B2 Gene Prediction & Annotation: Final Results

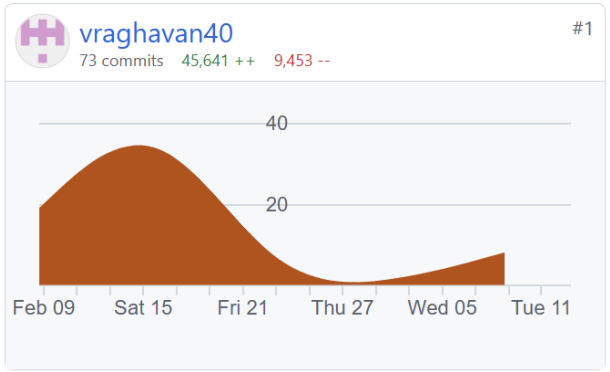
By: Celine Al-Noubani, Vishank Raghavan, Kyungbeom Kim, Sizhe Fang, Xuejiao (Jessica) Yuan

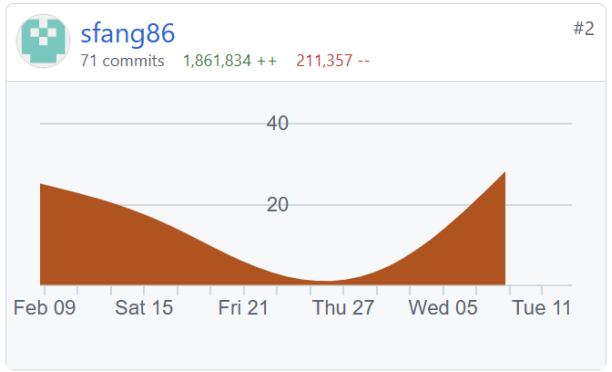
Georgia Tech

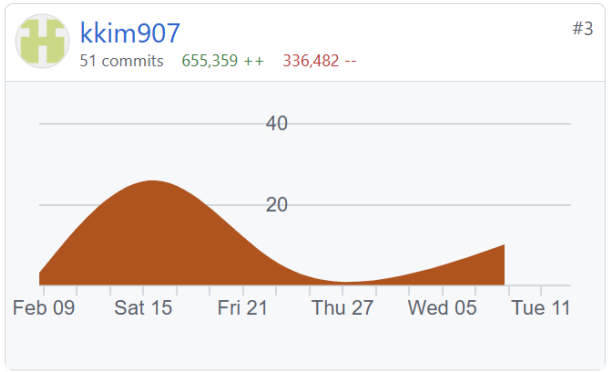
Breakdown & Student Roles

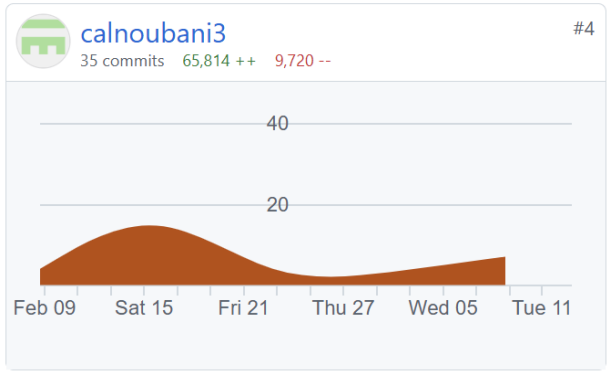
- Top Preliminary Results
  - Gene Prediction: Prodigal (ab initio) and GeMoMa (homology)
  - Gene Annotation: EggNog and InterPro
- Prodigal-EggNog Workflow (Kyungbeom)
- Prodigal-InterPro-Barrnap Workflow (Jessica)
- GeMoMa-EggNog Workflow (Sizhe)
- GeMoMa-InterPro Workflow (Celine)
- Final Pipeline Integration (Vishank)

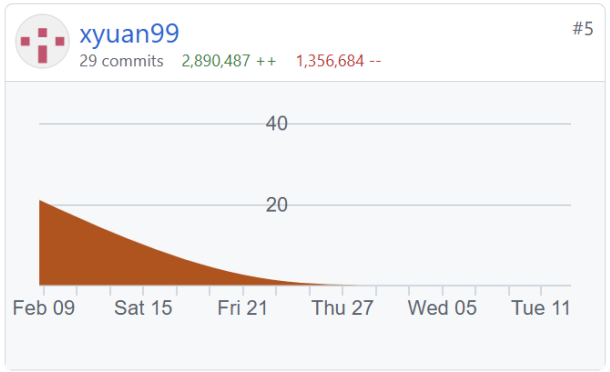
Github Activity





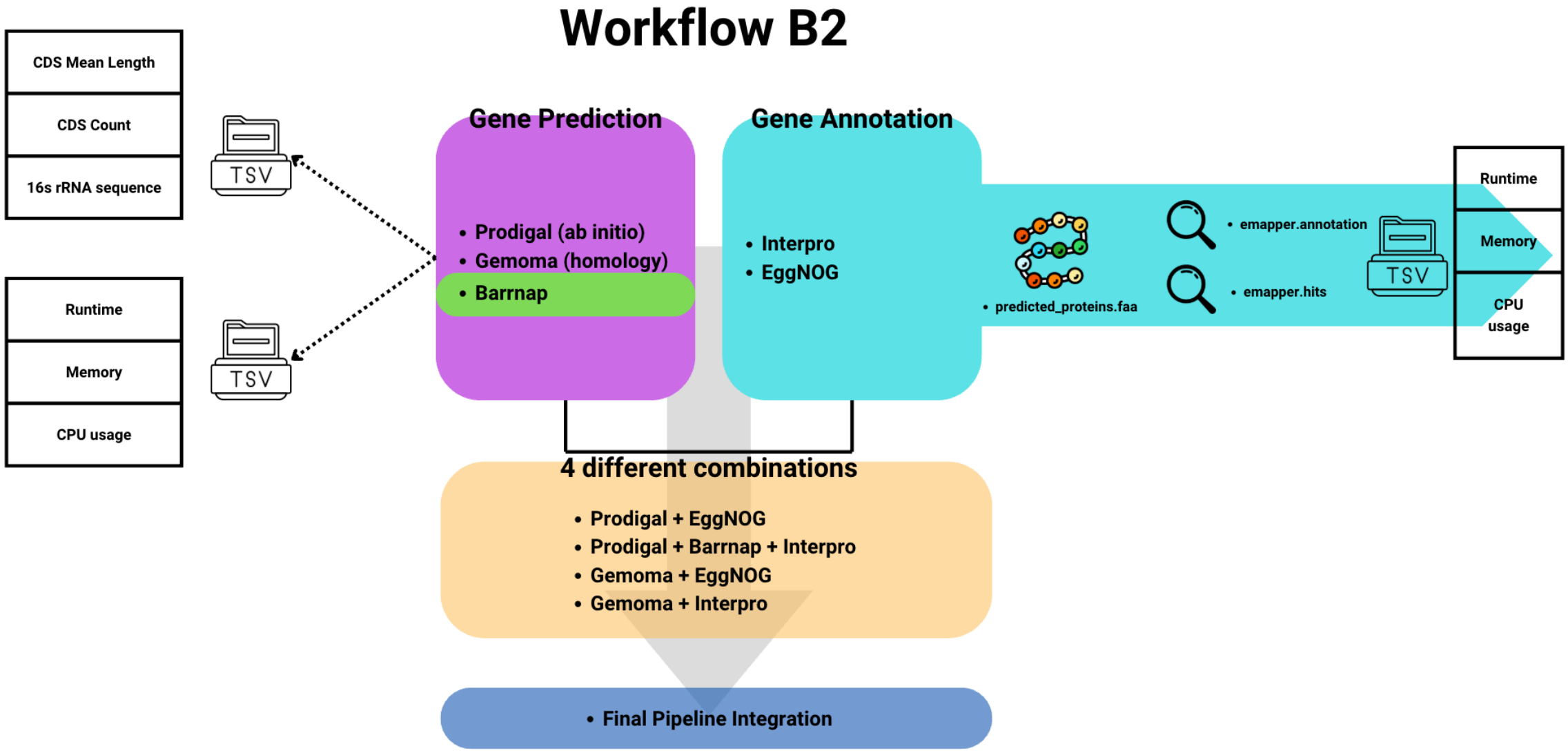






Georgia Tech

Workflow


Workflow B2
CDS Mean Length
CDS Count
16s rRNA sequence
TSV
Runtime
Memory
CPU usage
TSV
Gene Prediction
• Prodigal (ab initio)
• Gemoma (homology)
• Barrnap
Gene Annotation
• Interpro
• EggNOG
• predicted_proteins.faa
• emapper.annotation
• emapper.hits
TSV
Runtime
Memory
CPU usage
4 different combinations
• Prodigal + EggNOG
• Prodigal + Barrnap + Interpro
• Gemoma + EggNOG
• Gemoma + Interpro
• Final Pipeline Integration

Final Pipeline

```
$./gene_prediction_and_annnotation.py -h
usage: gene_prediction_and_annnotation.py [-h] [-p] [-g] -i [INPUT ...] [-c] [-pl PRED_LOGS] [-po PRED_OUTPUT] [-pm PRED_METRICS] [-pmn [METRIC_NAME]] [-pd]
                                          [-pp PRODIGAL_PARAMETERS] [-gp GEMOMA_PARAMETERS] [-gr GEMOMA_REF] [-ga GEMOMA_ANN] [-t [THREADS]] [-r] [-rp RRNA_PARAMETERS] [-ip]
                                          [-isp INTERPRO_SCRIPT] [-eg] [-ed EGGNOG_DB] [-ad] [-ipp INTERPRO_PARAMETERS] [-ep EGGNOG_PARAMETERS] [-al ANN_LOGS] [-ao ANN_OUTPUT]
                                          [-am ANN_METRICS] [-amn [ANN_METRIC_NAME]] [-op]

Runs gene prediction and annotation pipeline using prodigal/gemoma for gene prediction and interpro/eggnog for annotation

options:
  -h, --help            show this help message and exit
  -p                    Use prodigal for gene prediction
  -g                    Use gemoma for gene prediction
  -i [INPUT ...]        input contig assembly file(s)
  -c                    Indicates input files are compressed (.gz)
  -pl PRED_LOGS         Prediction logs directory
  -po PRED_OUTPUT       Prediction output (gff files + aa translations) directory
  -pm PRED_METRICS      Prediction metrics output directory
  -pmn [METRIC_NAME]    Prediction Metric File Name (String not path with no extension)
  -pd                   Use Default Prediction Parameters
  -pp PRODIGAL_PARAMETERS
                        Prodigal parameters (String containing prodigal arguments)
  -gp GEMOMA_PARAMETERS
                        Gemoma parameters (String containing gemoma arguments)
  -gr GEMOMA_REF        Path to ref genoma (fna) for gemoma
  -ga GEMOMA_ANN        Path to ref annotation (gff) for gemoma
  -t [THREADS]          (CPU) Threads for all multithreaded tools
  -r                    Do 16S rRNA prediction via barrnap
  -rp RRNA_PARAMETERS   16S rRNA parameters
  -ip                   Use InterPro for annotation
  -isp INTERPRO_SCRIPT  Path to interpro install(location of interpro.sh script + other intepro files)
  -eg                   Use eggnog for annotation
  -ed EGGNOG_DB         Path to eggnog data dir(eggnog database directory)
  -ad                   Use default annotation parameters
  -ipp INTERPRO_PARAMETERS
                        InterPro Parameters (String containing InterPro arguments)
  -ep EGGNOG_PARAMETERS
                        Eggnog Parameters (String containing Eggnog arguments)
  -al ANN_LOGS          Annotation logs directory
  -ao ANN_OUTPUT        Annotation output (annotated gff files) directory
  -am ANN_METRICS       Annotation metrics output directory
  -amn [ANN_METRIC_NAME]
                        Annotation Metric File Name (String not path with no extension)
  -op                   Only do gene prediction
```

```
$./gene_prediction_and_annnotation.py -p -i ./data/*.fa.gz -c -pd -t 24 -eg -ed ./eggnog_data -ad
Gene Prediction Completed! Logs verifying completion can be found in ./prediction_logs.
Gene Annotation Completed! Logs verifying completion can be found in ./annotation_logs.
```

Georgia Tech

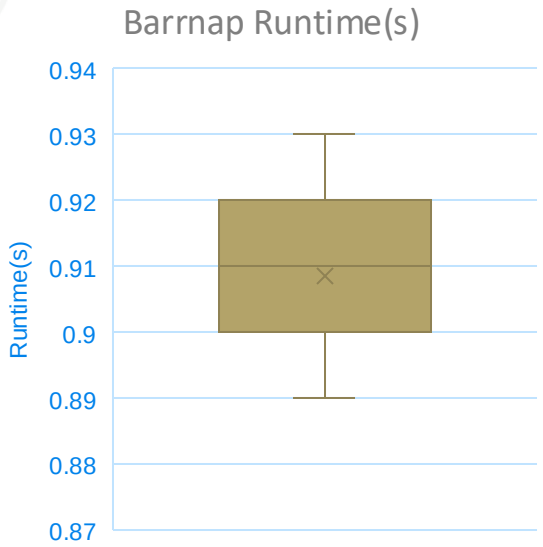
System Specifications (Standardization)

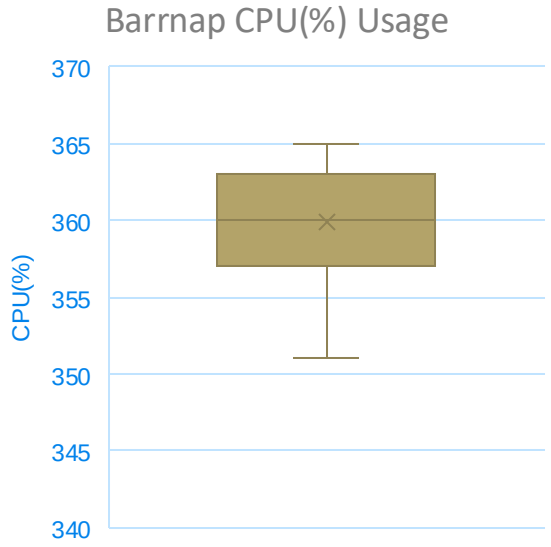
- **Cluster Environment**: PACE Cluster
- **Modules**: Anaconda 3 (2023.03)
- **Quality of Service**: coc-ice
- **Node Type**: AMD CPU
  - **Processor Model**: AMD EPYC 7513 32-Core Processor
  - **Allocated Cores**: 16
- **Memory**: 32 GB per node

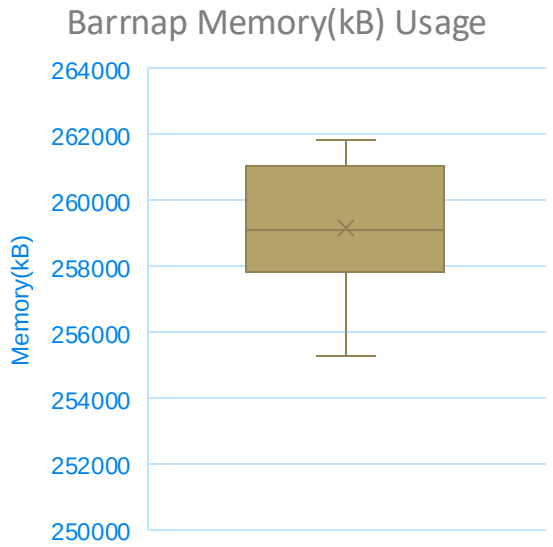
Barrnap: 16S

```
1    ##gff-version 3
2    contigs_49      barrnap:0.9     rRNA    266     1801    0       +       .       Name=16S_rRNA;product=16S ribosomal RNA
3    contigs_49      barrnap:0.9     rRNA    2409    5297    0       +       .       Name=23S_rRNA;product=23S ribosomal RNA
4    contigs_49      barrnap:0.9     rRNA    5397    5505
```

```
1    >16S_rRNA::contigs_49:265-1801(+)
2    TAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCTTTACACATGCAAGTCGGACGGCAGCACAGGGAAGCTTGCTTCTCGGGTGGCGAGTGGCGAACGGGTGAGTAATATATCGGAACGT
3    >23S_rRNA::contigs_49:2408-5297(+)
4    TCAAGTGAATAAGTGCATCAGGCGGATGCCTTGGCGATGATAGGCGACGAAGGACGTGTAAGCCTGCGAAAAGCGCGGGGGAGCTGGCAATAAAGCTATGATTCCGCGATGTCCGAATGGGGAAACCC
5    >5S_rRNA::contigs_49:5396-5505(+)
6    TGGCGGCCATAGCGAGTTGGTCCCACGCCTTCCCATCCCGAACAGGACCGTGAAACGACTCAGCGCCGATGATAGTGTGGTTCTTCCATGCGAAAGTAGGTCACTGCCA
```







Avg. Runtime (sec)	0.91
Avg. Memory Usage (kB)	259188
Avg. CPU Usage (%)	359.88

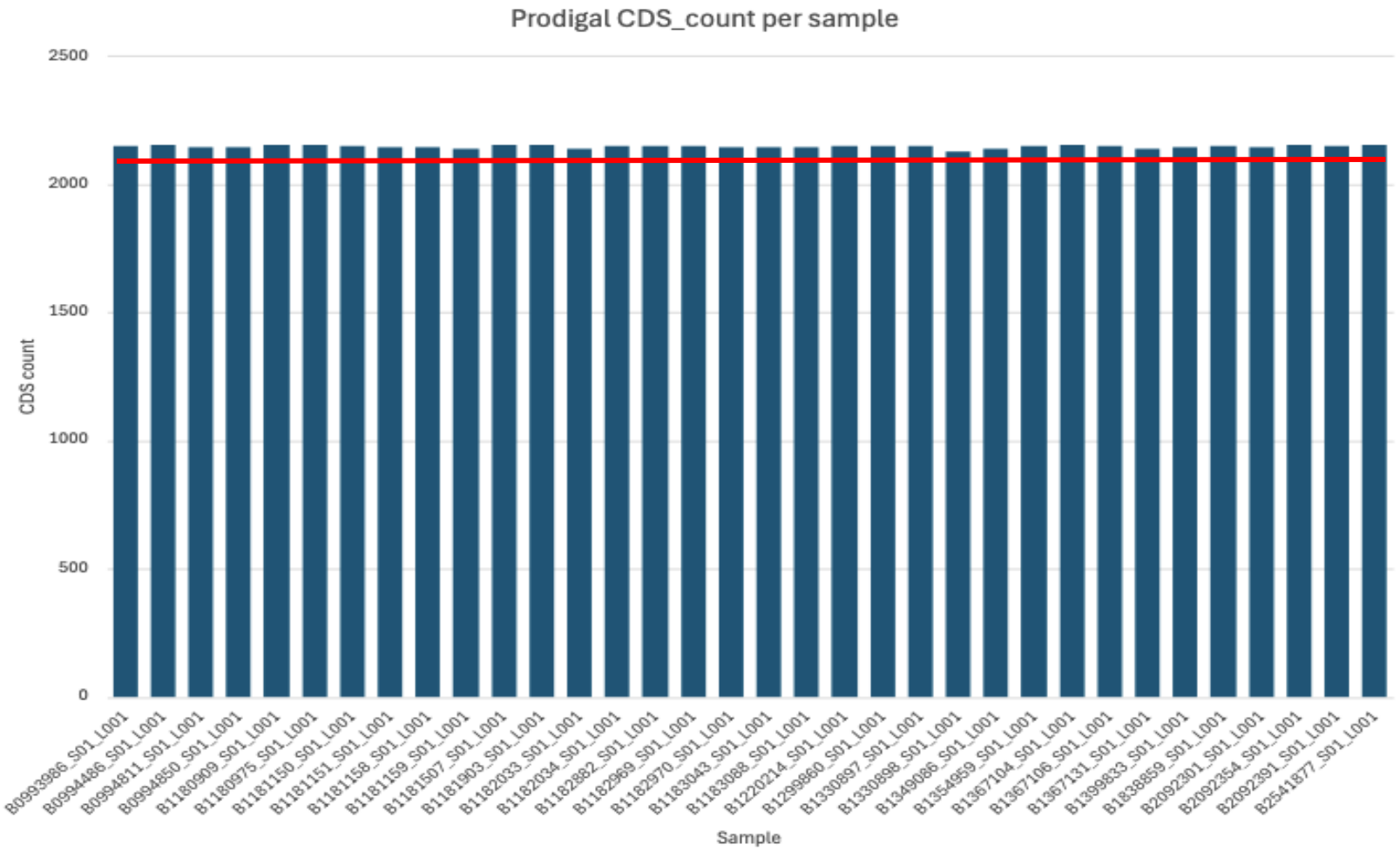
- Each sample contains 1 16s rRNA sequence.

Georgia Tech

Gene Prediction

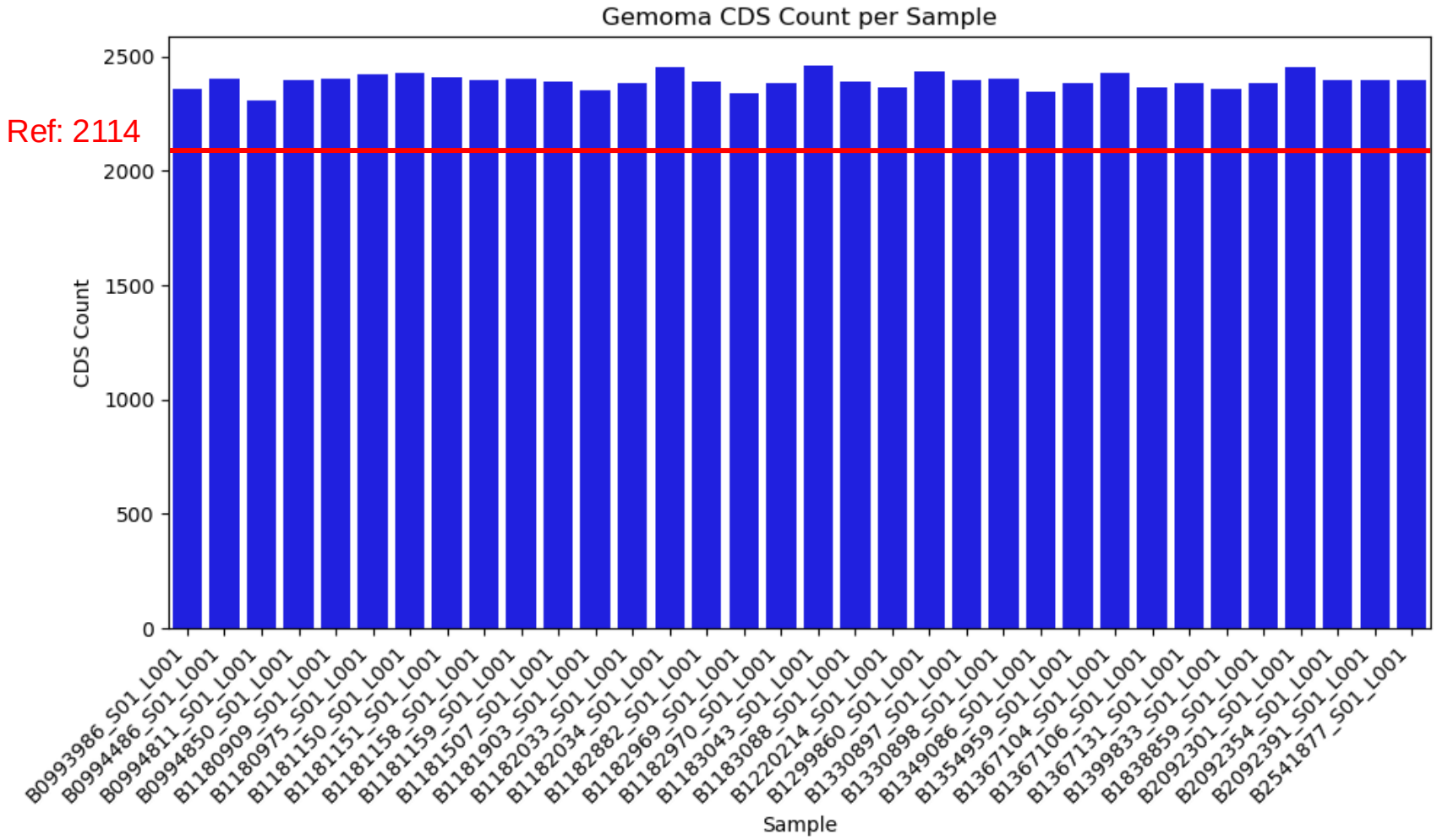
Prodigal: CDS Count

Ref: 2114



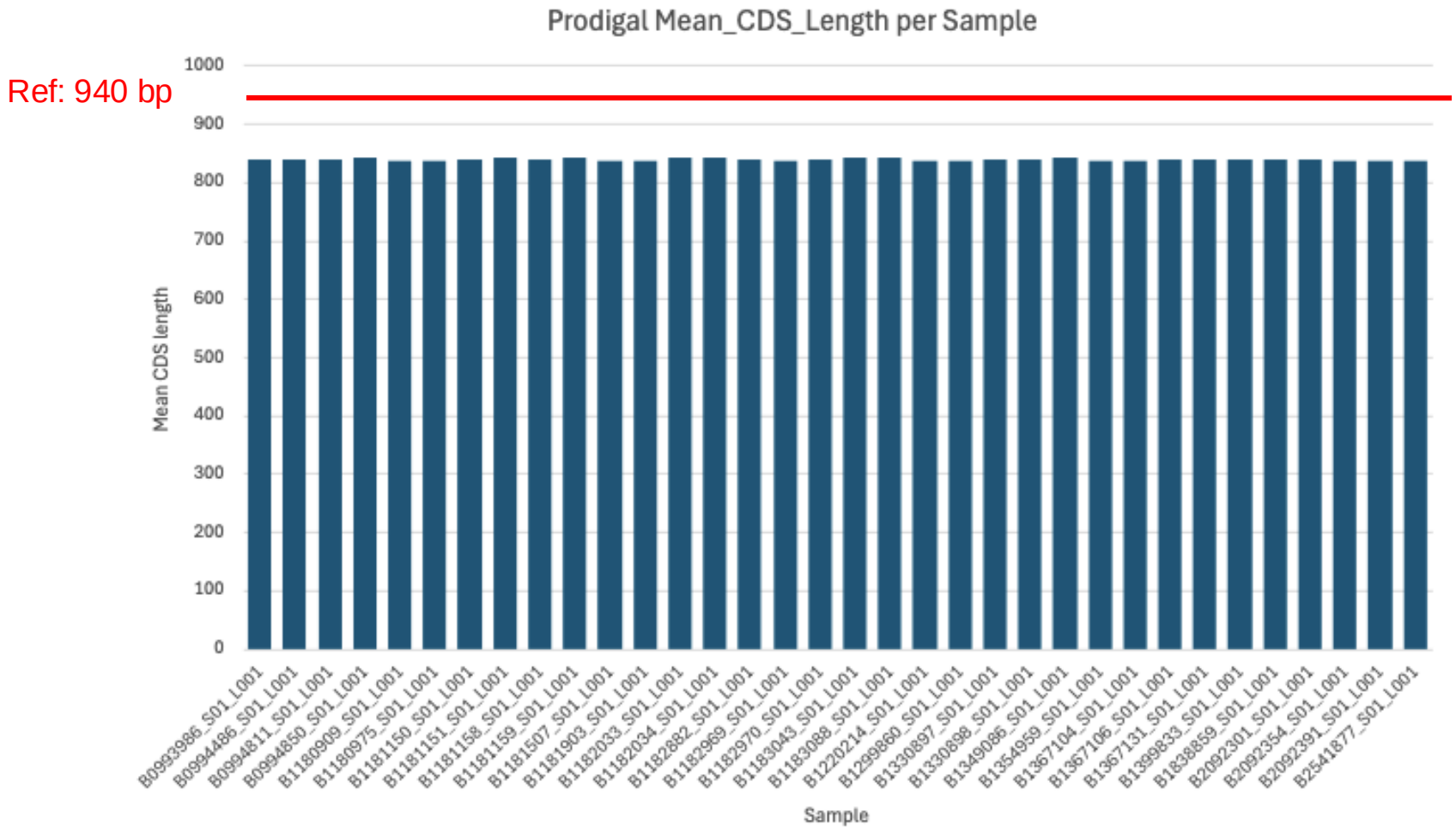
Georgia Tech

GeMoMa: CDS Count



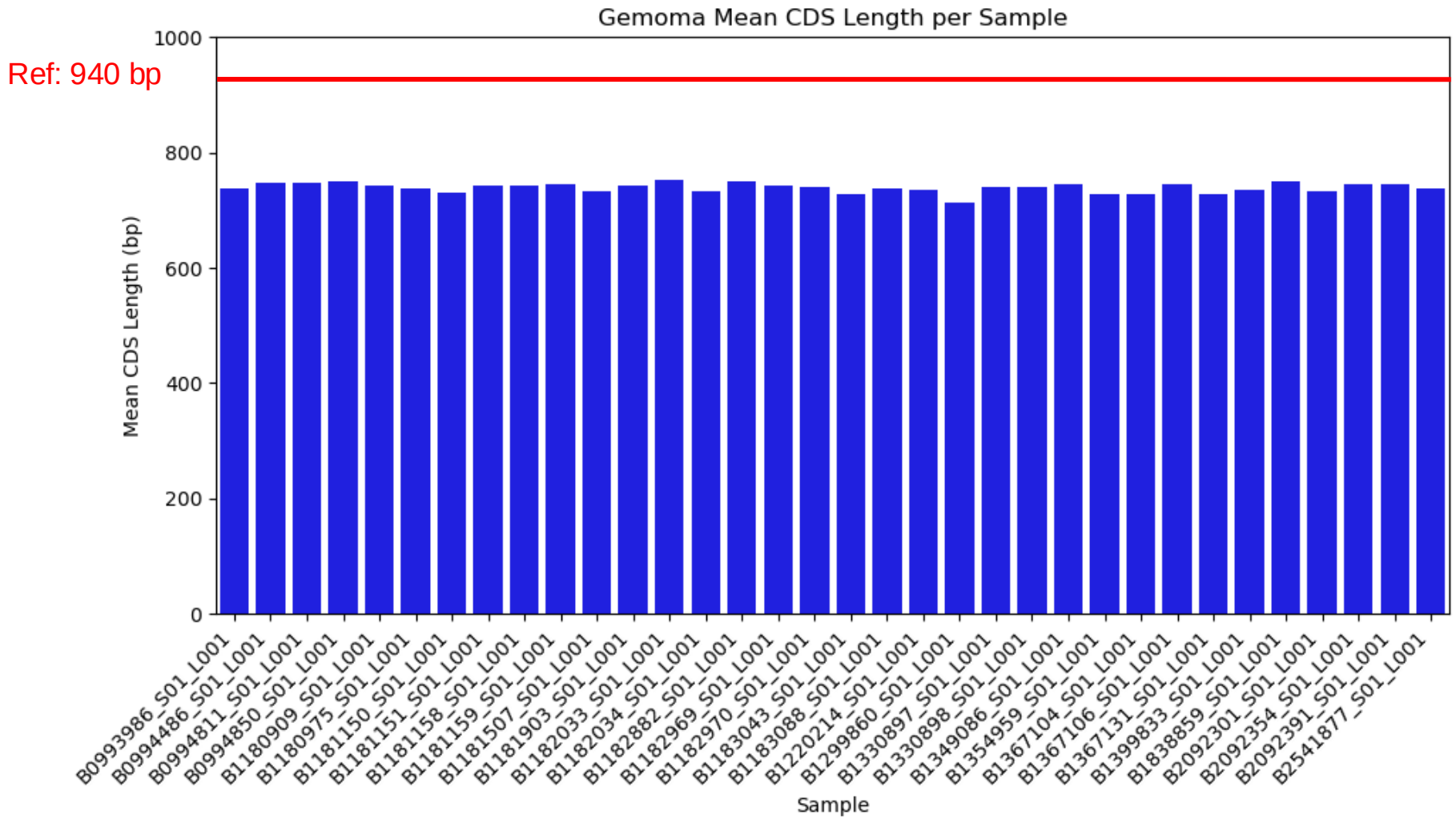
Georgia Tech

Prodigal: CDS Mean Length



Georgia Tech

GeMoMa: CDS Mean Length



Georgia Tech

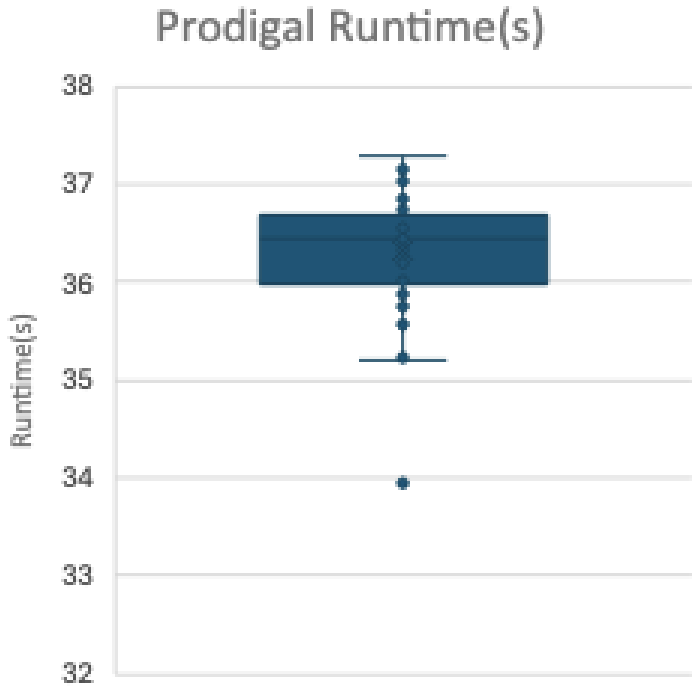
Prodigal vs GeMoMa: CDS Count Comparison

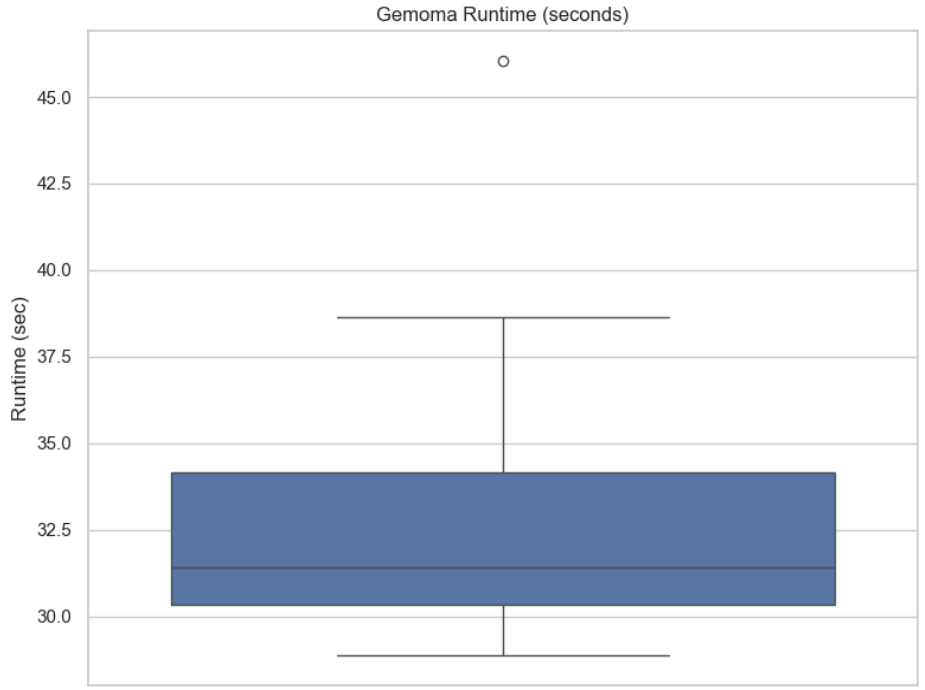
Tool	Total CDS	Mean CDS Length
Prodigal	**2147**	**838.70**
GeMoMa	2393	739
Ground Truth (Latest Ref. Genome)	**2114**	**~940**

Average across 34 runs

Georgia Tech

Prodigal vs GeMoMa: Runtime, Memory, CPU Usage



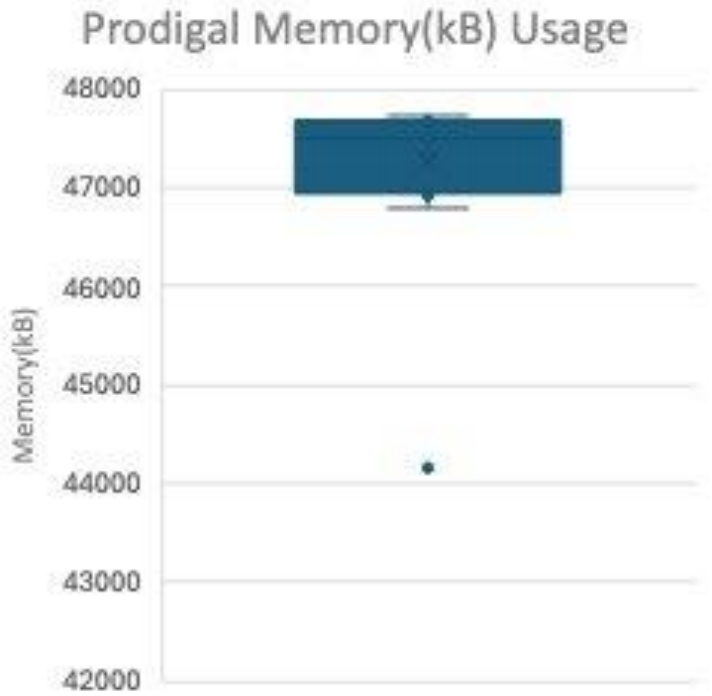


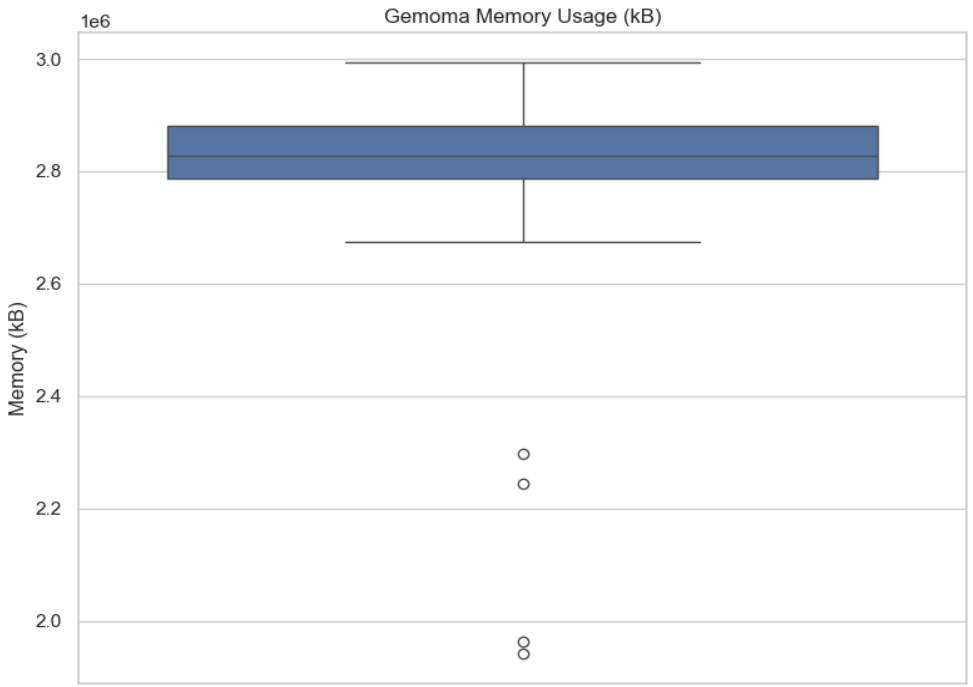
Prodigal	
Avg. Runtime (sec)	**36.32**
Avg. Memory Usage (kB)	47254
Avg. CPU Usage (%)	99

GeMoMa	
Avg. Runtime (sec)	**32.62**
Avg. Memory Usage (kB)	2760796.24
Avg. CPU Usage (%)	335.03

Georgia Tech

Prodigal vs GeMoMa: Runtime, Memory, CPU Usage



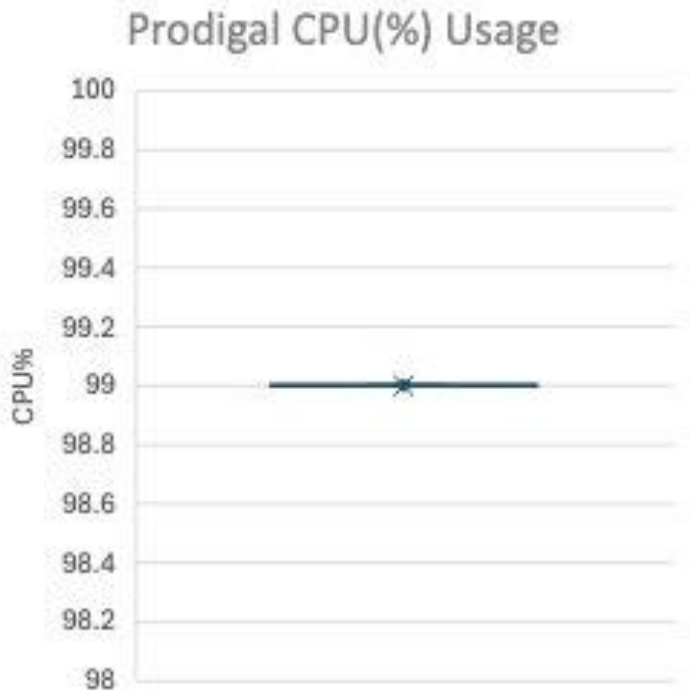


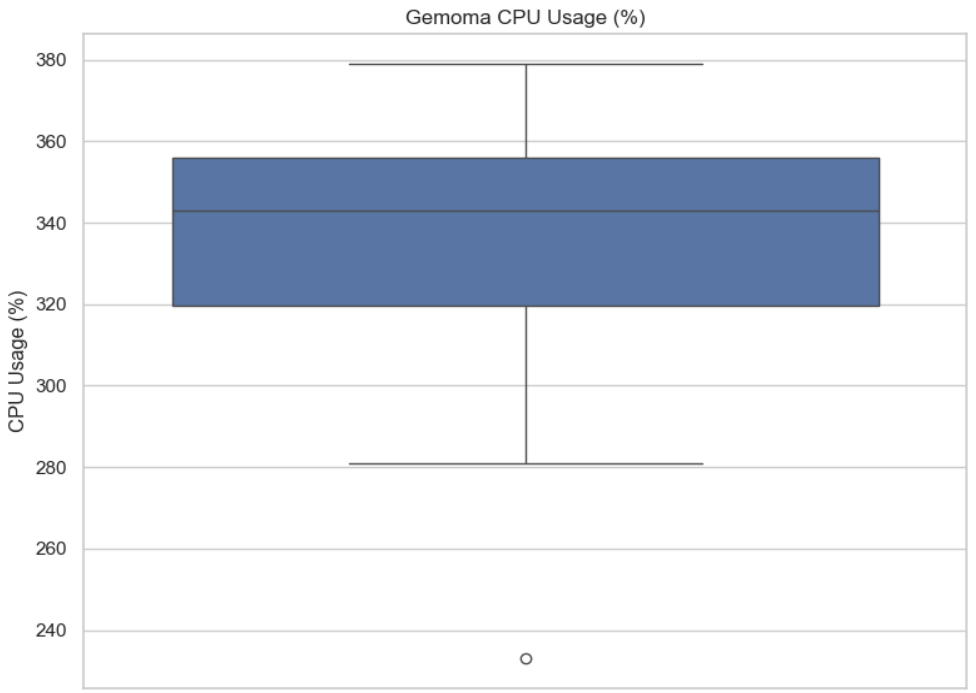
Prodigal	
Avg. Runtime (sec)	36.32
Avg. Memory Usage (kB)	**47,254**
Avg. CPU Usage (%)	99

GeMoMa	
Avg. Runtime (sec)	32.62
Avg. Memory Usage (kB)	**2,760,796**
Avg. CPU Usage (%)	335.03

Georgia Tech

Prodigal vs GeMoMa: Runtime, Memory, CPU Usage





Prodigal	
Avg. Runtime (sec)	36.32
Avg. Memory Usage (kB)	47254
Avg. CPU Usage (%)	**99**

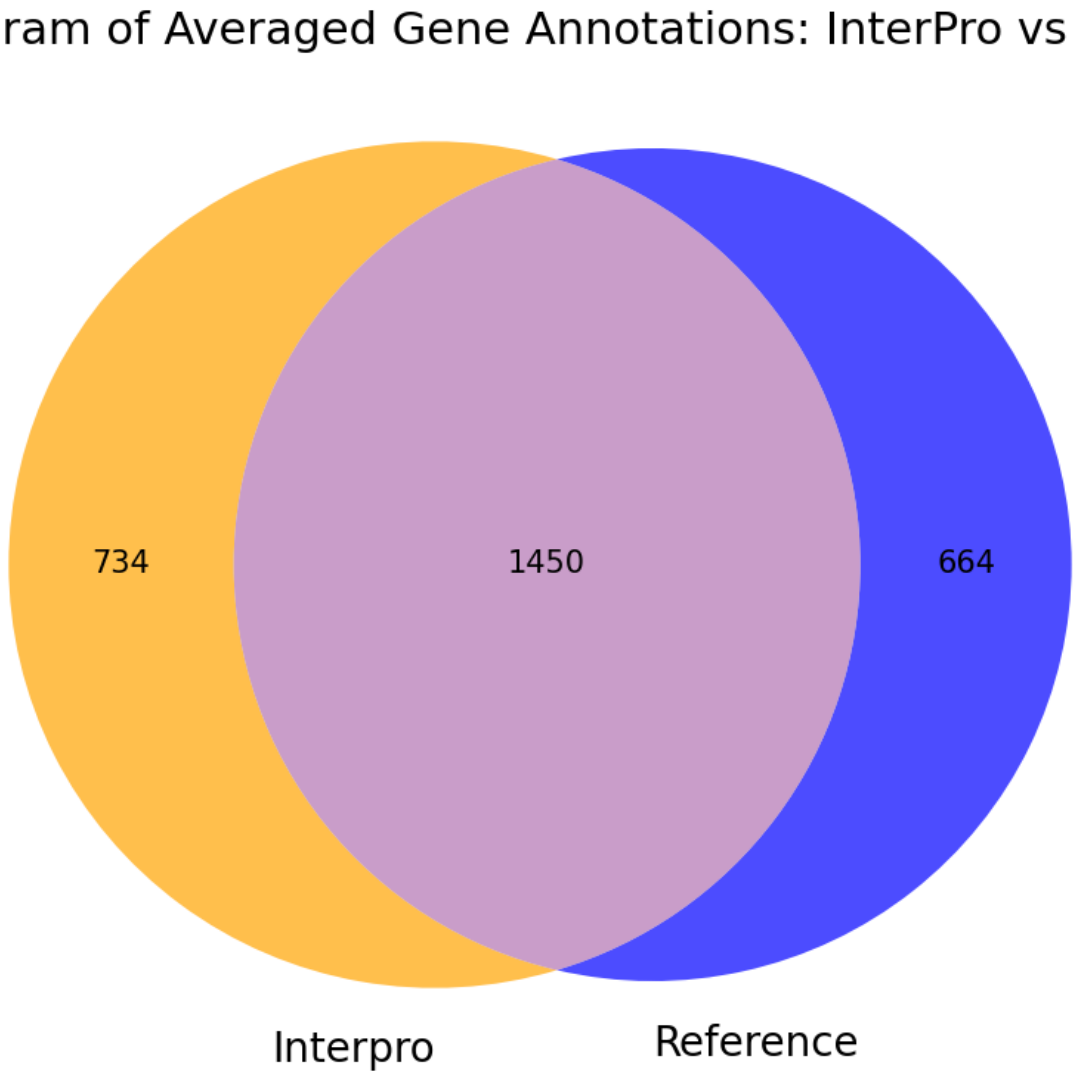
GeMoMa	
Avg. Runtime (sec)	32.62
Avg. Memory Usage (kB)	2760796.24
Avg. CPU Usage (%)	**335.03**

Georgia Tech

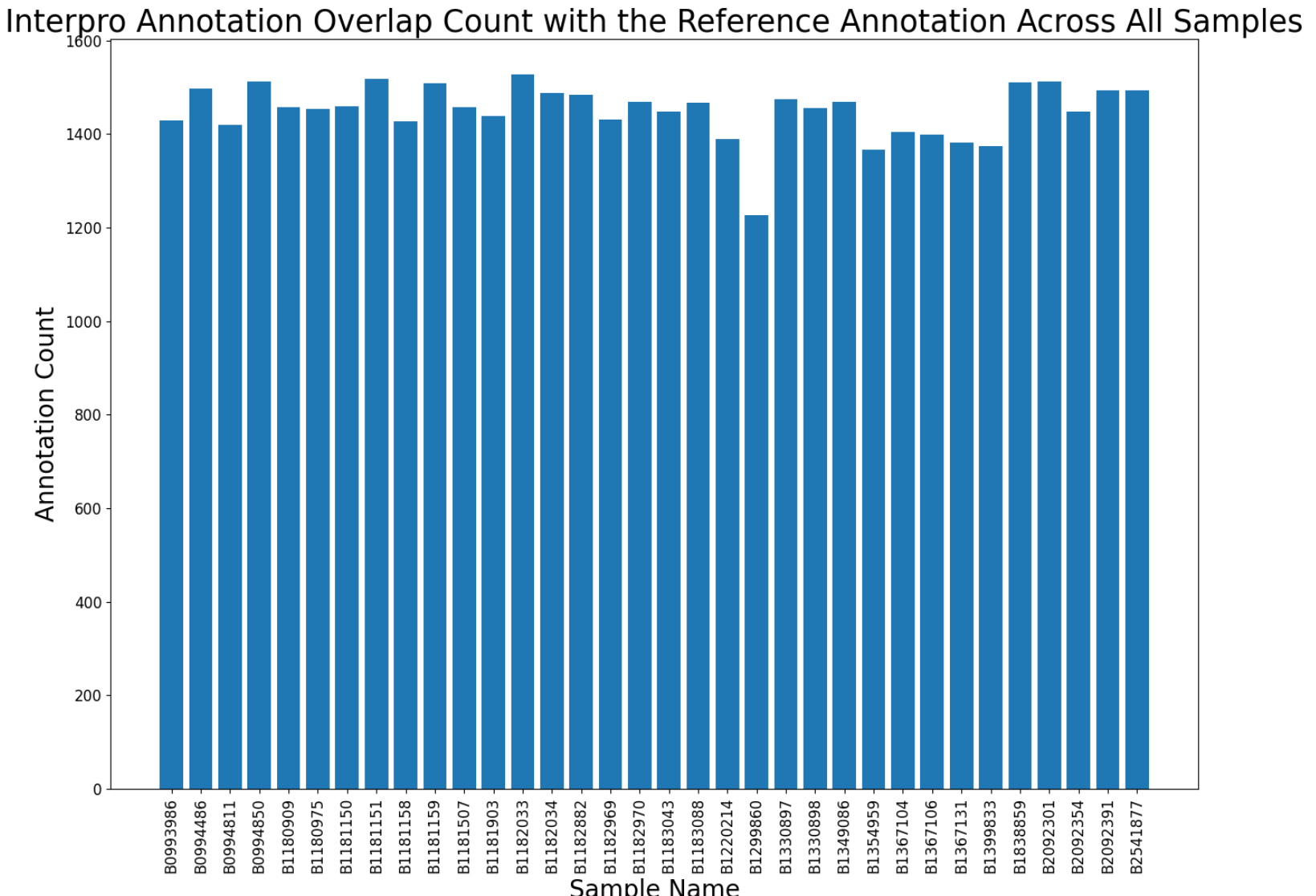
Gene Annotation

Georgia Tech

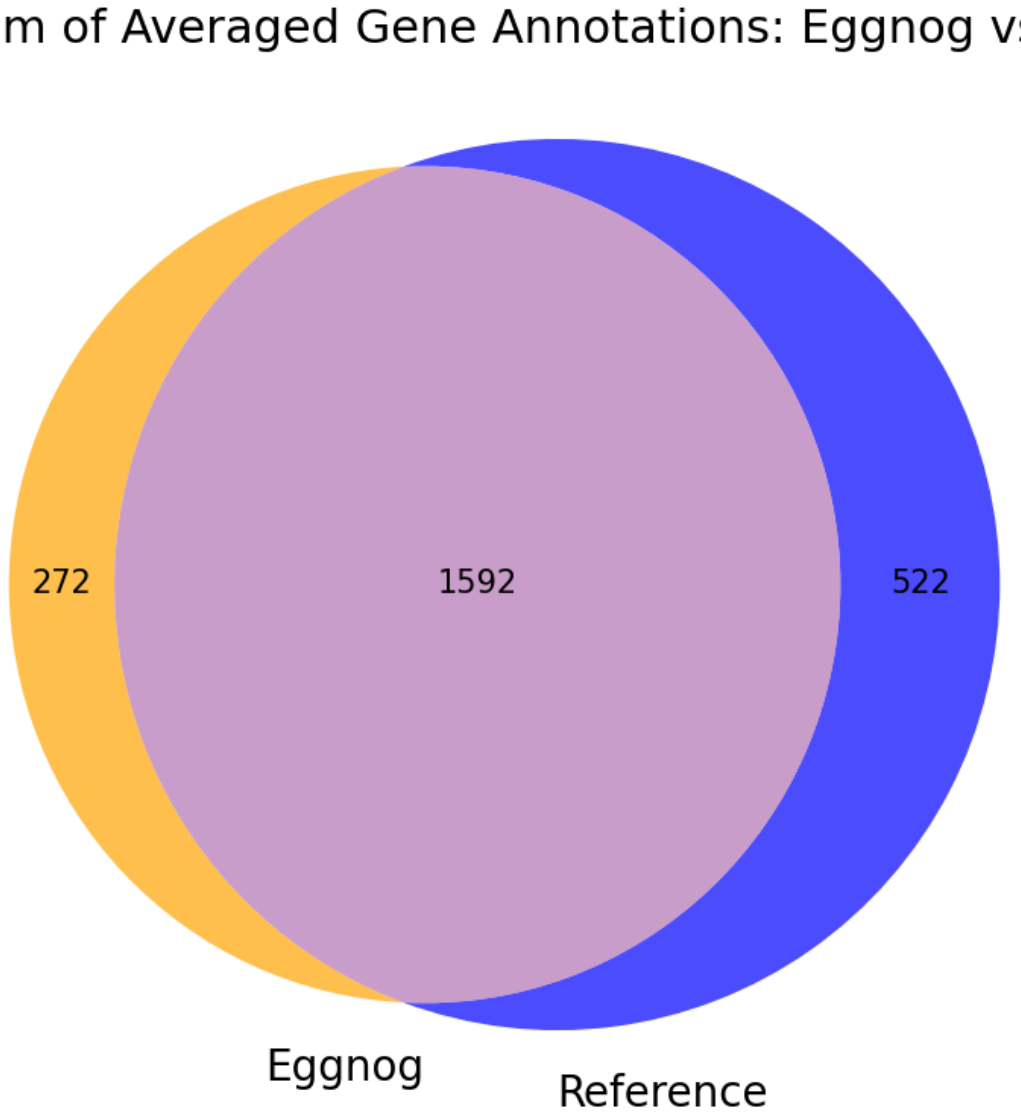
Interpro vs Reference: Averaged Annotation Overlap



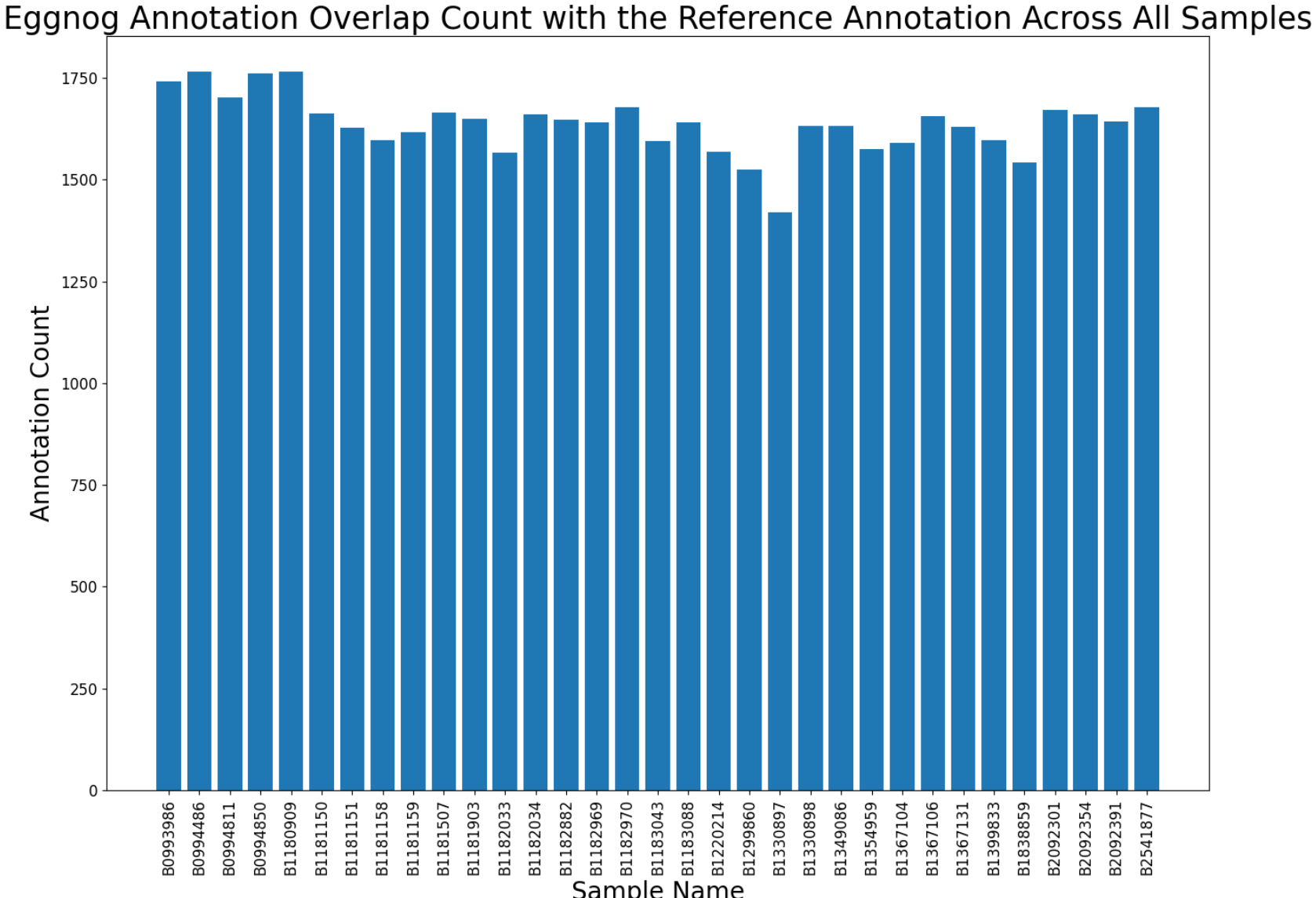
Interpro vs Reference: Annotation Overlap By Sample



Eggnog vs Reference: Averaged Annotation Overlap

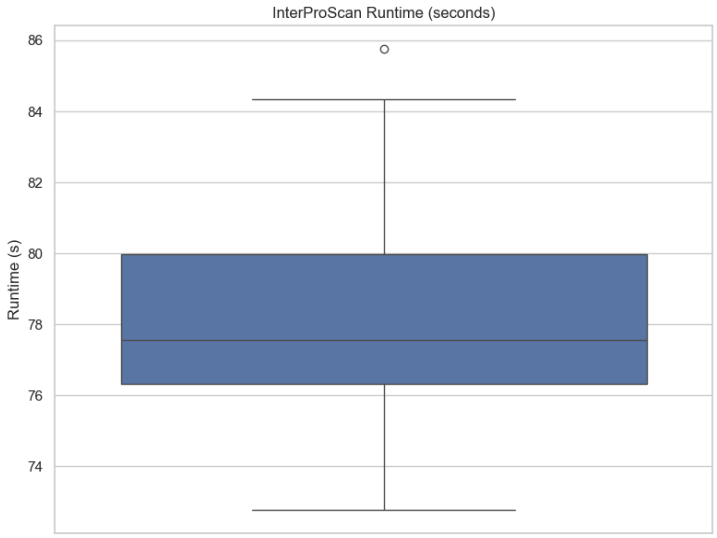


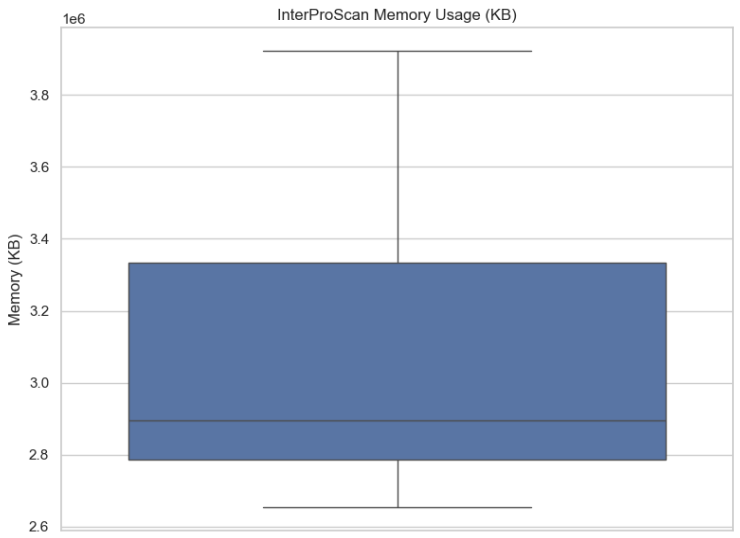
Eggnog vs Reference: Annotation Overlap By Sample

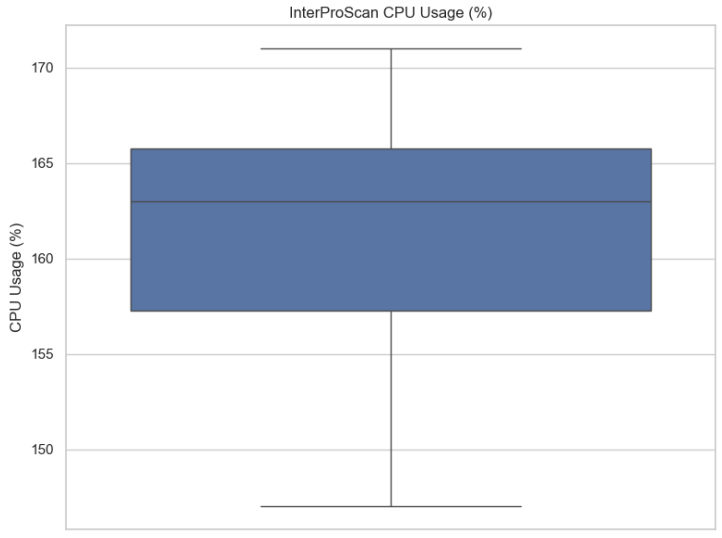


Georgia Tech

Interpro: Runtime, Memory, CPU Usage



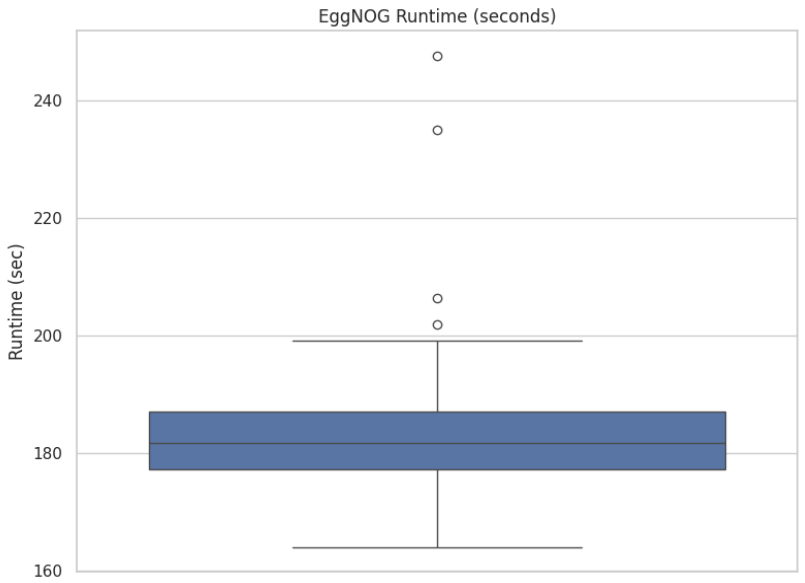


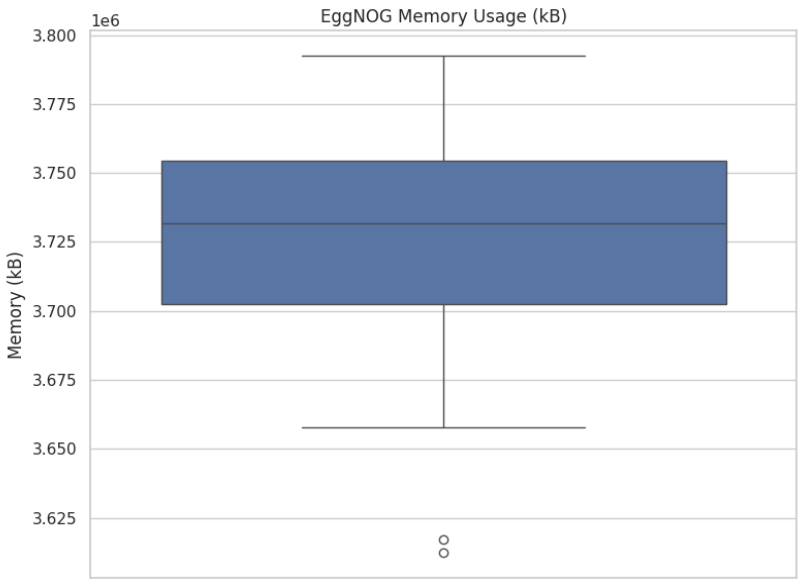


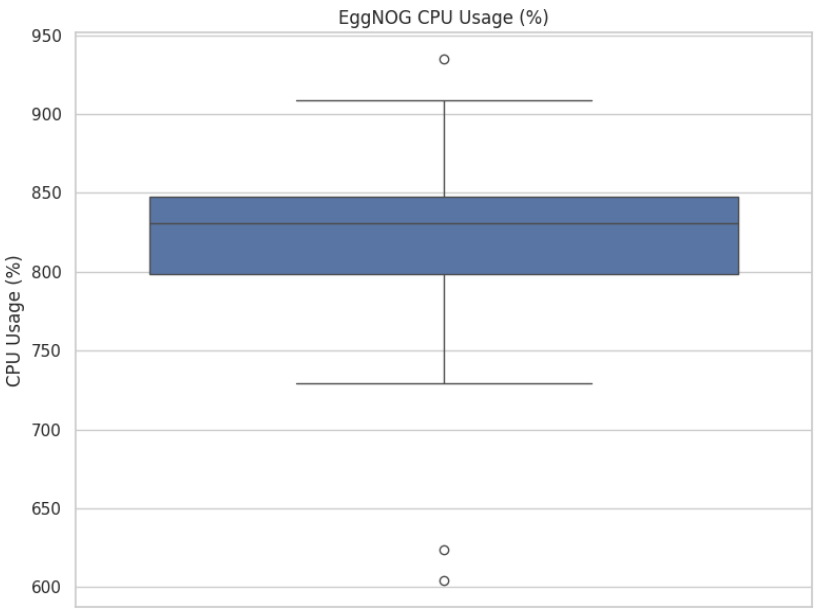
Avg. Runtime (sec)	**78.00**
Avg. Memory Usage (kB)	**3066533.06**
Avg. CPU Usage (%)	**161.44**

Georgia Tech

EggNog: Runtime, Memory, CPU Usage







Avg. Runtime (sec)	**184.99**
Avg. Memory Usage (kB)	**3725382.82**
Avg. CPU Usage (%)	**818.21**

Georgia Tech

Limitations

- Running on different architectures
- Collaborative challenges (splitting workflows)
- Interpro and EggNog have very large databases and the download took a lot of time for both (also storage)
- Using EggNog on command line was more difficult than the webserver version
- Generalization for main pipeline was difficult (inputs, outputs, accommodating many user input parameters)
- GeMoMa-InterPro annotation results were inflated
  - Our final results only include abinitio prediction results

Reflections & Conclusion

- For more robust comparison, we could have compared the gene prediction and annotation results to the reference
  - Ex. Using gff files to extract sequences from assemblies and comparing it to reference genome sequences
  - Ex. Comparing annotations to the reference annotation
- Use workflow languages like nextflow to be able to run multiple annotation/prediction methods in parallel for comparison
- Best gene prediction tool: Prodigal
- Best annotation tool: EggNog

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

Georgia Tech

Appendix

Running GeMoMa

```
# Reference files
GENOME="$REF_DIR/GCF_000006845.1_ASM684v1_genomic.fna"
GFF="$REF_DIR/GCF_000006845.1_ASM684v1_genomic.gff"
```

```
# ------------------------------------------------
# Run GeMoMa
# ------------------------------------------------
gemoma_time_log="$LOGS_DIR/${sample_name}_gemoma_time.log"
/usr/bin/time -v GeMoMa GeMoMaPipeline \
    threads=4 \
    outdir="$sample_gemoma_out" \
    GeMoMa.Score=ReAlign \
    AnnotationFinalizer.r=NO \
    o=true \
    t="$temp_contig_file" \
    i=1 \
    a="$GFF" \
    g="$GENOME" 2> "$gemoma_time_log"
```

Georgia Tech

Running Barrnap

```
# ---------- Global Configuration ----------
CONTIG_DIR="$HOME/xyuan99/biol7210/github/B2/final_contigs"
WORK_DIR="$HOME/xyuan99/biol7210/github/B2/Final_results/
Prodigal_InterPro_Barrnap_v2"
ENV_NAME="Prodigal_InterPro_Barrnap_v2"

# ---------- rRNA Extraction Module (Barrnap) ----------
run_barrnap() {
    local contig=$1
    local base=$(basename ${contig} .fa.gz)
    local temp_file="${WORK_DIR}/TempFiles/${base}.fa"
    local metrics_file="${WORK_DIR}/Metrics/barrnap_${base}.metrics"

    echo "[$(date +'%F %T')] Processing ${base} with Barrnap..."

    # Ensure temp file exists
    if [ ! -f ${temp_file} ]; then
        gzip -dc ${contig} > ${temp_file}
    fi

    # Run Barrnap for rRNA extraction
    /usr/bin/time -f "Time=%e\nCPU=%P\nMem=%M" -o ${metrics_file} \
    barrnap --kingdom bac \
        --threads ${THREADS} \
        --outseq ${WORK_DIR}/Barrnap/${base}_rRNA.fasta \
        --quiet ${temp_file} > ${WORK_DIR}/Barrnap/${base}_rRNA.gff 2> $
        {WORK_DIR}/Logs/barrnap_${base}.log
```

Georgia Tech

Running Prodigal

```
# ---------- Global Configuration ----------
CONTIG_DIR="$HOME/xyuan99/biol7210/github/B2/final_contigs"
WORK_DIR="$HOME/xyuan99/biol7210/github/B2/Final_results/
Prodigal_InterPro_Barrnap_v2"
ENV_NAME="Prodigal_InterPro_Barrnap_v2"

# ---------- Gene Prediction Module (Prodigal) ----------
run_prodigal() {
    local contig=$1
    local base=$(basename ${contig} .fa.gz)
    local temp_file="${WORK_DIR}/TempFiles/${base}.fa"
    local metrics_file="${WORK_DIR}/Metrics/prodigal_${base}.metrics"

    echo "[$(date +'%F %T')] Processing ${base} with Prodigal..."

    # Decompress contig file
    gzip -dc ${contig} > ${temp_file}

    # Run Prodigal for gene prediction
    /usr/bin/time -f "Time=%e\nCPU=%P\nMem=%M" -o ${metrics_file} \
    prodigal -i ${temp_file} \
        -o ${WORK_DIR}/Prodigal/${base}.gff \
        -a ${WORK_DIR}/Prodigal/${base}.faa \
        -d ${WORK_DIR}/Prodigal/${base}.fna \
        -p meta \
        -f gff > ${WORK_DIR}/Logs/prodigal_${base}.log 2>&1
```

Georgia Tech

Running InterProScan

```
# ---------- Global Configuration ----------
CONTIG_DIR="$HOME/xyuan99/biol7210/github/B2/final_contigs"
WORK_DIR="$HOME/xyuan99/biol7210/github/B2/Final_results/
Prodigal_InterPro_Barrnap_v2"
ENV_NAME="Prodigal_InterPro_Barrnap_v2"

# ---------- Gene Annotation Module (InterPro) ----------
run_interpro() {
    local faa=$1
    local base=$(basename ${faa} _clean.faa)
    local metrics_file="${WORK_DIR}/Metrics/interpro_${base}.metrics"

    echo "[$(date +'%F %T')] Processing ${base} with InterProScan..."

    # Run InterProScan for functional annotation
    /usr/bin/time -f "Time=%e\nCPU=%P\nMem=%M" -o ${metrics_file} \
    interproscan.sh -i ${faa} \
        -d ${WORK_DIR}/InterPro \
        -f TSV,GFF3 \
        -appl Pfam,SMART,TIGRFAM,CDD,PRINTS,SUPERFAMILY \
        -goterms \
        -pa > ${WORK_DIR}/Logs/interpro_${base}.log 2>&1

```

Georgia Tech

Running EggNOG

```
# Run EggNOG-mapper with the --decorate_gff parameter to get GFF output
/usr/bin/time -v emapper.py \
    -i "$clean_proteins" \
    --output "$sample_name" \
    --output_dir "$sample_eggnog_out" \
    --data_dir "$EGGNOG_DB_DIR" \
    --cpu 16 \
    --tax_scope bacteria \
    --go_evidence all \
    --decorate_gff "$gemoma_gff_output" \
    --override 2> "$eggnog_time_log"
```

Thank you!

Georgia Tech