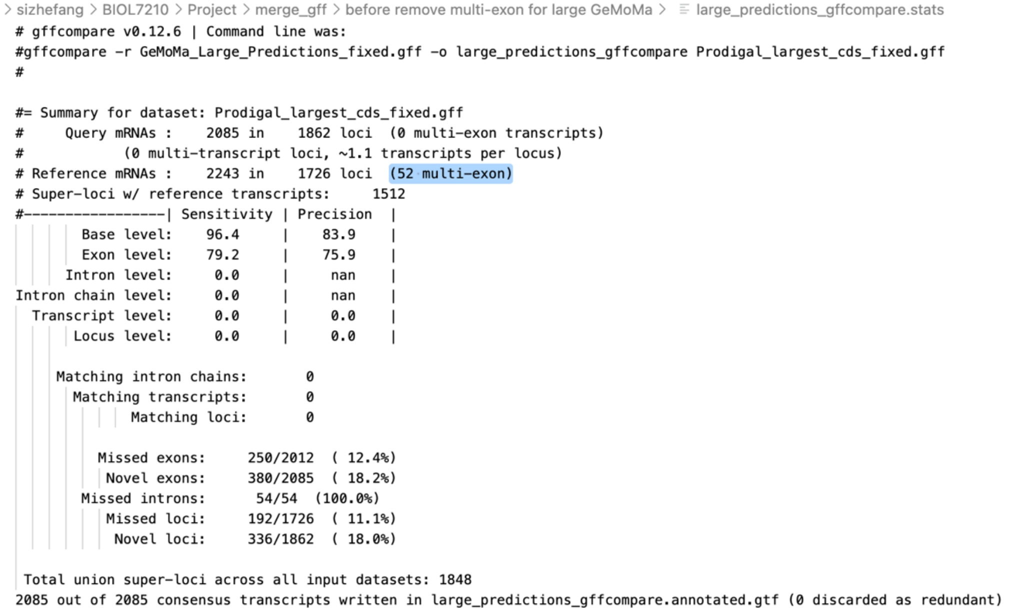
Since we already know that the optimal ab initio method is **Prodigal**, while the optimal homology-based method is **GeMoMa**, we could consider merging their prediction results together

- Use *gffcompare* to compare Prodigal with GeMoMa results. However GeMoMa results are reported to have 59 multi-exons (small) and 52 multi-exons (large). In contrast, Prodigal's results are all 0. That's weird, because prokaryotes rarely contain multi-exon.



> sizhefang > BIOL7210 > Project > merge_gff > gffcompare -r GeMoMa_Small_Predictions_fixed.gff -o small_predictions_gffcompare Prodigal_s

```
# gffcompare v0.12.6 | Command line was:
#gffcompare -r GeMoMa_Small_Predictions_fixed.gff -o small_predictions_gffcompare Prodigal_smallest_cds_fixed.gff
#

#= Summary for dataset: Prodigal_smallest_cds_fixed.gff
#     Query mRNAs :    2104 in    1880 loci  (0 multi-exon transcripts)
#            (0 multi-transcript loci, ~1.1 transcripts per locus)
# Reference mRNAs :    2297 in    1733 loci  (59 multi-exon)
# Super-loci w/ reference transcripts:     1526
#-----------------| Sensitivity | Precision  |
        Base level:    96.8     |    83.8    |
        Exon level:    79.7     |    75.9    |
      Intron level:     0.0     |     nan    |
Intron chain level:     0.0     |     nan    |
  Transcript level:     0.0     |     0.0    |
       Locus level:     0.0     |     0.0    |

     Matching intron chains:       0
       Matching transcripts:       0
              Matching loci:       0

          Missed exons:     237/2017  ( 11.8%)
           Novel exons:     383/2104  ( 18.2%)
        Missed introns:      61/61  (100.0%)
           Missed loci:     185/1733  ( 10.7%)
            Novel loci:     339/1880  ( 18.0%)

 Total union super-loci across all input datasets: 1865
2104 out of 2104 consensus transcripts written in small_predictions_gffcompare.annotated.gtf (0 discarded as redundant)
```

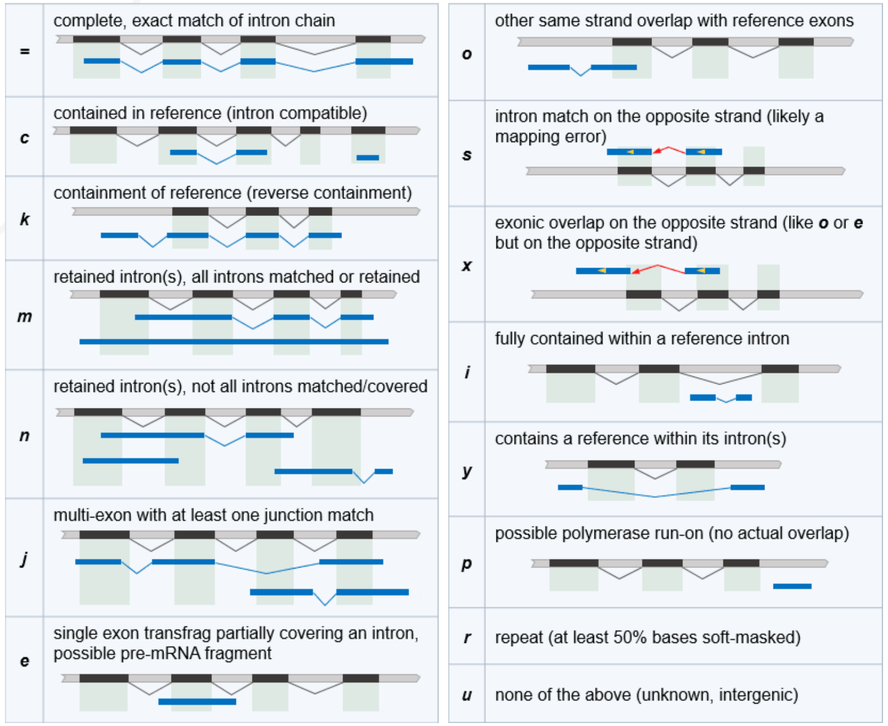


> sizhefang > BIOL7210 > Project > merge_gff > before remove multi-exon for large GeMoMa > large_predictions_gffcompare.stats

```
# gffcompare v0.12.6 | Command line was:
#gffcompare -r GeMoMa_Large_Predictions_fixed.gff -o large_predictions_gffcompare Prodigal_largest_cds_fixed.gff
#

#= Summary for dataset: Prodigal_largest_cds_fixed.gff
#     Query mRNAs :    2085 in    1862 loci  (0 multi-exon transcripts)
#            (0 multi-transcript loci, ~1.1 transcripts per locus)
# Reference mRNAs :    2243 in    1726 loci  (52 multi-exon)
# Super-loci w/ reference transcripts:     1512
#-----------------| Sensitivity | Precision  |
        Base level:    96.4     |    83.9    |
        Exon level:    79.2     |    75.9    |
      Intron level:     0.0     |     nan    |
Intron chain level:     0.0     |     nan    |
  Transcript level:     0.0     |     0.0    |
       Locus level:     0.0     |     0.0    |

     Matching intron chains:       0
       Matching transcripts:       0
              Matching loci:       0

          Missed exons:     250/2012  ( 12.4%)
           Novel exons:     380/2085  ( 18.2%)
        Missed introns:      54/54  (100.0%)
           Missed loci:     192/1726  ( 11.1%)
            Novel loci:     336/1862  ( 18.0%)

 Total union super-loci across all input datasets: 1848
2085 out of 2085 consensus transcripts written in large_predictions_gffcompare.annotated.gtf (0 discarded as redundant)
```

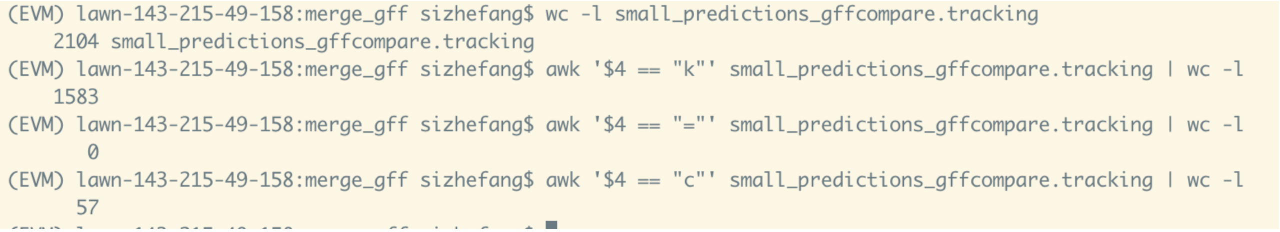
Decision on Merging

- Then we removed the multi-exon from the GeMoMa result, re-run gffcompare, and found that in the gffcompare result, most class_codes are '**k**', which means most CDS predicted by GeMoMa are subsets of Prodigal. However, as a homology-based method, GeMoMa CDSs should be larger than Prodigal.

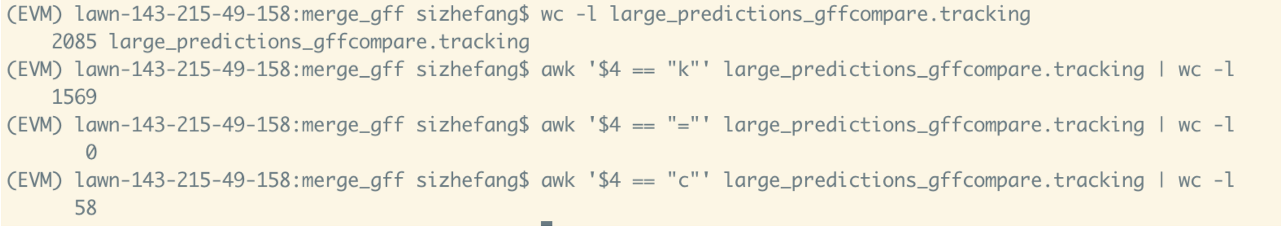


=	complete, exact match of intron chain
c	contained in reference (intron compatible)
k	containment of reference (reverse containment)
m	retained intron(s), all introns matched or retained
n	retained intron(s), not all introns matched/covered
j	multi-exon with at least one junction match
e	single exon transfrag partially covering an intron, possible pre-mRNA fragment
o	other same strand overlap with reference exons
s	intron match on the opposite strand (likely a mapping error)
x	exonic overlap on the opposite strand (like **o** or **e** but on the opposite strand)
i	fully contained within a reference intron
y	contains a reference within its intron(s)
p	possible polymerase run-on (no actual overlap)
r	repeat (at least 50% bases soft-masked)
u	none of the above (unknown, intergenic)

https://ccb.jhu.edu/software/stringtie/gffcompare.shtml



```
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ wc -l small_predictions_gffcompare.tracking
    2104 small_predictions_gffcompare.tracking
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "k"' small_predictions_gffcompare.tracking | wc -l
    1583
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "="' small_predictions_gffcompare.tracking | wc -l
       0
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "c"' small_predictions_gffcompare.tracking | wc -l
      57
```

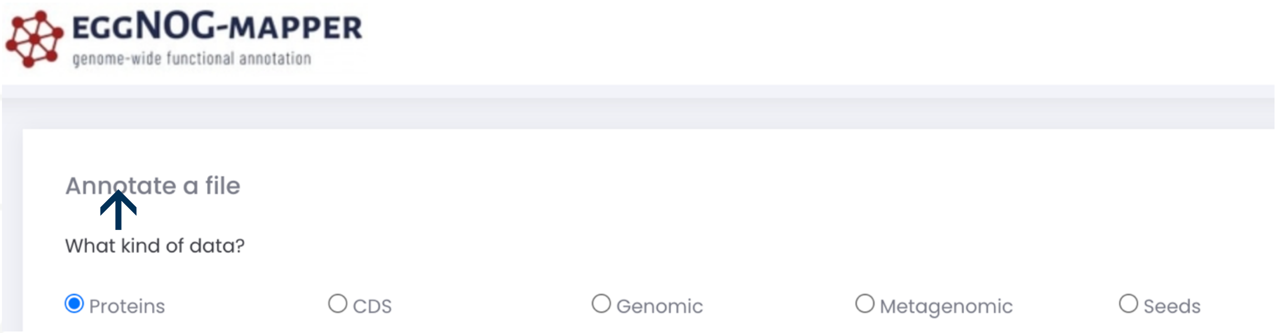


```
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ wc -l large_predictions_gffcompare.tracking
    2085 large_predictions_gffcompare.tracking
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "k"' large_predictions_gffcompare.tracking | wc -l
    1569
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "="' large_predictions_gffcompare.tracking | wc -l
       0
(EVM) lawn-143-215-49-158:merge_gff sizhefang$ awk '$4 == "c"' large_predictions_gffcompare.tracking | wc -l
      58
```

In summary, GeMoMa's performance on this dataset is not as favorable as Prodigal's. Thus, integrating GeMoMa with Prodigal's predictions is unlikely to be beneficial. **Therefore, we have decided to use only Prodigal's predictions for the annotation part.**

Gene Annotation: EggNog vs. InterPro

Annotation Tools

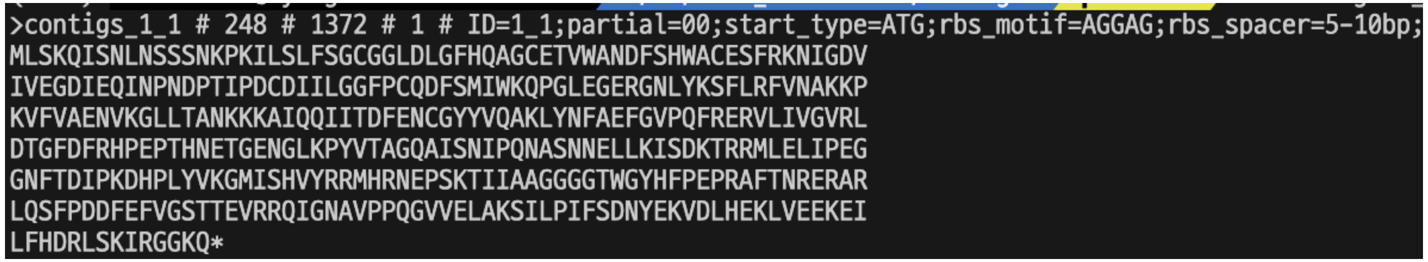


EGGNOG-MAPPER
genome-wide functional annotation

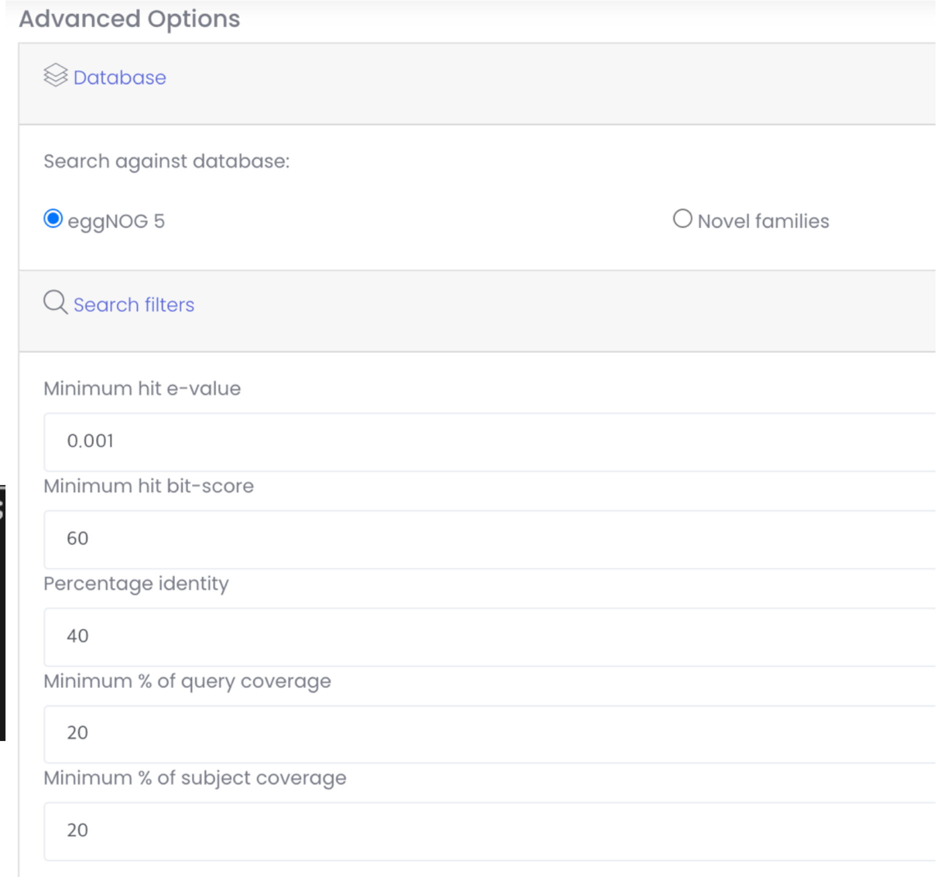
Annotate a file

What kind of data?

● Proteins ○ CDS ○ Genomic ○ Metagenomic ○ Seeds



```
>contigs_1_1 # 248 # 1372 # 1 # ID=1_1;partial=00;start_type=ATG;rbs_motif=AGGAG;rbs_spacer=5-10bp;
MLSKQISNLNSSSNKPKILSLFSGCGGLDLGFHQAGCETVWANDFSHWACESFRKNIGDV
IVEGDIEQINPNDPTIPDCDIILGGFPCQDFSMIWKQPGLEGERGNLYKSFLRFVNAKKP
KVFVAENVKGLLTANKKKAIQQIITDFENCGYYVQAKLYNFAEFGVPQFRERVLIVGVRL
DTGFDFRHPEPTHNETGENGLKPYVTAGQAISNIPQNASNNELLKISDKTRRMLELIPEG
GNFTDIPKDHPLYVKGMISHVYRRMHRNEPSKTIIAAGGGGTWGYHFPEPRAFTNRERAR
LQSFPDDFEFVGSTTEVRRQIGNAVPPQGVVELAKSILPIFSDNYEKVDLHEKLVEEKEI
LFHDRLSKIRGGKQ*
```



Advanced Options

Database

Search against database:

● eggNOG 5 ○ Novel families

Search filters

Minimum hit e-value: 0.001

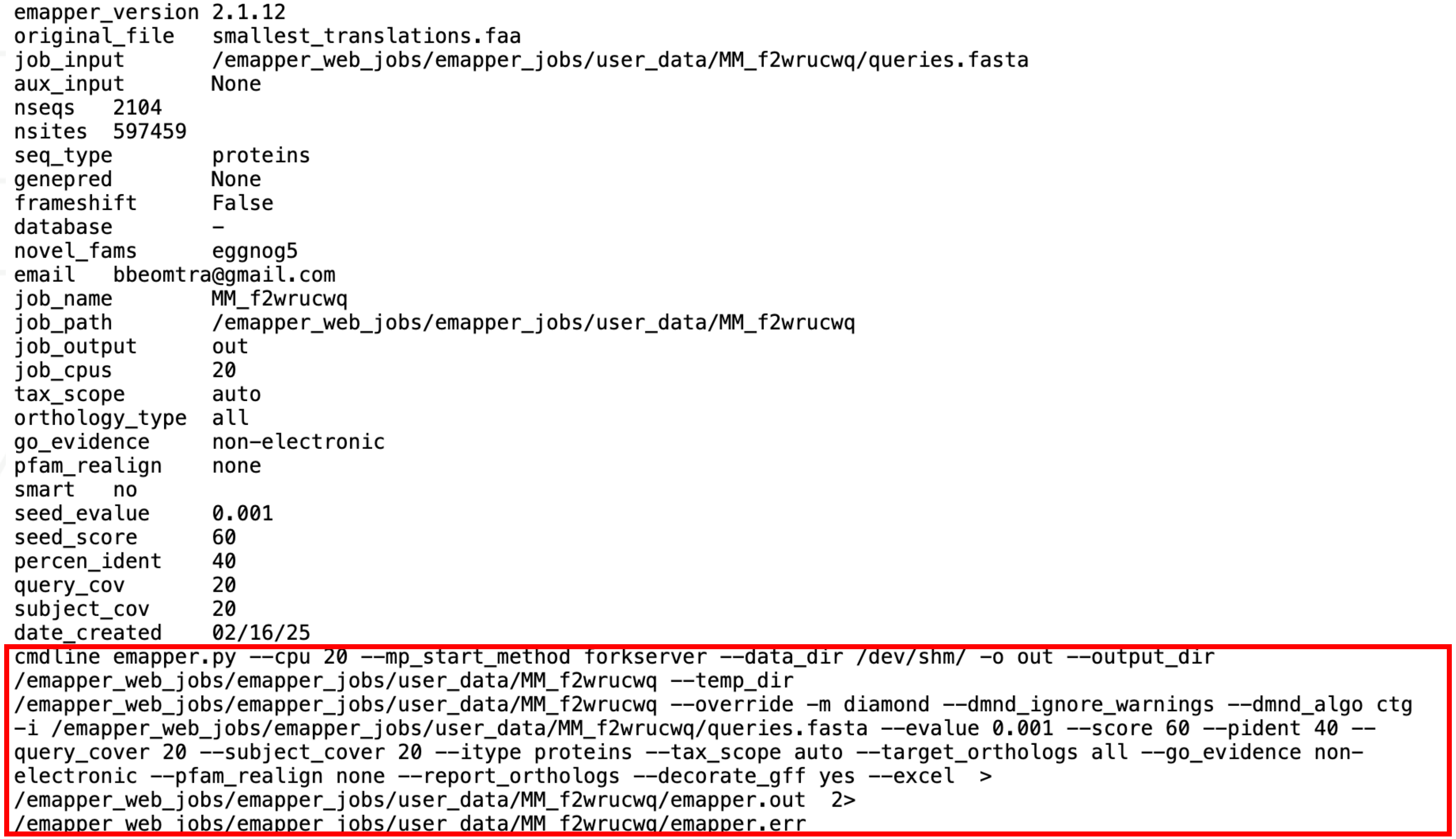
Minimum hit bit-score: 60

Percentage identity: 40

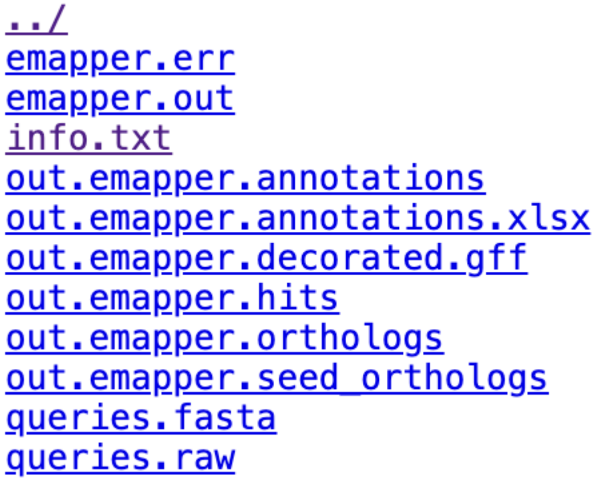
Minimum % of query coverage: 20

Minimum % of subject coverage: 20

Annotation Results: EggNog (small)

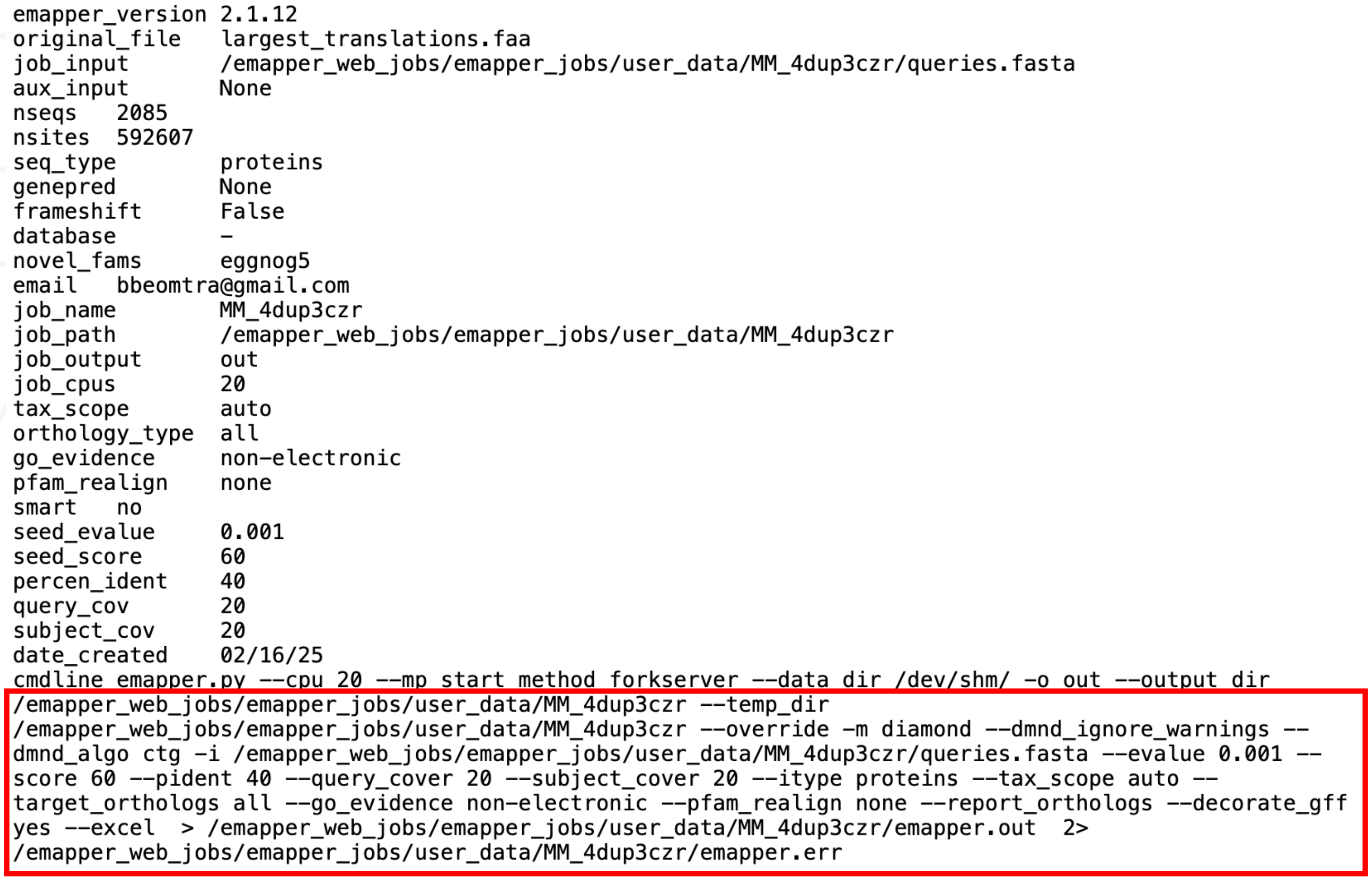


```
emapper_version 2.1.12
original_file   smallest_translations.faa
job_input       /emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq/queries.fasta
aux_input       None
nseqs   2104
nsites  597459
seq_type        proteins
genepred        None
frameshift      False
database        -
novel_fams      eggnog5
email   bbeomtra@gmail.com
job_name        MM_f2wrucwq
job_path        /emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq
job_output      out
job_cpus        20
tax_scope       auto
orthology_type  all
go_evidence     non-electronic
pfam_realign    none
smart   no
seed_evalue     0.001
seed_score      60
percen_ident    40
query_cov       20
subject_cov     20
date_created    02/16/25
cmdline emapper.py --cpu 20 --mp_start_method forkserver --data_dir /dev/shm/ -o out --output_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq --temp_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq --override -m diamond --dmnd_ignore_warnings --dmnd_algo ctg
-i /emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq/queries.fasta --evalue 0.001 --score 60 --pident 40 --
query_cover 20 --subject_cover 20 --itype proteins --tax_scope auto --target_orthologs all --go_evidence non-
electronic --pfam_realign none --report_orthologs --decorate_gff yes --excel  >
/emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq/emapper.out  2>
/emapper_web_jobs/emapper_jobs/user_data/MM_f2wrucwq/emapper.err
```

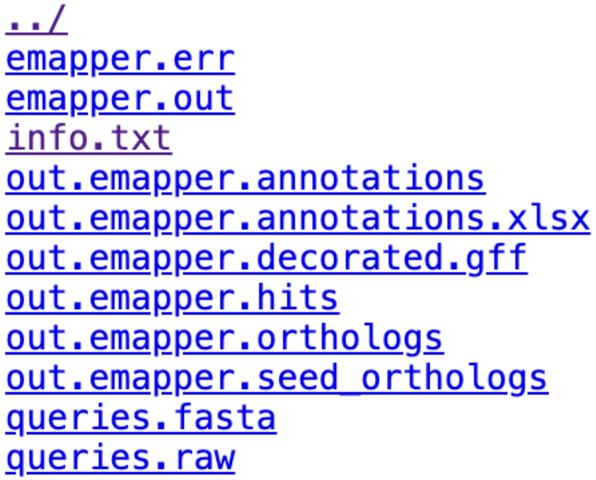


```
../
emapper.err
emapper.out
info.txt
out.emapper.annotations
out.emapper.annotations.xlsx
out.emapper.decorated.gff
out.emapper.hits
out.emapper.orthologs
out.emapper.seed_orthologs
queries.fasta
queries.raw
```

Annotation Results: EggNog (Large)

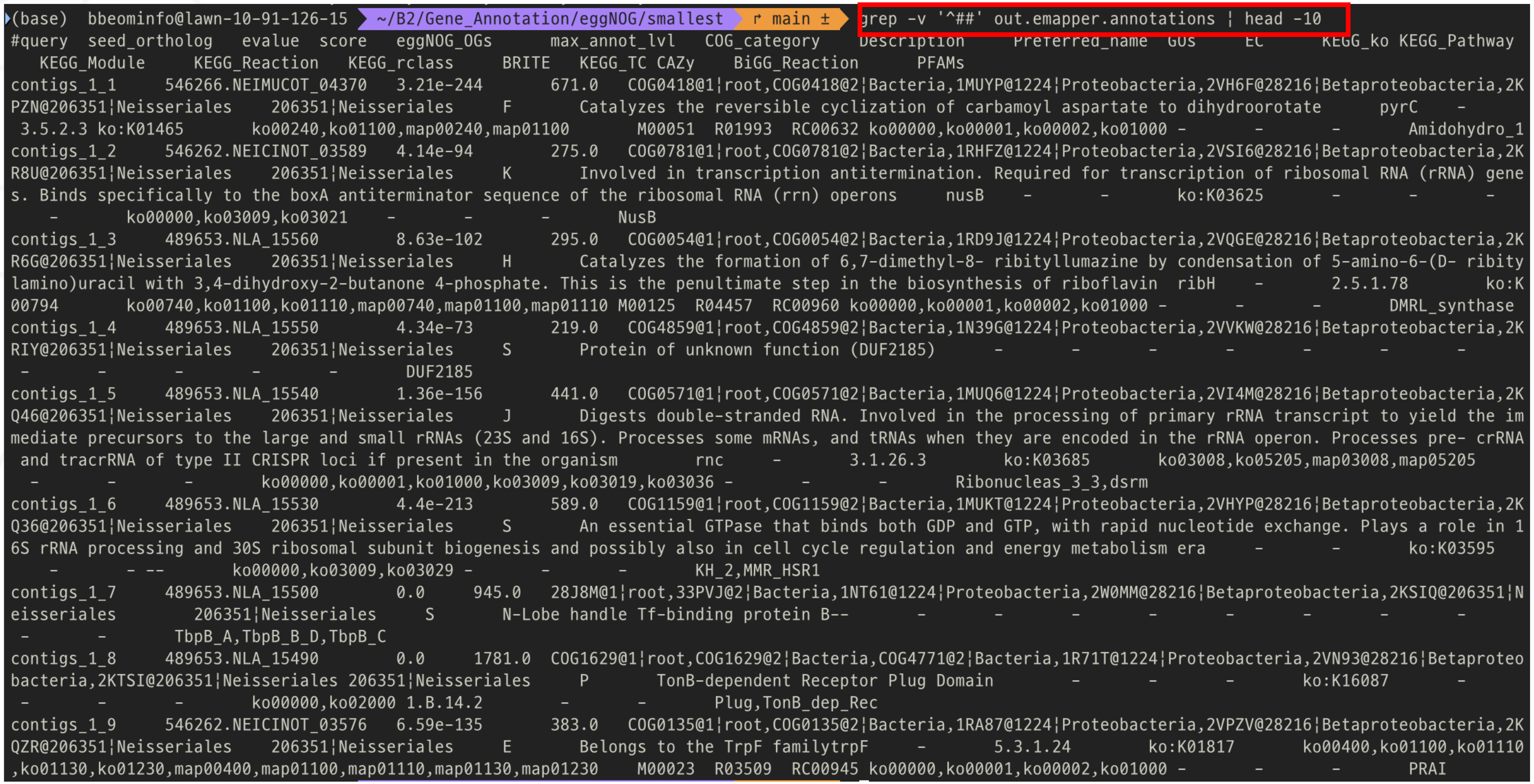


```
emapper_version 2.1.12
original_file   largest_translations.faa
job_input       /emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr/queries.fasta
aux_input       None
nseqs   2085
nsites  592607
seq_type        proteins
genepred        None
frameshift      False
database        -
novel_fams      eggnog5
email   bbeomtra@gmail.com
job_name        MM_4dup3czr
job_path        /emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr
job_output      out
job_cpus        20
tax_scope       auto
orthology_type  all
go_evidence     non-electronic
pfam_realign    none
smart   no
seed_evalue     0.001
seed_score      60
percen_ident    40
query_cov       20
subject_cov     20
date_created    02/16/25
cmdline emapper.py --cpu 20 --mp_start_method forkserver --data_dir /dev/shm/ -o out --output_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr --temp_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr --override -m diamond --dmnd_ignore_warnings --
dmnd_algo ctg -i /emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr/queries.fasta --evalue 0.001 --
score 60 --pident 40 --query_cover 20 --subject_cover 20 --itype proteins --tax_scope auto --
target_orthologs all --go_evidence non-electronic --pfam_realign none --report_orthologs --decorate_gff
yes --excel  > /emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr/emapper.out  2>
/emapper_web_jobs/emapper_jobs/user_data/MM_4dup3czr/emapper.err
```

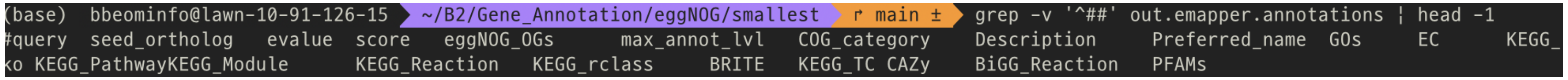


```
../
emapper.err
emapper.out
info.txt
out.emapper.annotations
out.emapper.annotations.xlsx
out.emapper.decorated.gff
out.emapper.hits
out.emapper.orthologs
out.emapper.seed_orthologs
queries.fasta
queries.raw
```

Annotation Results: EggNog



```
(base)  bbeominfo@lawn-10-91-126-15  ~/B2/Gene_Annotation/eggNOG/smallest  ↱ main ±  grep -v '^##' out.emapper.annotations ¦ head -10
#query  seed_ortholog   evalue  score   eggNOG_OGs      max_annot_lvl   COG_category    Description     Preferred_name  GOs     EC      KEGG_ko KEGG_Pathway
   KEGG_Module     KEGG_Reaction   KEGG_rclass     BRITE   KEGG_TC CAZy    BiGG_Reaction       PFAMs
contigs_1_1     546266.NEIMUCOT_04370   3.21e-244       671.0   COG0418@1¦root,COG0418@2¦Bacteria,1MUYP@1224¦Proteobacteria,2VH6F@28216¦Betaproteobacteria,2K
PZN@206351¦Neisseriales    206351¦Neisseriales     F       Catalyzes the reversible cyclization of carbamoyl aspartate to dihydroorotate      pyrC    -
 3.5.2.3 ko:K01465       ko00240,ko01100,map00240,map01100       M00051  R01993  RC00632 ko00000,ko00001,ko00002,ko01000 -       -       -       Amidohydro_1
contigs_1_2     546262.NEICINOT_03589   4.14e-94        275.0   COG0781@1¦root,COG0781@2¦Bacteria,1RHFZ@1224¦Proteobacteria,2VSI6@28216¦Betaproteobacteria,2K
R8U@206351¦Neisseriales    206351¦Neisseriales     K       Involved in transcription antitermination. Required for transcription of ribosomal RNA (rRNA) gene
s. Binds specifically to the boxA antiterminator sequence of the ribosomal RNA (rrn) operons     nusB    -       -       ko:K03625       -       -       -
    -       ko00000,ko03009,ko03021    -       -       -       NusB
contigs_1_3     489653.NLA_15560        8.63e-102       295.0   COG0054@1¦root,COG0054@2¦Bacteria,1RD9J@1224¦Proteobacteria,2VQGE@28216¦Betaproteobacteria,2K
R6G@206351¦Neisseriales    206351¦Neisseriales     H       Catalyzes the formation of 6,7-dimethyl-8- ribityllumazine by condensation of 5-amino-6-(D- ribity
lamino)uracil with 3,4-dihydroxy-2-butanone 4-phosphate. This is the penultimate step in the biosynthesis of riboflavin  ribH    -       2.5.1.78        ko:K
00794       ko00740,ko01100,ko01110,map00740,map01100,map01110 M00125  R04457  RC00960 ko00000,ko00001,ko00002,ko01000 -       -       DMRL_synthase
contigs_1_4     489653.NLA_15550        4.34e-73        219.0   COG4859@1¦root,COG4859@2¦Bacteria,1N39G@1224¦Proteobacteria,2VVKW@28216¦Betaproteobacteria,2K
RIY@206351¦Neisseriales    206351¦Neisseriales     S       Protein of unknown function (DUF2185)      -       -       -       -       -       -       -
 -       -       -       -       -       DUF2185
contigs_1_5     489653.NLA_15540        1.36e-156       441.0   COG0571@1¦root,COG0571@2¦Bacteria,1MUQ6@1224¦Proteobacteria,2VI4M@28216¦Betaproteobacteria,2K
Q46@206351¦Neisseriales    206351¦Neisseriales     J       Digests double-stranded RNA. Involved in the processing of primary rRNA transcript to yield the im
mediate precursors to the large and small rRNAs (23S and 16S). Processes some mRNAs, and tRNAs when they are encoded in the rRNA operon. Processes pre- crRNA
 and tracrRNA of type II CRISPR loci if present in the organism       rnc     -       3.1.26.3        ko:K03685       ko03008,ko05205,map03008,map05205
  -       -       -       ko00000,ko00001,ko01000,ko03009,ko03019,ko03036 -       -       -       Ribonucleas_3_3,dsrm
contigs_1_6     489653.NLA_15530        4.4e-213        589.0   COG1159@1¦root,COG1159@2¦Bacteria,1MUKT@1224¦Proteobacteria,2VHYP@28216¦Betaproteobacteria,2K
Q36@206351¦Neisseriales    206351¦Neisseriales     S       An essential GTPase that binds both GDP and GTP, with rapid nucleotide exchange. Plays a role in 1
6S rRNA processing and 30S ribosomal subunit biogenesis and possibly also in cell cycle regulation and energy metabolism era     -       -       ko:K03595
    -       - --      ko00000,ko03009,ko03029 -       -       -       KH_2,MMR_HSR1
contigs_1_7     489653.NLA_15500        0.0     945.0   28J8M@1¦root,33PVJ@2¦Bacteria,1NT61@1224¦Proteobacteria,2W0MM@28216¦Betaproteobacteria,2KSIQ@206351¦N
eisseriales        206351¦Neisseriales     S       N-Lobe handle Tf-binding protein B--      -       -       -       -       -       -       -       -
 -       -       TbpB_A,TbpB_B_D,TbpB_C
contigs_1_8     489653.NLA_15490        0.0     1781.0  COG1629@1¦root,COG1629@2¦Bacteria,COG4771@2¦Bacteria,1R71T@1224¦Proteobacteria,2VN93@28216¦Betaproteo
bacteria,2KTSI@206351¦Neisseriales 206351¦Neisseriales     P       TonB-dependent Receptor Plug Domain       -       -       -       ko:K16087       -
 -       -       -       ko00000,ko02000 1.B.14.2        -       -       Plug,TonB_dep_Rec
contigs_1_9     546262.NEICINOT_03576   6.59e-135       383.0   COG0135@1¦root,COG0135@2¦Bacteria,1RA87@1224¦Proteobacteria,2VPZV@28216¦Betaproteobacteria,2K
QZR@206351¦Neisseriales    206351¦Neisseriales     E       Belongs to the TrpF familytrpF    -       5.3.1.24        ko:K01817       ko00400,ko01100,ko01110
,ko01130,ko01230,map00400,map01100,map01110,map01130,map01230    M00023  R03509  RC00945 ko00000,ko00001,ko00002,ko01000 -       -       -       PRAI
```



```
(base)  bbeominfo@lawn-10-91-126-15  ~/B2/Gene_Annotation/eggNOG/smallest  ↱ main ±  grep -v '^##' out.emapper.annotations ¦ head -1
#query  seed_ortholog   evalue  score   eggNOG_OGs      max_annot_lvl   COG_category    Description     Preferred_name  GOs     EC      KEGG_
ko KEGG_PathwayKEGG_Module      KEGG_Reaction   KEGG_rclass     BRITE   KEGG_TC CAZy    BiGG_Reaction   PFAMs
```

#query

seed_ortholog

evalue

score

eggNOG_OGs

max_annot_lvl

COG_category

Description

Preferred_name

GOs

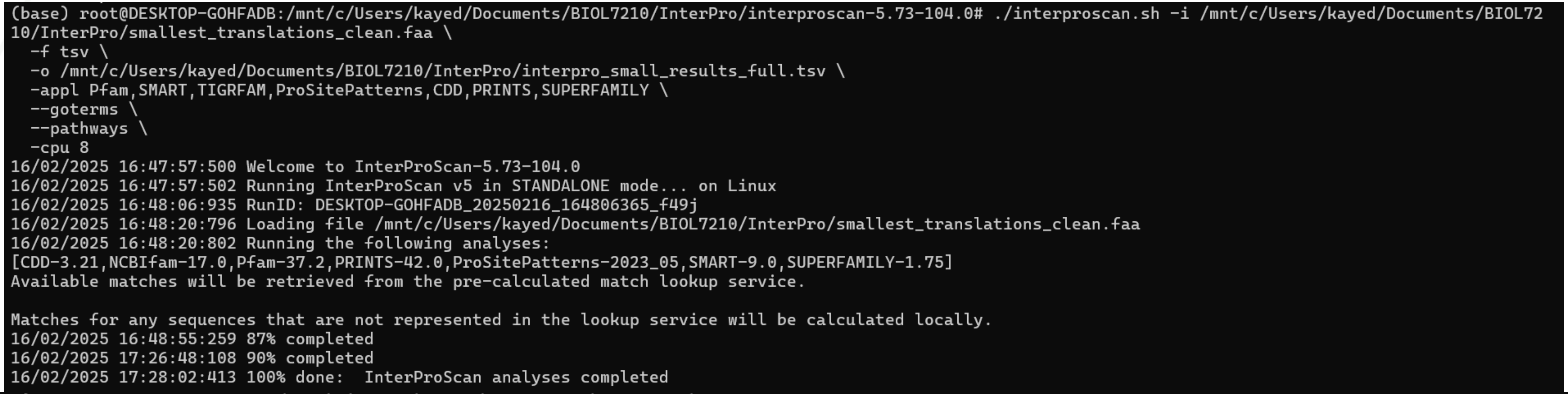
EC

KEGG_Pathway/Module/Reaction/rclass/Brite/TC

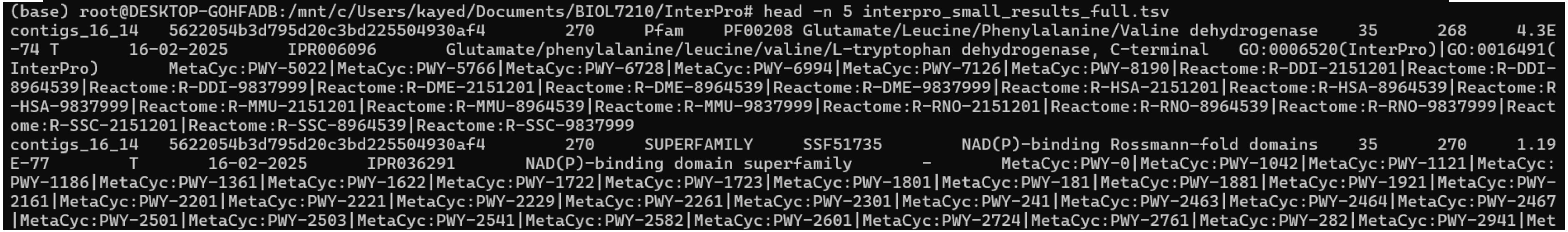
PFAMs

BiGG_Reaction

Annotation Results: InterPro (Small)



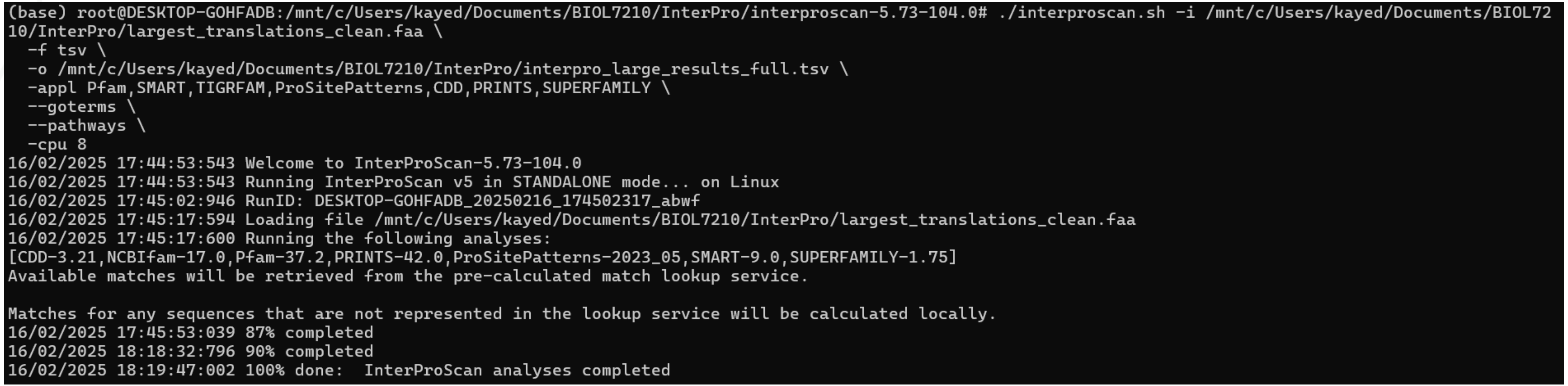
```
(base) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/InterPro/interproscan-5.73-104.0# ./interproscan.sh -i /mnt/c/Users/kayed/Documents/BIOL72
10/InterPro/smallest_translations_clean.faa \
  -f tsv \
  -o /mnt/c/Users/kayed/Documents/BIOL7210/InterPro/interpro_small_results_full.tsv \
  -appl Pfam,SMART,TIGRFAM,ProSitePatterns,CDD,PRINTS,SUPERFAMILY \
  --goterms \
  --pathways \
  -cpu 8
16/02/2025 16:47:57:500 Welcome to InterProScan-5.73-104.0
16/02/2025 16:47:57:502 Running InterProScan v5 in STANDALONE mode... on Linux
16/02/2025 16:48:06:935 RunID: DESKTOP-GOHFADB_20250216_164806365_f49j
16/02/2025 16:48:20:796 Loading file /mnt/c/Users/kayed/Documents/BIOL7210/InterPro/smallest_translations_clean.faa
16/02/2025 16:48:20:802 Running the following analyses:
[CDD-3.21,NCBIfam-17.0,Pfam-37.2,PRINTS-42.0,ProSitePatterns-2023_05,SMART-9.0,SUPERFAMILY-1.75]
Available matches will be retrieved from the pre-calculated match lookup service.

Matches for any sequences that are not represented in the lookup service will be calculated locally.
16/02/2025 16:48:55:259 87% completed
16/02/2025 17:26:48:108 90% completed
16/02/2025 17:28:02:413 100% done:  InterProScan analyses completed
```



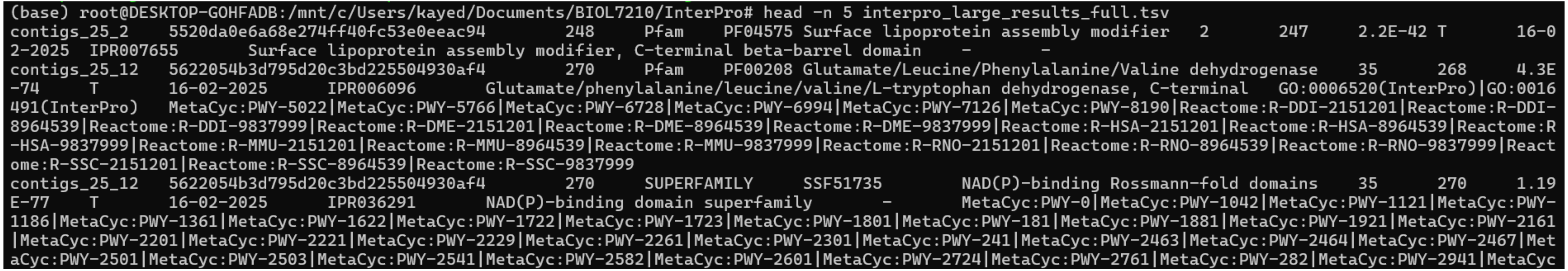
```
(base) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/InterPro# head -n 5 interpro_small_results_full.tsv
contigs_16_14   5622054b3d795d20c3bd225504930af4        270     Pfam    PF00208 Glutamate/Leucine/Phenylalanine/Valine dehydrogenase    35      268     4.3E
-74 T       16-02-2025      IPR006096       Glutamate/phenylalanine/leucine/valine/L-tryptophan dehydrogenase, C-terminal   GO:0006520(InterPro)|GO:0016491(
InterPro)       MetaCyc:PWY-5022|MetaCyc:PWY-5766|MetaCyc:PWY-6728|MetaCyc:PWY-6994|MetaCyc:PWY-7126|MetaCyc:PWY-8190|Reactome:R-DDI-2151201|Reactome:R-DDI-
8964539|Reactome:R-DDI-9837999|Reactome:R-DME-2151201|Reactome:R-DME-8964539|Reactome:R-DME-9837999|Reactome:R-HSA-2151201|Reactome:R-HSA-8964539|Reactome:R
-HSA-9837999|Reactome:R-MMU-2151201|Reactome:R-MMU-8964539|Reactome:R-MMU-9837999|Reactome:R-RNO-2151201|Reactome:R-RNO-8964539|Reactome:R-RNO-9837999|React
ome:R-SSC-2151201|Reactome:R-SSC-8964539|Reactome:R-SSC-9837999
contigs_16_14   5622054b3d795d20c3bd225504930af4        270     SUPERFAMILY     SSF51735        NAD(P)-binding Rossmann-fold domains    35      270     1.19
E-77        T       16-02-2025      IPR036291       NAD(P)-binding domain superfamily       -       MetaCyc:PWY-0|MetaCyc:PWY-1042|MetaCyc:PWY-1121|MetaCyc:
PWY-1186|MetaCyc:PWY-1361|MetaCyc:PWY-1622|MetaCyc:PWY-1722|MetaCyc:PWY-1723|MetaCyc:PWY-1801|MetaCyc:PWY-181|MetaCyc:PWY-1881|MetaCyc:PWY-1921|MetaCyc:PWY-
2161|MetaCyc:PWY-2201|MetaCyc:PWY-2221|MetaCyc:PWY-2229|MetaCyc:PWY-2261|MetaCyc:PWY-2301|MetaCyc:PWY-241|MetaCyc:PWY-2463|MetaCyc:PWY-2464|MetaCyc:PWY-2467
|MetaCyc:PWY-2501|MetaCyc:PWY-2503|MetaCyc:PWY-2541|MetaCyc:PWY-2582|MetaCyc:PWY-2601|MetaCyc:PWY-2724|MetaCyc:PWY-2761|MetaCyc:PWY-282|MetaCyc:PWY-2941|Met
```

Elapsed Time (seconds)	233.58
RAM (RSS)(kB)	4135020
CPU Usage (%)	66%

Annotation Results: InterPro (Large)



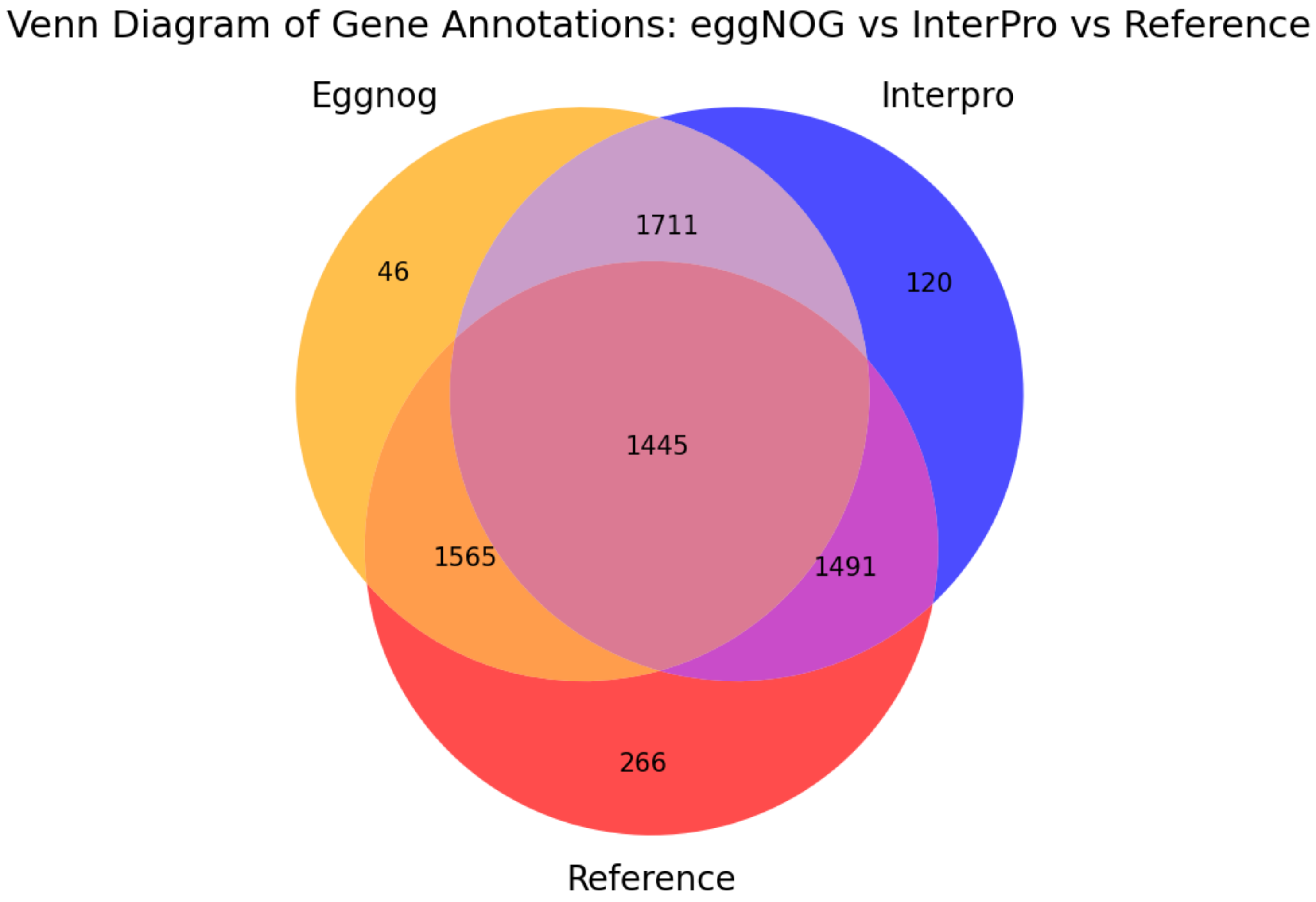
```
(base) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/InterPro/interproscan-5.73-104.0# ./interproscan.sh -i /mnt/c/Users/kayed/Documents/BIOL72
10/InterPro/largest_translations_clean.faa \
  -f tsv \
  -o /mnt/c/Users/kayed/Documents/BIOL7210/InterPro/interpro_large_results_full.tsv \
  -appl Pfam,SMART,TIGRFAM,ProSitePatterns,CDD,PRINTS,SUPERFAMILY \
  --goterms \
  --pathways \
  -cpu 8
16/02/2025 17:44:53:543 Welcome to InterProScan-5.73-104.0
16/02/2025 17:44:53:543 Running InterProScan v5 in STANDALONE mode... on Linux
16/02/2025 17:45:02:946 RunID: DESKTOP-GOHFADB_20250216_174502317_abwf
16/02/2025 17:45:17:594 Loading file /mnt/c/Users/kayed/Documents/BIOL7210/InterPro/largest_translations_clean.faa
16/02/2025 17:45:17:600 Running the following analyses:
[CDD-3.21,NCBIfam-17.0,Pfam-37.2,PRINTS-42.0,ProSitePatterns-2023_05,SMART-9.0,SUPERFAMILY-1.75]
Available matches will be retrieved from the pre-calculated match lookup service.

Matches for any sequences that are not represented in the lookup service will be calculated locally.
16/02/2025 17:45:53:039 87% completed
16/02/2025 18:18:32:796 90% completed
16/02/2025 18:19:47:002 100% done:  InterProScan analyses completed
```



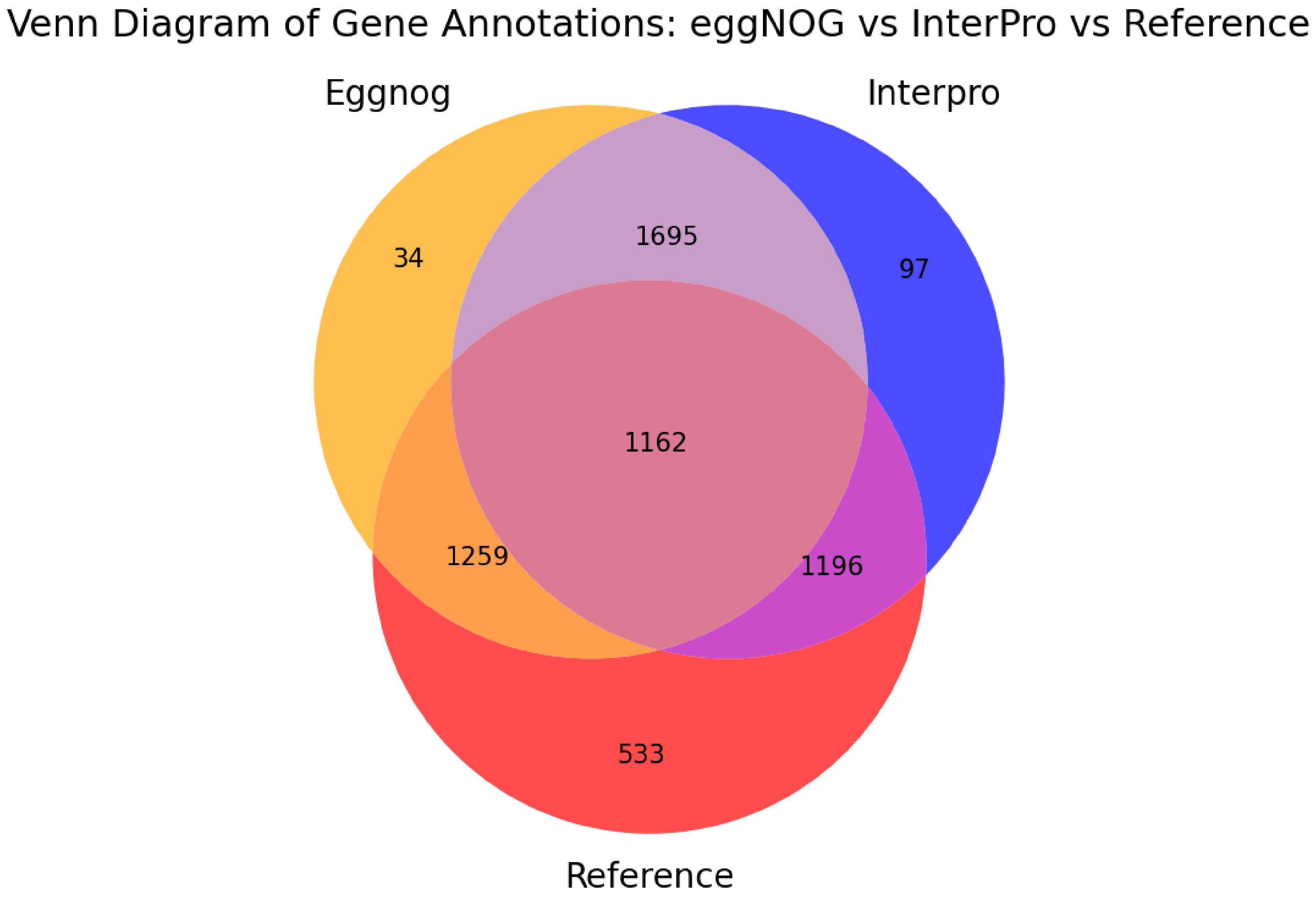
```
(base) root@DESKTOP-GOHFADB:/mnt/c/Users/kayed/Documents/BIOL7210/InterPro# head -n 5 interpro_large_results_full.tsv
contigs_25_2    5520da0e6a68e274ff40fc53e0eeac94        248     Pfam    PF04575 Surface lipoprotein assembly modifier   2       247     2.2E-42 T       16-0
2-2025  IPR007655       Surface lipoprotein assembly modifier, C-terminal beta-barrel domain    -       -
contigs_25_12   5622054b3d795d20c3bd225504930af4        270     Pfam    PF00208 Glutamate/Leucine/Phenylalanine/Valine dehydrogenase    35      268     4.3E
-74     T       16-02-2025      IPR006096       Glutamate/phenylalanine/leucine/valine/L-tryptophan dehydrogenase, C-terminal   GO:0006520(InterPro)|GO:0016
491(InterPro)   MetaCyc:PWY-5022|MetaCyc:PWY-5766|MetaCyc:PWY-6728|MetaCyc:PWY-6994|MetaCyc:PWY-7126|MetaCyc:PWY-8190|Reactome:R-DDI-2151201|Reactome:R-DDI-
8964539|Reactome:R-DDI-9837999|Reactome:R-DME-2151201|Reactome:R-DME-8964539|Reactome:R-DME-9837999|Reactome:R-HSA-2151201|Reactome:R-HSA-8964539|Reactome:R
-HSA-9837999|Reactome:R-MMU-2151201|Reactome:R-MMU-8964539|Reactome:R-MMU-9837999|Reactome:R-RNO-2151201|Reactome:R-RNO-8964539|Reactome:R-RNO-9837999|React
ome:R-SSC-2151201|Reactome:R-SSC-8964539|Reactome:R-SSC-9837999
contigs_25_12   5622054b3d795d20c3bd225504930af4        270     SUPERFAMILY     SSF51735        NAD(P)-binding Rossmann-fold domains    35      270     1.19
E-77    T       16-02-2025      IPR036291       NAD(P)-binding domain superfamily       -       MetaCyc:PWY-0|MetaCyc:PWY-1042|MetaCyc:PWY-1121|MetaCyc:PWY-
1186|MetaCyc:PWY-1361|MetaCyc:PWY-1622|MetaCyc:PWY-1722|MetaCyc:PWY-1723|MetaCyc:PWY-1801|MetaCyc:PWY-181|MetaCyc:PWY-1881|MetaCyc:PWY-1921|MetaCyc:PWY-2161
|MetaCyc:PWY-2201|MetaCyc:PWY-2221|MetaCyc:PWY-2229|MetaCyc:PWY-2261|MetaCyc:PWY-2301|MetaCyc:PWY-241|MetaCyc:PWY-2463|MetaCyc:PWY-2464|MetaCyc:PWY-2467|Met
aCyc:PWY-2501|MetaCyc:PWY-2503|MetaCyc:PWY-2541|MetaCyc:PWY-2582|MetaCyc:PWY-2601|MetaCyc:PWY-2724|MetaCyc:PWY-2761|MetaCyc:PWY-282|MetaCyc:PWY-2941|MetaCyc
```

Elapsed Time (seconds)	215.69
RAM (RSS)(kB)	4078484
CPU Usage (%)	72%

Annotation Comparison (Smallest File)



Annotation Comparison (Largest File)



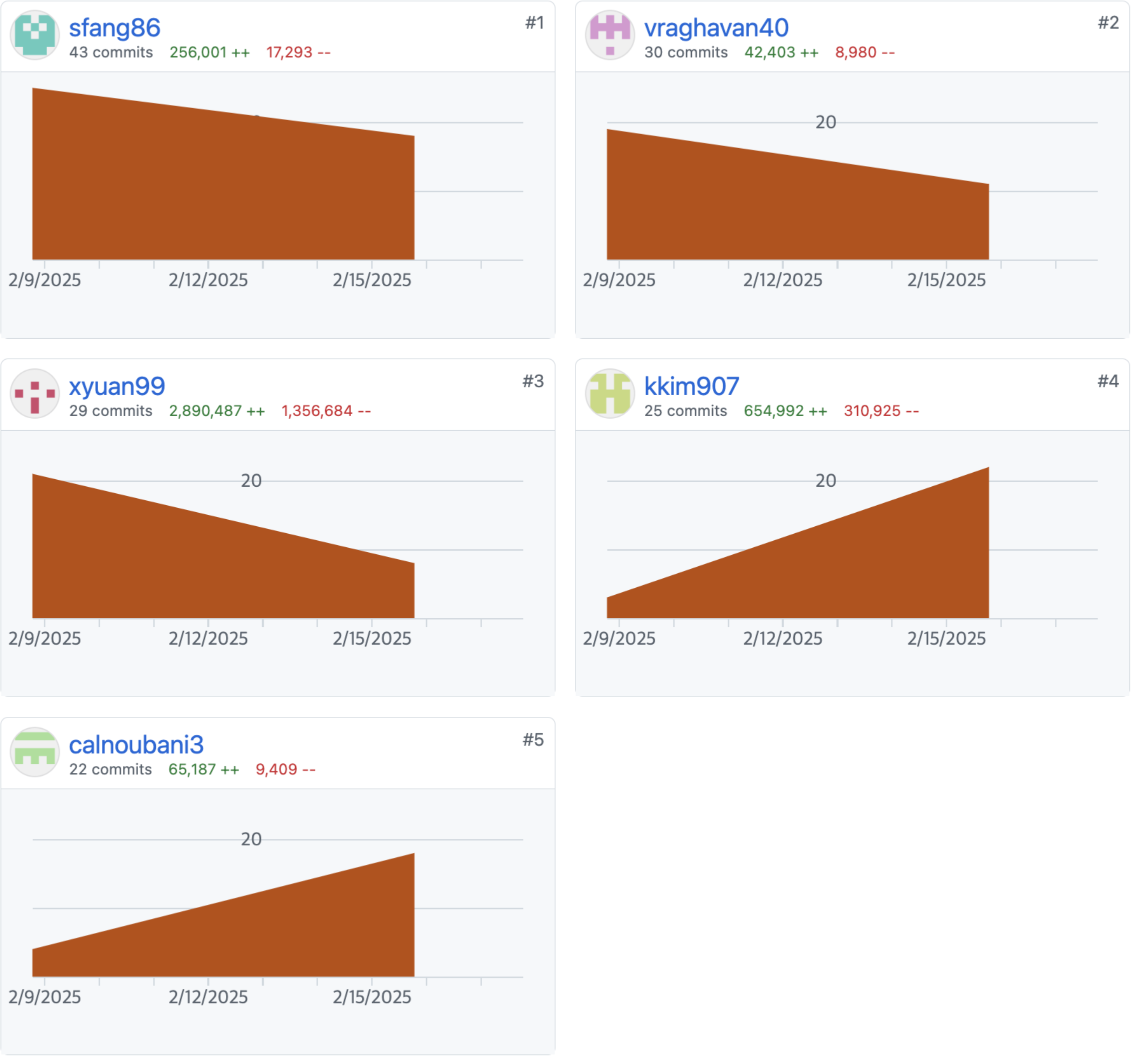
Conclusion

- Based on CDS and Mean CDS Length, Prodigal was the best gene prediction tool out of the Ab Initio tools and overall.
- GeMoMa was the best out of the homology-based gene prediction tools
- Prodigal had the shortest runtime and moderate RAM (RSS) consumption
- GeMoMa used far more RAM and CPU than BLAST
- EggNOG shared more annotations with the reference genome annotation
- Thus, the optimal workflow would involve **Prodigal** for prediction and **EggNOG** for annotation
- **Limitation:** Our workflow based on *Neisseria gonorrhoeae* . The validity for other more distantly related species still needs to be verified
- **Moving forward:** There is a tool based on prodigal that can do both gene prediction and annotation - Prokka - that seems to be very handy, and we'll try Prokka in the later process

References

- Stanke, Mario, and Stephan Waack. "Gene prediction with a hidden Markov model and a new intron submodel." *Bioinformatics-Oxford* 19.2 (2003): 215-225.
- Keilwagen, Jens, Frank Hartung, and Jan Grau. "GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data." *Gene prediction: Methods and protocols* (2019): 161-177.
- Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.
- Hyatt, Doug, et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* 11 (2010): 1-11.
- Delcher, Arthur L., et al. "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics* 23.6 (2007): 673-679.
- Lomsadze, Alexandre, et al. "Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes." *Genome research* 28.7 (2018): 1079-1089.
- Pertea, Geo, and Mihaela Pertea. "GFF utilities: GffRead and GffCompare." *F1000Research* 9 (2020).
- Nicholas J Dimonaco, Wayne Aubrey, Kim Kenobi, Amanda Clare, Christopher J Creevey, No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study, Bioinformatics, Volume 38, Issue 5, March 2022, Pages 1198–1207, https://doi.org/10.1093/bioinformatics/btab827

GitHub Activity



sfang86 #1
43 commits 256,001 ++ 17,293 --

vraghavan40 #2
30 commits 42,403 ++ 8,980 --

xyuan99 #3
29 commits 2,890,487 ++ 1,356,684 --

kkim907 #4
25 commits 654,992 ++ 310,925 --

calnoubani3 #5
22 commits 65,187 ++ 9,409 --

Thank you!

--Team B Group 2

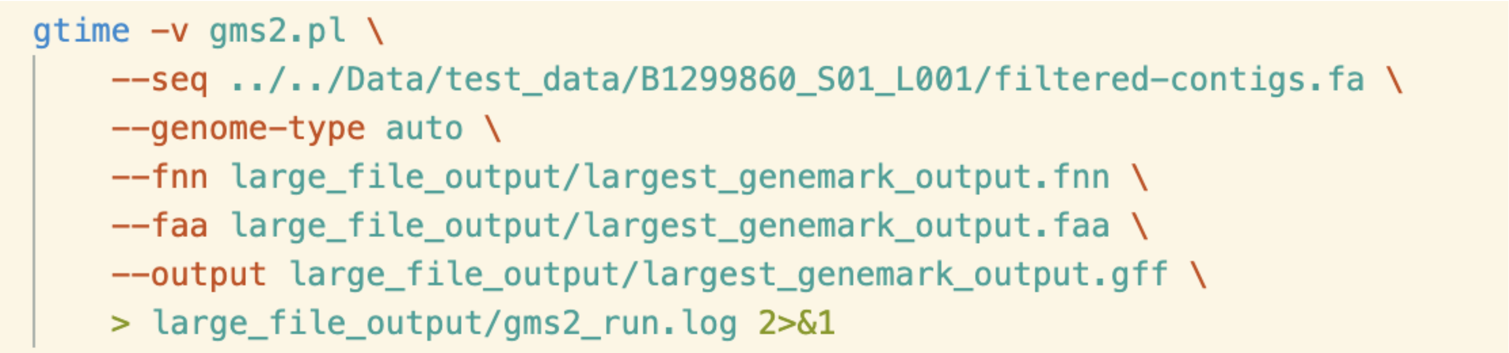
Appendix

Command and output of other tools that we researched

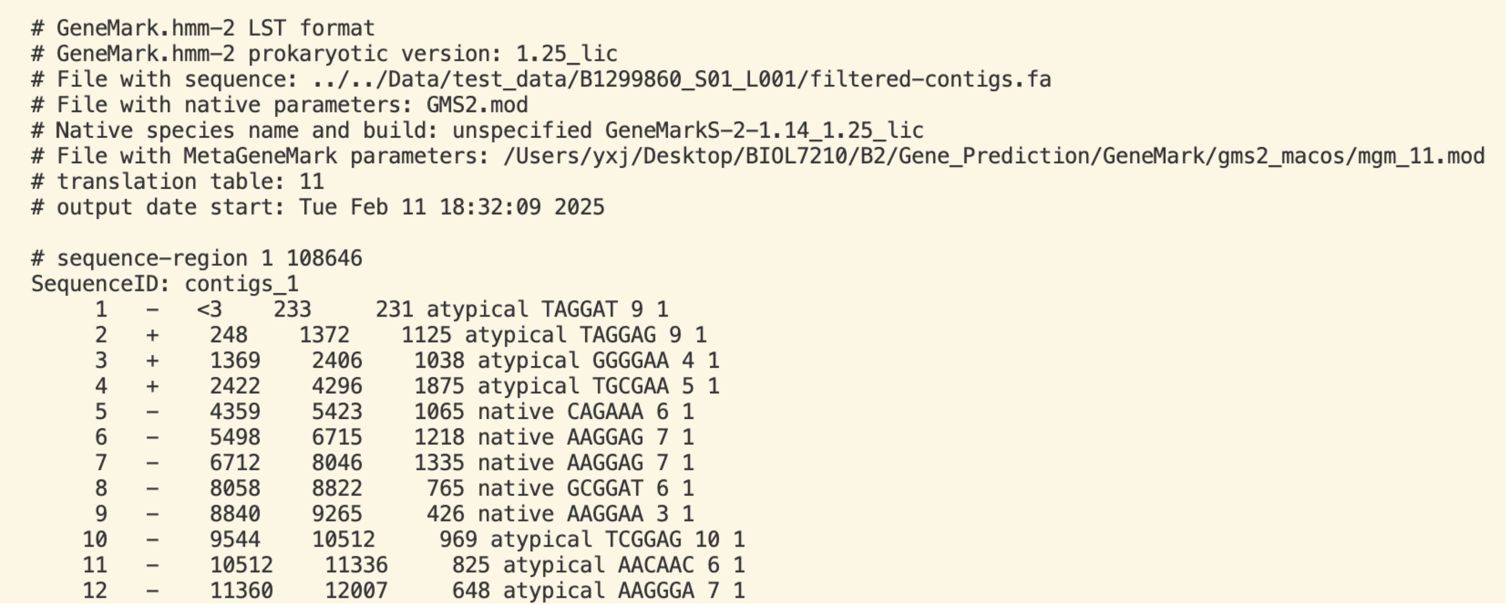
GeneMarkS-2

- Suited for prokaryotic genomes, predicts CDSs with high accuracy.
- Output format: GFF file showing predicted CDSs. FNN and FAA file contains nucleotide and protein sequences of predicted genes

	# CDS Predicted
Largest File	2317
Smallest File	2318



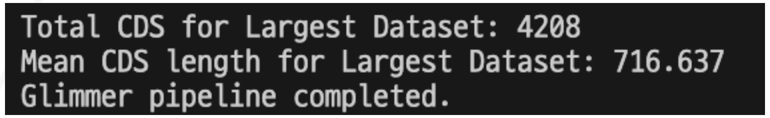
```
gtime -v gms2.pl \
    --seq ../../Data/test_data/B1299860_S01_L001/filtered-contigs.fa \
    --genome-type auto \
    --fnn large_file_output/largest_genemark_output.fnn \
    --faa large_file_output/largest_genemark_output.faa \
    --output large_file_output/largest_genemark_output.gff \
    > large_file_output/gms2_run.log 2>&1
```



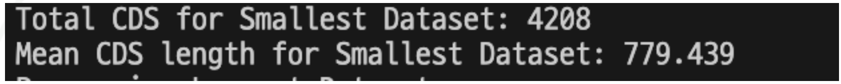
```
# GeneMark.hmm-2 LST format
# GeneMark.hmm-2 prokaryotic version: 1.25_lic
# File with sequence: ../../Data/test_data/B1299860_S01_L001/filtered-contigs.fa
# File with native parameters: GMS2.mod
# Native species name and build: unspecified GeneMarkS-2-1.14_1.25_lic
# File with MetaGeneMark parameters: /Users/yxj/Desktop/BIOL7210/B2/Gene_Prediction/GeneMark/gms2_macos/mgm_11.mod
# translation table: 11
# output date start: Tue Feb 11 18:32:09 2025

# sequence-region 1 108646
SequenceID: contigs_1
     1   -   <3    233     231 atypical TAGGAT 9 1
     2   +    248    1372    1125 atypical TAGGAG 9 1
     3   +    1369    2406    1038 atypical GGGGAA 4 1
     4   +    2422    4296    1875 atypical TGCGAA 5 1
     5   -    4359    5423    1065 native CAGAAA 6 1
     6   -    5498    6715    1218 native AAGGAG 7 1
     7   -    6712    8046    1335 native AAGGAG 7 1
     8   -    8058    8822     765 native GCGGAT 6 1
     9   -    8840    9265     426 native AAGGAA 3 1
    10   -    9544    10512     969 atypical TCGGAG 10 1
    11   -    10512    11336     825 atypical AACAAC 6 1
    12   -    11360    12007     648 atypical AAGGGA 7 1
```

Glimmer: Parameters & Output

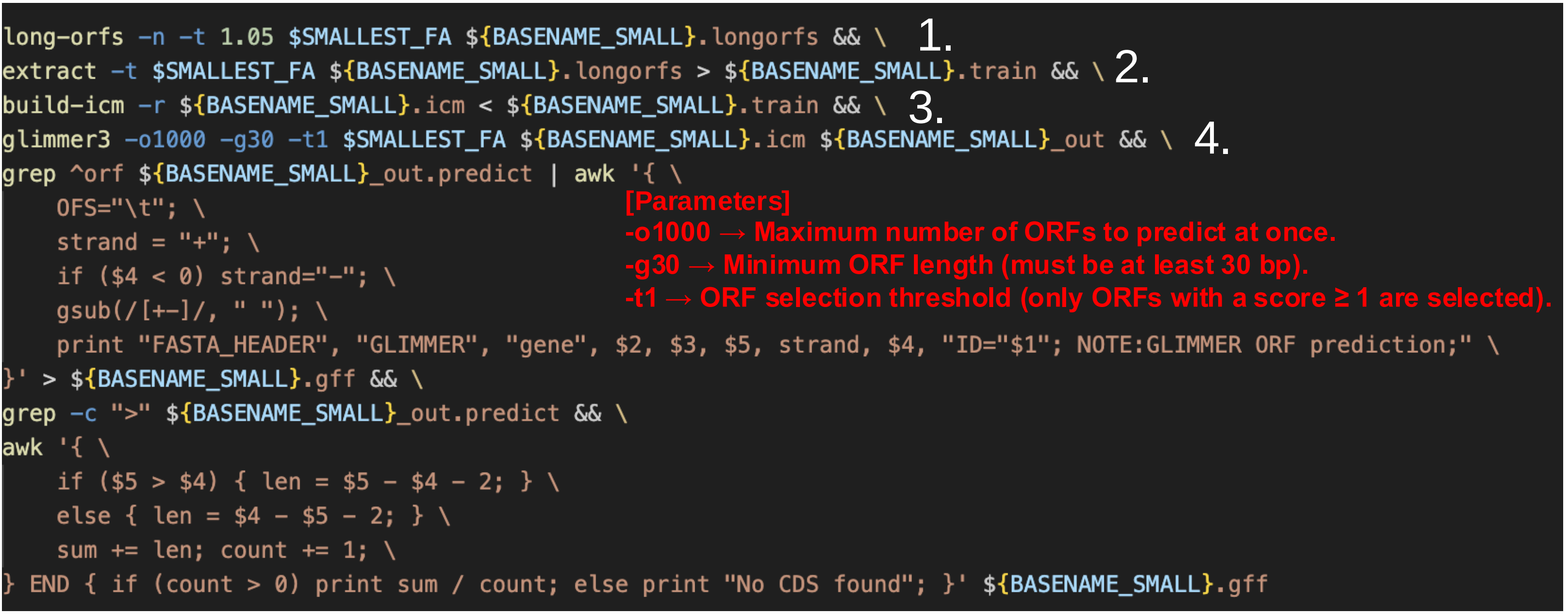
- Uses interpolated Markov models, sensitive to short genes.
- Effective for bacterial genomes but may overpredict in some cases.



```
Total CDS for Largest Dataset: 4208
Mean CDS length for Largest Dataset: 716.637
Glimmer pipeline completed.
```



```
Total CDS for Smallest Dataset: 4208
Mean CDS length for Smallest Dataset: 779.439
```



```
long-orfs -n -t 1.05 $SMALLEST_FA ${BASENAME_SMALL}.longorfs && \  1.
extract -t $SMALLEST_FA ${BASENAME_SMALL}.longorfs > ${BASENAME_SMALL}.train && \ 2.
build-icm -r ${BASENAME_SMALL}.icm < ${BASENAME_SMALL}.train && \ 3.
glimmer3 -o1000 -g30 -t1 $SMALLEST_FA ${BASENAME_SMALL}.icm ${BASENAME_SMALL}_out && \ 4.
grep ^orf ${BASENAME_SMALL}_out.predict | awk '{ \
    OFS="\t"; \
    strand = "+"; \
    if ($4 < 0) strand="-"; \
    gsub(/[+-]/, " "); \
    print "FASTA_HEADER", "GLIMMER", "gene", $2, $3, $5, strand, $4, "ID="$1"; NOTE:GLIMMER ORF prediction;" \
}' > ${BASENAME_SMALL}.gff && \
grep -c ">" ${BASENAME_SMALL}_out.predict && \
awk '{ \
    if ($5 > $4) { len = $5 - $4 - 2; } \
    else { len = $4 - $5 - 2; } \
    sum += len; count += 1; \
} END { if (count > 0) print sum / count; else print "No CDS found"; }' ${BASENAME_SMALL}.gff
```

[Parameters]
-o1000 → Maximum number of ORFs to predict at once.
-g30 → Minimum ORF length (must be at least 30 bp).
-t1 → ORF selection threshold (only ORFs with a score ≥ 1 are selected).

	# CDS Predicted
Largest File	2574
Smallest File	2550

1. Find long-orfs — Initial Filtering for CDS Predicion

2. Extract — Generating training data

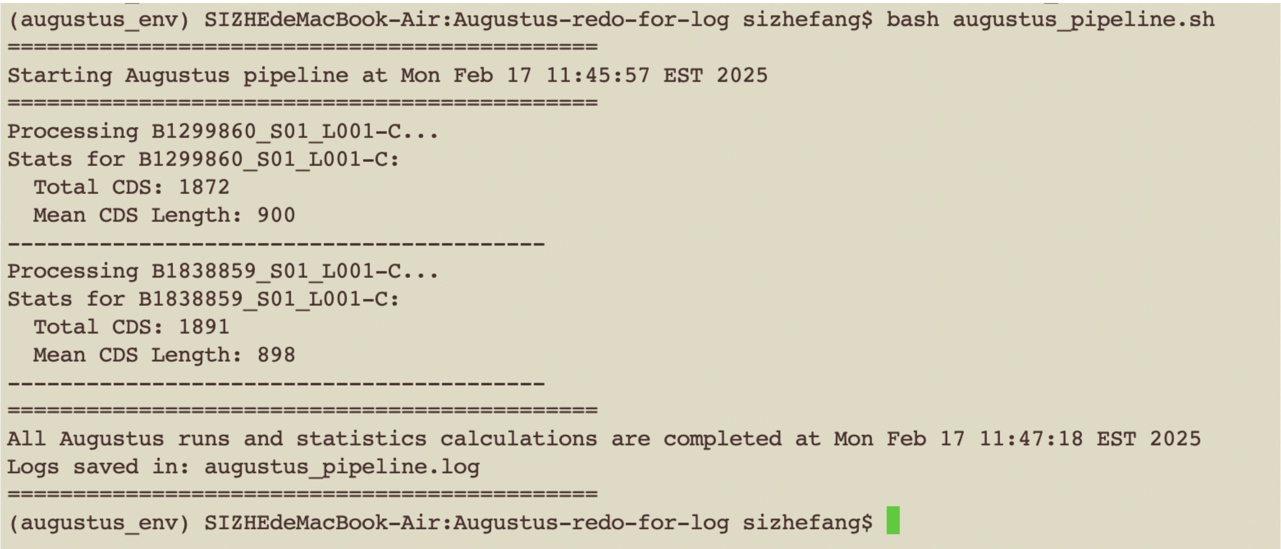
3. Build-icm — ICM(Interpolated Context Model)

4. Glimmer3 — * needed to transform .predict format to .gff format

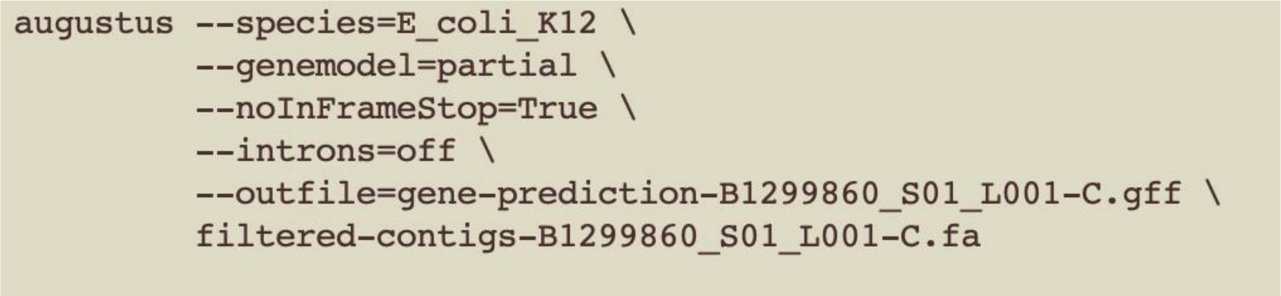
Augustus: Parameters & Output

- A software tool for gene prediction in eukaryotes based on a Generalized Hidden Markov Model
- Has been pre-trained for some species and can perform gene prediction for similar species
- Performs best on eukaryotes, but works on prokaryotes as well
- *Neisseria gonorrhoeae* is Gram-negative bacteria, can try to predict with pre-trained model for *E.coli*

	# CDS Predicted
Largest File	1872
Smallest File	1891



```
(augustus_env) SIZHEdeMacBook-Air:Augustus-redo-for-log sizhefang$ bash augustus_pipeline.sh
=============================================
Starting Augustus pipeline at Mon Feb 17 11:45:57 EST 2025
=============================================
Processing B1299860_S01_L001-C...
Stats for B1299860_S01_L001-C:
  Total CDS: 1872
  Mean CDS Length: 900
-----------------------------------------
Processing B1838859_S01_L001-C...
Stats for B1838859_S01_L001-C:
  Total CDS: 1891
  Mean CDS Length: 898
-----------------------------------------
=============================================
All Augustus runs and statistics calculations are completed at Mon Feb 17 11:47:18 EST 2025
Logs saved in: augustus_pipeline.log
=============================================
(augustus_env) SIZHEdeMacBook-Air:Augustus-redo-for-log sizhefang$
```

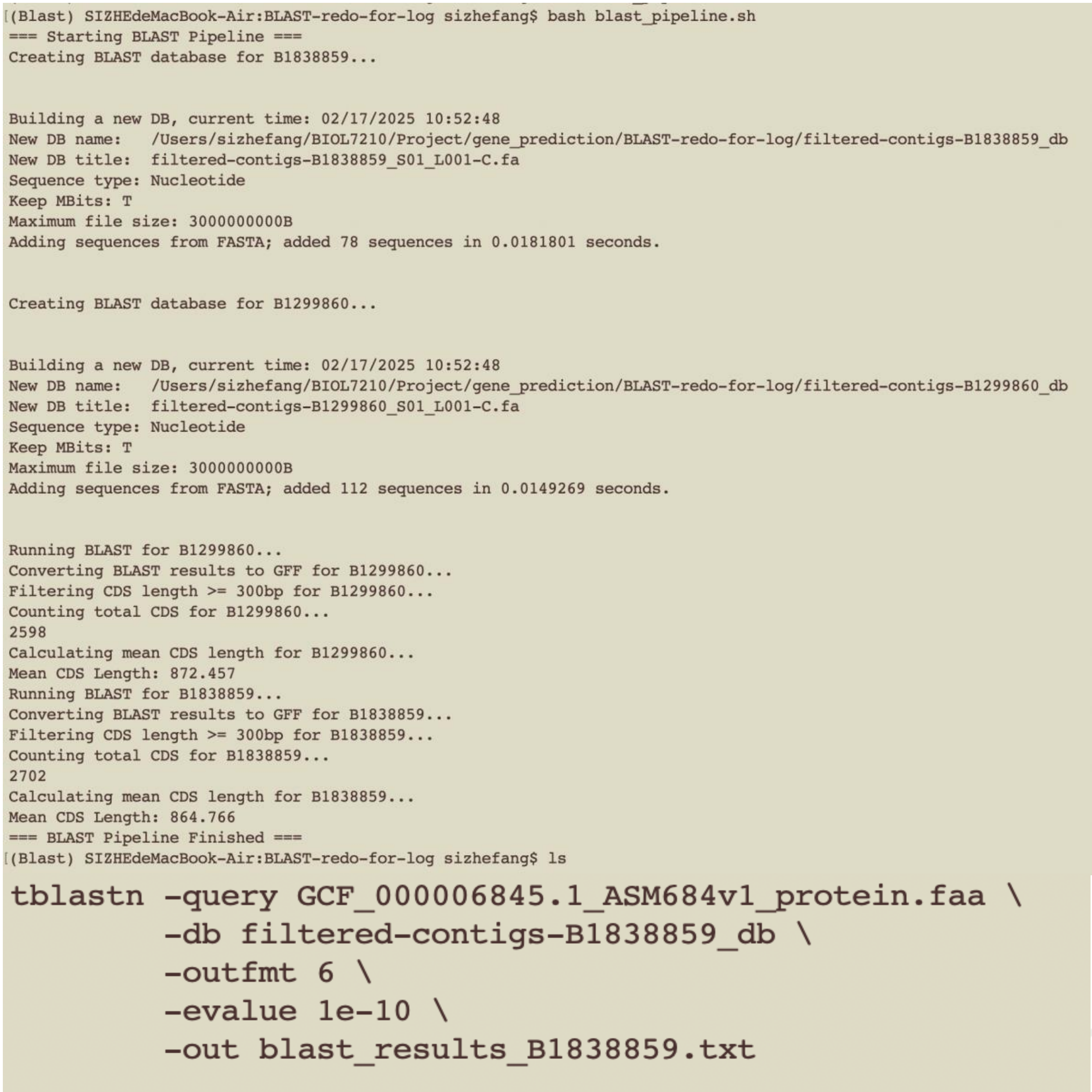


```
augustus --species=E_coli_K12 \
         --genemodel=partial \
         --noInFrameStop=True \
         --introns=off \
         --outfile=gene-prediction-B1299860_S01_L001-C.gff \
         filtered-contigs-B1299860_S01_L001-C.fa
```

BLAST Parameters & Output

- Homology-based: Used tblastn to identify CDS by comparing reference genome proteins to the target genomes
- Converted BLAST results to GFF format, correcting start > end regions and assigning strand orientation
- Retained CDS ≥ 300bp, then calculated Total CDS and Mean CDS Length for quality assessment

	# CDS Predicted
Largest File	2598
Smallest File	2702



```
(Blast) SIZHEdeMacBook-Air:BLAST-redo-for-log sizhefang$ bash blast_pipeline.sh
=== Starting BLAST Pipeline ===
Creating BLAST database for B1838859...


Building a new DB, current time: 02/17/2025 10:52:48
New DB name:   /Users/sizhefang/BIOL7210/Project/gene_prediction/BLAST-redo-for-log/filtered-contigs-B1838859_db
New DB title:  filtered-contigs-B1838859_S01_L001-C.fa
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 3000000000B
Adding sequences from FASTA; added 78 sequences in 0.0181801 seconds.


Creating BLAST database for B1299860...


Building a new DB, current time: 02/17/2025 10:52:48
New DB name:   /Users/sizhefang/BIOL7210/Project/gene_prediction/BLAST-redo-for-log/filtered-contigs-B1299860_db
New DB title:  filtered-contigs-B1299860_S01_L001-C.fa
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 3000000000B
Adding sequences from FASTA; added 112 sequences in 0.0149269 seconds.


Running BLAST for B1299860...
Converting BLAST results to GFF for B1299860...
Filtering CDS length >= 300bp for B1299860...
Counting total CDS for B1299860...
2598
Calculating mean CDS length for B1299860...
Mean CDS Length: 872.457
Running BLAST for B1838859...
Converting BLAST results to GFF for B1838859...
Filtering CDS length >= 300bp for B1838859...
Counting total CDS for B1838859...
2702
Calculating mean CDS length for B1838859...
Mean CDS Length: 864.766
=== BLAST Pipeline Finished ===
(Blast) SIZHEdeMacBook-Air:BLAST-redo-for-log sizhefang$ ls
```

```
tblastn -query GCF_000006845.1_ASM684v1_protein.faa \
        -db filtered-contigs-B1838859_db \
        -outfmt 6 \
        -evalue 1e-10 \
        -out blast_results_B1838859.txt
```