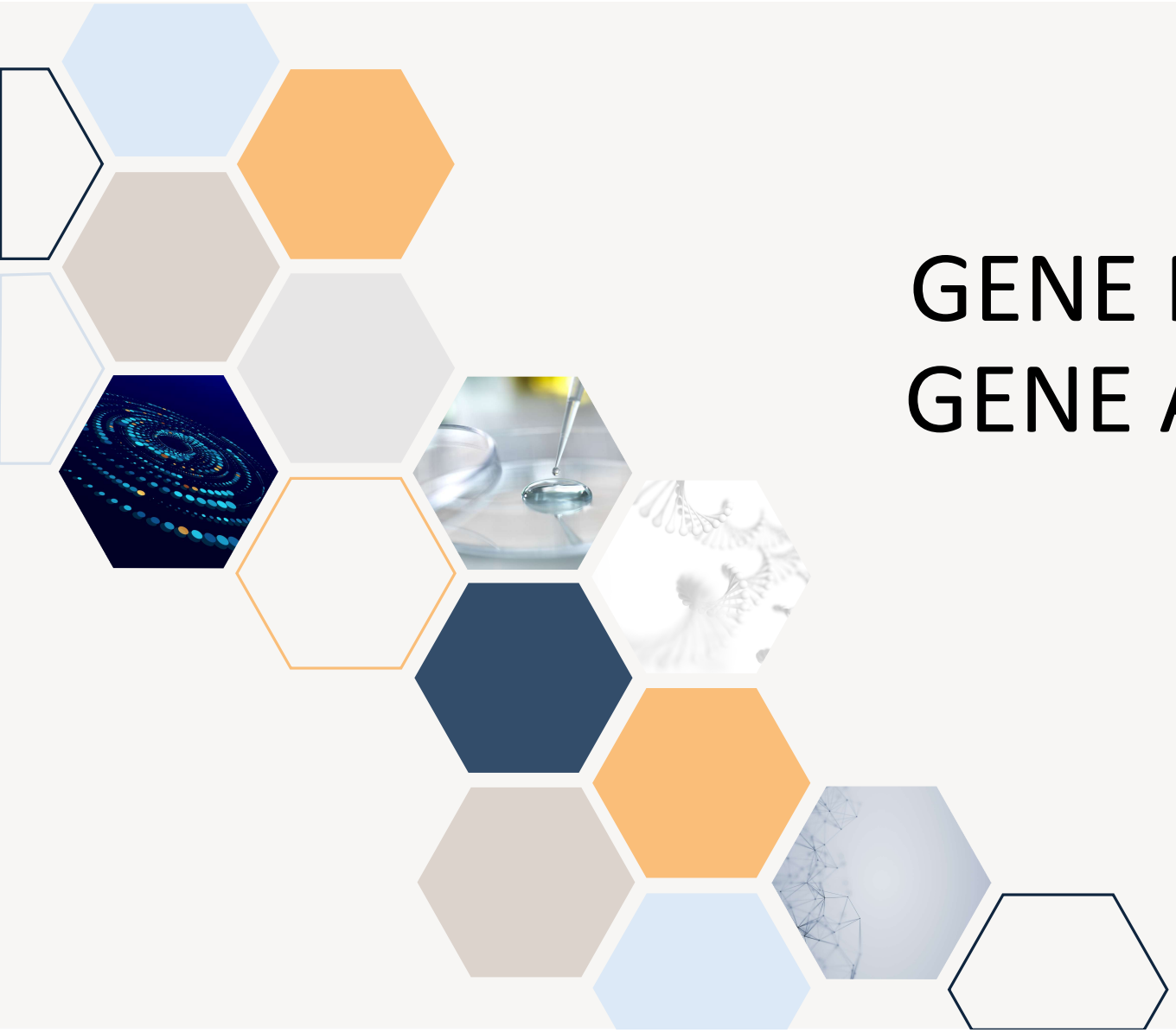
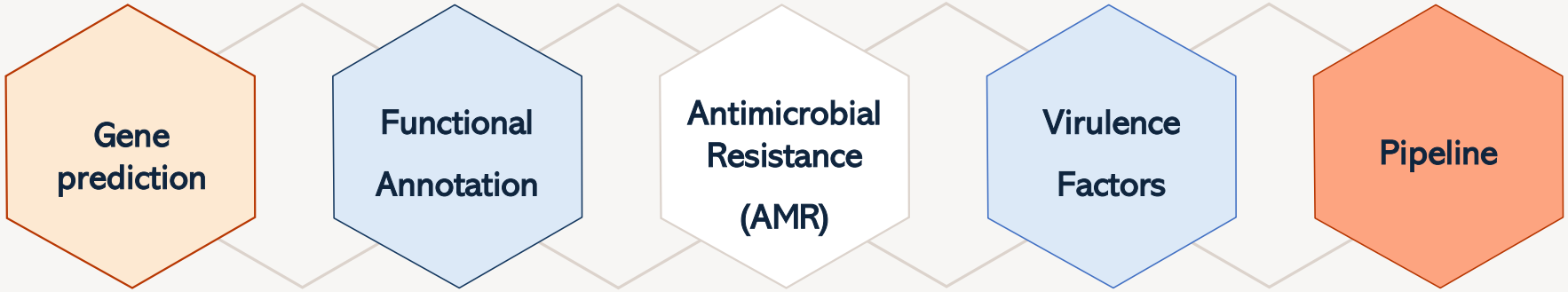


GENE PREDICTION& GENE ANNOTATION: RESULTS

Team F Group 2

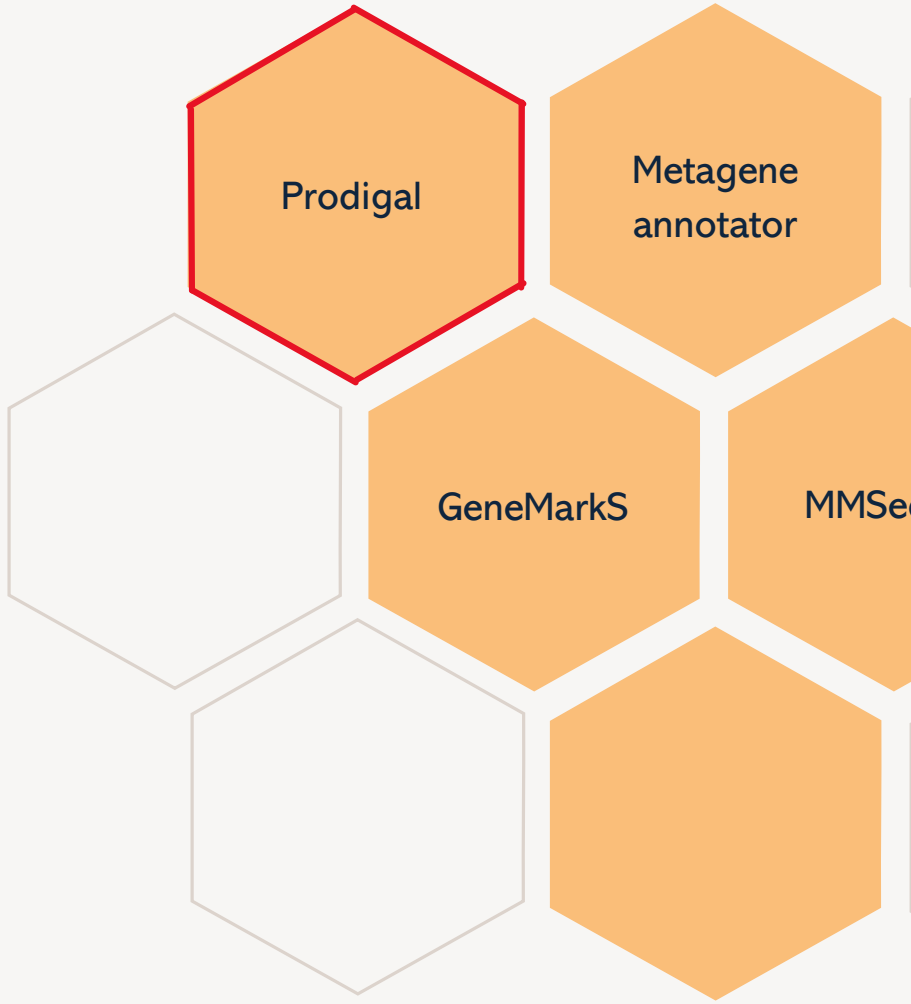
Review workflow



Benchmark with 10 sample files
Select best preformed tools

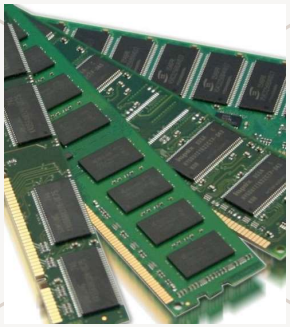
Use selected tools for all 30 samples files
Build the final pipeline

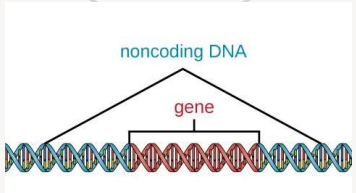
Gene Prediction Tools

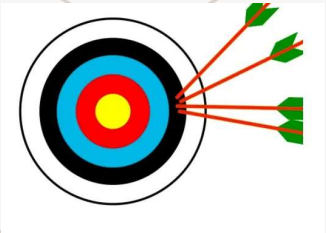


*Homology-based; MMseqs database from ensembl reference sequence *Listeria innocua*

Prediction Benchmarks







Time per file

/usr/bin/time -f
"Time: %e seconds"
<command>

Max RAM

/usr/bin/time -f
"Memory: %M kB"
<command>

Number of Predicted Genes

Precision*

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

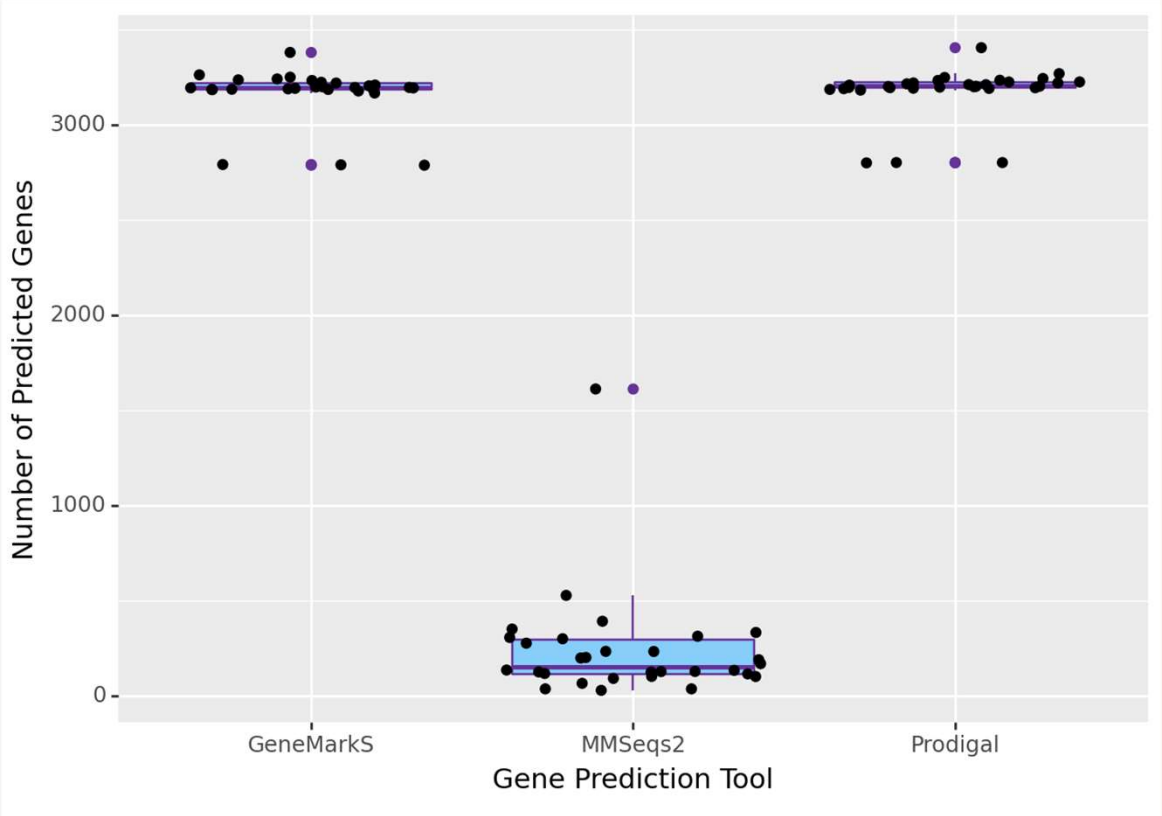
Recall*

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

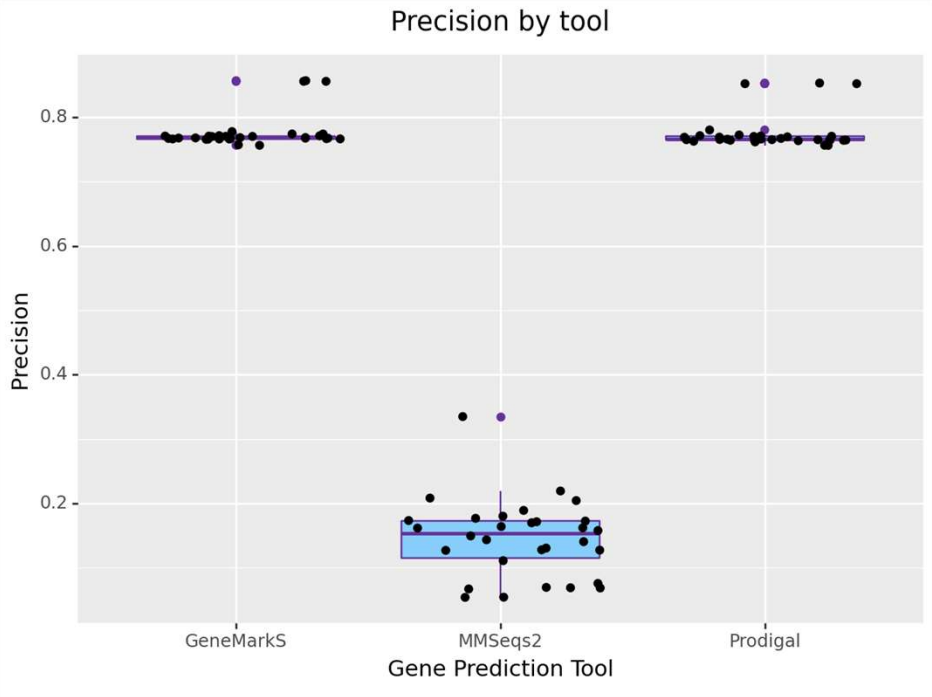
- Based on comparison with *Listeria innocua* reference sequence
- TP: >80% sequence identity, alignment >90% length of query

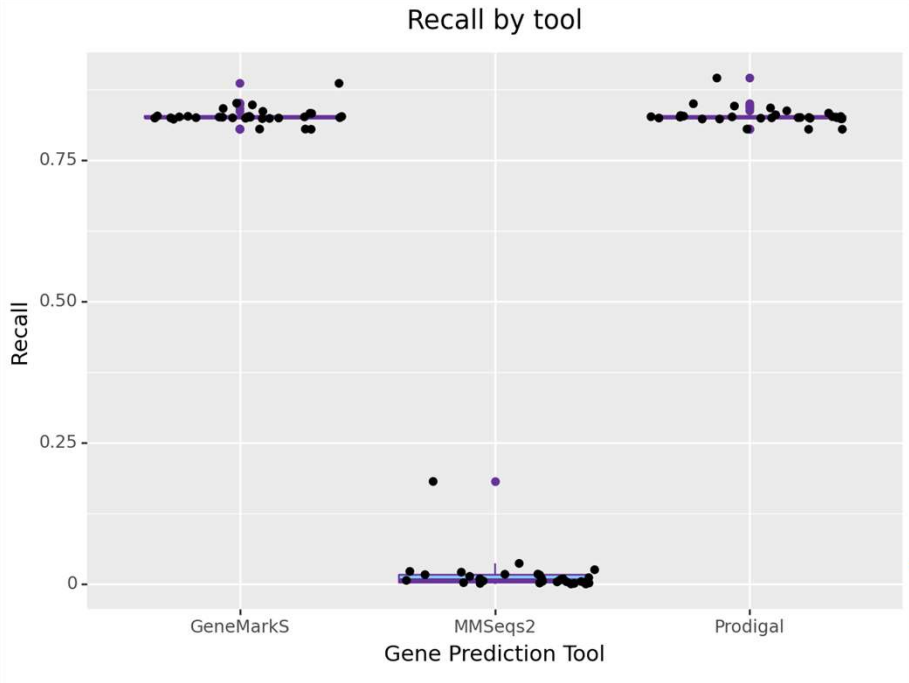
Tool	Time per file (sec)	Max RAM (kb)
GeneMarkS	42	127,276
Metageneannotator	1	32,604
MMSeqs2	8	8,489,580
Prodigal	2	211,220

Number of Predicted Genes by Tool

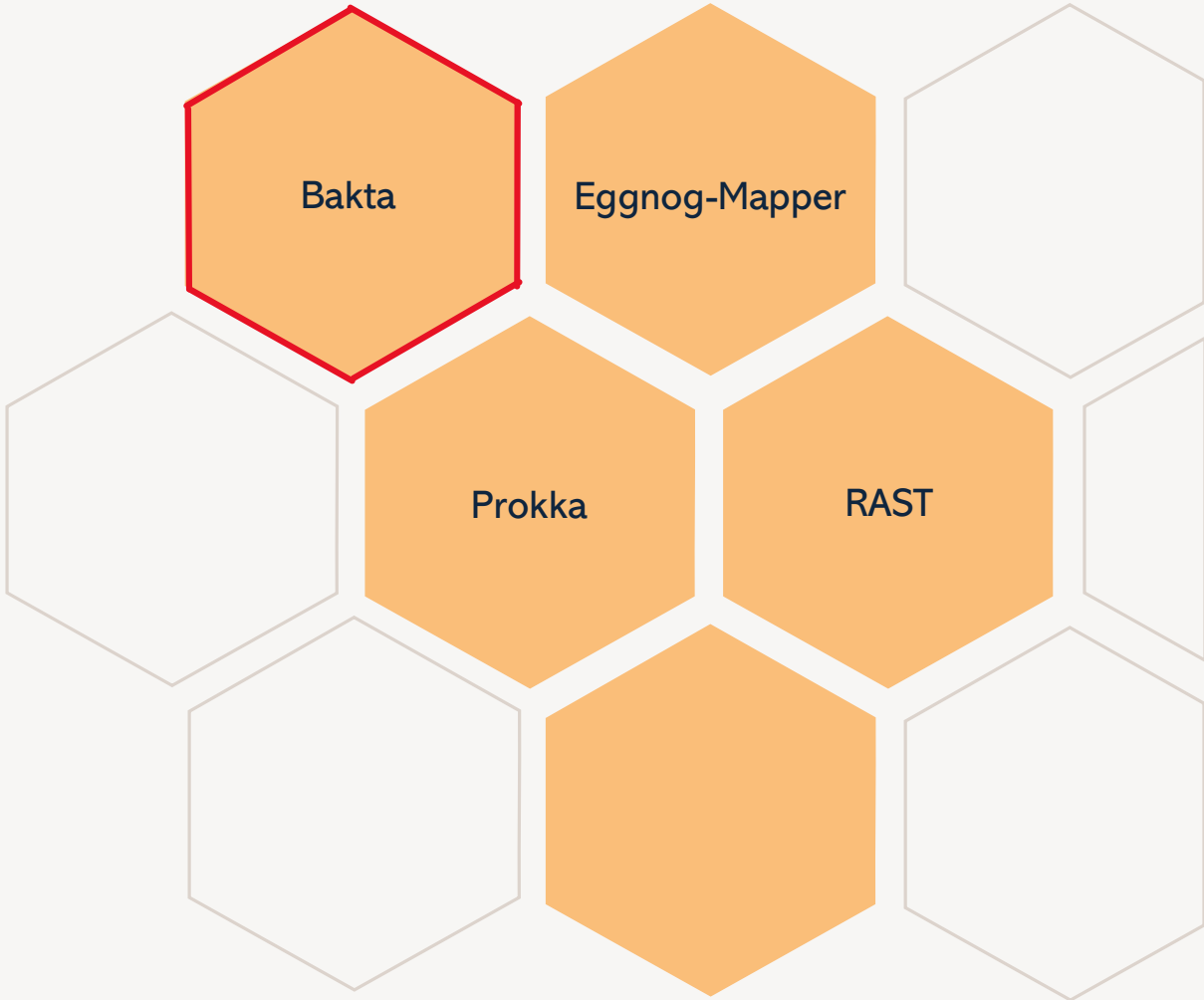


Precision and Recall by Gene Prediction Tool



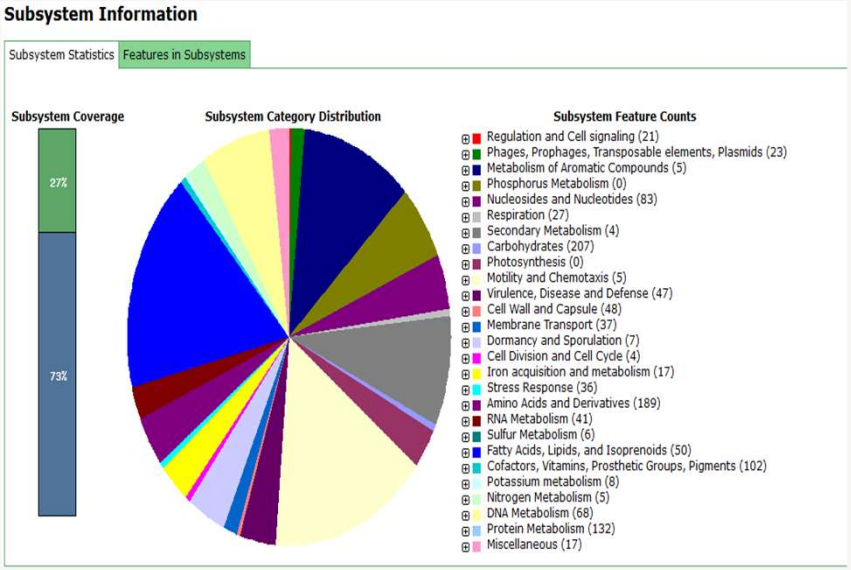


Gene Annotation Tool comparison



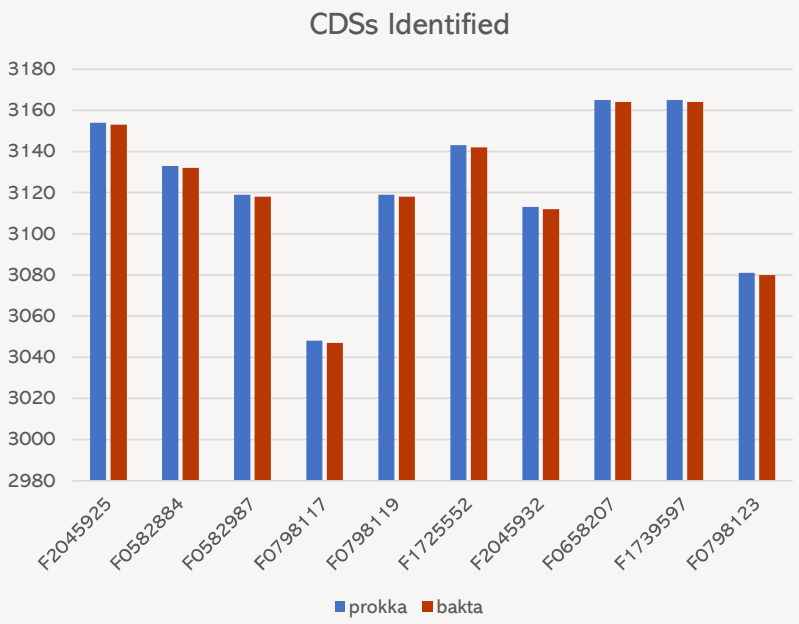
Comparison of RAST and EggNOG-Mapper

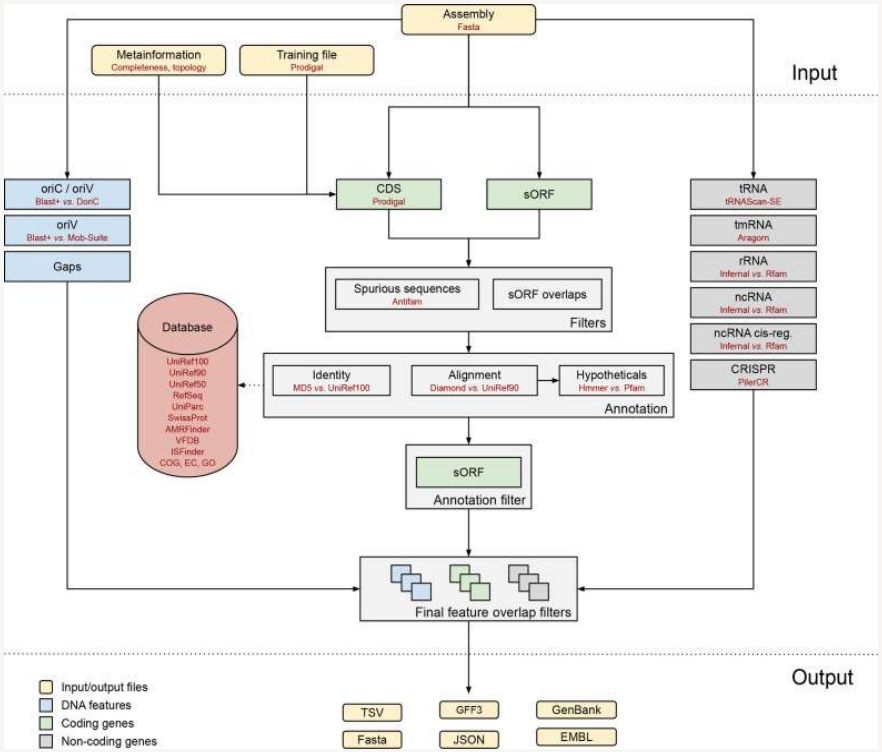
- **RAST:**
- Web-based service.
- Detailed annotation and visualization: Offers comprehensive insights.
- Not considered for final pipeline: Due to limitations in integration and automation.
- **EggNOG Mapper:**
- Web interface used: Command line setup was challenging.
- Average annotation time: 4 minutes per isolate.
- Assigns functions based on orthology, can not always provide the fine-scale functional annotations



Comparison of Prokka and Bakta

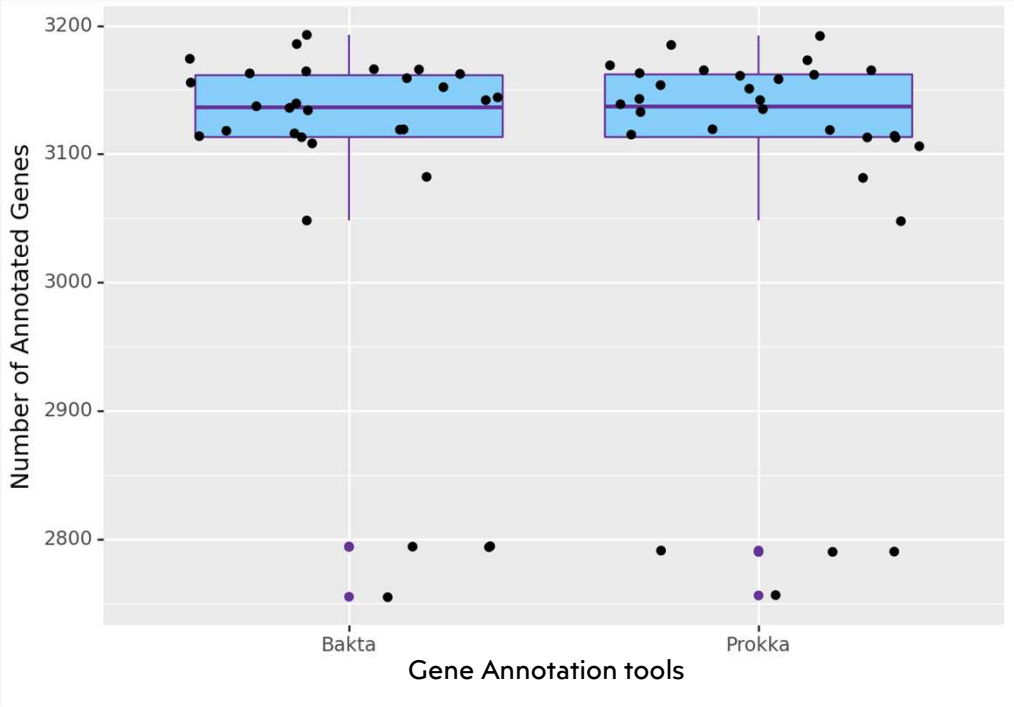
- Prokka provides a straightforward and efficient way to annotate bacterial and archaeal genomes.
- Bakta is a newer command-line software tool for the robust, taxon-independent, thorough annotation of bacterial genomes.
- Regarding tRNAs, tmRNAs, rRNAs and CRISPR arrays, both prokka and bakta predicted equal or comparable numbers of features.
- Bakta is able to distinguish between ncRNA genes and ncRNA regulatory regions.
- Bakta can also predict origins of replication and sORF. It also has the ability to assign GO terms.



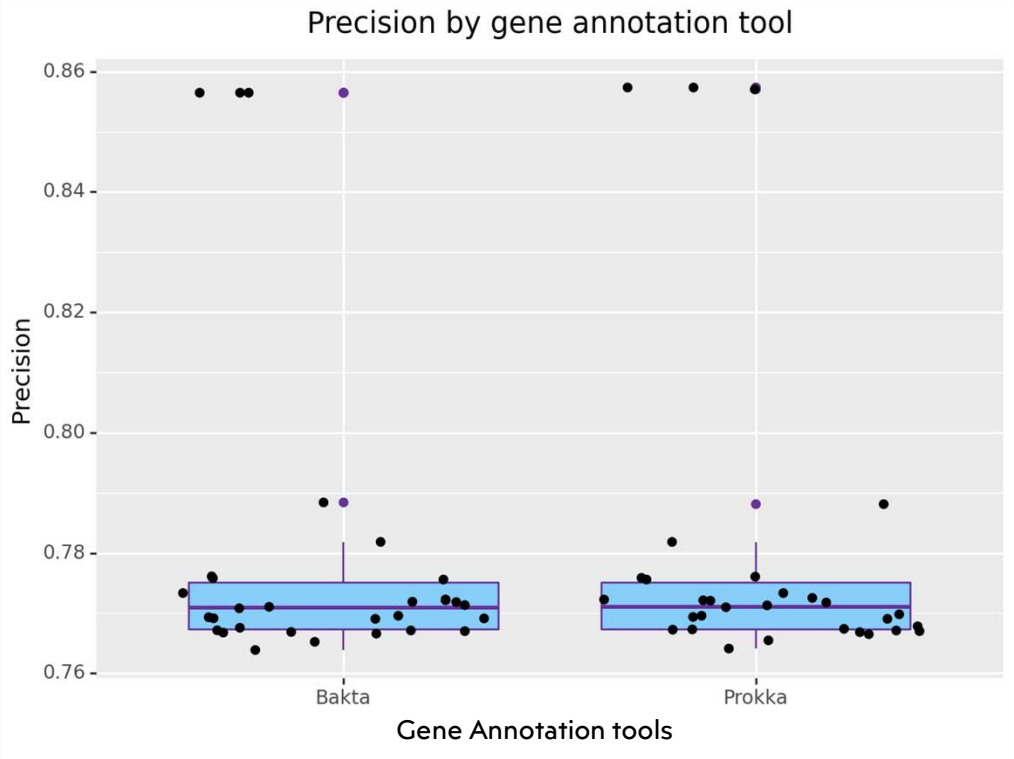

Assembly
Fasta
Metainformation
Completeness, topology
Training file
Prodigal
Input
oriC / oriV
Blast+ vs. DoriC
oriV
Blast+ vs. Mob-Suite
Gaps
CDS
Prodigal
sORF
tRNA
tRNAScan-SE
tmRNA
Aragorn
rRNA
Infernal vs. Rfam
ncRNA
Infernal vs. Rfam
ncRNA cis-reg.
Infernal vs. Rfam
CRISPR
PlerCR
Spurious sequences
Antifam
sORF overlaps
Filters
Database
UniRef100
UniRef90
UniRef50
RefSeq
UniParc
SwissProt
AMRFinder
VFDB
ISFinder
COG, EC, GO
Identity
MD5 vs. UniRef100
Alignment
Diamond vs. UniRef90
Hypotheticals
Hmmer vs. Pfam
Annotation
sORF
Annotation filter
Final feature overlap filters
Output
Input/output files
DNA features
Coding genes
Non-coding genes
TSV
Fasta
GFF3
JSON
GenBank
EMBL

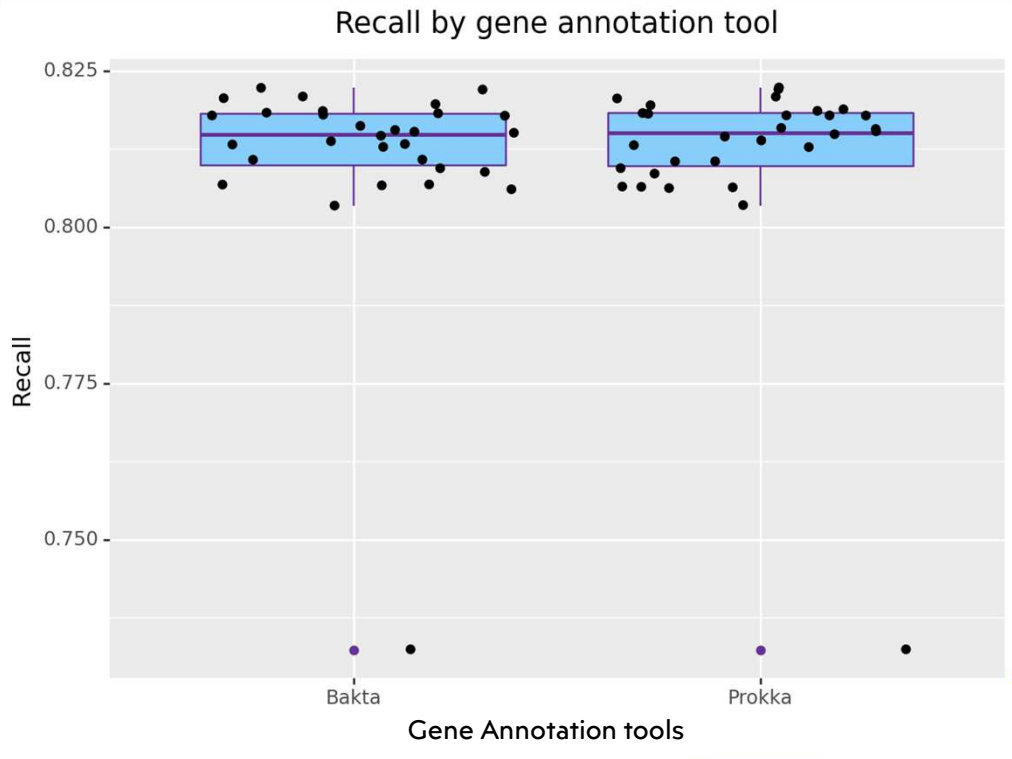

kb
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
700 kb
CDS/sORF
tRNA/tmRNA/rRNA
Arg(tct)
tRNA-Gly(tcc)
tRNA-Ser(cga)
ncRNA
Listeria sRNA rliH
Listeria LhrC
Listeria sRNA rli37
Listeria sRNA rli34
Listeria sRNA rli33
Listeria sRNA rli32
Listeria sF
ncRNA-region
T-box leader
Listeria sRNA rli54
Listeria sRNA rli53
Lysine riboswitch
SAM riboswitch (S box leader)
CRISPR
Gap
oriC/oriV/oriT

Number of Annotated Genes comparison



Precision and Recall of Gene Annotation Tools

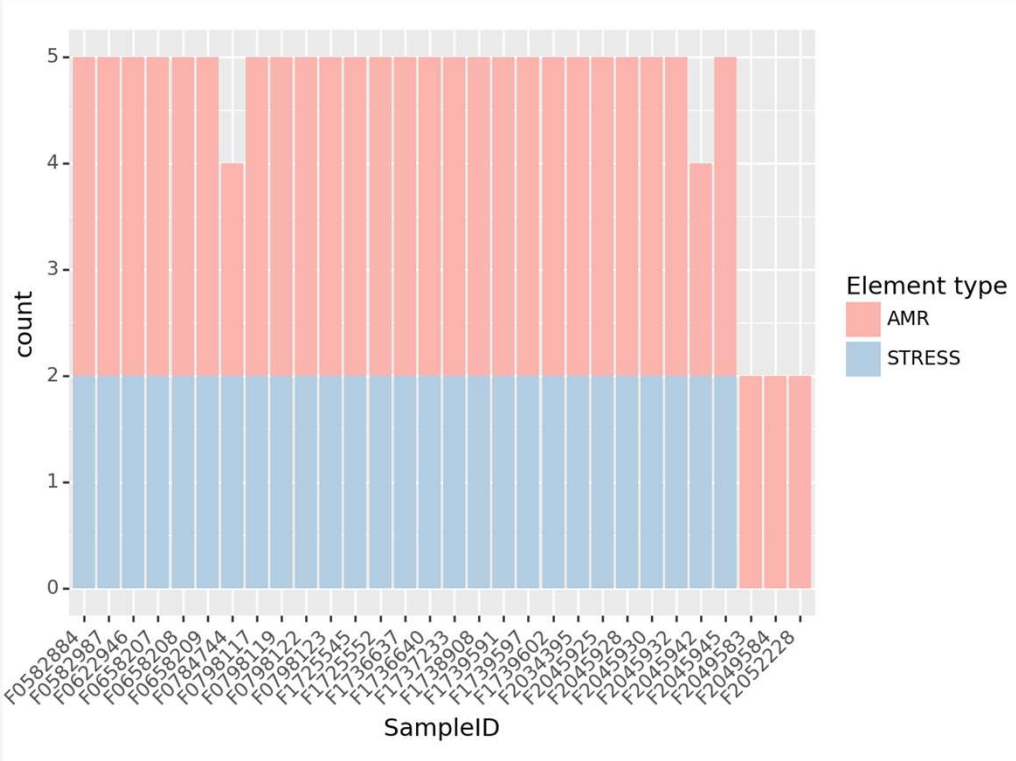


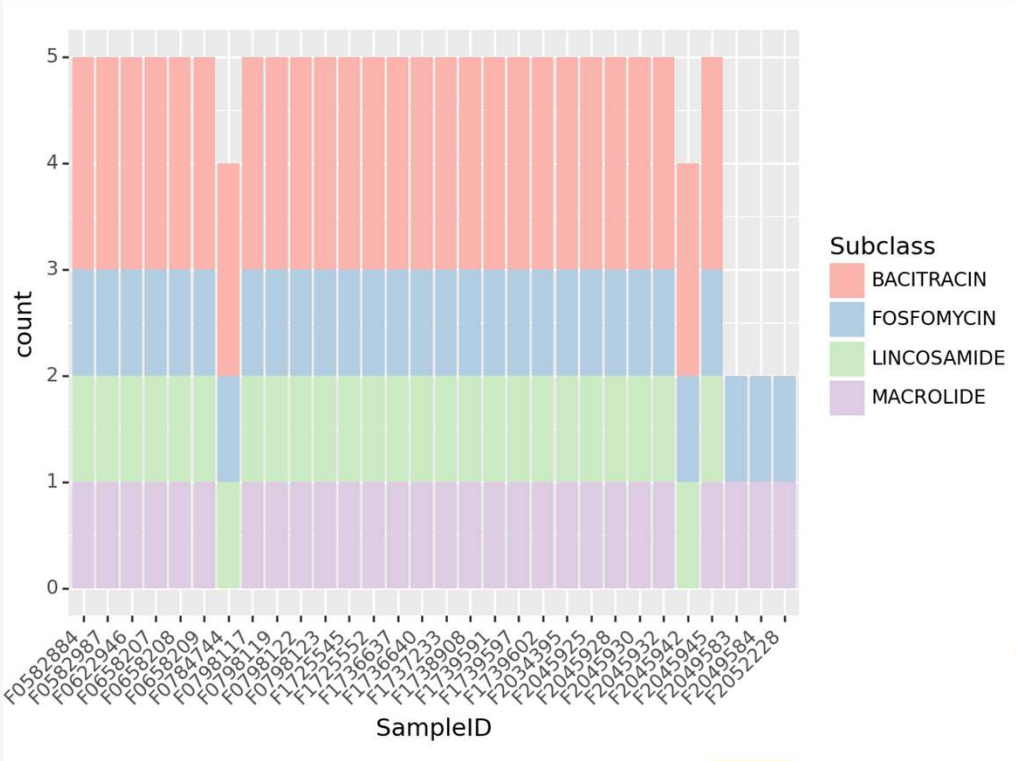


Gene Annotation Tools

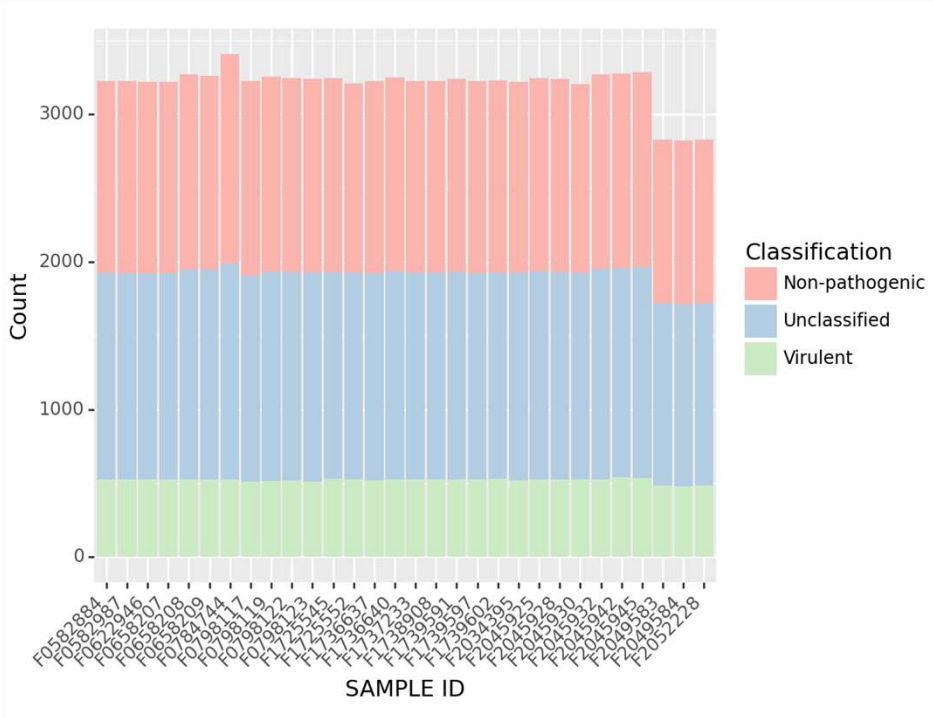
TOOL	Time_taken/isolate	Avg. % Memory used	Database (for CDS)	Usability	Database download required	Database size
Prokka	3 mins 27 secs	3.17	UniProtKB/Swiss-Prot	CLI easy to use	Yes	0.6 GB
Egg-nog Mapper	4 mins	25.65	eggNOG database	CLI interface not used too large sized database; Web interface easy to use	No (web interface used)	NA
RAST	4 mins 16 secs	14.2	relies on its own curated SEED database	Easy to use Webpage	No; Web-based	NA
Bakta	10 mins 35 secs	44.02	RefSeq, UniRef100, and UniParc	CLI interface easy to use; also available as a webpage	Yes	40 GB

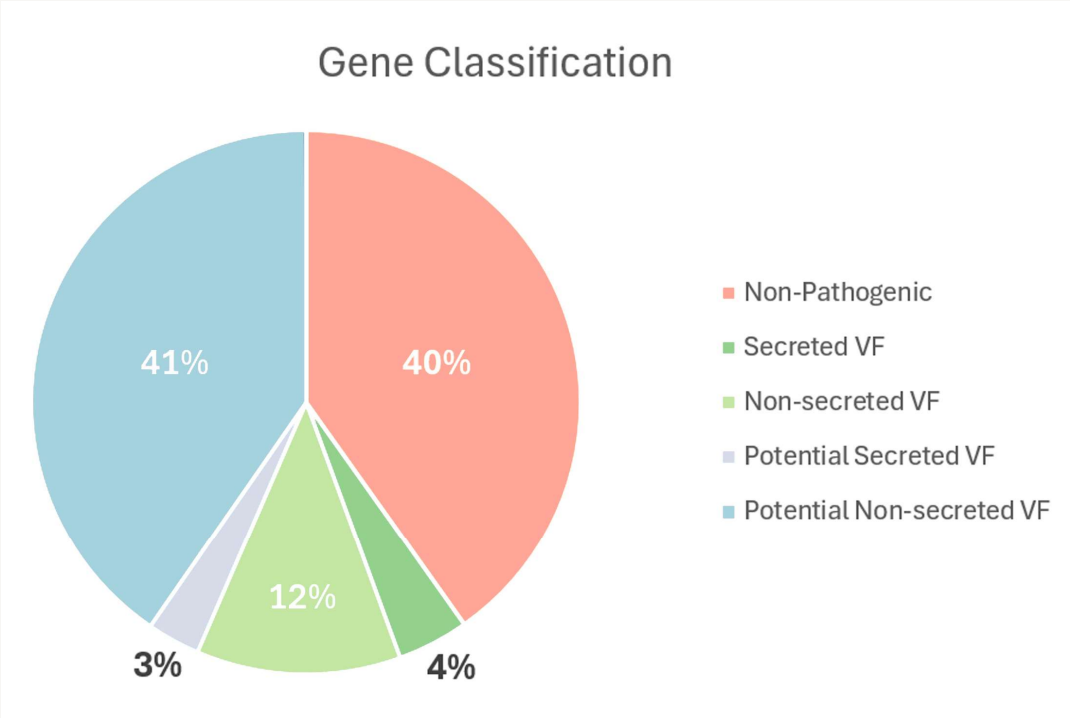
Antimicrobial Resistance: AMRFinder



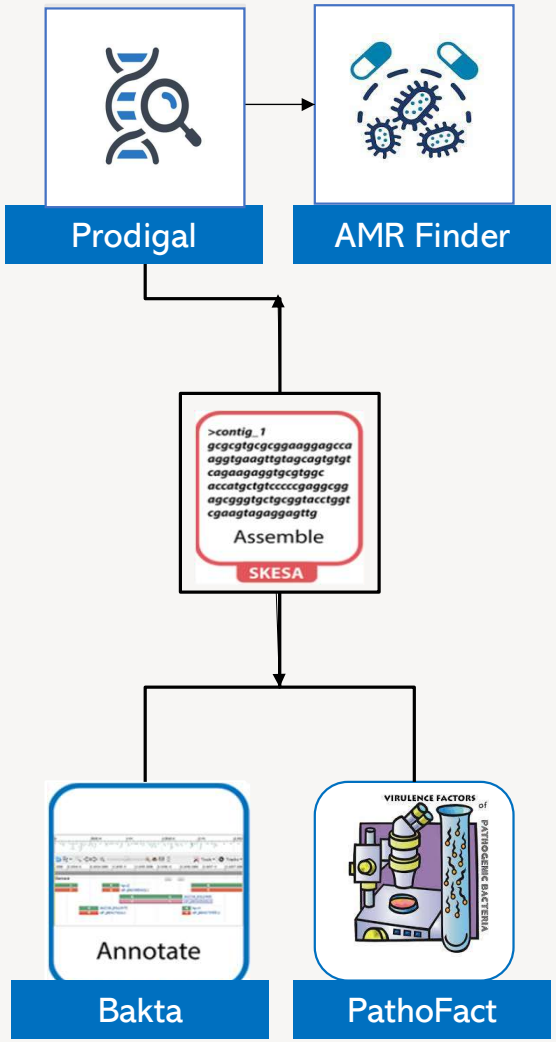


Pathofact


Count
3000
2000
1000
0
SAMPLE ID
Classification
Non-pathogenic
Unclassified
Virulent


Gene Classification
41%
40%
12%
3%
4%
Non-Pathogenic
Secreted VF
Non-secreted VF
Potential Secreted VF
Potential Non-secreted VF

FINAL PIPELINE



Pipeline.sh

Command line usage:
sh pipeline.sh [input_dir] [output_dir]

```
# env for amrfinder
conda create -n amrfinder
conda activate amrfinder
mamba install -c bioconda ncbi-amrfinderplus
conda deactivate

#env for pathofact
git lfs install
git clone -b master --recursive https://git-r3lab.uni.lu/laura.denies/
cd PathoFact
conda env create -f=envs/PathoFact.yaml
cd ..

# commandline usage: sh pipeline.sh [input_dir] [output_dir]
input_dir="$1"
output_dir="$2"

for file in "$input_dir"/*; do
    # isolate name saved as $isolate
    isolate=$(basename "$file")

    # create output dirs for fastp, bbduk, skesa and quast
    mkdir -p "$output_dir"/{prodigal, bakta, amrfinder}

    #gene prediction with prodigal
    conda activate gen_pred
    prodigal \
    -i "$isolate/filtered_contigs.fa" \
    -o "$output_dir/prodigal/${isolate}_gene.coords.gff" \
    -f gff \
    -a "$output_dir/prodigal/${isolate}_protein.translations.faa" \
    -d "$output_dir/prodigal/${isolate}_gene.predictions.fasta" \
    > "$output_dir/prodigal/${isolate}_logfile.log" 2>&1
    conda deactivate

    #annotation with bakta
    conda activate bakta
    bakta --db ~/bakta/db/ \
    "$isolate/filtered_contigs.fa" \
    --threads 8 \
    > "$output_dir/bakta/bakta_output.log" \
    2> "$output_dir/bakta/bakta_error.log"
    conda deactivate

    #amr
    conda activate amrfinder
    amrfinder \
    -p "$output_dir/prodigal/${isolate}_gene.predictions.fasta" \
    > "$output_dir/amrfinder/${isolate}_amr"
    conda deactivate

    #pathofact
    conda activate PathoFact
    sh ./pathofact.sh
    conda deactivate
done
```

Citations

- Hyatt, D., Chen, GL., LoCascio, P.F. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org/10.1186/1471-2105-11-119
- Git: https://github.com/hyattpd/Prodigal
- Schwengers, Oliver, et al. "Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification." *Microbial genomics* 7.11 (2021): 000685.
- Git: https://github.com/oschwengers/bakta
- AMRFinder+: Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728. doi: 10.1038/s41598-021-91456-0. PMID: 34135355; PMCID: PMC8208984.
- Git: https://github.com/ncbi/amr
- PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data: de Nies, L., Lopes, S., Busi, S.B. et al. Microbiome 9, 49 (2021). https://doi.org/10.1186/s40168-020-00993-9
- Git: https://git-r3lab.uni.lu/laura.denies/PathoFact/-/tree/v1.0

Thank you