



# How to run the RGI tool on MetaPro Outputs

Lesson Date	@November 8, 2023
-------------	-------------------

## Background

The Resistance Gene Identifier (RGI) is used to predict antibiotic resistance(s) from protein or nucleotide data based on homology and SNP models. The application uses reference data from the Comprehensive Antibiotic Resistance Database (CARD). For more in-depth documentation, see link here: <https://github.com/arpacard/rgi>

## Running the RGI on SciNet

### 1. Installing RGI tool

```
#go on download node
ssh nia-datamover1

#install through the download node onto my home directory
singularity pull docker://finlaymaguire/rgi:latest

#download reference database on download node
wget https://card.mcmaster.ca/latest/data
tar -xvf data ./card.json #extract

#exit download node
exit
```

### 2. Running the RGI tool (example commands)

```
#set path to RGI tool
RGI="/path_to_RGI_tool"
#set path to reference database
DB="/path/card.json"
#set input path
input="/scratch_path/GA_FINAL_MERGE/final_results/all_proteins.faa"
#set output path
output="/scratch_path/${SAMPLE}_rgiout"

#print help msg to check if RGI is installed and works
singularity run -B $PWD:/home -B $SCRATCH -B $RGI ${RGI}/rgi_latest.sif rgi -h

#load database
singularity run -B $PWD:/home -B $SCRATCH -B $RGI ${RGI}/rgi_latest.sif rgi load --card_json $DB --local

#run rgi
#-t protein flag: if protein FASTA sequences are submitted, RGI skips ORF prediction and uses the protein sequences directly
#takes me 1+ hour to run 400k protein sequences
singularity run -B $PWD:/home -B $SCRATCH -B $RGI ${RGI}/rgi_latest.sif rgi main --input_sequence $input --output_file $output --local

#clear previous database load if you wish, not necessary
singularity run -B $PWD:/home -B $SCRATCH -B $RGI ${RGI}/rgi_latest.sif rgi clean --local
```

### 3. RGI will output 2 files (.JSON and .TXT). The text file will have the following columns:

- ORF\_ID → protein header from your original all\_proteins.faa file
- Contig → NA
- Start → NA

- Stop → NA
- Orientation → NA
- Cut\_Off → Strict or Perfect
- Pass\_Bitscore
- Best\_Hit\_Bitscore
- Best\_Hit\_ARO
- Best\_Identities
- ARO
- Model\_type
- SNPs\_in\_Best\_Hit\_ARO → NA
- Other\_SNPs → NA
- Drug Class → e.g. glycopeptide antibiotic
- Resistance Mechanism → e.g. antibiotic target alteration
- AMR Gene Family → e.g. glycopeptide resistance gene cluster; vanT
- Predicted\_DNA → NA
- Predicted\_Protein → protein sequence
- CARD\_Protein\_Sequence → protein sequence
- Percentage Length of Reference Sequence → percentage number
- ID
- Model\_ID
- Nudged → NA
- Note → NA