Final Results

Team 1 Gene Prediction

Genevieve Brandt, Victoria Caban, Yuntian He, Junyu Li, Yiqiuyi Liu, Yihao Ou, Wenyi Qiu, Casey Smith, Mohit Thakur, Stephen Wist, Qinyu Yue

Content

Protein Coding Gene Prediction

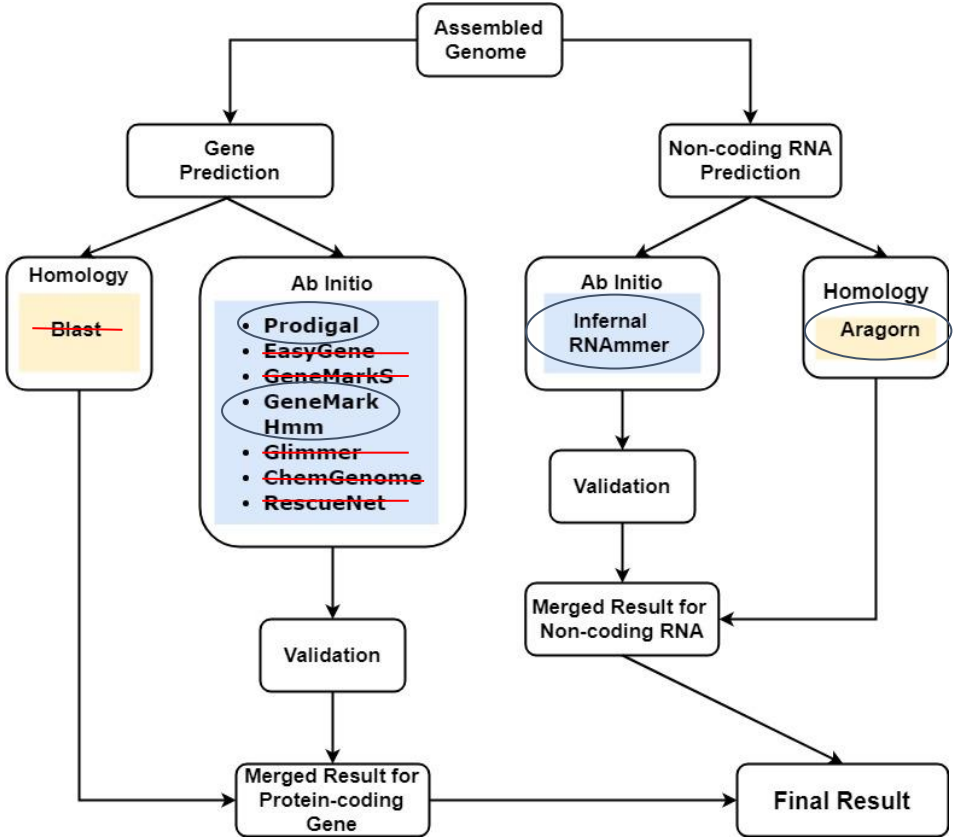
Protein Coding Gene Validation

RNA Region Prediction

Final Pipeline

Content

Overview of tools and pipeline



Genemark HMM - Background

Average runtime: 1.71 seconds/assembly (after training file)
Sensitivity: 93.11%
PPV: 93.10%

- Input: assembly file (Fasta)
- Output: gene coordinates (GFF), nucleotide file (Fasta), protein file (Fasta)

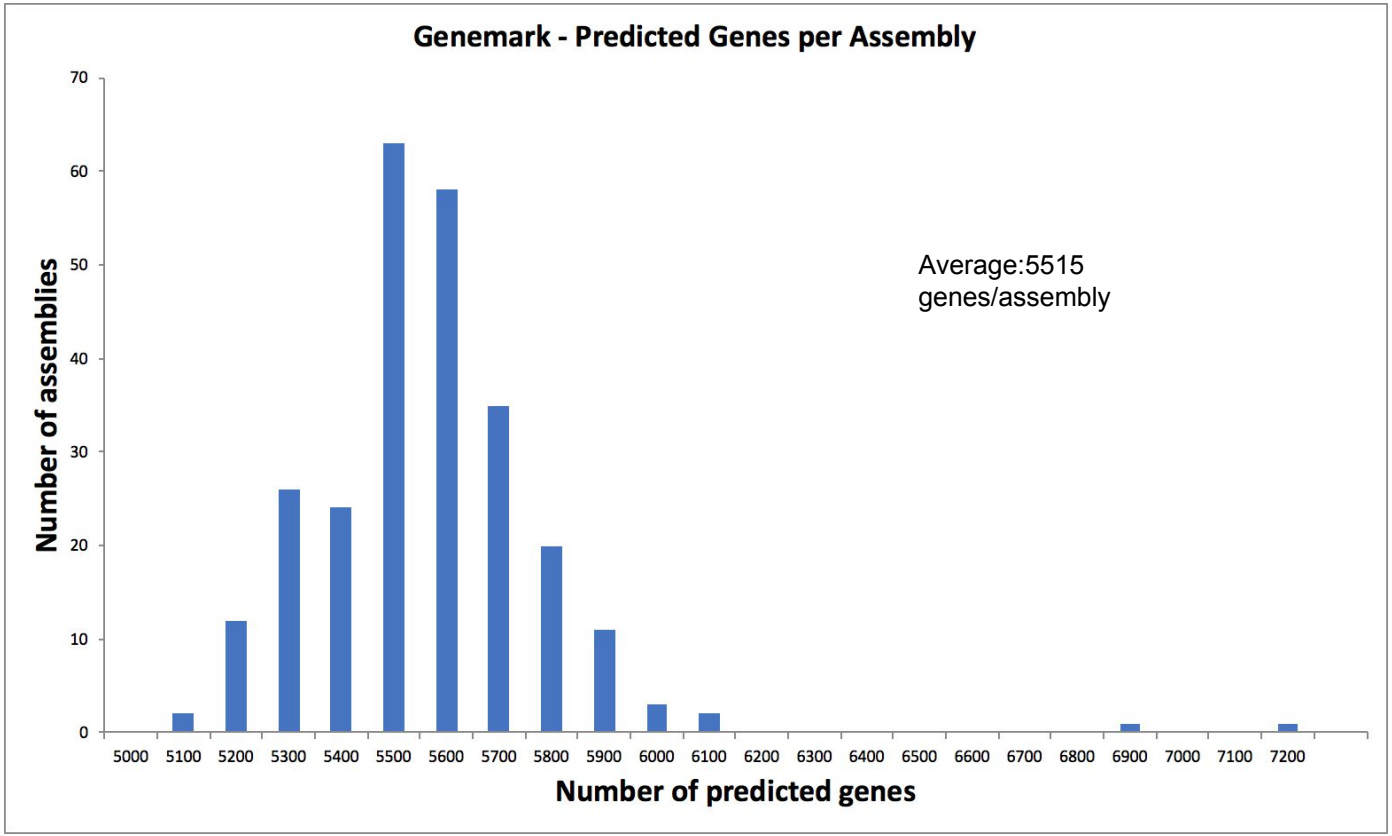
- Use a trained model generated by GeneMarkS
- After training, runs very quickly

Genemark HMM - Script (run_gmhmm.pl)

```
use strict;
#define variables
my $filename = ();
my @SRRname = ();
#input file
$filename = $ARGV[0];
#check file if exit and open file
unless (-e $filename){
    print "This file \"$filename\" do not exit! Please check it!";
    exit;
}
unless (open FILENAME, $filename){
    print "Cannot open this file!!";
    exit;
}

@SRRname = <FILENAME>;
chomp @SRRname;
close FILENAME;
#run GeneMarkhmm, you can replace the command line in ``
foreach my $i (@SRRname){
    `gmhmmp -o $i.HMM.gff -f G -m /projects/data/team1_GenePrediction/bin/genema
    rk_suite_linux_64/gmsuite/GeneMark_hmm.mod $i`
}
```

Genemark HMM - Summary Histogram



Prodigal - Background

Average runtime: 17 seconds/assembly
Sensitivity: 94.71%
PPV: 94.07%

- Input: assembly file (Fasta)
- Output: gene coordinates (GFF), nucleotide file (Fasta), protein file (Fasta)
- Uses a preset training file it creates

- Simple to install, simple to use

Prodigal - Script (Run_Prodigal.sh)

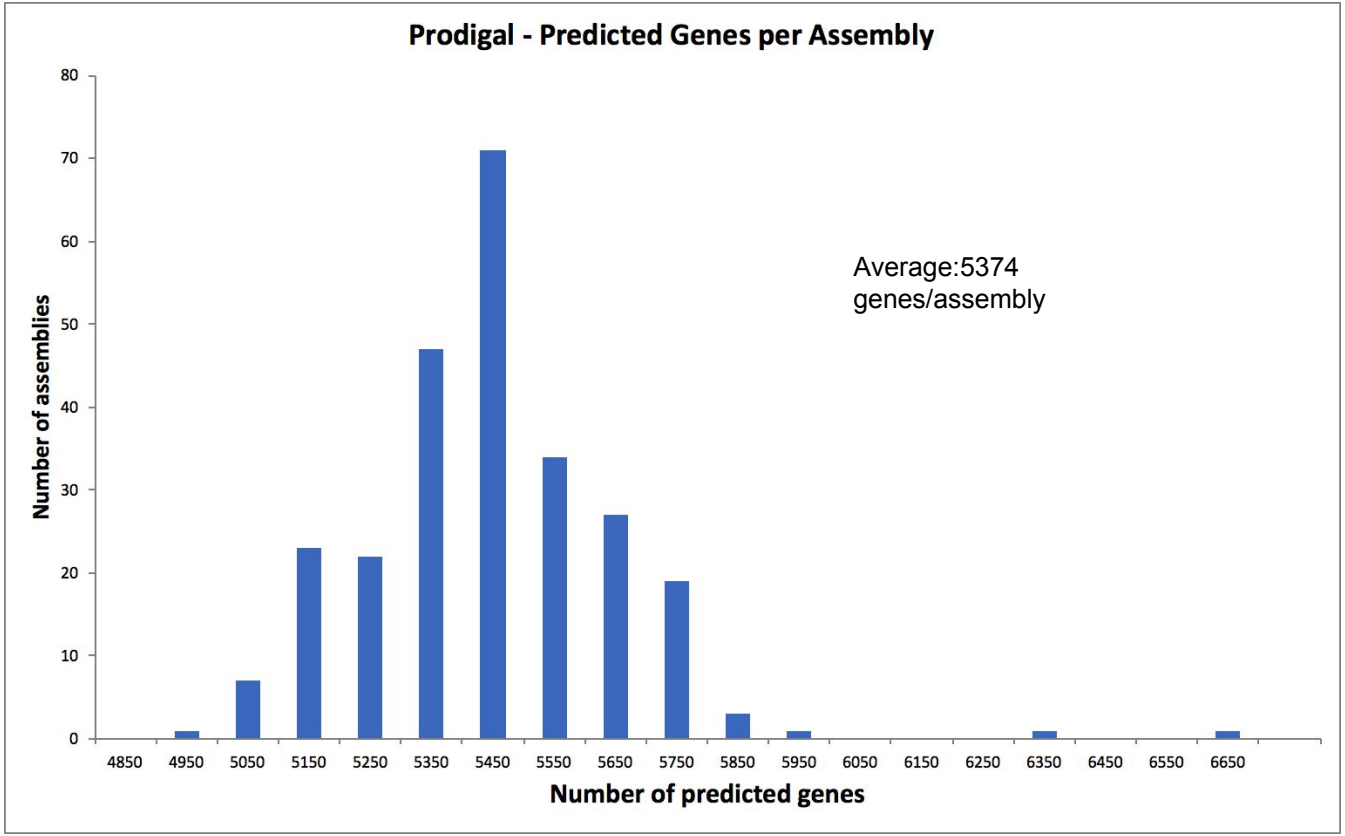
```
# Currently requires assemblies to be in a directory called "assemblies".  Place script next to the assemblies directory to run.


mkdir output nucleotide protein log 2> /dev/null;


for file in assemblies/*; do
   base=`echo $file | awk -F'[/.]' '{print $2}'`;
   echo "Running Prodigal on $base";
   Prodigal -i $file -f gff -o output/"$base"_Prodigal.gff -d nucleotide/"$base"_Prodigal.nucleotide.fa -a
protein/"$base"_Prodigal.protein.fa 2> log/"$base"_Prodigal.txt;
   echo "Finished $base on `date`";
done


echo "done!";
exit
```

Prodigal - Summary histogram



Server Files

Final Prodigal output path: /projects/data/team1_GenePrediction/Prodigal_output_all/

- ./output contains the gene coordinates in gff format
- ./nucleotide contains the nucleotide sequences in fasta format
- ./protein contains the protein sequences in fasta format

- Extra:
  - ./log contains the log files for each run of Prodigal
  - ./assemblies is a symlink of the final assemblies
  - ./archive contains all the above for an older version of the assemblies
  - ./graph is a histogram of predicted number of genes for all assemblies with Prodigal

Final GeneMark HMM output path: /projects/data/team1_GenePrediction/GeneMark_HMM_output/

- Contains the gene coordinates in gff format

Content

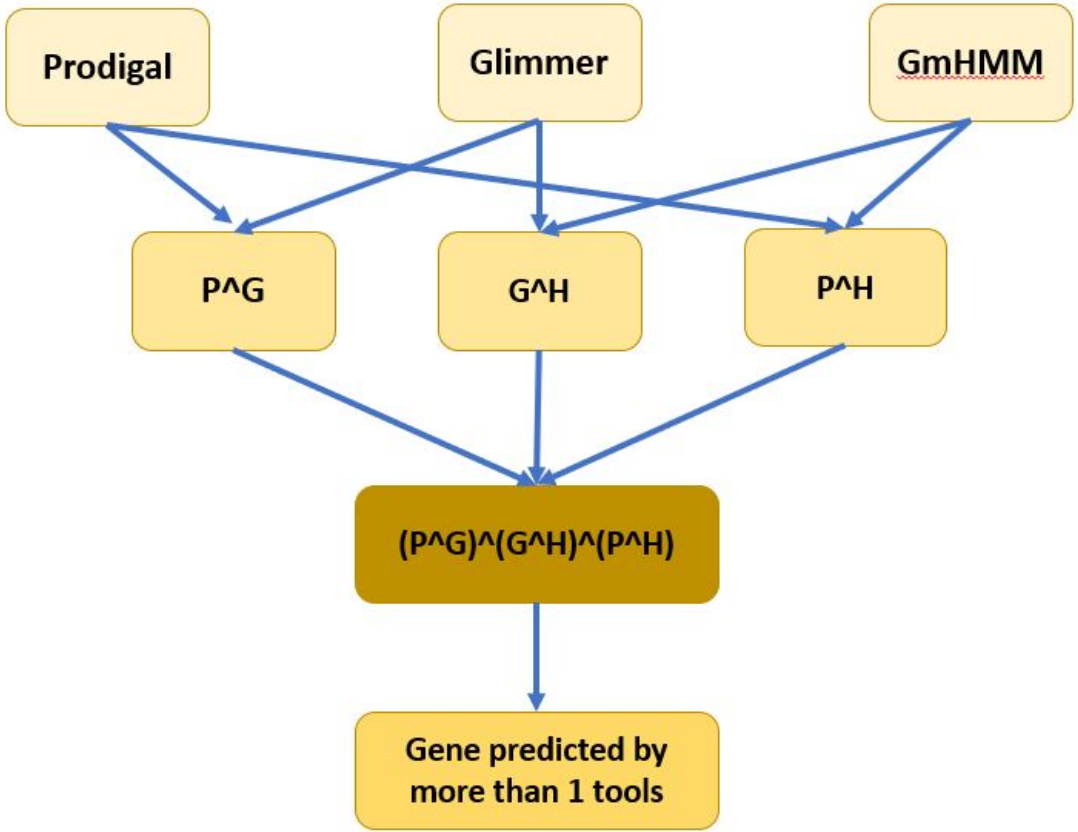
Protein Coding Gene Prediction

Protein Coding Gene Validation

RNA Region Prediction

Final Pipeline

Workflow

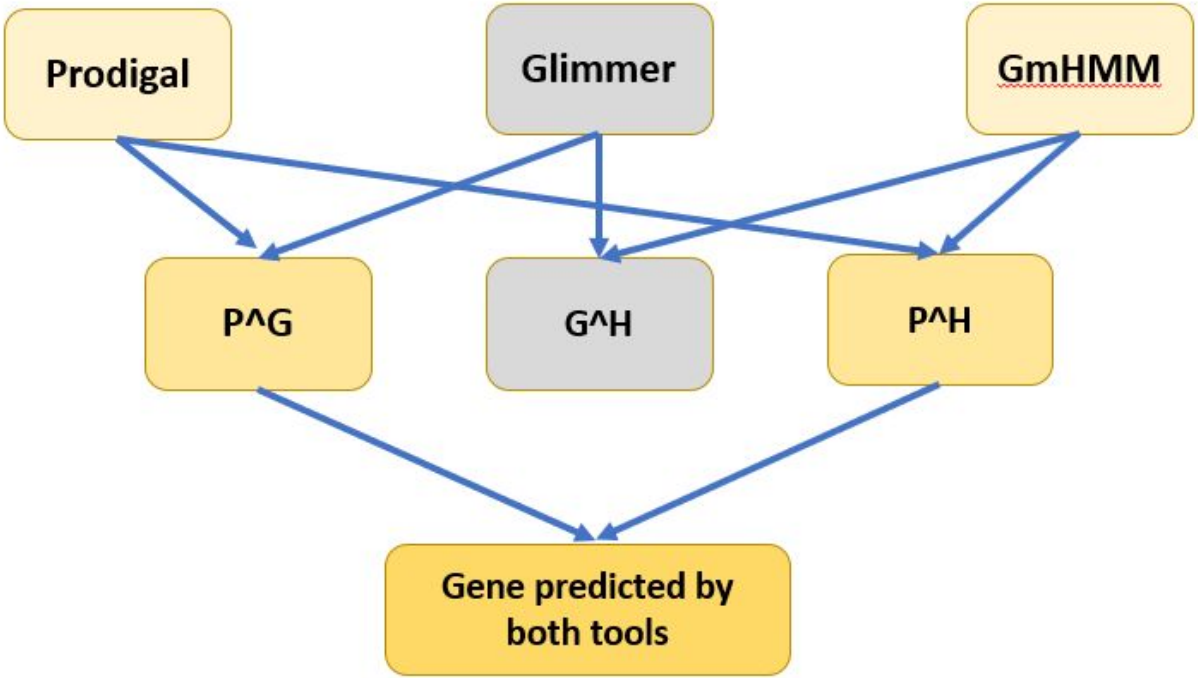


Compare

Combine

Merge

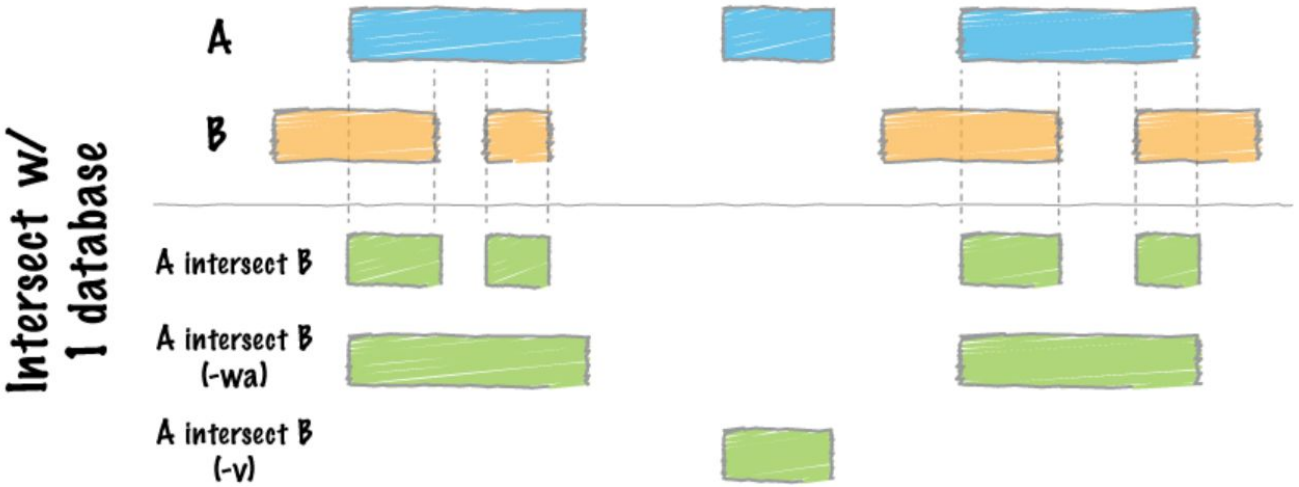
Workflow



Compare

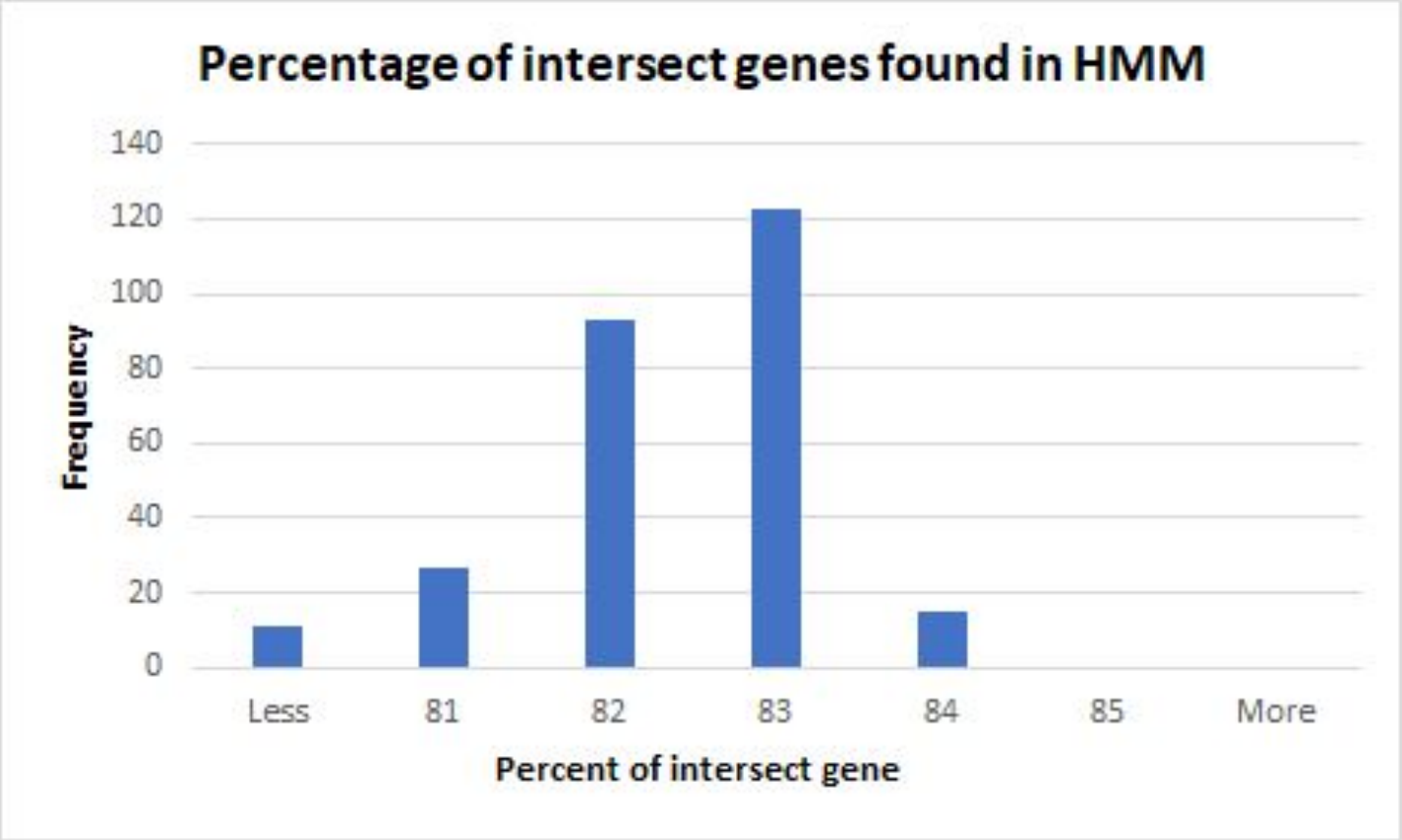
Combine

Tool

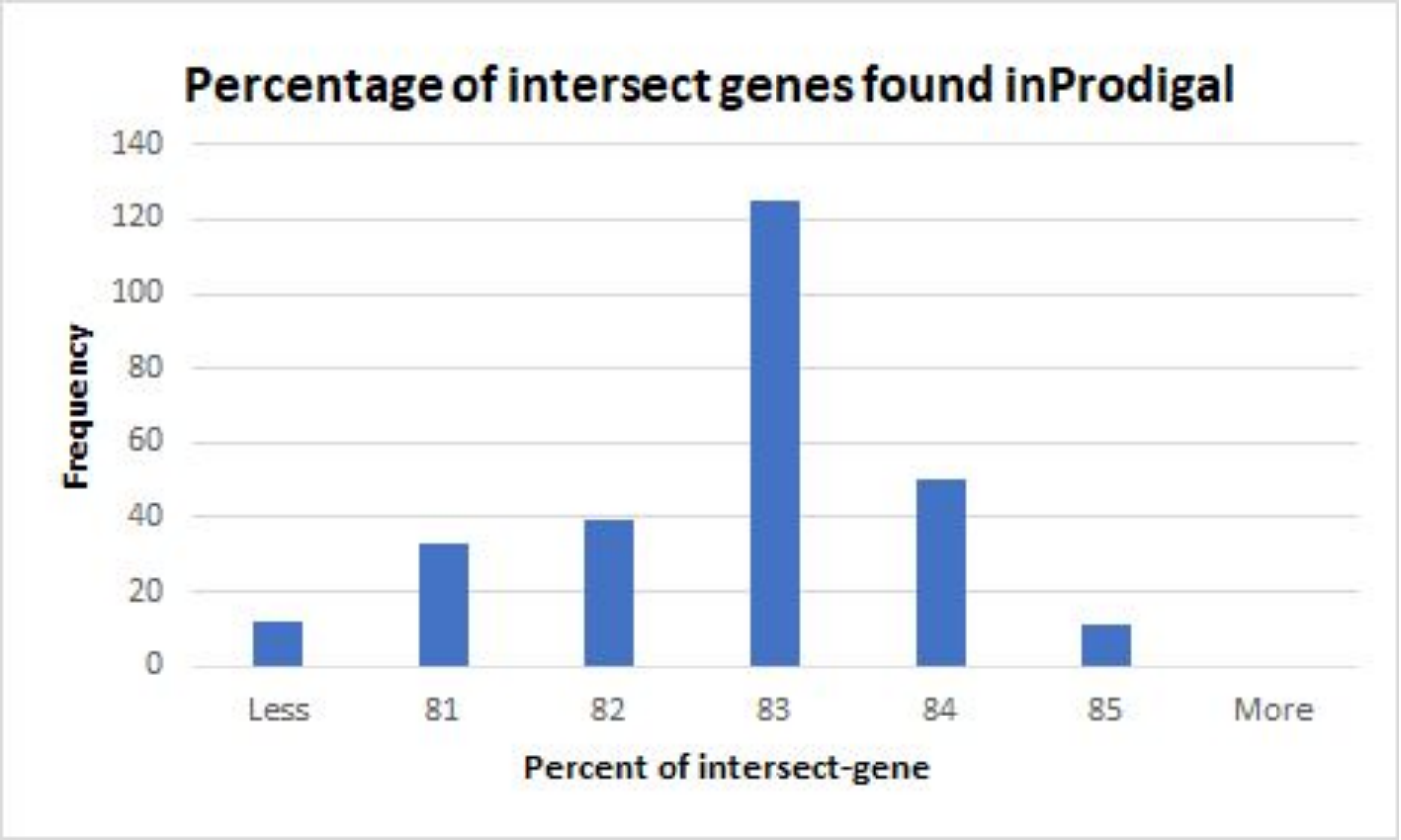


Code: `bedtools intersect -f 0.99 -r -a $f1 -b $f2`

Final Result



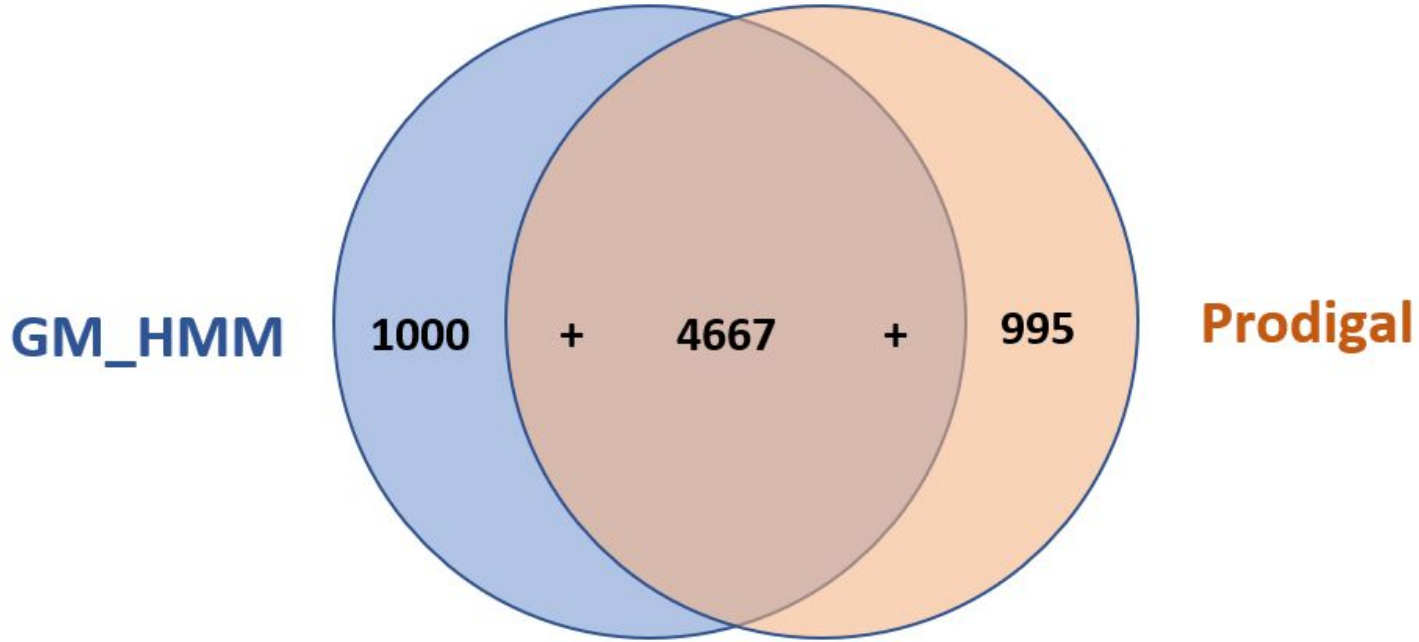
Final Result



Checking our Method

Method	True Positives	False Positives	False negatives	Sensitivity	PPV
Prodigal	5015.8	437.7	480.6	91.2	92.0
GeneMark HMM	5061.5	507.1	456.4	91.7	91.1
Intersect	4383.4	323.6	1096.4	80.0	93.1
Union	5693.9	618.5	423.1	93.1	90.25

Final Result



Final method

Keeping the union

- Use bedtools intersect to get the unique genes for each tool and combine with the overlap

Code:

```
bedtools intersect -f 0.99 -r -wa -v -a $f1 -b $f2 > complement.gff
```

```
bedtools intersect -f 0.99 -r -a $f1 -b $f2
```

```
concatenate files
```

Content

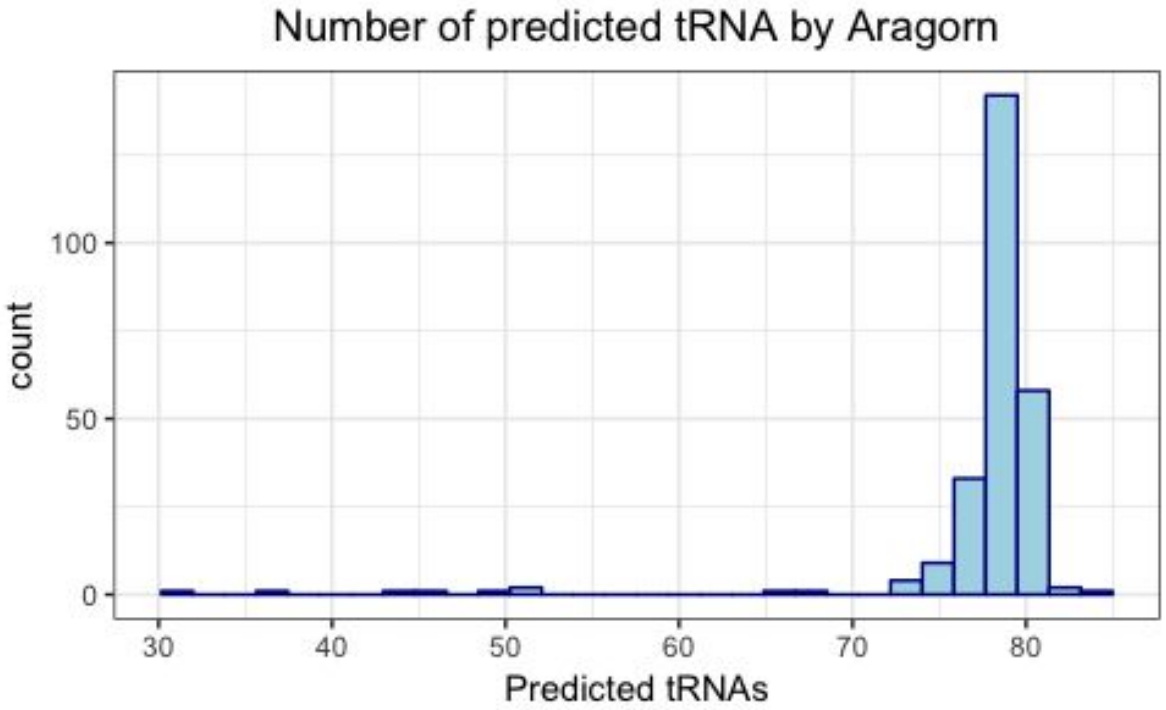
Protein Coding Gene Prediction

Protein Coding Gene Validation

RNA Region Prediction

Final Pipeline

Aragorn



Convenience: Already part of Prokka

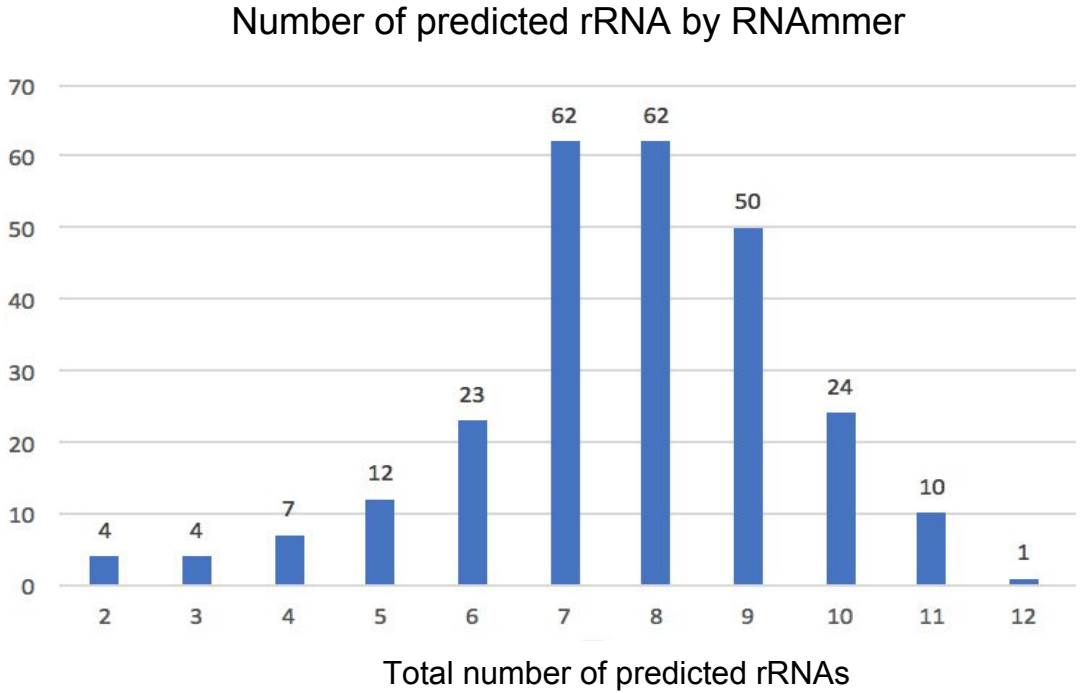
Average running time: seconds per genome

Output: fasta sequences

Average tRNA predicted: 77

RNAmmer

- Average running time: 1 min / genome (compared to ~3 mins without scaffold)



Average rRNA: 7

Infernal & Rfam

Infernal

Ab initio tool based on hidden markov models
Uses primary and secondary structure information for greater accuracy

\+ Detection of remote homologs
\- Slow (for our purposes)

Rfam

Database of RNA family covariance models (and other RNA information)
RNA family selection:

1. Filter in families with sequences reported for all Rfam Klebsiella pneumoniae species
2. Choose RNA family whose function may contribute to heteroresistance

Result → **istR** (Rfam ID: RF01400)

Infernal & Rfam

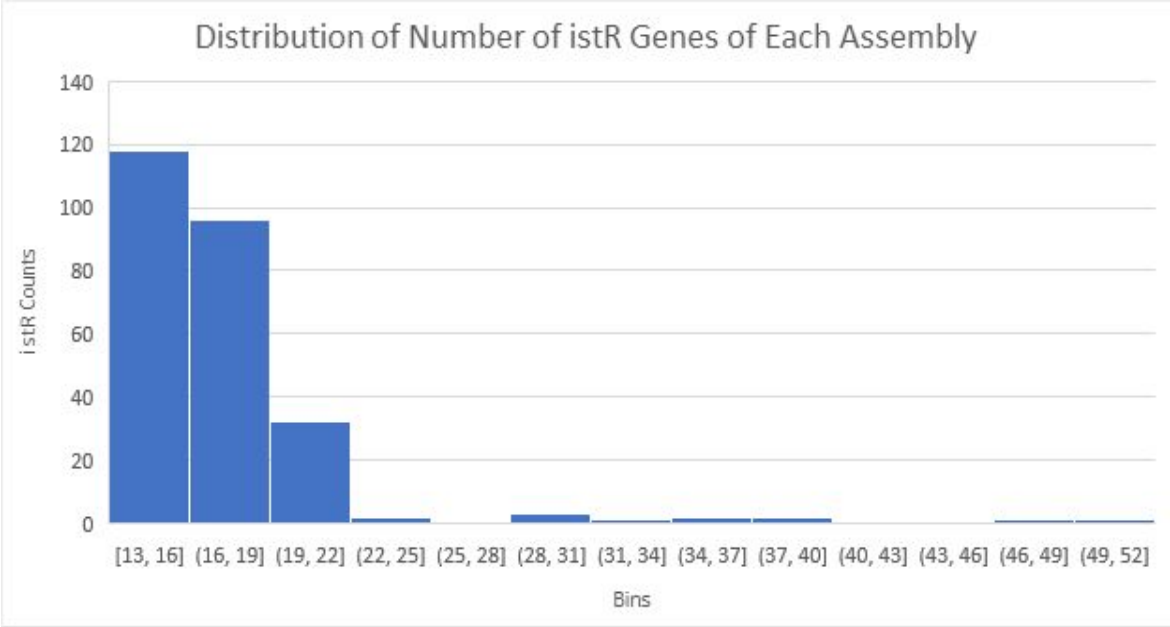
istR
Inhibitor of SOS-induced toxicity by RNA

istR is an antitoxin against TisB

Normal physiological conditions → istR inhibits TisB toxicity

DNA Damage (SOS response) → TisB overexpression depletes istR, cells grow slowly rather than die

Possible role in Klebsiella pneumoniae spp. heteroresistance.



ncRNA Validation

Difference from validation of protein coding genes:

1. Method: sRNA -- Reference based method, hard to validate
2. Query sequence: tRNA & rRNA -- highly conserved
3. Focus: Second structure > sequence
4. Purpose: confirm existence > predict gene
5. Assembly level: scaffolds vs complete

ncRNA Validation

complete level assembly

Type	RefSeq	Size (Mb)	GC%	rRNA	tRNA
Chr	NZ_CP009775.1	5.52	57.3	25	88
Chr	NZ_CP011976.1	5.39	57.4	25	87
Chr	NZ_CP011976.1	5.39	57.4	25	87
Chr	NZ_CP015822.1	5.45	57.3	25	88
Chr	NZ_CP022573.1	5.39	57.4	25	88

scaffolds level assembly

Scaffolds	RefSeq	Size (Mb)	GC%	rRNA	tRNA
3	NZ_AMLM00000000.1	5.75	56.9	22	89
30	NZ_AMRH00000000.2	5.77	56.9	6	72
51	NZ_ACZD00000000.1	5.45	57.2	3	62
72	NZ_JRGE00000000.1	3.07	57	2	45
120	NZ_LEZX00000000.1	5.85	56.8	18	92

Our results (average):

rRNA: 7

tRNA: 77

Content

Protein Coding Gene Prediction

Protein Coding Gene Validation

RNA Region prediction

Final Pipeline

Final Pipeline