Integrated Rapid Infectious Disease Analysis: A comprehensive platform for public health bioinformatics and AMR surveillance using genomic data

F Maguire[1], B Alcock[2], AR Raphenya[2], B Jia[3], EJ Griffiths[3], TC Matthews[4], J Adam[4], A Petkau[4], GL Winsor[3], **IRIDA Consortium,** RG Beiko[1], FSL Brinkman[3], WWL Hsiao[5], G Van Domselaar[4], AG McArthur[2]

[1] Faculty of Computer Science, Dalhousie University, NS. [2] Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON. [3] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC. [4] National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB. [5] British Columbia Centre for Disease Control Public Health Laboratory, Vancouver, BC.

IRIDA

CARD

Abstract

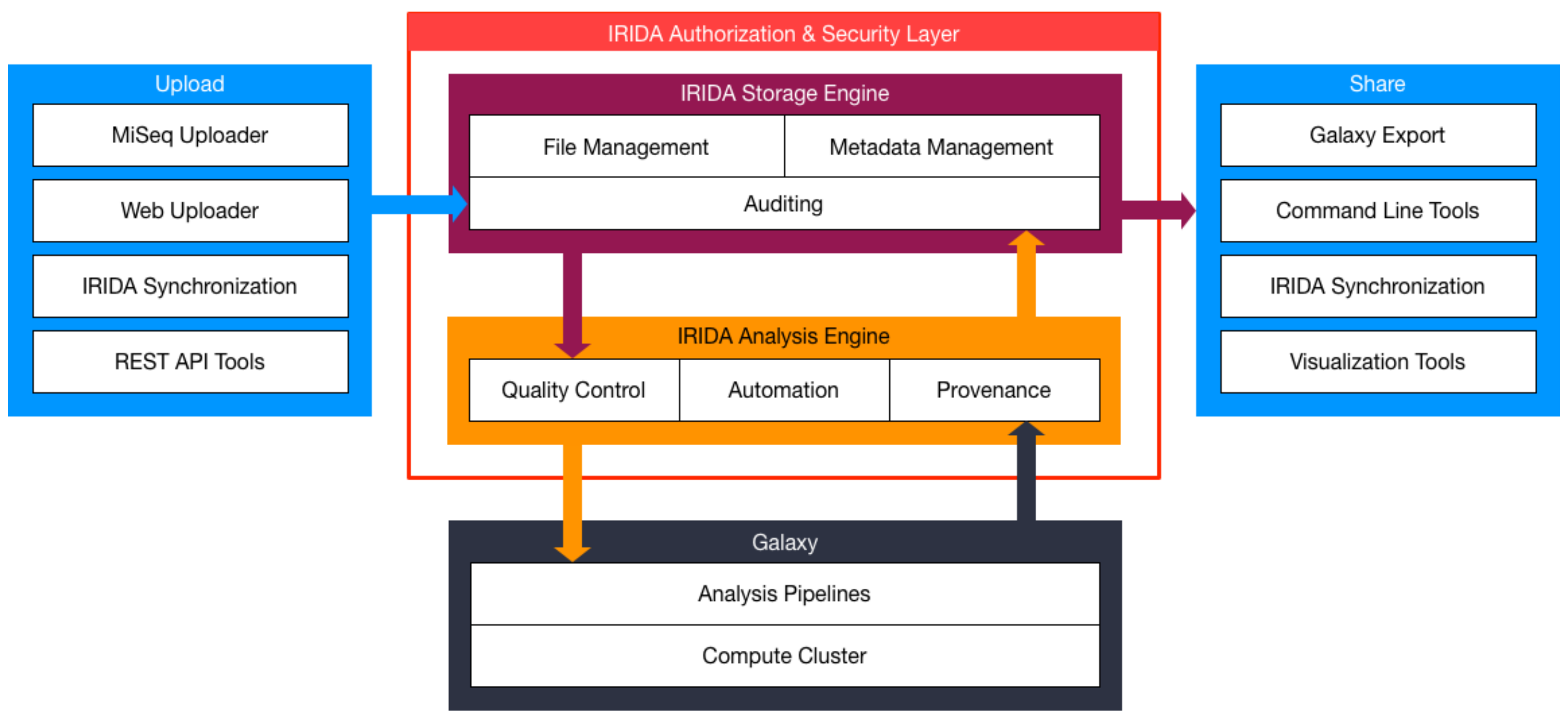
Antimicrobial resistance (AMR) is a major and growing global health crisis. Development of comprehensive AMR surveillance and rapid diagnostics capabilities with genomic and metagenomic data is fundamental to improving understanding of AMR transmission and stewardship of our limited and dwindling antibiotic arsenal. Unfortunately, effectively using genomic data for public health applications such as this is impeded by a scarcity of easy-to-use automated and semi-automated pipelines and robust data sharing tools. To address these issues, we developed the Integrated Rapid Infectious Disease Analysis (IRIDA) platform (*irida.ca*), a user-friendly, decentralized, open-source bioinformatics and analytical web platform to support multi-jurisdictional infectious disease outbreak investigations using genomic sequencing data. IRIDA incorporates quality control, genomics assembly and annotation, *in silico* serotyping, multi-locus sequence typing, and outbreak phylogenetics. This platform also incorporates direct visualization of results to assist with hypothesis generation in epidemiological investigations. IRIDA allows users to perform secure and local analysis, while also enabling controlled data sharing with trusted partners and public sequence repositories.

Additionally, via built-in Galaxy support it integrates high-quality curated AMR data from the Comprehensive Antibiotic Resistance Database (CARD; *card.mcmaster.ca*) with state-of-the-art AMR detection methods for genomes and metagenomes such as the Resistance Gene Identifier (RGI) and, in the future, AMRtime. CARD is an ontology-centric database and bioinformatics resource on the molecular and genetic basis of antimicrobial resistance (AMR) that has recently been greatly expanded with Resistomes and Variants surveillance data for over 81,000 isolates, plus extensions to its underlying Antibiotic Resistance Ontology. These improvements have been leveraged to extend RGI to enable rapid read-mapping for detection of AMR genes from metagenomic data, with associated *k*-mer classifiers for pathogen-of-origin prediction. They have also served as the basis for the development of AMRtime, a highly sensitive and accurate machine-learning based metagenomic AMR classification tool.

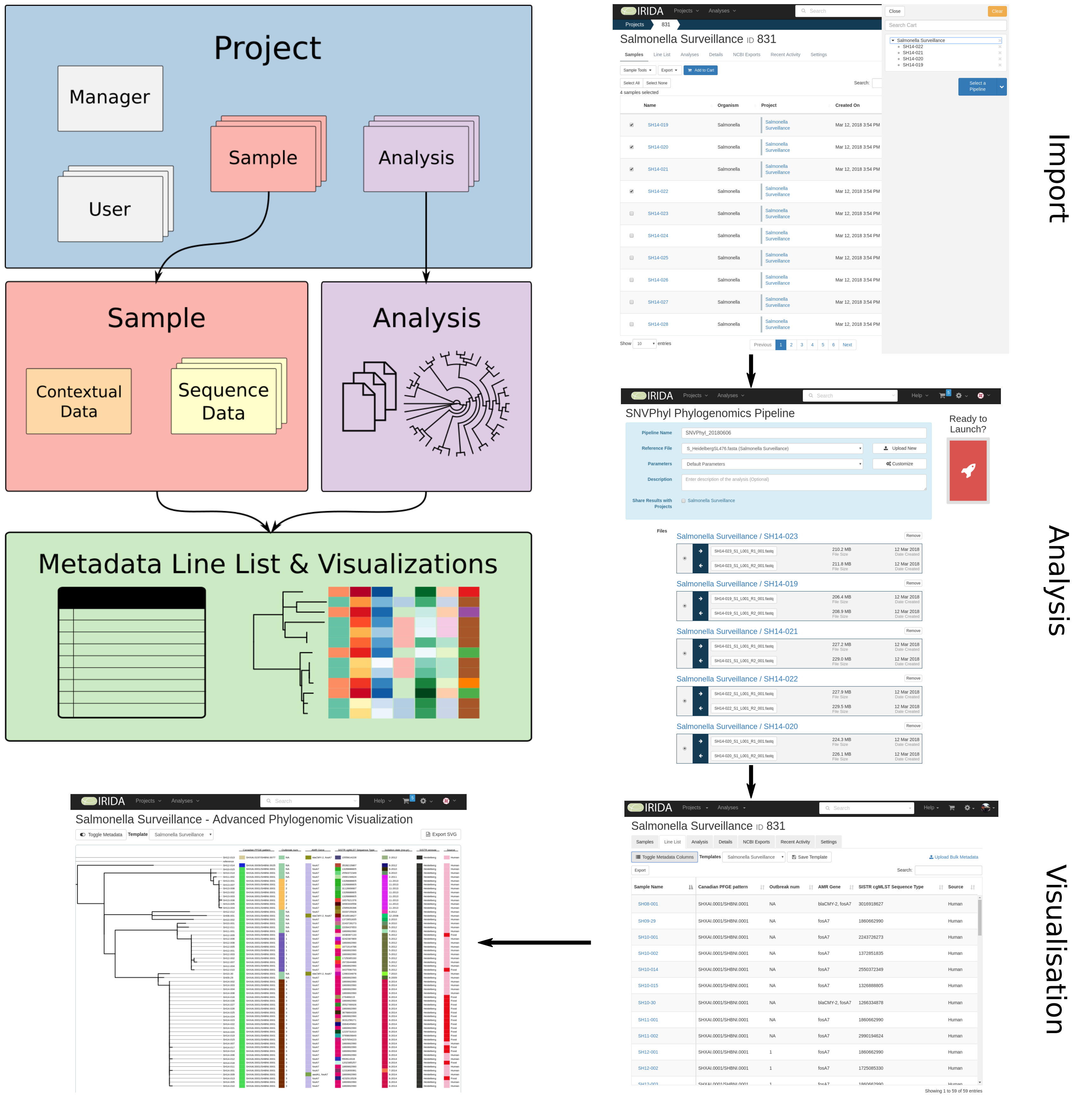
This extensive set of genomic epidemiology focused analysis methods and robust data management tools are completely open-source and available to all. Currently, IRIDA is being used by the Public Health Agency of Canada as a primary tool for infectious disease outbreak investigations and is used/installed by other government and academic groups internationally across four continents (including the US, UK, South Africa, and Singapore). IRIDA is freely available at *github.com/phac-nml/irida* and *www.irida.ca*

IRIDA

IRIDA is a web-based application that can be locally deployed as a stand-alone platform, designed to provide public health, clinical microbiology, and food regulatory authorities with the capability to incorporate next-generation sequencing into their surveillance, diagnostics, reference typing, and research programs. IRIDA's core functionality encompasses four main areas: 1) data management, 2) user management and data sharing, 3) data analysis, and 4) reporting and visualization.



IRIDA performs all aspects of sequence data and metadata management, including data import, export, storage, organization, tracking, and sharing. IRIDA's data structure is rooted by a Project. Projects contain a collection of Samples, which contain a collection of Sequence Data. Projects also contain Analyses generated from the sequence data within that project. Metadata can be associated with Projects, Samples, Sequence Data, and Analyses. This project-centric model incorporates robust auditing, security and sharing tools within the same installation or across different installations/institutions. IRIDA's data structure is modeled after, and fully compatible with, the BioProject data structure used by INSDC databases (NCBI, EMBL-EBI and DDBJ). IRIDA's data structure facilitates deposition of reads, assemblies, and annotated contigs to the various public sequence repositories.



Salmonella Surveillance ID 831

Samples | Line List | Analyses | Details | NCBI Exports | Recent Activity | Settings

Sample Tools | Export | Add to Cart

Select All | Select None

4 samples selected

Name	Organism	Project	Created On
SH14-019	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-020	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-021	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-022	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-023	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-024	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-025	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-026	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-027	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
SH14-028	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM

Show 10 entries

Salmonella Surveillance: SH14-022, SH14-021, SH14-020, SH14-019

SNVPhyl Phylogenomics Pipeline

Pipeline Name: SNVPhyl_20180606
Reference File: S_HeidelbergSL476.fasta (Salmonella Surveillance) | Upload New
Parameters: Default Parameters | Customize
Description: Enter description of the analysis (Optional)
Share Results with Projects: Salmonella Surveillance

Files

Salmonella Surveillance / SH14-023
SH14-023_S1_L001_R1_001.fastq 210.2 MB File Size 12 Mar 2018 Date Created
SH14-023_S1_L001_R2_001.fastq 211.8 MB File Size 12 Mar 2018 Date Created

Salmonella Surveillance / SH14-019
SH14-019_S1_L001_R1_001.fastq 206.4 MB File Size 12 Mar 2018 Date Created
SH14-019_S1_L001_R2_001.fastq 208.9 MB File Size 12 Mar 2018 Date Created

Salmonella Surveillance / SH14-021
SH14-021_S1_L001_R1_001.fastq 227.2 MB File Size 12 Mar 2018 Date Created
SH14-021_S1_L001_R2_001.fastq 229.0 MB File Size 12 Mar 2018 Date Created

Salmonella Surveillance / SH14-022
SH14-022_S1_L001_R1_001.fastq 227.9 MB File Size 12 Mar 2018 Date Created
SH14-022_S1_L001_R2_001.fastq 229.5 MB File Size 12 Mar 2018 Date Created

Salmonella Surveillance / SH14-020
SH14-020_S1_L001_R1_001.fastq 224.3 MB File Size 12 Mar 2018 Date Created
SH14-020_S1_L001_R2_001.fastq 226.1 MB File Size 12 Mar 2018 Date Created

Salmonella Surveillance ID 831

Samples | Line List | Analysis | Details | NCBI Exports | Recent Activity | Settings

Toggle Metadata Columns | Templates: Salmonella Surveillance | Save Template | Upload Bulk Metadata

Export

Sample Name	Canadian PFGE pattern	Outbreak num	AMR Gene	SISTR cgMLST Sequence Type	Source
SH08-001	SHXAI.0001/SHBNI.0001	NA	blaCMY-2, fosA7	3016918627	Human
SH09-29	SHXAI.0001/SHBNI.0001	NA	fosA7	1860662990	Human
SH10-001	SHXAI.0001/SHBNI.0001	NA	fosA7	2243726273	Human
SH10-002	SHXAI.0001/SHBNI.0001	NA	fosA7	1372851835	Human
SH10-014	SHXAI.0001/SHBNI.0001	NA	fosA7	2550372349	Human
SH10-015	SHXAI.0001/SHBNI.0001	NA	fosA7	1326888805	Human
SH10-30	SHXAI.0001/SHBNI.0001	NA	blaCMY-2, fosA7	1266334878	Human
SH11-001	SHXAI.0001/SHBNI.0001	NA	fosA7	1860662990	Human
SH11-002	SHXAI.0001/SHBNI.0001	NA	fosA7	2990194624	Human
SH12-001	SHXAI.0001/SHBNI.0001	1	fosA7	1860662990	Human
SH12-002	SHXAI.0001/SHBNI.0001	1	fosA7	1725085330	Human

Showing 1 to 59 of 59 entries

Salmonella Surveillance - Advanced Phylogenomic Visualization

IRIDA provides the ability to automatically analyze genomic sequence data and metadata using its collection of analysis pipelines. These currently include quality control, genome assembly, annotation, genome distance calculations, single nucleotide variant detection, phylogenomic inference, serotype prediction, and multi-locus sequence typing.

IRIDA also provides support for visualizing metadata and analysis results with metadata viewable in a standardised interactive tabular "line-list" format. Similarly, generated phylogenies can be viewed and manipulated in browser using a modified version of PhyloCanvas (*phylocanvas.org*). This includes optional annotation of useful metadata directly only the phylogeny.

Via it's tight integration with Galaxy, IRIDA also supports state-of-the art analysis and annotation of AMR genes from metagenomic data. This is mainly based on the Comprehensive Antibiotic Resistance Database and it's associated Resistance Gene Identifier (RGI) and the RGI inspired AMRtime tool.

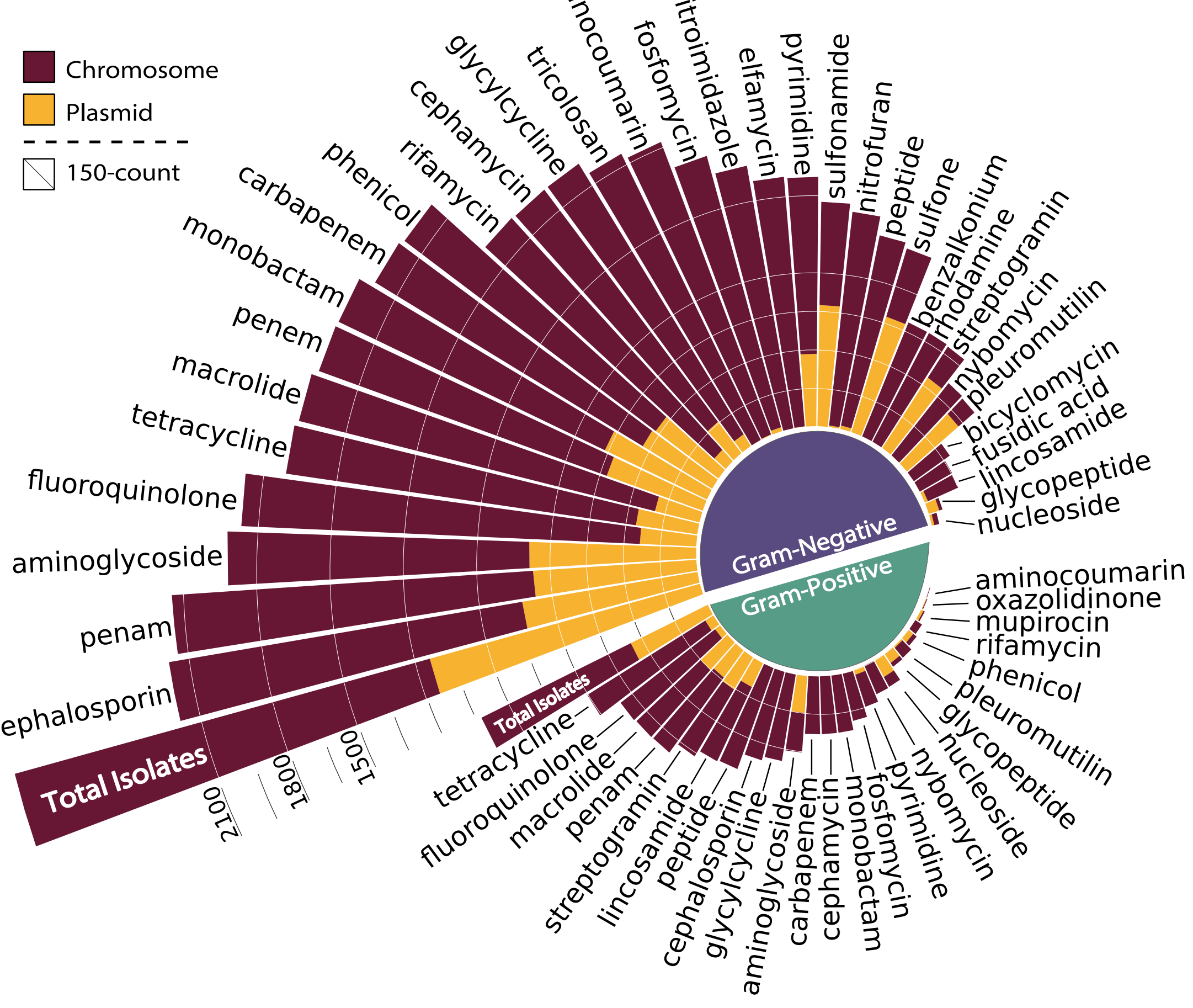
CARD

CARD (card.mcmaster.ca) is an open-source, expertly curated, regularly updated, ontology-centric database that catalogues the molecular determinants underpinning AMR. Automated literature searches via the CARD*Shark PubMED text mining algorithms are performed at regular intervals and manually reviewed by domain experts for inclusion into the database. Uniquely, CARD is built upon the ARO (Antibiotic Resistance Ontology) which contains terms related to antimicrobial resistance genes and mechanisms, as well as antibiotics and their targets. This ontology and associated classification tags allows rapid classification and summary resistome predictions.
Within the CARD are a range of AMR determinants including gene and 16S rRNA SNPs associated with resistance phenotypes, variants associated with protein over-expression, and meta-models describing multi-component systems such as efflux pumps and glycopeptide resistance clusters.
CARD is freely available at *card.mcaster.ca* and openly curated via *github.com/arpcard/amr_curation*

CARD Resistomes & Variants

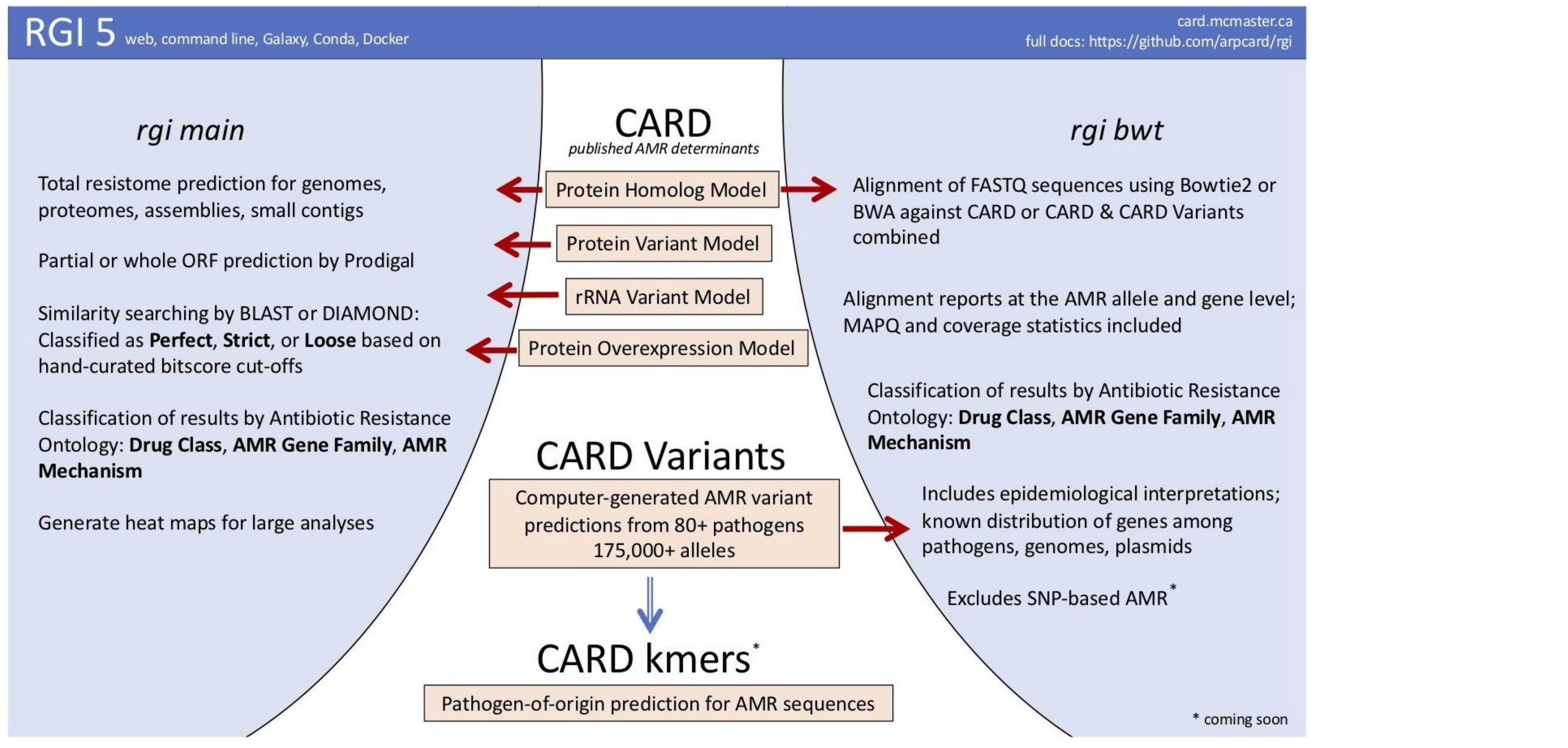
For inclusion in CARD, an AMR gene must have a nucleotide sequence available in GenBank and appear in a published peer-reviewed journal with a demonstrably elevated minimum inhibitory concentration. While this paradigm promotes a verifiable and concise database, it omits valuable information pertaining to AMR gene prevalence, resistome prediction, and unpublished AMR sequence variants, including those predicted in silico.

To overcome this limitation, CARD has recently been expanded to include predicted AMR gene variants from plasmid and genomic assemblies from a wider range of pathogens. These include 75,446 genomes, plasmid and/or contig assemblies from 79 ESKAPE and WHO priority pathogens. Using the ARO these can then be summarised to identify the relative distribution of resistance to certain antibiotic classes in mobile genetic elements such as plasmids (below). These data allow us to better understand genetic AMR diversity, monitor mobile AMR elements, detect AMR determinant co-occurrence, provide better reference data for metagenomic analyses, and improve AMR phenotype prediction. CARD Resistomes & Variants are open-access and periodically updated via *card.mcmaster.ca/download*



Resistance Gene Identifier

The Resistance Gene Identifier (RGI; card.mcmaster.ca/analyze/rgi or github.com/arpcard/rgi) was developed to predict AMR determinants from genomic data and has been recently expanded to support metagenomic datasets. RGI can predict homolog and variant related resistance in both protein and rRNA sequences from either metagenomic/genomic assemblies or raw sequencing reads. For genome sequences and assemblies, RGI uses a combination of prodigal gene prediction, BLAST/DIAMOND against the CARD database, and mutation mapping to predict resistance determinants. These results are classified into Perfect (exact match to reference), Strict (likely functional variants with a match above a manually curated gene specific bit-score), and Loose hits (putative distant homologs or new emergent threats).


RGI 5 web, command line, Galaxy, Conda, Docker
card.mcmaster.ca
full docs: https://github.com/arpcard/rgi
rgi main
CARD
published AMR determinants
rgi bwt
Protein Homolog Model
Protein Variant Model
rRNA Variant Model
Protein Overexpression Model
CARD Variants
CARD kmers*
* coming soon

rgi main
- Total resistome prediction for genomes, proteomes, assemblies, small contigs
- Partial or whole ORF prediction by Prodigal
- Similarity searching by BLAST or DIAMOND: Classified as **Perfect**, **Strict**, or **Loose** based on hand-curated bitscore cut-offs
- Classification of results by Antibiotic Resistance Ontology: **Drug Class**, **AMR Gene Family**, **AMR Mechanism**
- Generate heat maps for large analyses

rgi bwt
- Alignment of FASTQ sequences using Bowtie2 or BWA against CARD or CARD & CARD Variants combined
- Alignment reports at the AMR allele and gene level; MAPQ and coverage statistics included
- Classification of results by Antibiotic Resistance Ontology: **Drug Class**, **AMR Gene Family**, **AMR Mechanism**
- Includes epidemiological interpretations; known distribution of genes among pathogens, genomes, plasmids
- Excludes SNP-based AMR*

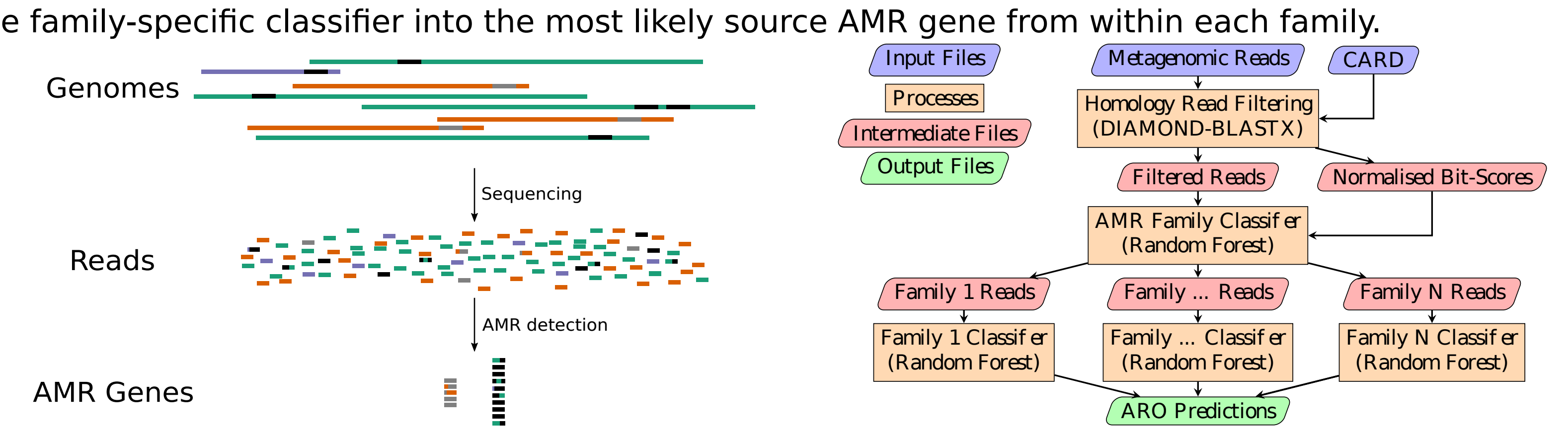
CARD Variants: Computer-generated AMR variant predictions from 80+ pathogens 175,000+ alleles

CARD kmers*: Pathogen-of-origin prediction for AMR sequences

RGI for raw reads uses the Burrows-Wheeler Transform (through BWA or bowtie 2 software) to align reads to CARD reference sequences and *in silico* predicted allelic variants from CARD's Resistomes & Variants dataset. In addition, RGI uses AMR specific *k*-mers mined from CARD Resistomes & Variants to predict pathogen-of-origin for detected AMR alleles and mobile AMR genes. Lastly, this tool includes additional algorithms to support analysis of AMR in metagenomic samples via bait capture technologies by helping with probe design and validation.
RGI is open-source and can be accessed via either a web-interface (*card.mcmaster.ca/analyze/rgi*) or installed locally from source (*github.com/arpcard/rgi*), conda, or Docker (*quay.io/repository/biocontainers/rgi*).

AMRtime

K-mer and read-mapping methods, while rapid and controlled in terms of false positives are relatively conservative and liable to a greater rate of missed AMR-related reads than other homology search methods.
AMRtime is an alternative complementary approach for accurately profiling AMR related reads present in metagenomic data that provides much greater sensitivity at the expense of increased complexity. AMRtime achieves this by first filtering the raw metagenomic dataset using DIAMOND-BLASTX searches against CARD with an optimised bitscore cut-off to remove non-AMR reads. The filtered reads are then more precisely classified using a series of hierarchical series of Random Forest classifiers using features derived from the filtering stage. Reads are classified into one of 271 AMR gene families via a top-level classifier. Each read is then classified by a separate family-specific classifier into the most likely source AMR gene from within each family.



AMRtime is open source (*github.com/beiko-lab/AMRtime*) with Galaxy bindings for IRIDA and wider tool-sheds planned for the near future.

Interested in this work?

Interested in bioinformatics, AMR, and machine learning? We are hiring: arete-amr.ca

Public Health Agency of Canada | Agence de santé publique du Canada
CIHR IRSC Canadian Institutes of Health Research | Instituts de recherche en santé du Canada
GenomeCanada
BC Centre for Disease Control
SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD
McMaster University
DALHOUSIE UNIVERSITY Inspiring Minds