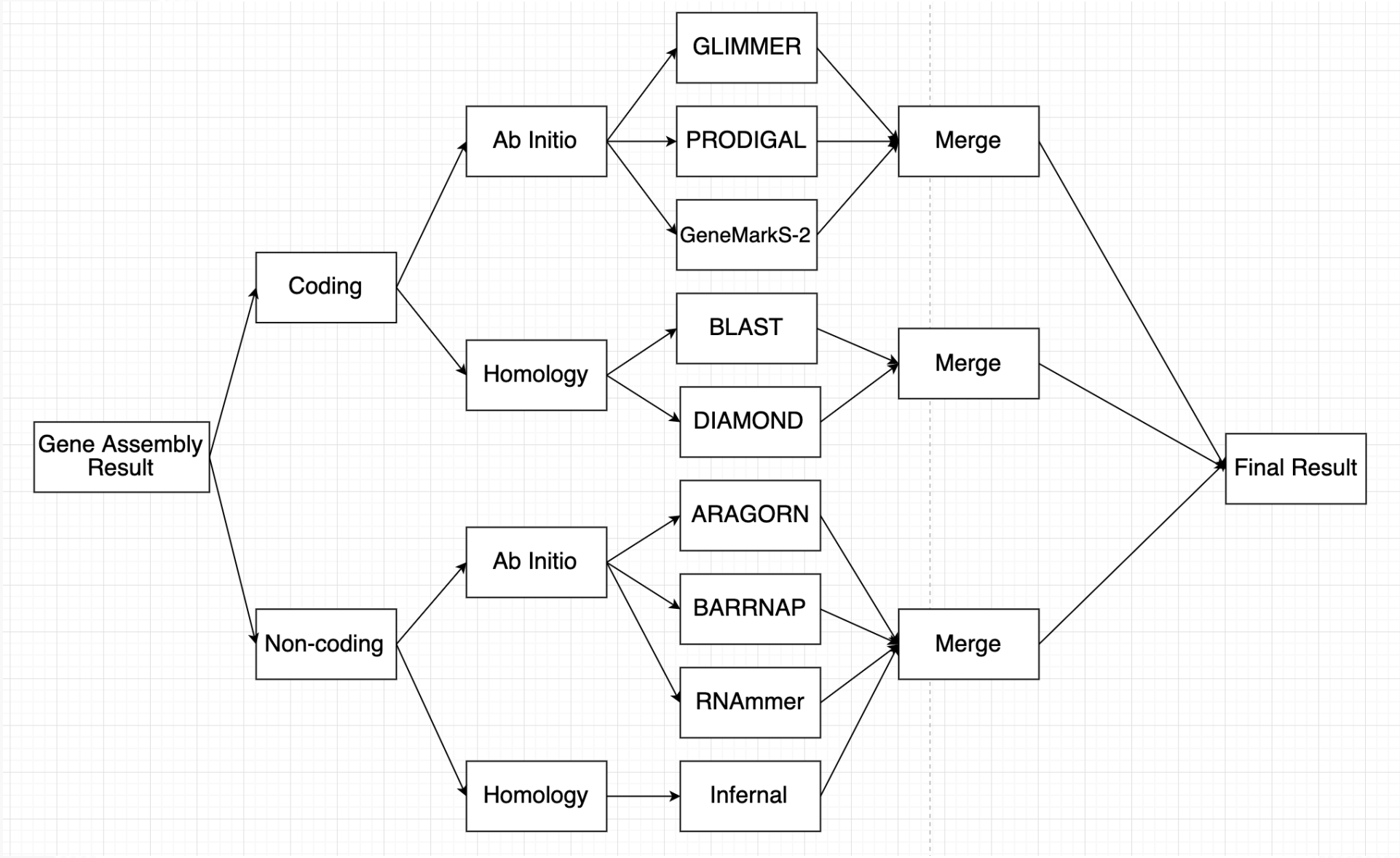
Gene Prediction Team 3: Final Result

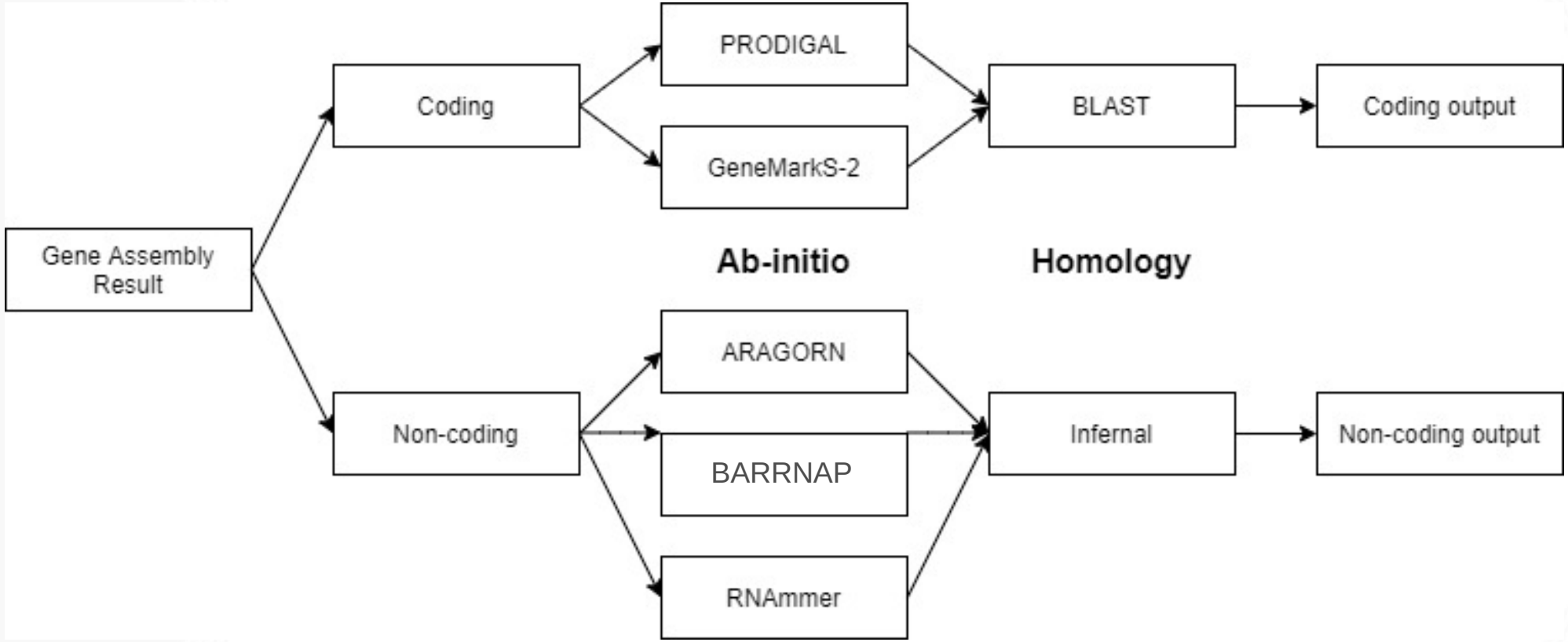
Pallavi Misra
Sonali Gupta
Ahish Melkote Sujay
Shen-Yi Cheng
Jie Zhou

March 5, 2020

Previous workflow



Final workflow



Commands Used

- PRODIGAL

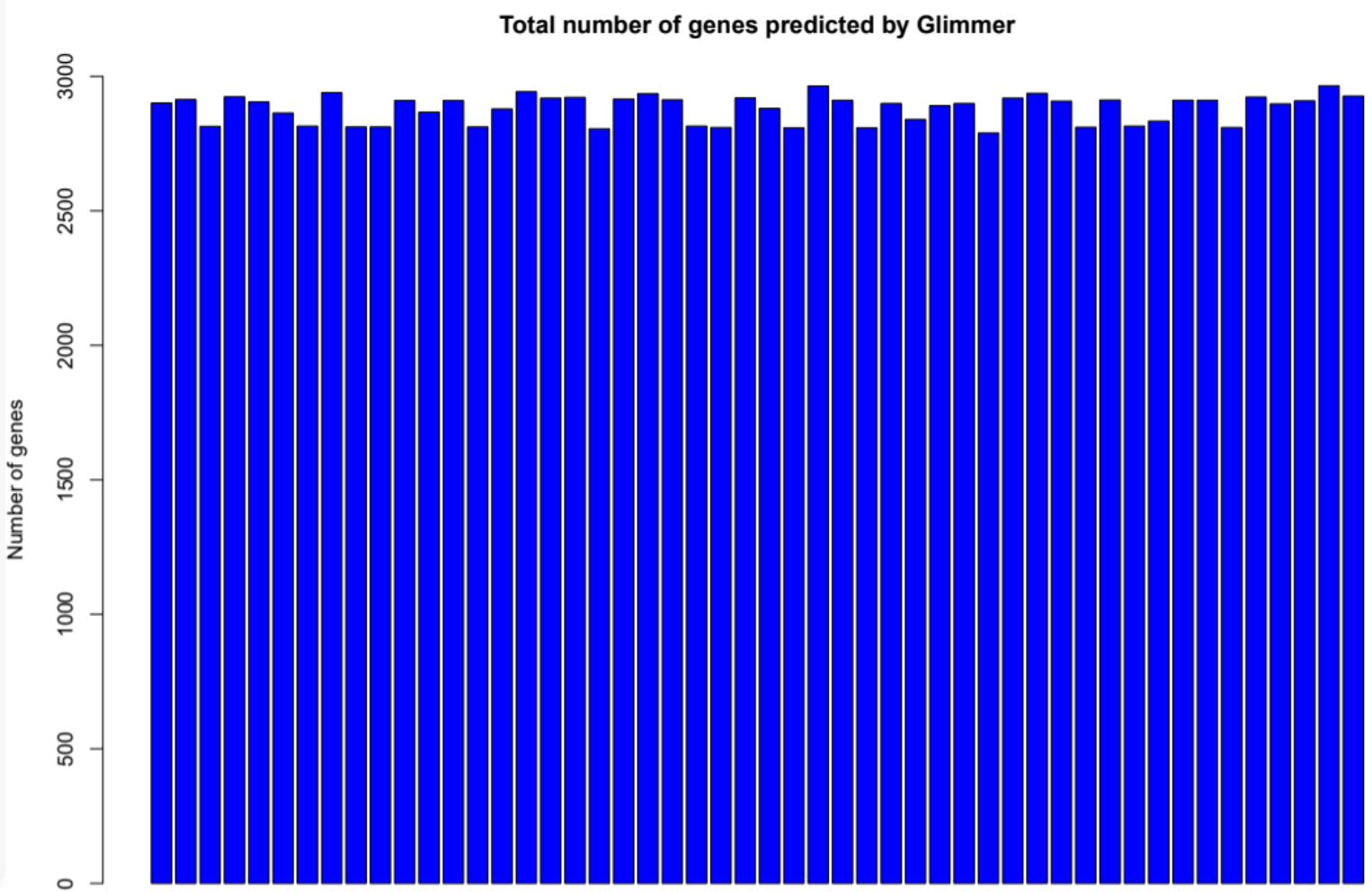
```
prodigal -i contig_file -d fasta_file -f gff -o gff_file
```

- GeneMarkS-2

```
perl gms2.pl --seq contig_file --genome-type bacteria -fnn nucl_file --output lst_file
```

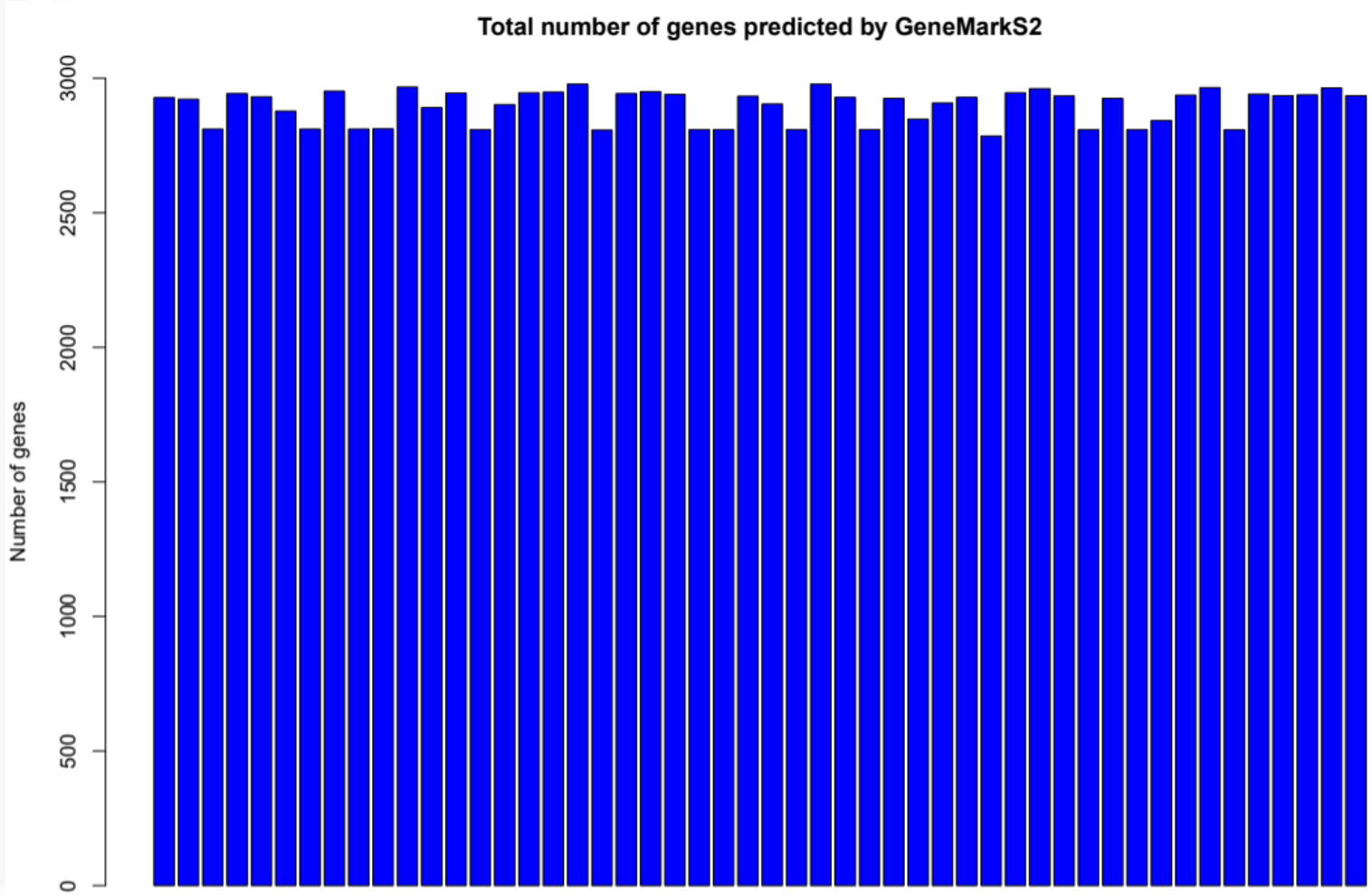
```
perl gms2.pl --seq contig_file --genome-type bacteria --output gff_file --format gff
```

Total Predictions - Glimmer



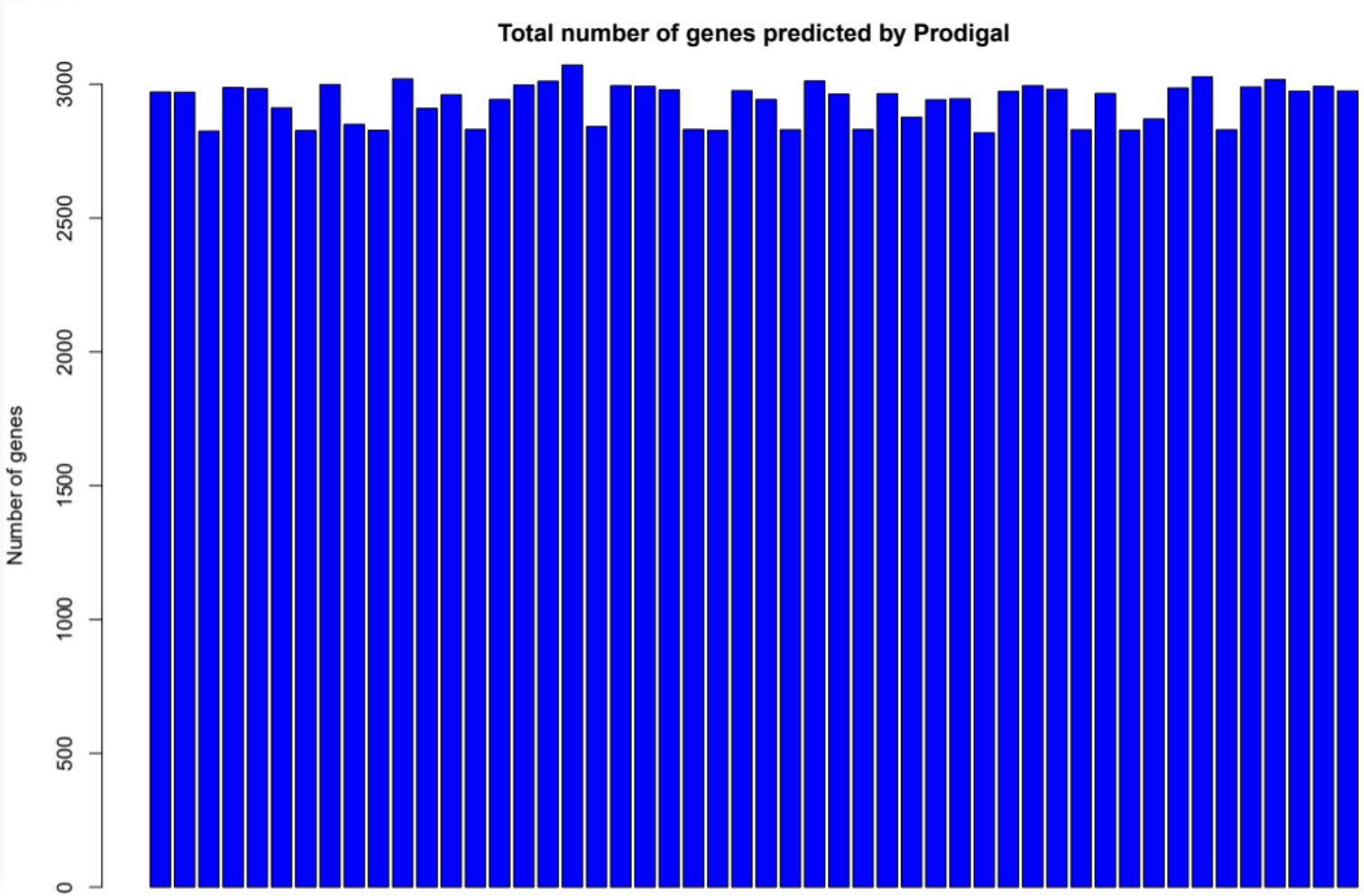
Average = 2934

Total Predictions – GeneMarkS2



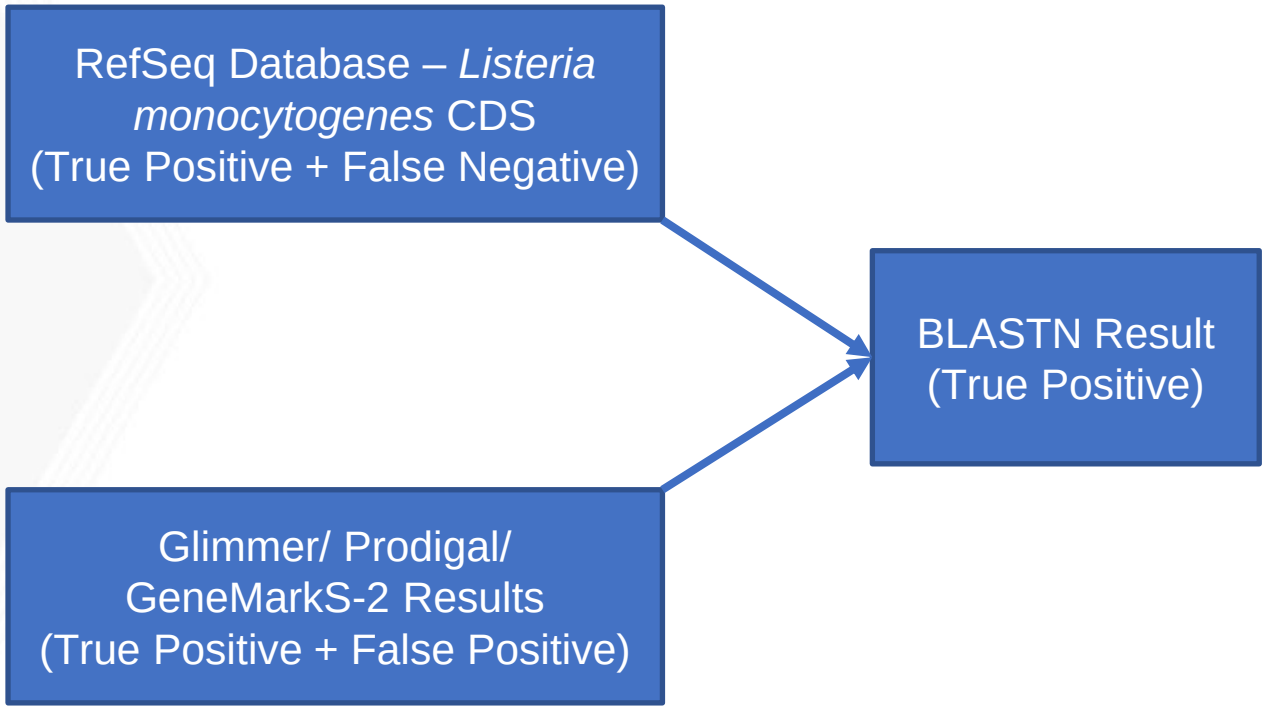
Average = 2897

Total Predictions – Prodigal

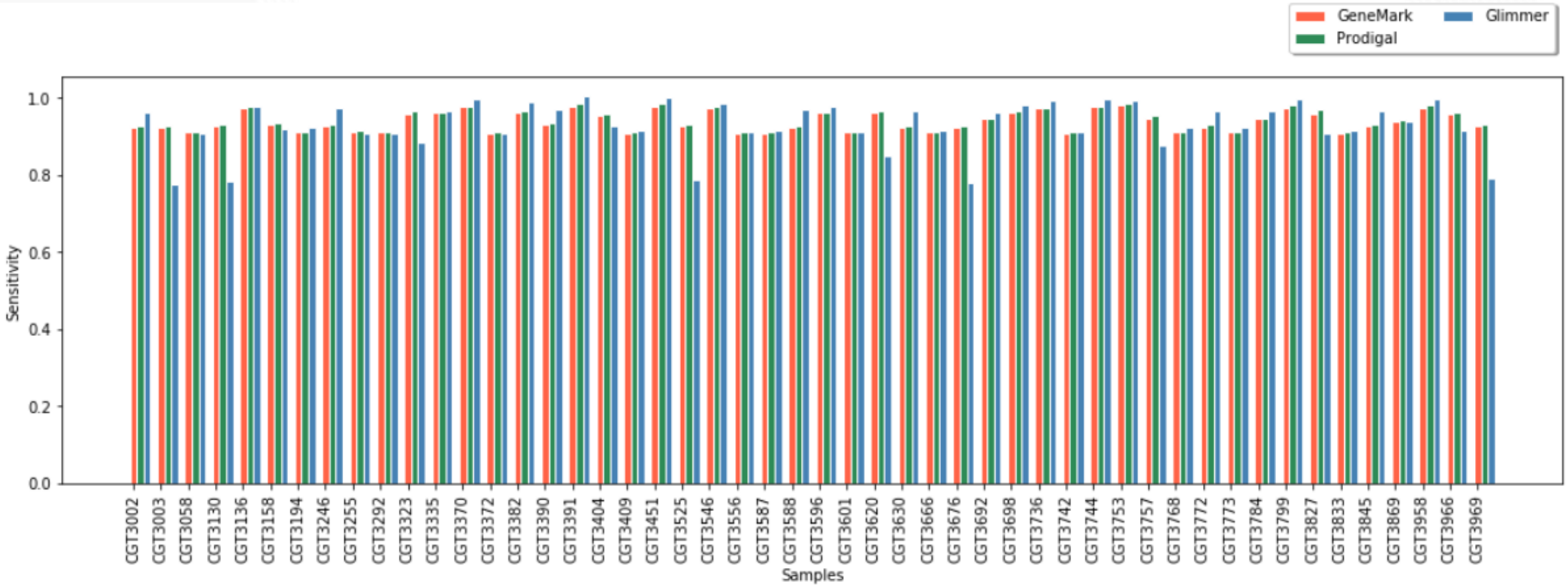


Average = 2881

Workflow – Sensitivity, False Discovery Rate, Precision

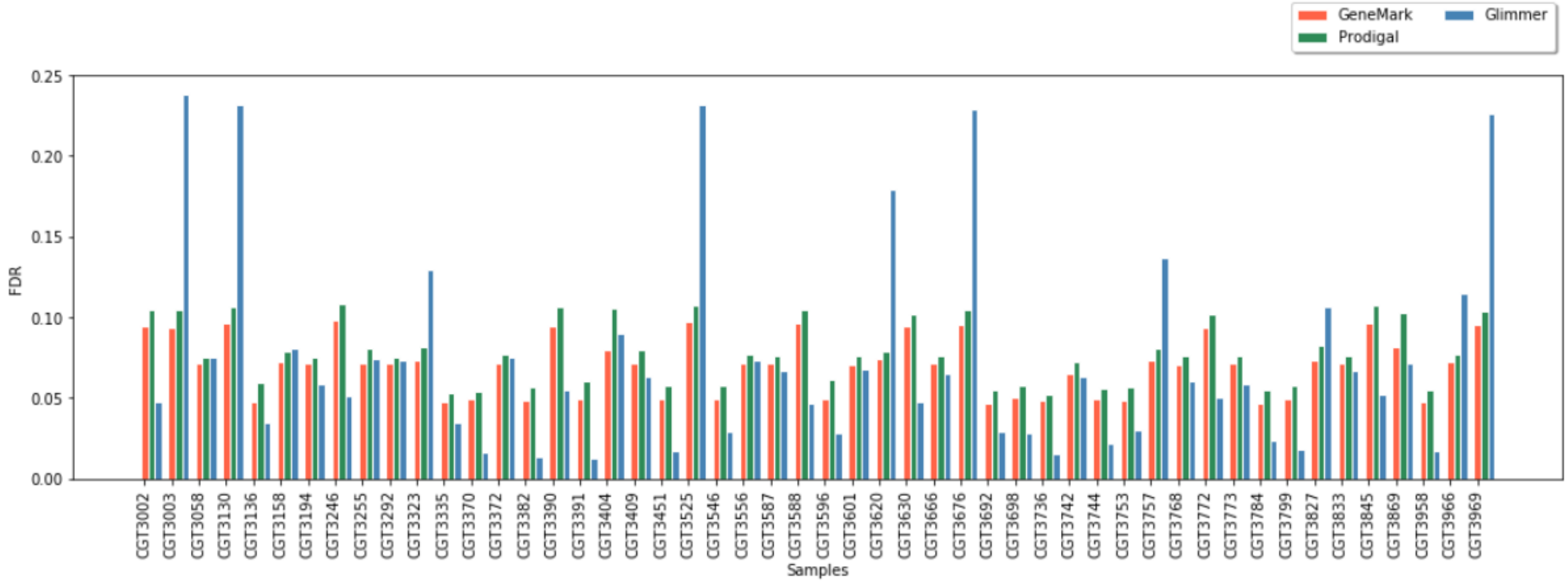


Sensitivity



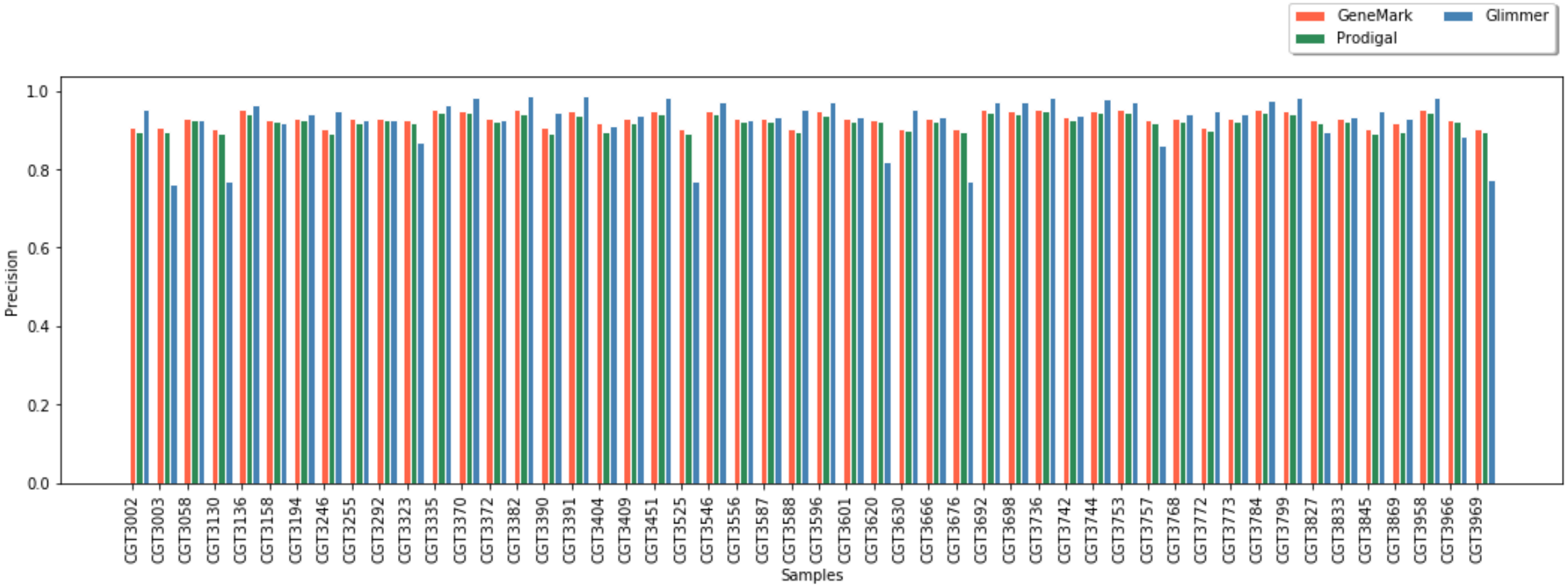
Sensitivity: $\frac{True\ Positive}{True\ Positive + False\ Negative}$

False Discovery Rate (FDR)



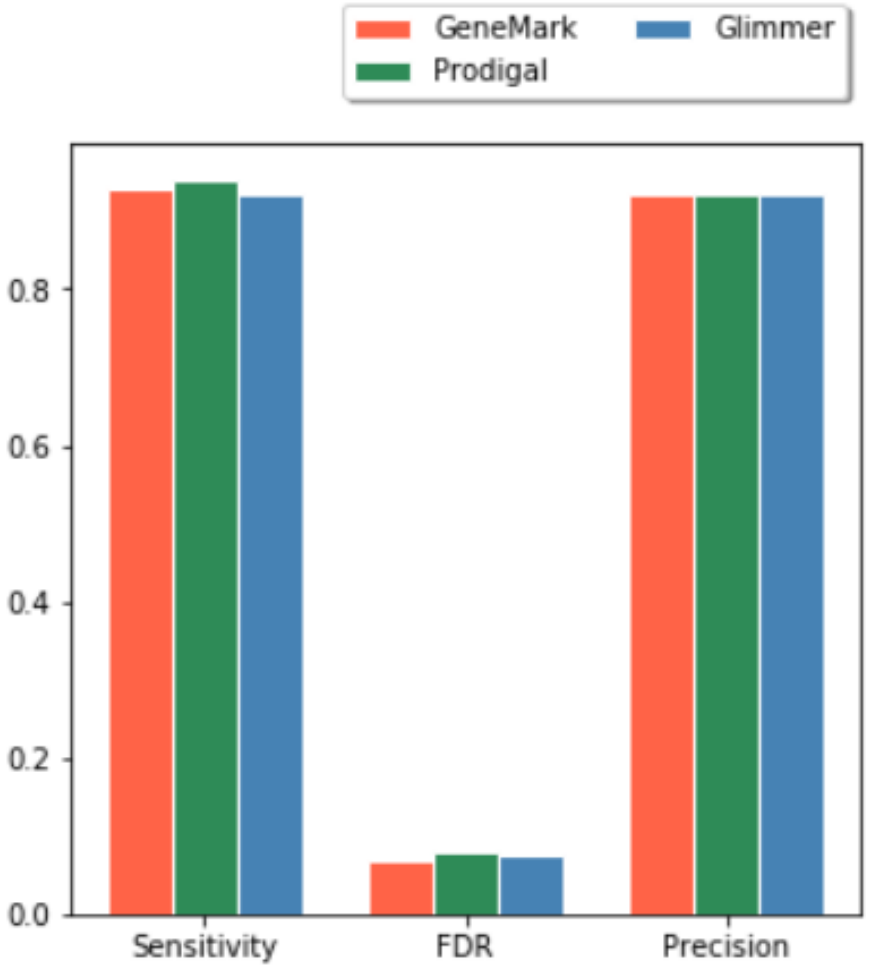
False Discovery Rate: $\frac{False\ Positive}{True\ Positive + False\ Positive}$

Precision

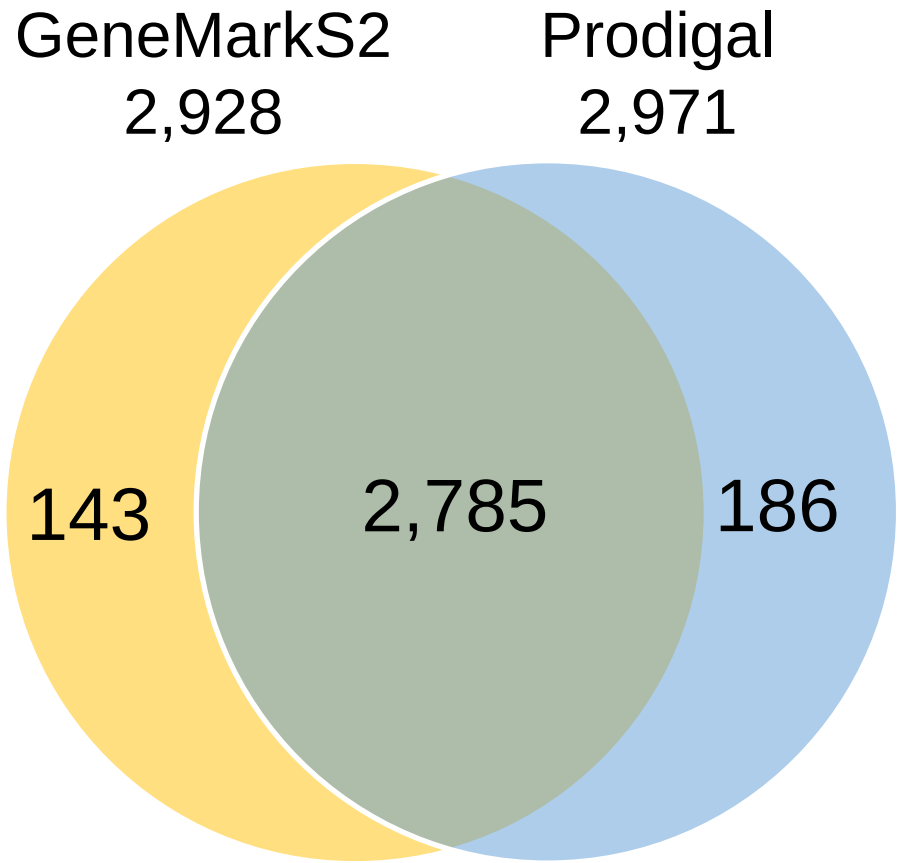


Precision: $\frac{True\ Positive}{True\ Positive + False\ Positive}$

Average Measures



Merging ab initio results



Representative of Contig File 1

Commands for merging

- Reciprocal Overlap

```
bedtools intersect -f 1.0 -r -a gms2.out -b prodigal.out > Intersection
```

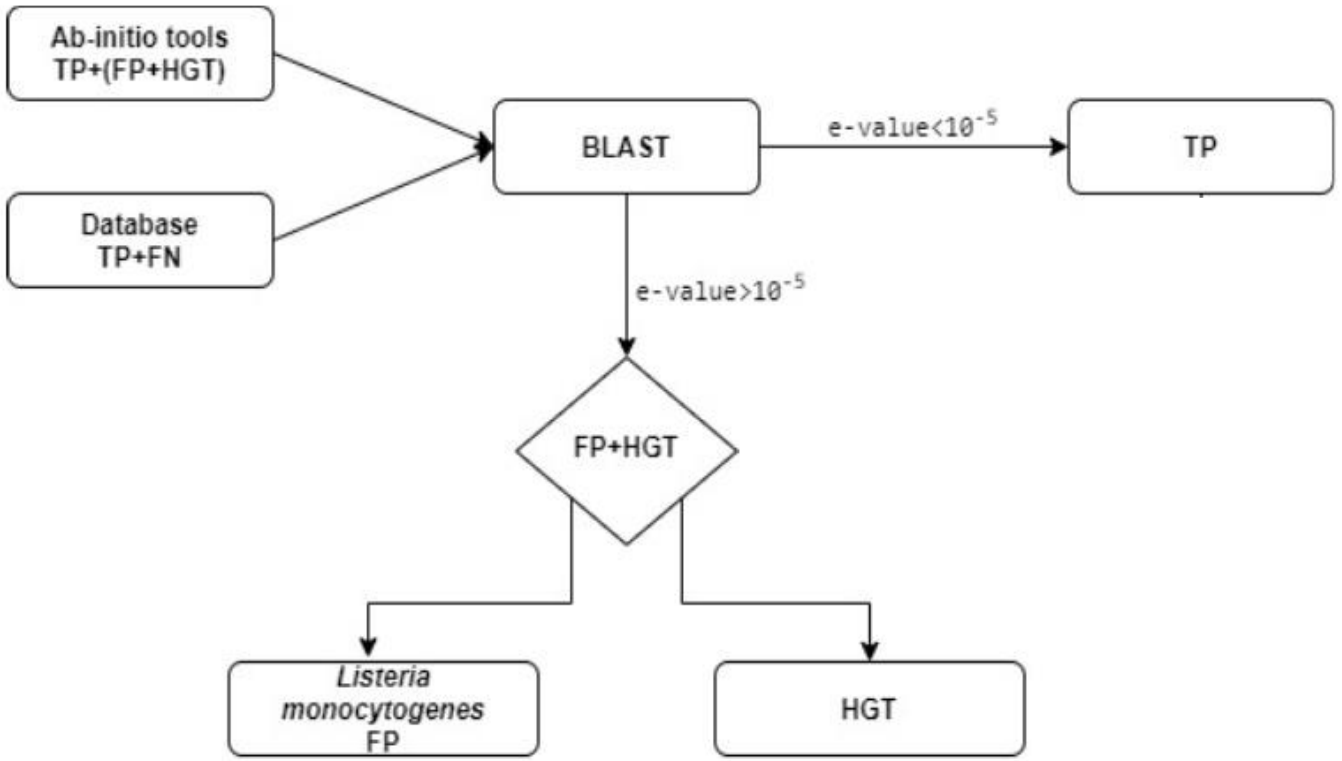
- Only in GeneMarkS2

```
bedtools intersect -f 1.0 -v -r -a gms2.out -b prodigal.out > GeneMark_only
```

- Only in Prodigal

```
bedtools intersect -f 1.0 -v -r -a prodigal.out -b gms2.out > Prodigal_only
```

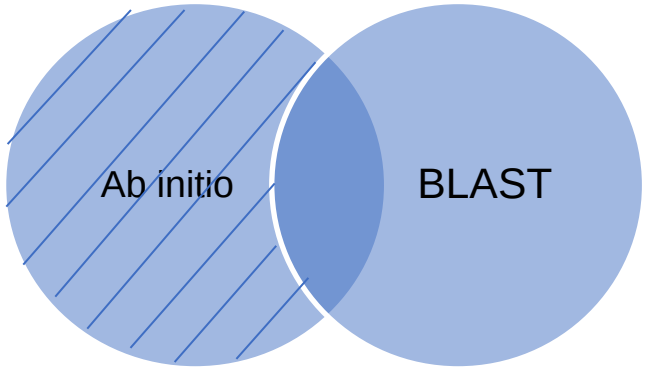
Pipeline for HGT identification



Identifying FPs and HGTs

1. (Ab Initio) – (BLAST)

HGT + FP



2. Assembly contigs BLAST

```
Query= 1 NODE_1_length_103861_cov_22.749292 2 211 + gene_type=native
partial=10

Length=210

                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

NC_003210.1_cds_NP_466232.1_2718  [locus_tag=lmo2710] [db_xref=Ge...  350     2e-97


>NC_003210.1_cds_NP_466232.1_2718 [locus_tag=lmo2710] [db_xref=GeneID:987111] [protein=hypothetical
protein] [protein_id=NP_466232.1] [location=complement(2783466..2784107)]
```

Ab Initio BLAST

```
Query= NODE_1_length_103861_cov_22.749292

Length=103861

                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

NC_003210.1  Listeria monocytogenes EGD-e chromosome, complete ge...  54333   0.0


>NC_003210.1 Listeria monocytogenes EGD-e chromosome, complete genome
Length=2944528
```

Assembly contigs BLAST

```
Query= NODE_147_length_402_cov_34.391931

Length=402

                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

NZ_LT727813.1  Oceanobacillus sojae strain SK-2, whole genome sho...  355     7e-95
```

Assembly contigs BLAST

FP

HGT

Non-Coding Sequence

- Aragorn: Detect tRNA and tmRNA

```
aragorn -l -t -gc1 -w input.fasta -o output.fasta
```

```
aragorn -l -m -gc1 -w input.fasta -o output.fasta
```

- Transfer Aragorn output from .fasta to .gff format

```
./cnv_aragorn2gff.pl -i output.fasta > output.gff
```

- Barrnap: Detect rRNA

```
barrnap --quiet input.fasta > output.gff
```

RNAmmer

- Run RNAmmer

```
rnammer -S bac -m 'tsu,ssu,lsu' -gff $output/$(basename $filename .fasta).gff2 $filename
```

- Transfer it to GFF version3

```
perl convert_RNAmmer_to_gff3.pl --input $filename > ./gff/$(basename $filename .gff2).gff
```

- Count numbers of rRNA

```
awk '{print $3}' $filename | grep $RNAname| wc -l
```

Infernal

- Run Infernal

```
cmscan --cut_ga --rfam --nohmmonly --tblout $output/$(basename $filename .fasta).tblout      \
 --fmt 2 --clanin Rfam.clanin Rfam.cm $filename > $output/$(basename $filename .fasta).cmscan
```

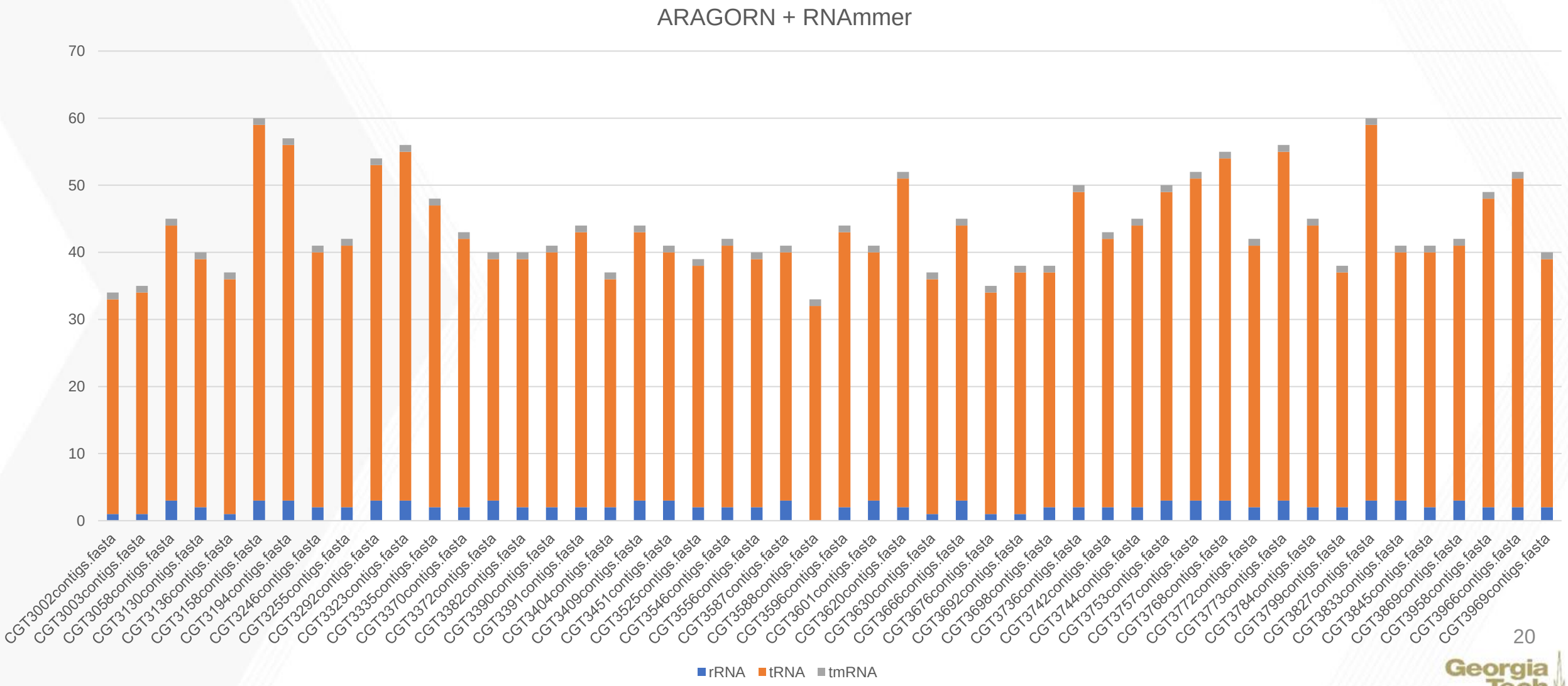
- Transfer results to GFF version3

```
perl infernal-tblout2gff.pl --cmscan --fmt2 $filename > ./gff/$(basename $filename .tblout).gff
```

- Count different kinds of non-coding RNA

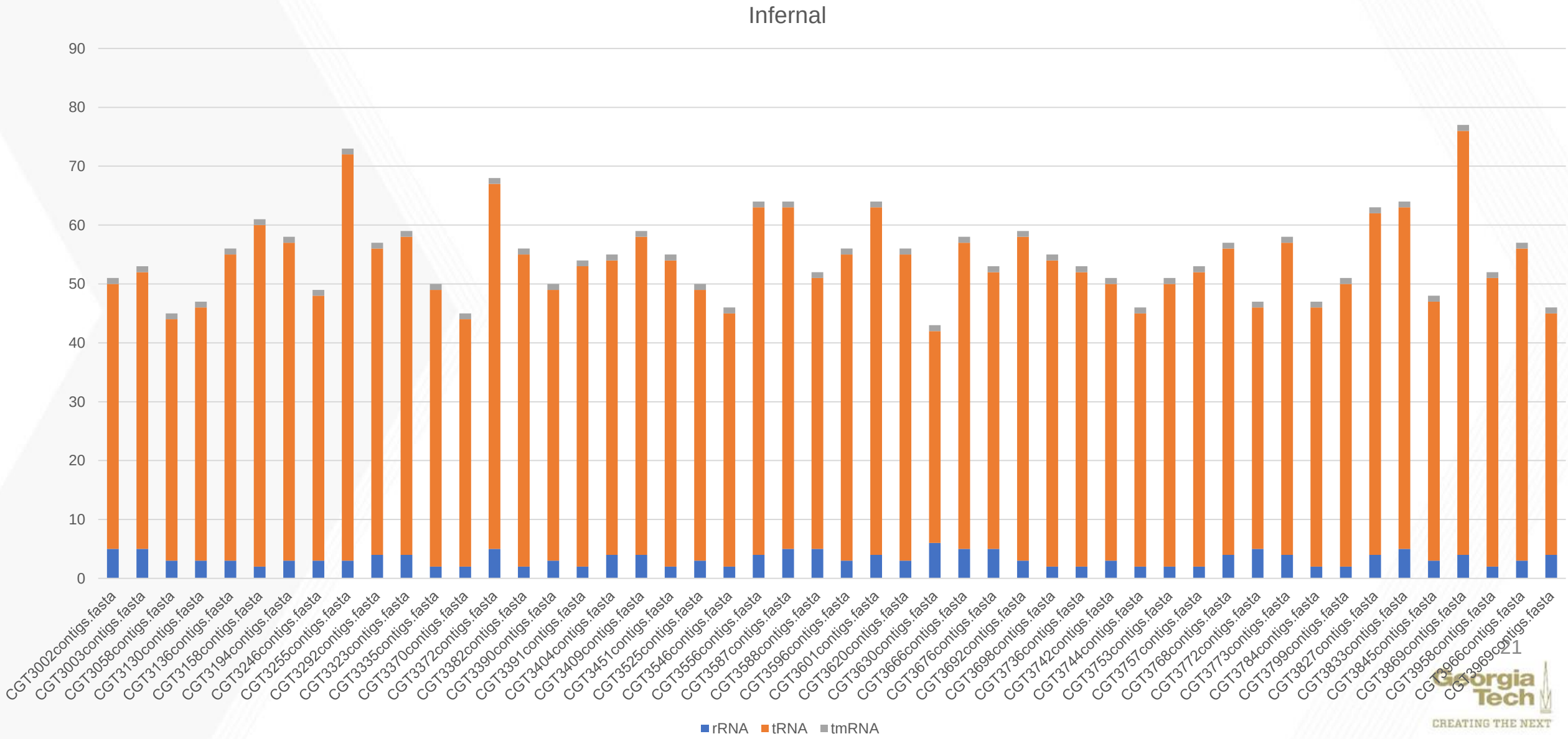
```
awk '{print $3}' $filename | grep $RNAname| wc -l
```

Aragorn + RNAmmer



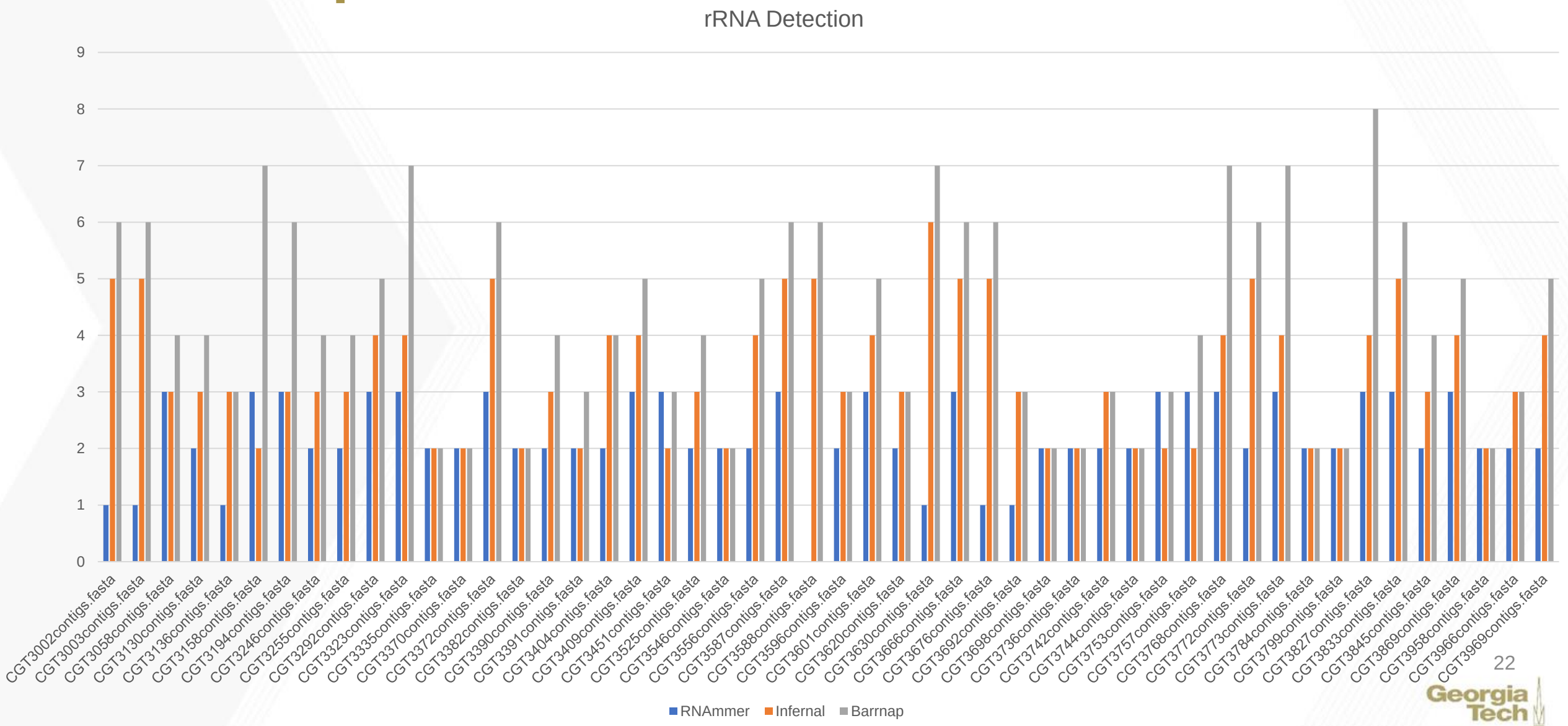
Georgia Tech
CREATING THE NEXT

Infernal



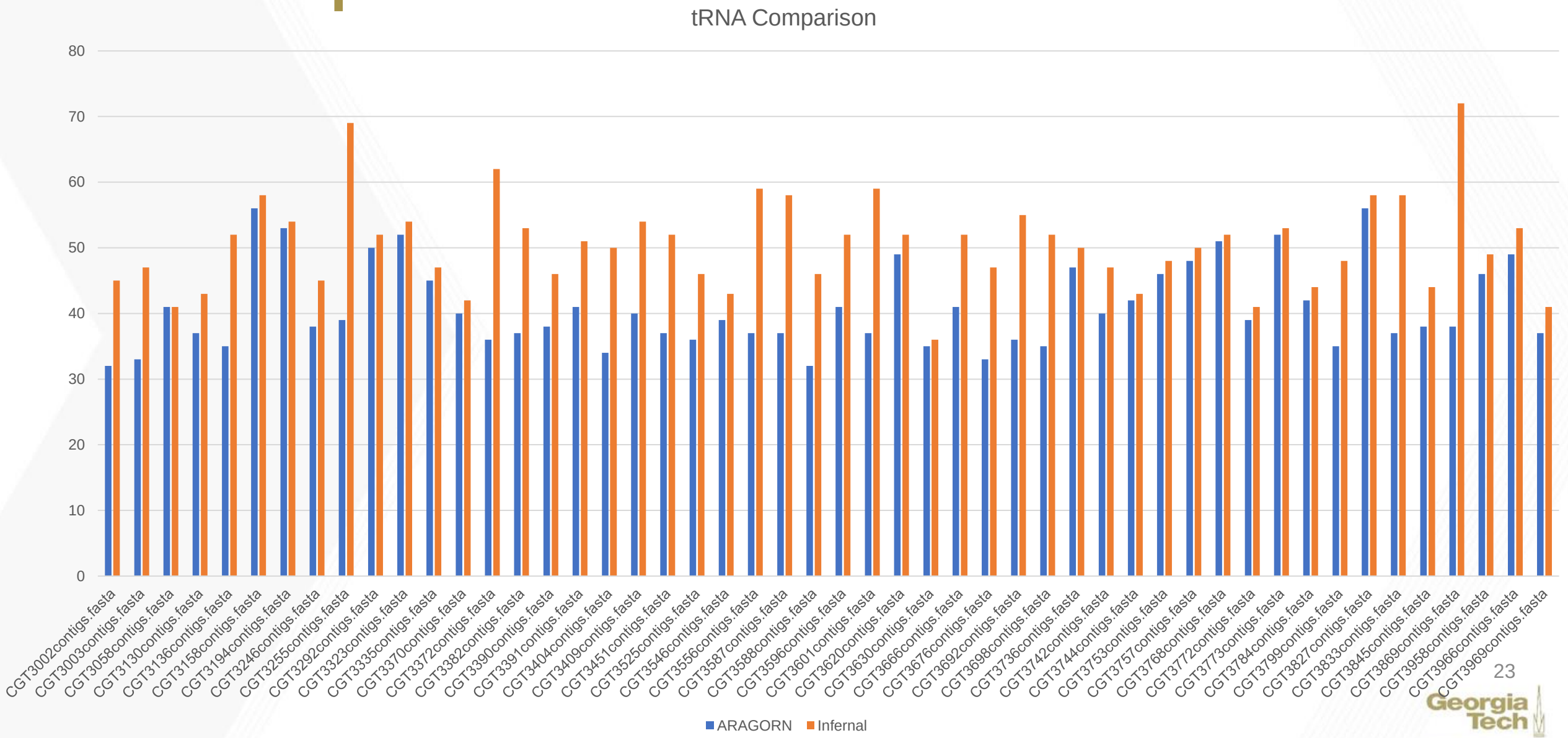
Georgia Tech
CREATING THE NEXT

rRNA Comparison



Georgia Tech
CREATING THE NEXT

tRNA Comparison



Georgia Tech
CREATING THE NEXT

Final workflow