Gene Prediction

Final Pipeline & Results

Beatriz E Saldana, Parisa Y Zowj, Ayush Semwal, Siu Lung Ng, Sini Nagpal, Sarthak Sharma, Rong Jin, Jiani Long, Qi Zhang

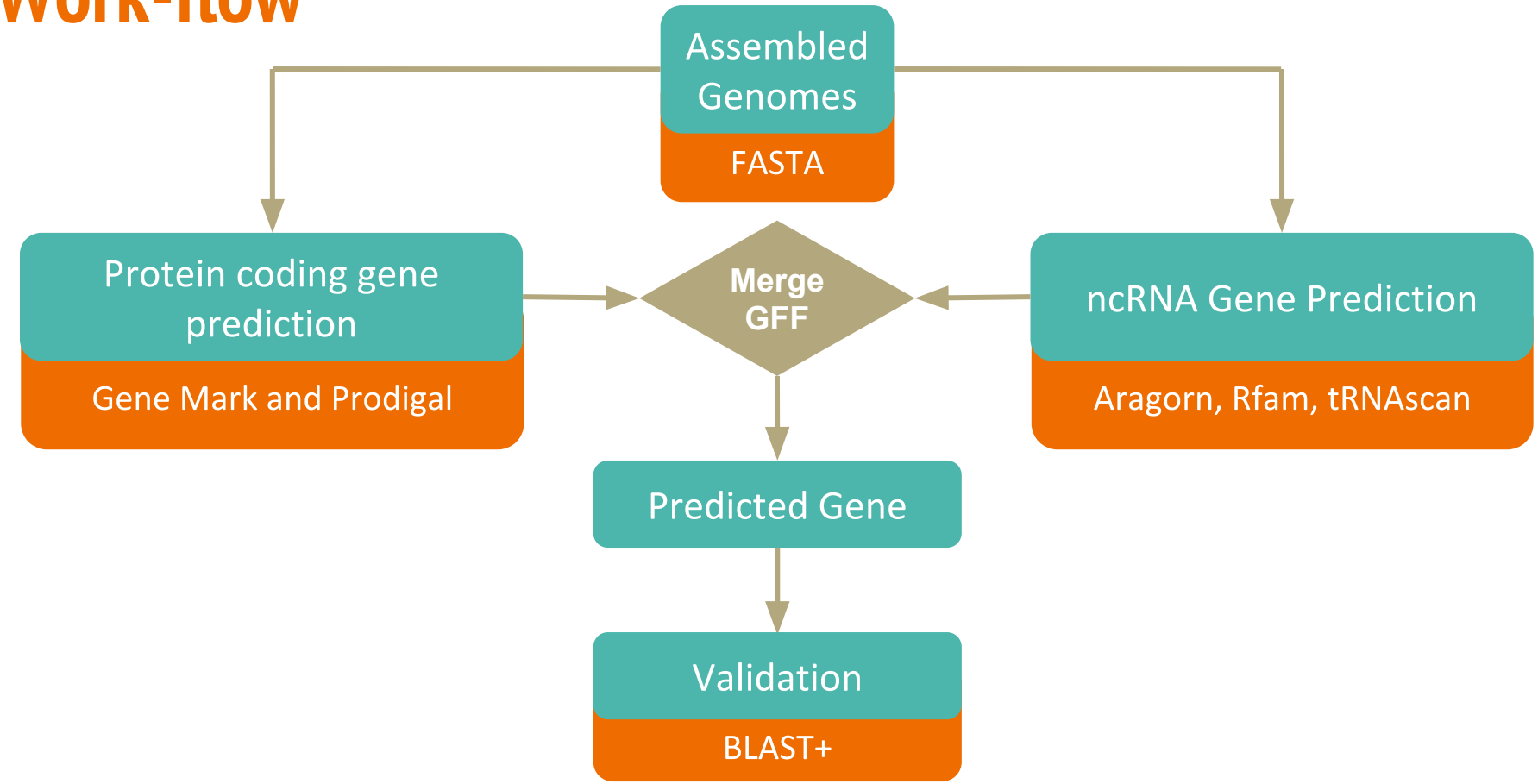
Introduction

Project Goal : Understand heteroresistance using genetic determinants of antibiotic resistance

- Genome Assembly
- **Gene Prediction**
- Functional Annotation
- Comparative Genomics
- Webserver

Problem Statement: Gene Prediction Group

Work-flow


Assembled Genomes
FASTA
Protein coding gene prediction
Gene Mark and Prodigal
Merge GFF
ncRNA Gene Prediction
Aragorn, Rfam, tRNAscan
Predicted Gene
Validation
BLAST+

Tools for Gene Prediction

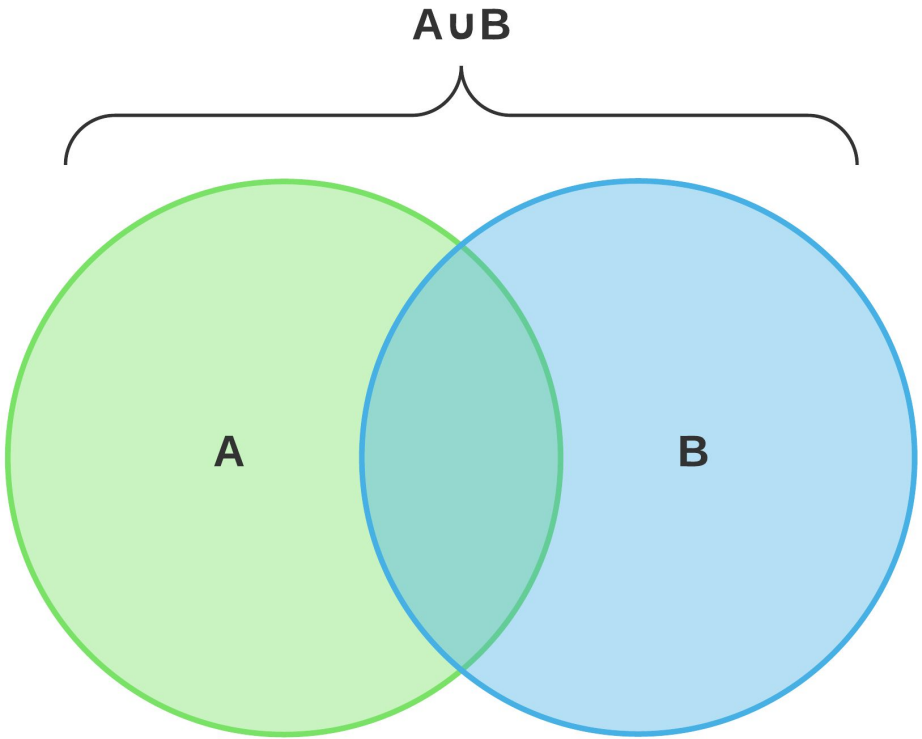
Coding region tools

- Prodigal
  - Good balance between speed and accuracy
  - Performs well with high GC-Content
- GeneMark S
  - Most up-to-date
  - Creates one model per genome

Non-coding RNA tools

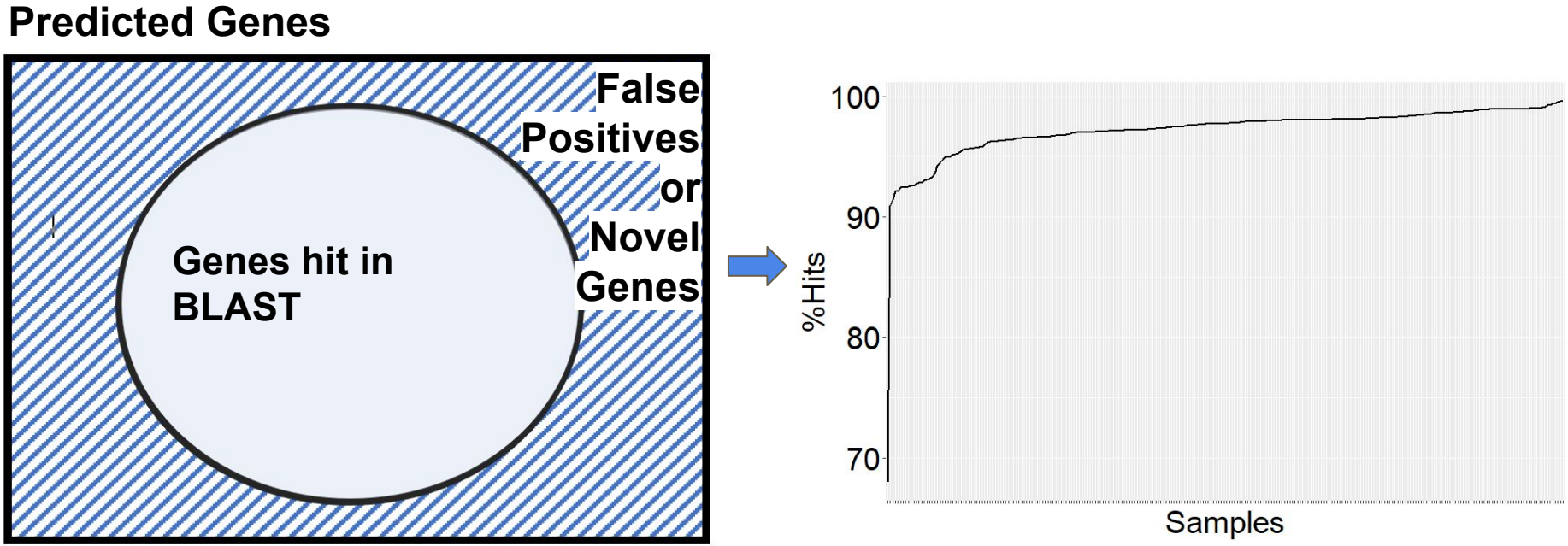
- Rfam
  - Predicts tRNA, tmRAN, rRNA, and sRNA
- Aragon
  - Predicts tRNA and tmRNA
- tRNAscan
  - Predicts tRNA

Merging Results - Bedtools Intersect



$$A \cup B = A + B - (A \cap B)$$

Validation with BLAST+



Results

Non-Coding RNA

ncRNA Results - Average Predictions per Genome

Rfam

- tRNA: 77
- tmRNA: 1
- rRNA: 12
- sRNA: 124

Aragorn

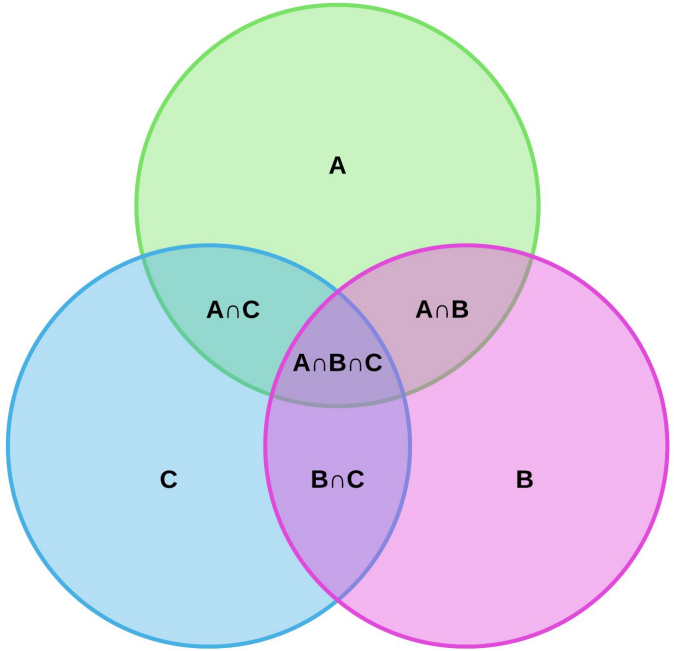
- tRNA: 76
- tmRNA: 1

tRNAscan

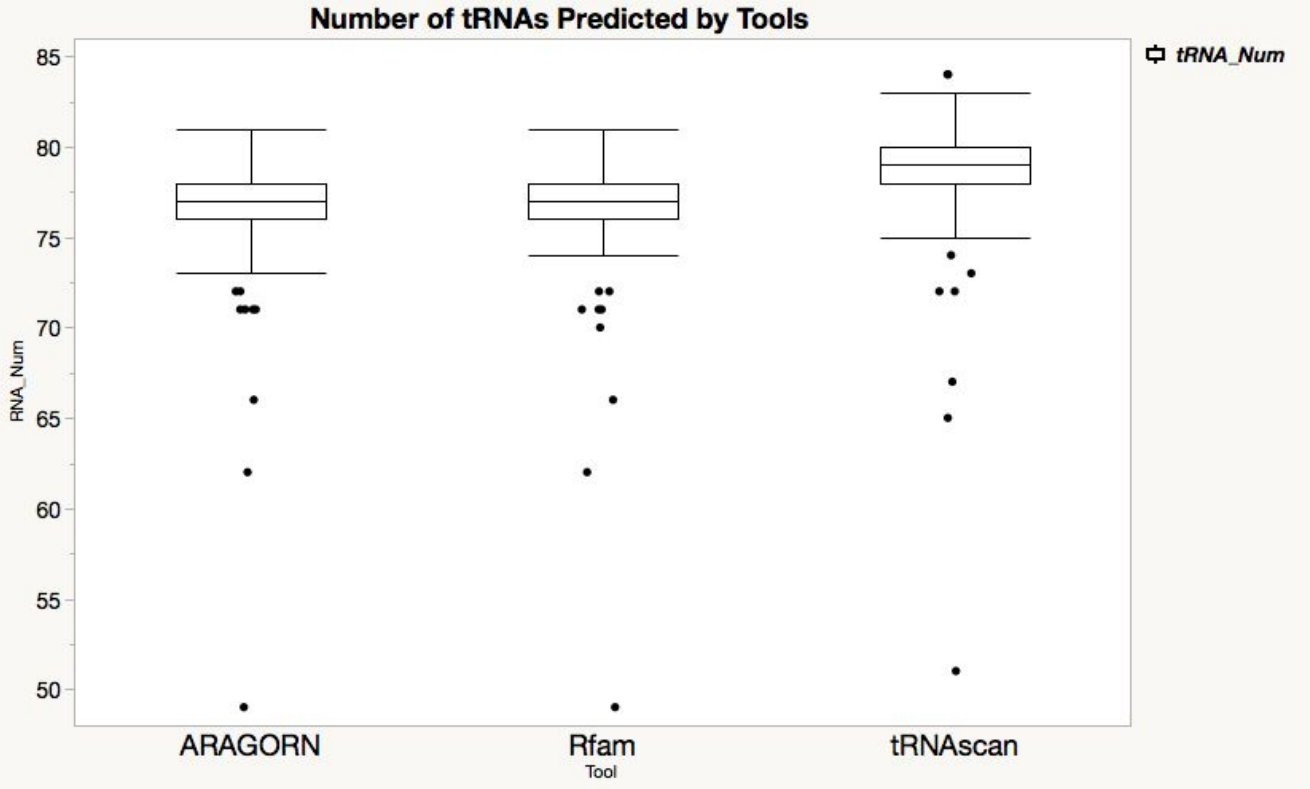
- tRNA: 79

Merging Results

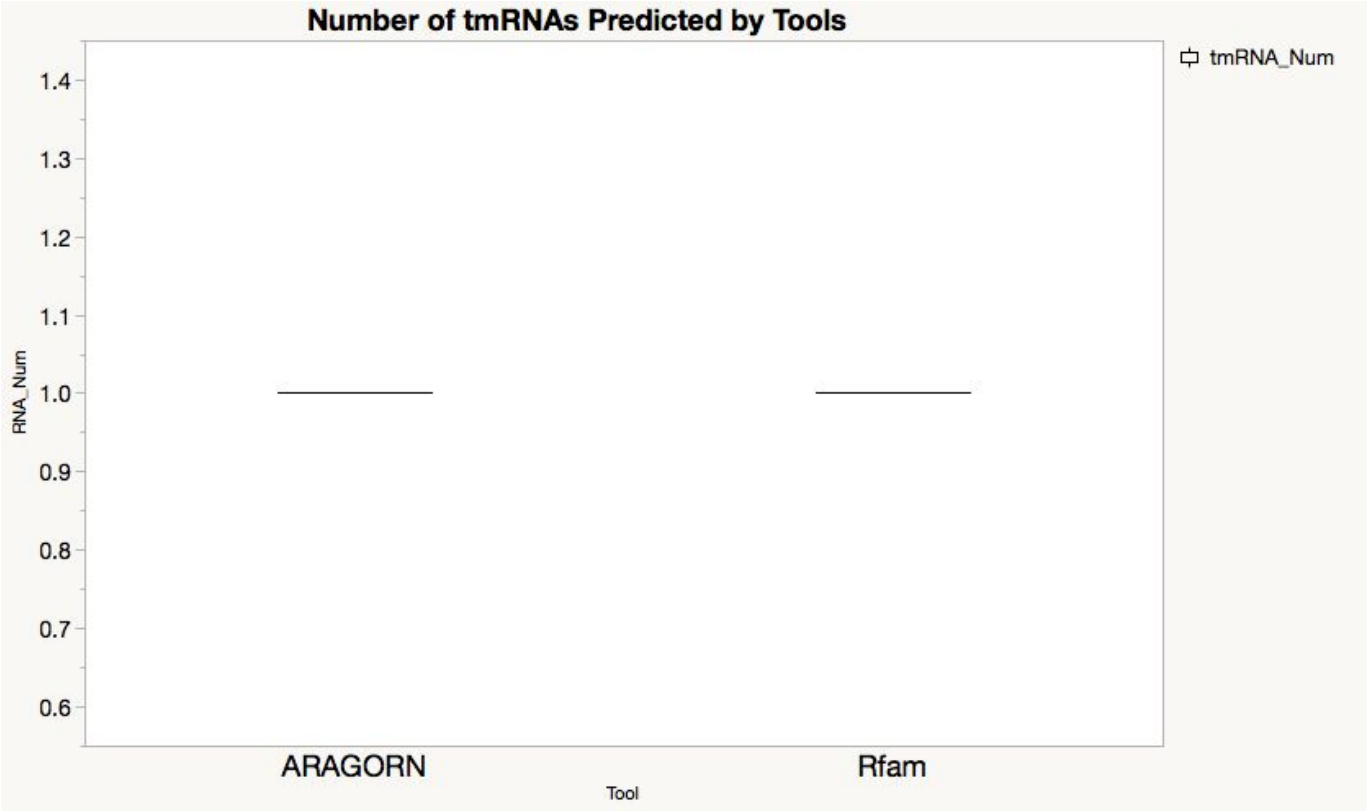
- UNION
  - tRNA: 81
  - tmRNA: 1
- INTERSECTION
  - tRNA: 77
  - tmRNA: 1



Prediction Results -tRNA



Prediction Results -tmRNA

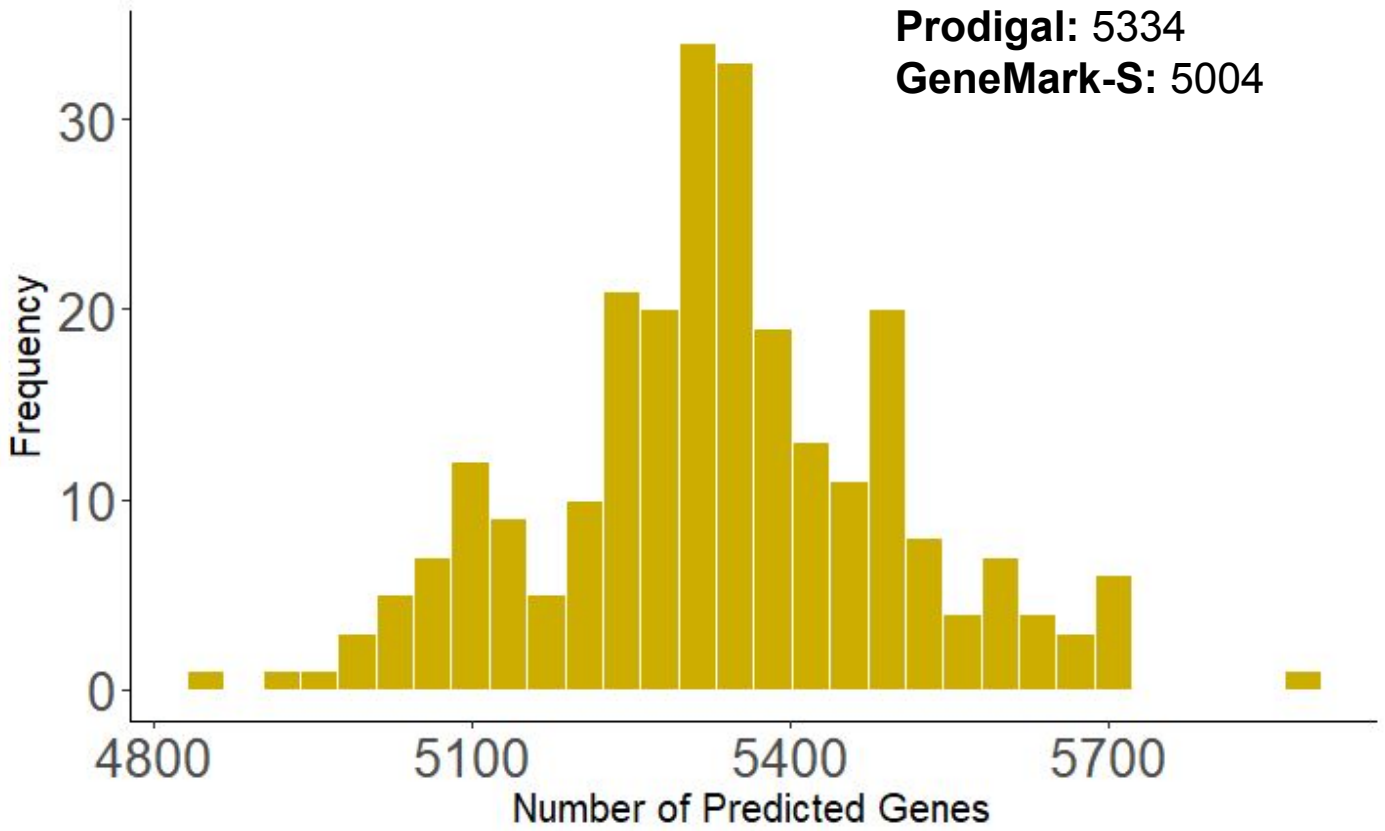


Results

Coding Regions

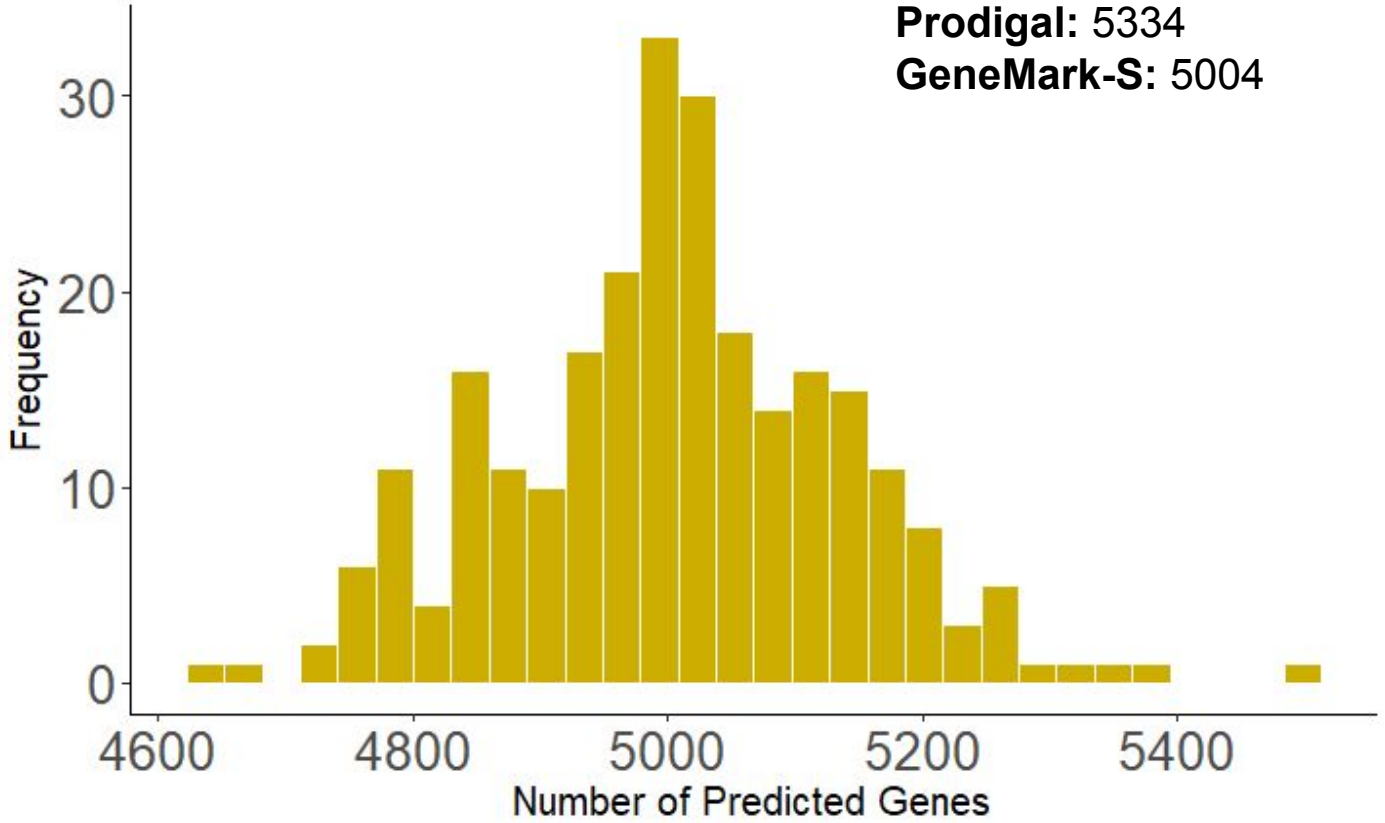
Prodigal Results

Average Predictions Per Genome
Prodigal: 5334
GeneMark-S: 5004

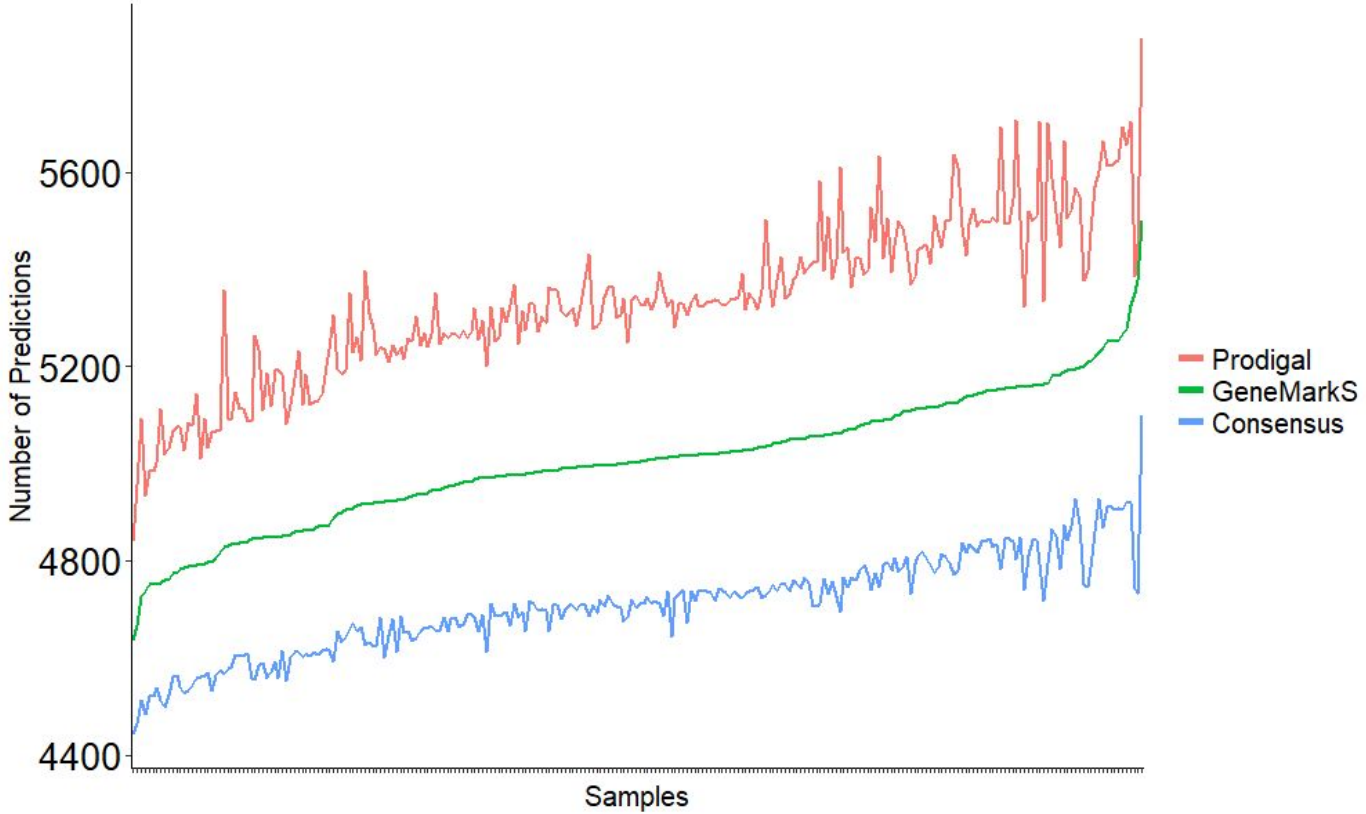


GeneMark S Results

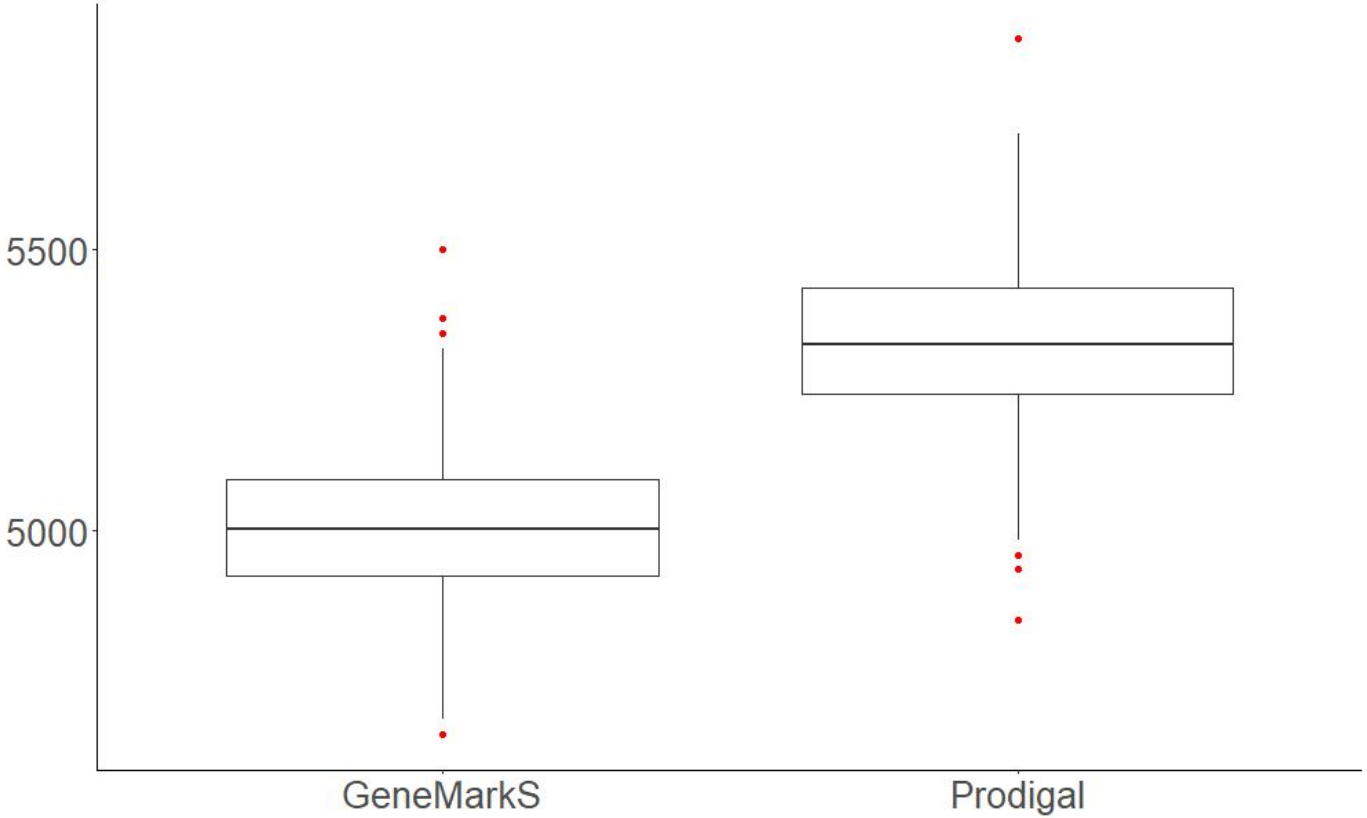
Average Predictions Per Genome
Prodigal: 5334
GeneMark-S: 5004



Final Coding Region Results



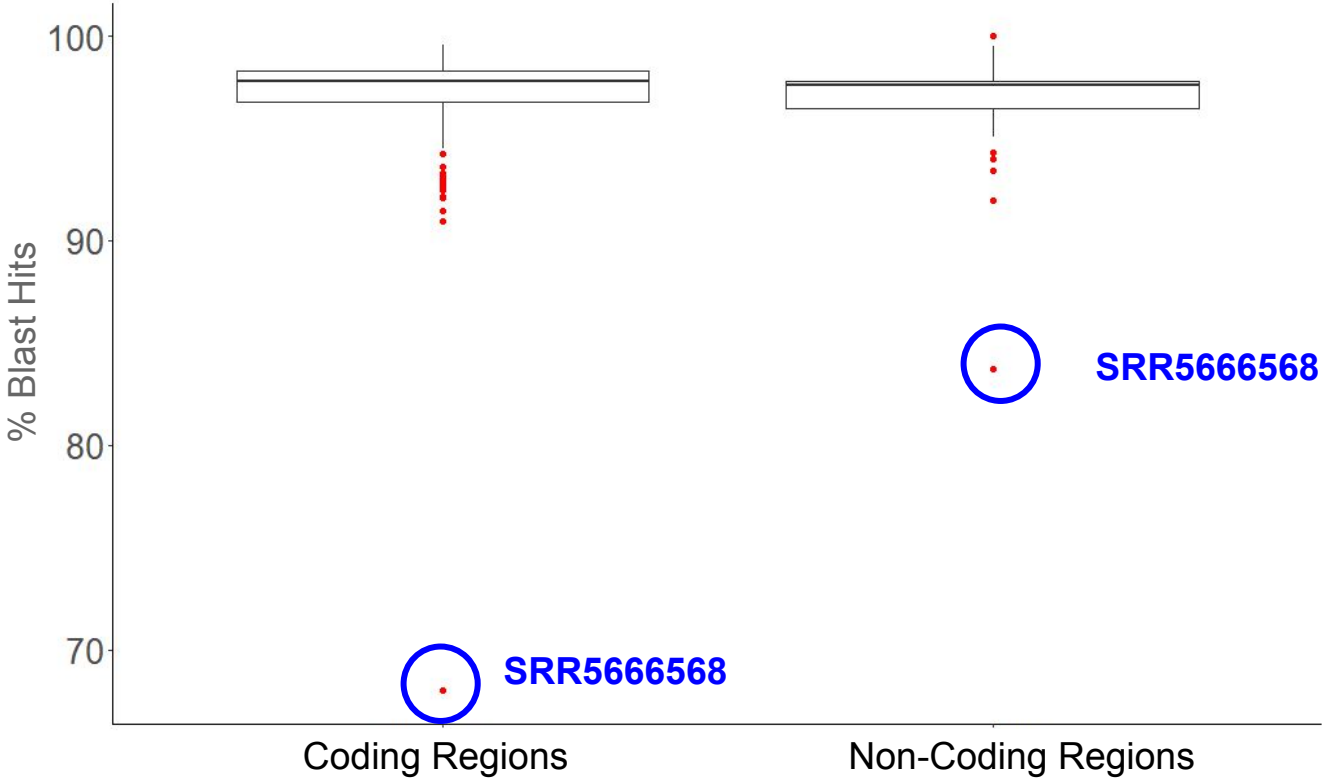
Final Coding Region Results



Final Results

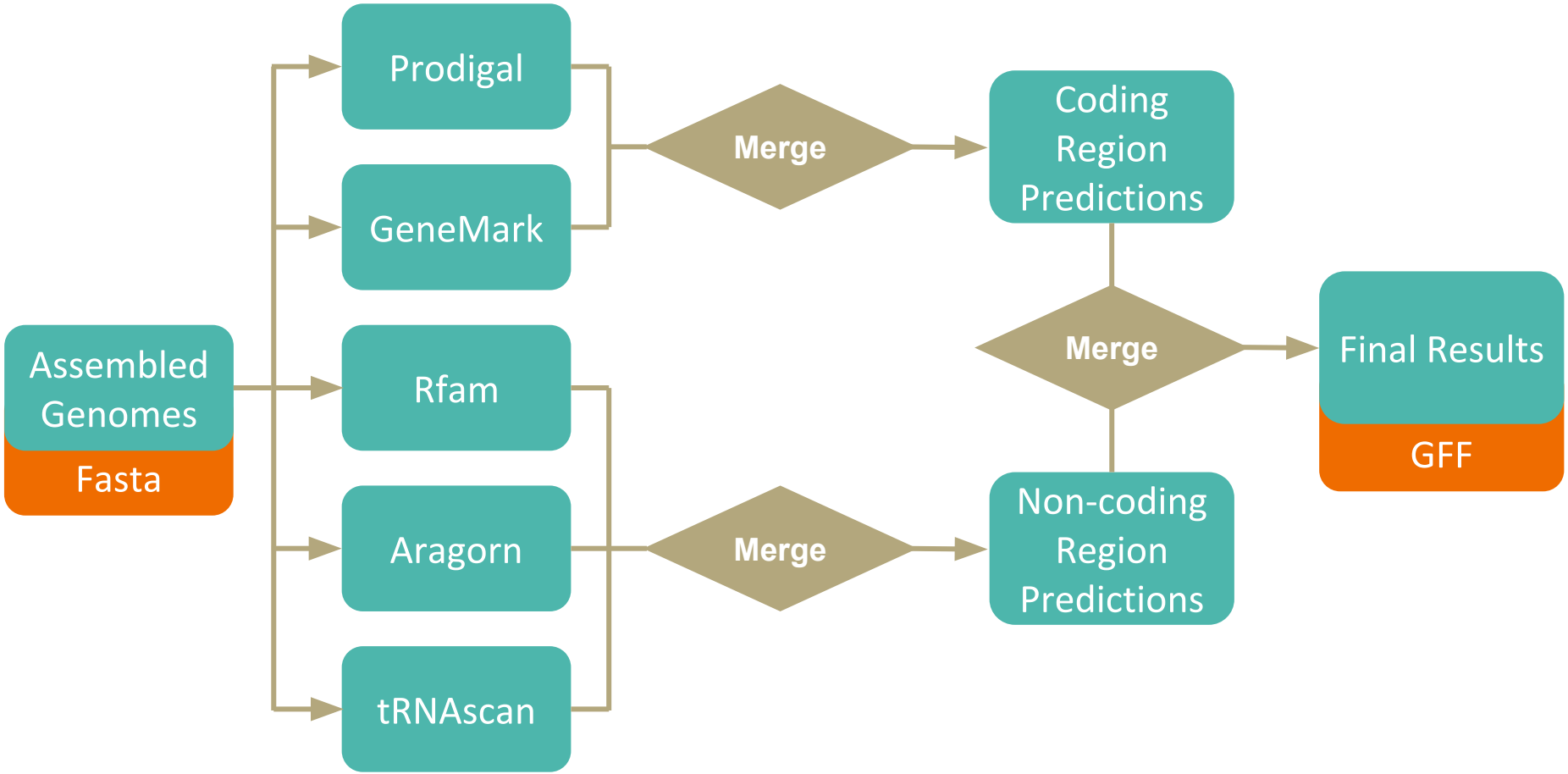
Coding + Non-coding Predictions

Final Results - BLAST Hits



Pipeline

Description



./genePredictionPipelineT2G2.sh -p path/to/assembled/genomes/

Pipeline Options

- -f The fast version of the pipeline will be run
  - GeneMark-S will not be run → GeneMark HMM will be run instead
    - GeneMark-S: Creates a model for every genome → More accurate!
    - GeneMark-HMM: Creates one model and uses it for all genomes
  - Rfam will not be run (could affect sRNA predictions)

- -t The temporary files will be kept
  - Results from all tools
  - Results from all merging steps

- -v Verbose mode

Wiki Documentation

Team II Gene Prediction Group

Contents [hide]

Questions?