

CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database

Brian P. Alcock^{1,2,3}, William Huynh^{1,2,3}, Romeo Chalil^{1,2,3}, Keaton W. Smith^{1,2,3}, Amogelang R. Raphenya^{1,2,3}, Mateusz A. Włodarski^{1,2,3}, Arman Edalatmand^{1,2,3}, Aaron Petkau^{4,5}, Sohaib A. Syed^{1,2,3}, Kara K. Tsang^{1,2,3}, Sheridan J.C. Baker^{1,2,3}, Mugdha Dave^{1,2,3}, Madeline C. McCarthy^{1,2,3}, Karyn M. Mukiri^{1,2,3}, Jalees A. Nasir^{1,2,3}, Bahar Golbon^{1,2,3}, Hamna Imtiaz^{1,2,3}, Xingjian Jiang^{1,2,3}, Komal Kaur^{1,2,3}, Megan Kwong^{1,2,3}, Zi Cheng Liang^{1,2,3}, Keyu C. Niu^{1,2,3}, Prabakar Shan^{1,2,3}, Jasmine Y.J. Yang^{1,2,3}, Kristen L. Gray⁶, Gemma R. Hoad⁷, Baofeng Jia⁶, Timsy Bhando^{1,2,3}, Lindsey A. Carfrae^{1,2,3}, Maya A. Farha^{1,2,3}, Shawn French^{1,2,3}, Rodion Gordzevich^{1,2,3}, Kenneth Rachwalski^{ID 1,2,3}, Megan M. Tu^{1,2,3}, Emily Bordeleau^{1,2,3}, Damion Dooley⁸, Emma Griffiths^{ID 8}, Haley L. Zubyk^{1,2,3}, Eric D. Brown^{ID 1,2,3}, Finlay Maguire^{9,10,11}, Robert G. Beiko^{9,10}, William W.L. Hsiao^{6,8}, Fiona S.L. Brinkman⁶, Gary Van Domselaar^{ID 5,12} and Andrew G. McArthur^{ID 1,2,3,*}

¹David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada, ²Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada, ³Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada, ⁴Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada, ⁵National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada, ⁶Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, ⁷Research Computing Group, Simon Fraser University, Burnaby, British Columbia, Canada, ⁸Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada, ⁹Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada, ¹⁰Institute for Comparative Genomics, Dalhousie University, Halifax, Nova Scotia, Canada, ¹¹Department of Community Health & Epidemiology, Dalhousie University, Halifax, Nova Scotia, Canada and ¹²Department of Medical Microbiology and Infectious Diseases, Max Rady College of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada

Received September 15, 2022; Revised October 03, 2022; Editorial Decision October 04, 2022; Accepted October 11, 2022

ABSTRACT

The Comprehensive Antibiotic Resistance Database (CARD; card.mcmaster.ca) combines the Antibiotic Resistance Ontology (ARO) with curated AMR gene (ARG) sequences and resistance-conferring mutations to provide an informatics framework for annotation and interpretation of resistomes. As of version 3.2.4, CARD encompasses 6627 ontology terms, 5010 reference sequences, 1933 mutations, 3004 publications, and 5057 AMR detection models that can be used by the accompanying Resistance Gene Identifier (RGI) software to annotate ge-

nomic or metagenomic sequences. Focused curation enhancements since 2020 include expanded β-lactamase curation, incorporation of likelihood-based AMR mutations for *Mycobacterium tuberculosis*, addition of disinfectants and antiseptics plus their associated ARGs, and systematic curation of resistance-modifying agents. This expanded curation includes 180 new AMR gene families, 15 new drug classes, 1 new resistance mechanism, and two new ontological relationships: *evolutionary_variant_of* and *is_small_molecule_inhibitor*. In silico prediction of resistomes and prevalence statis-

*To whom correspondence should be addressed. Tel: +1 905 525 9140 (Ext 21663); Email: mcarthua@mcmaster.ca
Present address: Kara K. Tsang, Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom.

tics of ARGs has been expanded to 377 pathogens, 21,079 chromosomes, 2,662 genomic islands, 41,828 plasmids and 155,606 whole-genome shotgun assemblies, resulting in collation of 322,710 unique ARG allele sequences. New features include the CARD:Live collection of community submitted isolate resistome data and the introduction of standardized 15 character CARD Short Names for ARGs to support machine learning efforts.

INTRODUCTION

In the decade since the initial publication of the Comprehensive Antibiotic Resistance Database (CARD) (1), the public health threat of antimicrobial resistance (AMR) has grown significantly, prompting a collective global initiative to combat the spread of AMR through collaborative research, molecular and phenotypic surveillance, and antimicrobial stewardship (2–5). However, the complexity and global burden of the AMR crisis (6) continues to expand with the continuous discovery of novel antibiotic resistance genes (ARGs), including those using novel mechanisms such as MCR-mediated colistin resistance via change in the charge of the cell membrane (7) or ejection of rifamycin from its target RNA polymerase by helicase-like protein HelR (8), as well as contributing factors from environmental, veterinary, and food-related sources (9,10). Furthermore, emerging evidence details a complex relationship between AMR and the COVID-19 pandemic, in part due to prolonged hospitalizations, relaxed antibiotic stewardship in overwhelmed hospital settings, and increased household and topical biocide use (11,12).

As previously described (13), CARD is an ontology-centric knowledgebase on the molecular determinants of antibiotic resistance. By integrating a detailed controlled vocabulary, the Antibiotic Resistance Ontology (ARO), with ARG molecular sequences and mutation data, CARD is able to provide both reference set and software tools for guiding AMR research, particularly for ARG annotation and discovery from genomic and metagenomic data. One such analytical framework is the Resistance Gene Identifier (RGI), CARD's algorithm for computational AMR genotype and phenotype prediction from genomic data using the bioinformatics detection models curated in CARD. To ensure accuracy, CARD is manually curated by a team of biocurators and researchers, following a core curation paradigm. Additionally, CARD works towards harmonization with other AMR databases, specifically including those provided by the National Center for Biotechnology Information (NCBI), such as the Pathogen Detection Reference Gene Catalog (<https://www.ncbi.nlm.nih.gov/pathogens/>).

Here we provide an update on the state of CARD and the ARO, with a review of the available data and updates since our 2020 publication. We discuss in detail (i) general and specific improvements to the ARO and curation of AMR in CARD, (ii) improved integration and harmonization with existing resources for specific AMR gene families (such as β-lactamase resistance genes) and pathogens (such as *Mycobacterium tuberculosis*), (iii) curation of resistance modifying agents, (iv) gene name standardization

in support of machine learning efforts, (v) improvements to CARD's Model Ontology in support of bioinformatic annotation of genomes or metagenomes, (vi) description of CARD's curation paradigms, protocols and quality assurance algorithms and (vii) expansion of the previously introduced CARD Resistomes & Variants (CARD-R) resource.

EXPANSION OF CARD CURATION

Current state of CARD and the ARO

CARD's primary curation paradigm continues as previously described: 'to be included in CARD an AMR determinant must be described in a peer-reviewed scientific publication, with its DNA sequence available in GenBank, including clear experimental evidence of elevated minimum inhibitory concentration (MIC) over controls. AMR genes predicted by *in silico* methods, but not experimentally characterized, are not included in CARD's primary curation' (13). Data in CARD is organized using four core ontologies: the ARO describing all aspects of AMR, the CARD Model Ontology (MO) defining bioinformatics parameters for annotating ARGs in genomic data, the CARD Relations Ontology (RO) for association of terms within ontologies, and NCBITaxon for tracking the source organisms and taxonomic distribution of ARG sequences. The ARO is the primary ontology for organization and interpretation of data within CARD and it contains seven major branches: antibiotic molecule (ARO:1000003), mechanism of antibiotic resistance (ARO:1000002), determinant of antibiotic resistance (ARO:3000000), resistance-modifying agents (ARO:0000076), antibiotic target (ARO:3000708), antibiotic biosynthesis (ARO:3000082), and component of AMR genotypic or phenotypic terminology (ARO:3000045). Each curated ARG must include an ontological connection to the antibiotic molecule, mechanism of antibiotic resistance, and determinant of antibiotic resistance branches of the ARO, as outlined below. Notably, since our last publication the ARO terminology branch has been updated to include terms for phenotypic testing standards for the National Antimicrobial Resistance Monitoring System (NARMS) and National Committee for Clinical Laboratory Standards (NCCLS) to support clinical data harmonization efforts. Connections among terms within the ARO are supported by a set of relation terms from CARD's Relations Ontology, some of which are generic (e.g. *is_a* or *part_of*) while others reflect key aspects of antimicrobial resistance (e.g. *confers_resistance_to_antibiotic*) (Table 1). With use of these relations, terms within the ARO form a knowledge network for AMR and combined with curation rules support computational approaches for annotation of genome sequences, interpretation of these annotations, and development of powerful curation quality control algorithms, as outlined below.

In 2020, CARD v3.0.3 included 4336 terms from the ARO, up from 3567 in our 2017 publication (14), and covering 2923 AMR determinants encoded within CARD detection models (13). As of August 2022, CARD v3.2.4 has expanded to 6627 ARO terms and now includes 5057 AMR detection models covering 5010 reference sequences for resistance determinants (including 1933 curated resistance-associated variant mutations). While the ARO provides a

Table 1. Ontological relationships in use by CARD within the Antibiotic Resistance Ontology (ARO). New terms since the CARD 2020 publication are in bold

Relationship Label	Accession	Definition ^d
<i>is_a</i>	n/a	An axiomatic relationship wherein the subject class A is a subclass of class B. All instances of class A are also instances of class B.
<i>part_of</i>	BFO ^a :0000050	A relationship wherein a subject class A is but a part of class B.
<i>has_part</i>	BFO:0000051	A core relation that holds between a part and its whole.
<i>participates_in</i>	RO ^b :0000056	A relationship wherein a subject class A has a part class B (inverse of <i>part_of</i>).
<i>confers_resistance_to_drug_class</i>	Pending ^c	A relationship between continuant A and process B wherein A is somehow involved in B.
<i>confers_resistance_to_antibiotic</i>	Pending ^c	A relationship wherein the subject class A confers or contributes to antibiotic resistance to drug class B.
<i>targeted_by</i>	Pending ^c	A relationship wherein molecule A is targeted by drug class B.
<i>targeted_by_antibiotic</i>	Pending ^c	A relationship wherein molecule A is targeted by antibiotic B.
<i>regulates</i>	RO:0002211	A relationship wherein the subject class A regulates the activity of class B.
<i>derives_from</i>	RO:0001000	A relationship between class A and class B wherein B inherits many properties from A.
<i>evolutionary_variant_of</i>	RO:0002312	A relationship wherein gene or protein A is a paralogous or orthologous variant of gene or protein B.
<i>is_small_molecule_inhibitor</i>	RO:0012006	A relationship between a continuant A and a process B, in which A is a small molecule that inhibits B.

^aBasic Formal Ontology (<http://purl.obolibrary.org/obo/bfo>).^bRelations Ontology (<http://purl.obolibrary.org/obo/ro>).^cCustom relationships for CARD included in ARO but not yet in the official Relations Ontology.^dParaphrased, see CARD website for full descriptive text.

detailed framework for describing antibiotics and AMR, CARD also applies ARO Classification Tags where curators manually ‘tag’ specific ARO terms as particularly informative for interpretation (13): primary tags AMR Gene Family, Drug Class, Resistance Mechanism, Antibiotic and secondary tags Efflux Component, Efflux Regulator, Adjuvant, Antibiotic + Adjuvant (Table 2). CARD’s secondary curation paradigm requires that ‘every curated AMR determinant must have an ontological path including each of the four primary ARO classification tags, i.e. the [AMR Gene Family] to which that determinant belongs, the [Resistance Mechanism], the [Drug Class] to which resistance is conferred, and the specific [Antibiotic] with a demonstrably elevated MIC’ (13). Overall, the curated sequence data in CARD v3.2.4 spans 458 AMR Gene Families (180 new since v3.0.3), 64 Drug Classes (15 new since v3.0.3), and 8 AMR Resistance Mechanisms (1 new since v.3.0.3) and the ARO contains additional information on 367 antimicrobial molecules and 33 adjuvants. Curation of impact of individual ARGs upon individual antibiotics via the *confers_resistance_to_antibiotic* relationship is ongoing (3386 curated to date from published MIC data), but CARD is also actively developing a MIC module to curate the quantitative MIC values reported in the scientific literature.

The ontology terms, detection models, and reference sequences in CARD v3.2.4 have all been sourced from 3004 publications, with the exception of many β-lactamases, as outlined below. Several core aspects of CARD have been reviewed, updated and/or corrected where necessary. These include: (i) updates to partial gene sequences, (ii) correction of translation frame errors for curated nucleotide se-

quences, (iii) taxonomic review and updates to over 40 pathogen species names (i.e. *Staphylococcus sciuri* updated to *Mammaliicoccussciuri*), (iv) systematic review of ARO term names, definitions, and synonyms, (v) review of the antibiotic molecule branch of the ARO with addition of disinfecting agents and antiseptics (and their ARGs where applicable) and (vi) review of the nomenclature of aminoglycoside acetyltransferases. Larger initiatives have also been undertaken by the CARD biocuration team and are described in more detail below.

Updates to the Model Ontology

While CARD’s curated sequence data can be browsed, downloaded, or searched via BLAST at the CARD website, in our 2017 update (14) we introduced CARD’s Model Ontology, which allows AMR sequence and mutation reference data to be organized by the underlying specific mechanisms of resistance. These bioinformatics models are subsequently used by CARD’s RGI software to annotate genome or metagenome sequences. Descriptions of the model types and parameters in the MO have been updated to more accurately reflect their usage, with some removed due to their non-use. CARD currently curates molecular sequences under nine possible detection model types (each with a variety of parameters): protein homolog models (PHM), protein variant models (PVM), protein overexpression models (POM), rRNA gene variant models (RVM), protein knockout models (PKM), nonfunctional insertion models (NFI), gene cluster meta-models (GCM), efflux pump system meta-models (EPS), and protein domain meta-models (PDM). Supplementary Table S1 gives an overview of

Table 2. ARO classification tags used to provide easy interpretation of genome annotations

Classification tag	Requirement ^a	Annotated ARO terms	ARO example ^b
AMR Gene Family	Primary	458	NDM β-lactamase (ARO:300057)
Drug Class	Primary	64	Aminoglycoside (ARO:0000016)
Resistance Mechanism	Primary	8	Modification to cell morphology (ARO:3007115) ^c
Antibiotic	Primary	367	Streptomycin (ARO:0000040)
Adjuvant	Secondary	33	Tazobactam (ARO:0000077)
Antibiotic + Adjuvant ^c	Secondary	20	Ceftolozane-Tazobactam (ARO:3004724)
Efflux Component	Secondary	2	Efflux pump complex or subunit (ARO:3000159)
Efflux Regulator	Secondary	1	Two-component regulatory system modulating efflux (ARO:3000451)

^aPrimary tags are required for all CARD AMR determinants where applicable; secondary tags apply only rarely and can be omitted at the curator's discretion.

^bExample names are abbreviated, see ARO accession in CARD for the complete description.

^cNew since CARD 2020 publication.

CARD's MO-encoded detection model types and parameters.

Curation at CARD is routinely ahead of RGI software development, so not all MO parameters or models curated in CARD will be annotated in sequences analyzed using RGI. For example, RGI does not currently support the PKM, PDM, GCM or EPS model and meta-model types. In addition, while the PVM, POM and RVM model types are currently supported by RGI, mutation screening currently only supports annotation of resistance-conferring SNPs via the single resistance variant parameter. Supplementary Table S1 outlines all model types and parameters curated in CARD v.3.2.4, with indications of which are supported by RGI. Full descriptions of each model type and parameter in the Model Ontology are available at the CARD website and their use by RGI is outlined in the RGI documentation (<https://github.com/arpcard/rgi>).

Ongoing cross-database harmonization of AMR determinants

A significant portion of AMR determinants in CARD represent β-lactamase (*bla*) genes or gene variants (~71%). Among the 2124 determinants added since our 2020 publication, 1599 (75.3%) of these are described as β-lactam determinants. This is due, in part, to a large effort to coordinate and harmonize *bla* genes in CARD with those in the NCBI Pathogen Detection Reference Gene Catalog (15). NCBI is leading development of β-lactamase nomenclature rules and the naming of new β-lactamases (16), many of which lack corresponding published functional validation due to the rapid pace of sequencing-based discovery. As such, β-lactamases represent an exception to CARD's primary curation paradigm: β-lactamase genes listed in the NCBI Reference Gene Catalog are included in CARD even if lacking experimental evidence of elevated MIC relative to a control or a peer-reviewed scientific publication. Overall, 72 new β-lactamase gene families have been added to CARD since 2020, along with general ontological improvements. For example, through review we discovered *bla*_{ADC-68}, a published carbapenem-hydrolyzing class C β-lactamase (17) missing from CARD. This led to a restructuring of the *bla*_{ADC} branch of the ARO to differentiate *bla*_{ADC} genes with or without carbapenemase activity (see ARO:3005459), as well as the identification of several other missing *bla*_{ADC} lacking experimental examination of carbapenemase activity. Similarly, CARD curators

performed literature reviews for β-lactamases with historically unclear nomenclature, such as the *Bacillus cereus* and *B. anthracis* BcI, BcII, BcIII, Bla1, and Bla2 class A and class B1 (metallo-) β-lactamases, harmonizing nomenclature with NCBI while updating the ARO to reflect the underlying complexity of function and evolutionary history. Yet, challenges continue for curation of unnamed, experimentally validated environmental β-lactamases discovered via functional metagenomics, such as found in wastewater, oil spills, and well water by Berglund *et al.* (18).

The mobile colistin resistance (MCR) phosphoethanolamine transferase (ARO:3004268) AMR gene family is another focus of ARO harmonization efforts. First detected in 2015 as MCR-1 (7), emergence of plasmid-mediated colistin resistance has spread and diversified rapidly (19) with over a hundred variants currently reported. MCR variants are grouped by gene family (MCR-1, MCR-2, etc.) with named individual alleles (20), similar to β-lactamase naming conventions. To reflect this in ARO, we expanded use of the semantic relationship *evolutionary_variant_of* (RO:0002312) which exists between MCR alleles of the same gene family (Table 1). For example, the MCR alleles MCR-1.2 (ARO:3004194), MCR-1.3 (ARO:3004514), and MCR-1.4 (ARO:3004515) are each described as *evolutionary_variant_of* MCR-1.1 (ARO:3003689). This is a relatively new addition to CARD's Relationship Ontology, with the MCR gene family serving as a trial before potential further implementation for other AMR gene families, including particularly β-lactam resistance determinants, e.g. V240G variants of *bla*_{KPC-3} (21).

General cross-database harmonization efforts remain ongoing (both manual and software-assisted), with periodic database and literature surveys to identify problematic or absent ARO curation. A particular emphasis in the next year will be review of the current and historical nomenclature of aminoglycoside modifying enzymes, with an eye towards standardization.

Integration of likelihood-based mutations for *Mycobacterium tuberculosis*

In 2020, CARD adopted novel model parameters for *Mycobacterium tuberculosis* ARG data as unlike other organisms examined, culture and antibiotic susceptibility testing for *M. tuberculosis* is very challenging. The TB community instead developed a likelihood framework, based on statis-

tical association of SNPs with observed phenotypic resistance across a large number of sequenced genomes, which was collated in the Relational Sequencing TB Data Platform (ReSeqTB) (22). This database has since ceased operation, but fortunately CARD captured their last available data set from their 2019 Resistance Report totalling 424 mutations with significant likelihood scores for AMR. Resistance in *M. tuberculosis* is almost exclusively SNPs, but CARD could not use its Single Resistance Variant parameter from the Model Ontology as it is reserved for data with underlying experimental evidence of elevated MIC relative to controls. As such, we created new MO parameters based on the ReSeqTB framework to curate these data: no association with resistance TB (MO:0000053), indeterminate confidence TB (MO:0000054), minimal confidence TB (MO:0000052), moderate confidence TB (MO:0000051), and high confidence TB (MO:0000050) (Supplementary Table S1). These parameters are based on ReSeqTB's likelihood ratio test (LR+) statistic that is used to evaluate whether mutations are positively or negatively associated with phenotypic resistance based on a drug susceptibility test, testing sensitivity, and specificity. Under the null hypothesis of no association, the LR value is expected to be 1, but deviations from this can be due to an association with resistance or a low number of available isolate samples. The LR+ measures the strength of association between the presence of a mutation and the drug resistance phenotype. For the purpose of CARD's curation efforts mutations found to be non-significant ($p > 0.05$) were excluded, synonymous mutations were excluded, and mutations graded as no association with resistance or indeterminate have been archived but excluded from being displayed on the CARD website, included in download files, or RGI reporting. Future curation efforts will include incorporation of the large volume of information available in the recent CRyPTIC project (23,24).

Expanded curation of resistance-modifying agents

One of the seven major branches of the ARO is resistance-modifying agents (ARO:0000076), representing antibiotic adjuvants and other molecules which modify AMR either directly, e.g. β -lactamase inhibitors, or indirectly through other cellular mechanisms, e.g. enhancing antibiotic entry or modifying cell physiology. Adjuvant molecules, as defined here, are not intrinsically growth-inhibitory on their own. This branch of the ARO was, however, historically given little attention and was not subject to definitive curation reviews. In our most recent update, we reviewed the existing ARO entries and the literature on adjuvants and other resistance modifying molecules, leading to significant ontological modifications, clarification of definitions, and expansion of included molecules. This led to reorganization of this branch of the ARO to include five novel ARO adjuvant categories: (i) inhibitor of antibiotic resistance mechanism (ARO:3007222), (ii) adjuvants enhancing antibiotic entry (ARO:3007223), (iii) adjuvants inhibiting antibiotic removal (ARO:3007224), (iv) adjuvants which alter physiology (ARO:3007225) and (v) host-related antibiotic adjuvants (ARO:3007226). In addition, β -lactamase inhibitors are now separated into serine β -lactamase in-

hibitor (ARO:3007128) and metallo- β -lactamase inhibitor (ARO:3007128) sub-branches, with further subcategorization based on chemical structure and composition, e.g. tazobactam as a boronic acid β -lactamase inhibitor. Lastly, β -lactamase inhibitors are now tagged as Adjuvants using ARO Classification Tags (Table 2), with the newly added ontological relationship *is_small_molecule_inhibitor* (Table 1) used to encode the relationship between β -lactamase inhibitors and their target β -lactamase, e.g. tazobactam (a serine β -lactamase inhibitor) *is_small_molecule_inhibitor* of TEM-1, TEM-2, etc. While we acknowledge that antibiotics can act as adjuvants in clinically relevant antibiotic-antibiotic combinations (trimethoprim-sulfamethoxazole, quinupristin-dalfopristin, etc.), their intrinsic lethality on their own excludes them from the adjuvant definition presented herein.

SUPPORTING MACHINE LEARNING WITH CARD SHORT NAMES

While some ARGs have established nomenclature rules, standardization of ARG names is not universal and some names curated into the ARO are very long. A common request from our data-focused users has been the addition of a curated and abbreviated naming convention for ARO vocabulary terms used for individual ARGs. Long ARO names are cumbersome or prohibited when dealing with commonly used data formats or software in bioinformatics and machine learning, e.g. *Escherichia coli gyrA conferring resistance to fluoroquinolones* (ARO:3003294). However, descriptive ARO names are also necessary for precluding inaccuracies and ambiguities among AMR determinants, particularly when managing point mutation-based resistance that often varies between taxa, e.g. the above term versus *Shigella flexneri gyrA conferring resistance to fluoroquinolones* (ARO:3003940). These two examples are necessarily descriptive, a requirement of ontologies, to both indicate the species from which the *gyrA* sequence was obtained and to differentiate point mutations specific to *E. coli* or *S. flexneri* which contribute to phenotypic fluoroquinolone resistance. Yet, at 63 and 64 characters these names are a bane for programmers, data scientists, and for concise visualization. To address this issue, we here introduce CARD Short Names, a CARD-specific abbreviation for ARG names associated with Antibiotic Resistance Ontology terms, often not based on the literature but instead designed by CARD curators.

If the original ARG name is 15 characters or less, the CARD short name is identical; if the gene name is >15 characters, the CARD Short Name has been abbreviated by CARD curators specifically to identify the proper gene or protein name. All CARD Short Names are unique and have whitespace characters replaced by underscore characters. The convention for pathogen names is capitalized first letter of the genus followed by the lowercase first three letters of the species name, for example 'Abau' for *Acinetobacter baumannii*. The antibiotic abbreviations are from abbreviations listed at the scientific journal *Antimicrobial Agents and Chemotherapy* (<https://journals.asm.org/journal/aac/abbreviations>) plus some custom abbreviations from the CARD curators, for example 'AMG' for

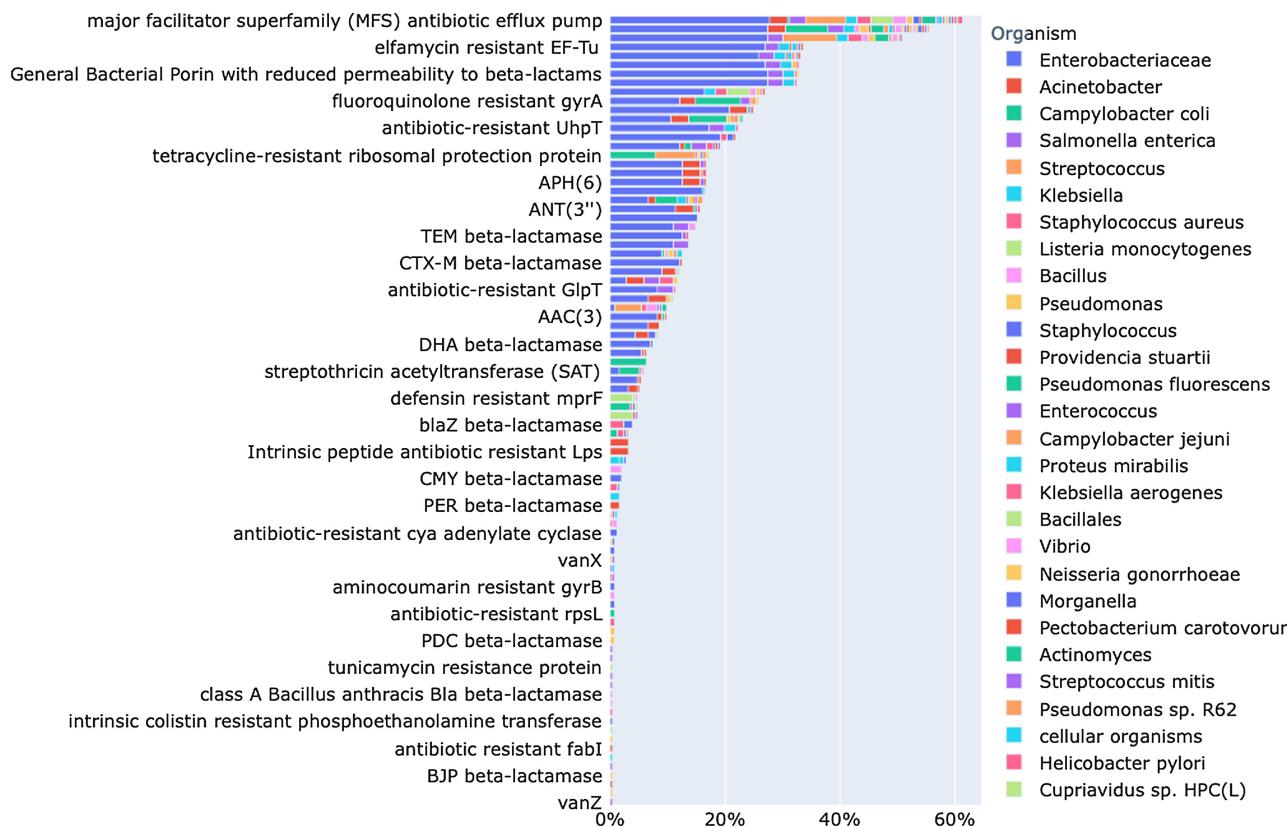


Figure 1. Prevalence of individual AMR Gene Families in data collected by CARD:Live, broken down by predicted pathogen of the submitted genome sequences. Approximately 60% of submitted genome sequences have an ARG from the major facilitator superfamily (MFS) antibiotic efflux pumps.

aminoglycosides or ‘AZM’ for azithromycin. Simple CARD Short Names often do not involve either, e.g. CTX-M-15, but where applicable the CARD Short Names follow pathogen_gene or pathogen_gene_drug, e.g. *EcoI-gyrA_FLO* and *Sfle-gyrA_FLO* for the above examples. For ARGs conferring resistance to multiple drug classes, we use the generic term *MULT*, e.g. *MexEF-OprN* with *MexS mutations conferring resistance to chloramphenicol, ciprofloxacin, and trimethoprim* (ARO:3004068) becomes *Paer_MexS_MULT*, where ‘Paer’ indicates the pathogen *Pseudomonas aeruginosa*. The full lists of abbreviations and CARD Short Names (5045 as of CARD v3.2.4) can be obtained at the CARD website (<https://card.mcmaster.ca/download>).

ONLINE RESISTOME GENE IDENTIFIER (RGI) AND CARD:LIVE

The CARD website allows users to predict ARGs from genome sequences using RGI, providing interactive visualization of the results organized by ARO Classification Tags and ARG annotation (<https://card.mcmaster.ca/analyze/rgi>). Usage of RGI is global and a genome sequence from a drug resistant infection is processed by RGI approximately every 4 min. In late 2020, we launched CARD:Live, an initiative to collect pathogen identification, MLST, annotated ARG names, date, and geographical region for genome sequences submitted to RGI online, providing a dynamic view of antibiotic resistant isolates from around the world being analyzed on RGI online (Figure 1). All infor-

mation collected is voluntary, based on consent, and no sequence information is retained to protect the research programs of contributors. To date, over 13 000 RGI results have been collected and can be visualized online. Tools for versioning, downloading, or performing complex analyses of these data are forthcoming.

IN SILICO PREDICTION OF CARD RESISTOMES AND VARIANTS

In our 2020 publication, we introduced CARD Resistomes & Variants (13), a new database module for computationally predicted resistomes and ARG variant sequences to supplement canonical CARD curation of functionally characterized ARGs from the literature. Briefly, CARD-R uses CARD’s curated reference data, detection models, and RGI software to predict ARG sequence variants *in silico* from available genomic data for a targeted list of pathogens. This process thus predicts a putative resistome for each included genome or plasmid sequence, genomic island, or whole-genome shotgun assembly based on the AMR detection models curated in CARD and RGI’s ‘Perfect’ and ‘Strict’ annotations, where the ‘Perfect’ algorithm detects perfect matches to the curated reference sequences and mutations in CARD while the ‘Strict’ algorithm predicts variants of known ARGs, including screens for key mutations, using CARD’s curated bit-score cut-offs (Figure 2). At its inception, CARD-R consisted of a relatively small slate of pathogens (82 in total) chosen to encompass

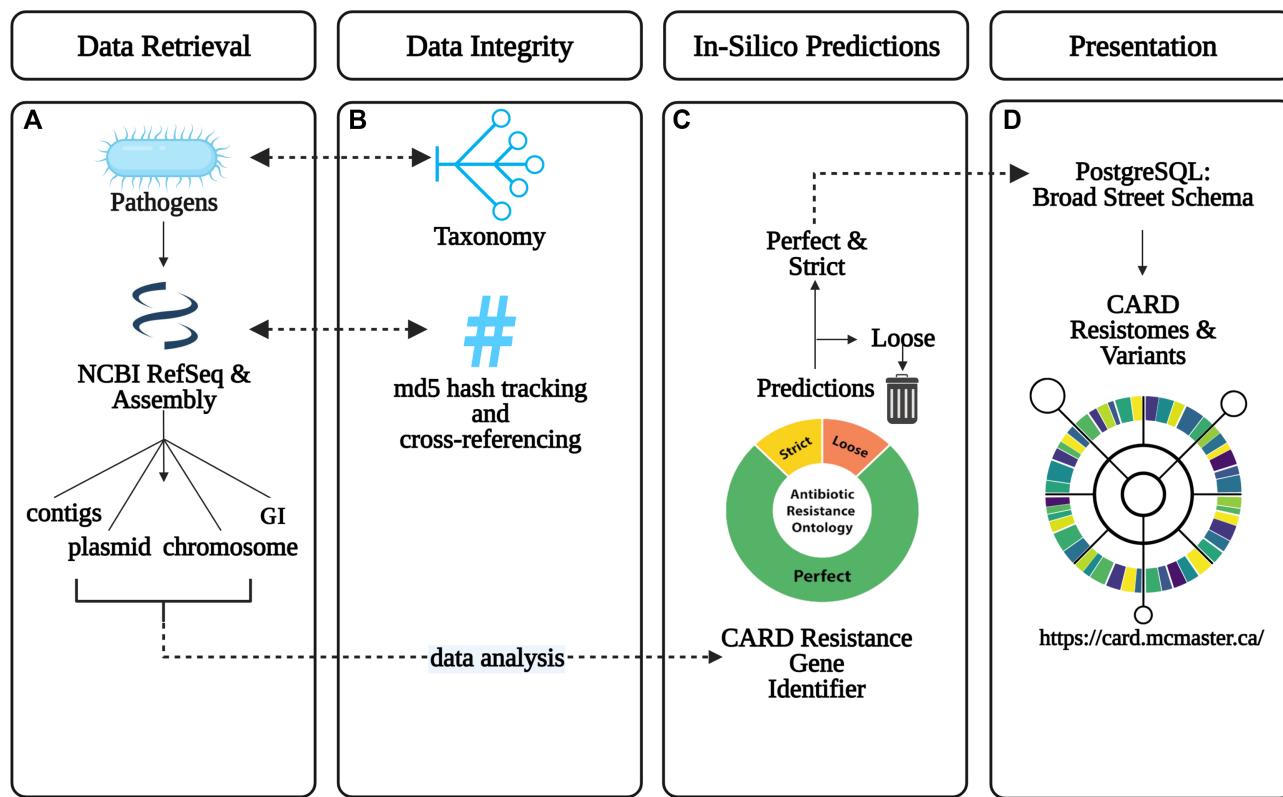


Figure 2. Schematic workflow for CARD Resistomes and Variants. (A) Data is retrieved by downloading available genome assemblies for a list of pathogens from NCBI's RefSeq database. Retrieved genome assemblies are subsequently parsed to identify plasmid- and chromosome-originating sequences, while genomic islands are identified separately through IslandViewer 4. (B) Each pathogen is specified using the NCBI Taxonomy Database ID and the genome data integrity is verified with an MD5 checksum. (C) Each genome is analyzed with RGI to generate a predicted resistome, which is then filtered to remove divergent or otherwise dissimilar homologs below the respective CARD model threshold for RGI ('Loose'). (D) The remaining high-fidelity predictions (RGI 'Perfect' and 'Strict' categories) are parsed to integrate predicted resistome information and genomic sequences into the Broad Street PostgreSQL schema underlying CARD, where the data is contextualized through CARD and the ARO. These data are then uploaded to the CARD website where they are available publicly for viewing, download, or use by RGI's metagenomic read mapping algorithms.

those of particular public health and research significance, e.g. the World Health Organization's (WHO) priority list of antibiotic resistant bacteria (2,25). Since then, CARD-R has expanded significantly, both in the volume of accumulated data (through additional pathogens and genomes) and the scale of collected AMR metadata. CARD-R v4.0.0 encompasses 377 pathogens covering 165 bacterial genera and 21,079 complete chromosomes, 41,828 complete plasmids, and 155,606 whole-genome shotgun assemblies from NCBI's RefSeq catalog (26). An additional 50 bacterial pathogens are included in this *in silico* analysis, but lack prediction of any 'Strict' or 'Perfect' ARGs by RGI. A recent addition to CARD-R is the integration of 356,325 genomic island (GI) sequence predictions from IslandViewer 4 (27) for AMR determinant prediction, resulting in 2,662 GIs predicted to encode ARGs (across 1,084 genomes).

Overall CARD-R's *in silico* prediction of resistomes yielded 322,710 unique nucleotide ARG allele sequences, each associated with ARO Classification Tags and covering 247 of CARD's 461 AMR Gene Families and 31 of CARD's 64 Drug Classes. Expansion of the list of pathogens included in CARD-R reflects harmonization with IslandViewer plus a targeted curation of pathogens associated with sepsis. The CARD-R data have several uses: (i) provid-

ing statistics on prevalence of individual ARGs within and among pathogens (e.g. *bla_{NDM-1}* is found in 47 pathogens), (ii) determining association of ARGs with mobile genetic elements as a step towards risk assessment, (iii) providing a more diverse reference sequence set for RGI's metagenomic read alignment algorithms than available in canonical CARD (which is strongly biased towards common clinical pathogens due to its dependence upon the published literature) and (iv) providing data for RGI's k-mer based ARG pathogen-of-origin prediction algorithms. Supplemental Table S2 details the breakdown of assemblies into chromosomal, plasmid, and assembly contig sequence origins and indicates the proportion of ARG-containing genomes for each pathogen.

CONTINUOUS CURATION AND EXPANDED QUALITY CONTROL

As described previously, CARD has developed the custom 'Broad Street' schema for information storage and organization (13), broken down into six modules: controlled vocabularies, AMR reference sequences & detection models, resistomes & variants & prevalence; publications; external references; and administrative. Detailed de-

scription of this schema is now available online (https://github.com/arpard/broadstreet_schema). The schema and data are managed with PostgreSQL 9.5.24 and the public CARD website and curator tools are designed with the Laravel 5.2.45 PHP framework, PHP 7.0.33, Apache 2.4.18 and PostgreSQL 9.5.24. The website, software, data, and curation issue tracking are all version-controlled using GitLab CE version 15.2.0.

CARD curation occurs continuously, with updates released every 1–3 months by a team of biocurators. Curation involves regular review of the scientific literature, supplemented by custom CARD*Shark text mining tools to identify high value publications in PubMED via machine learning. CARD curation generally starts with addition of a term for a new AMR gene. This will include entering the name of the gene, recording synonyms, writing a description, creating a novel CARD Short Name, and entering of the related publications. Curation may additionally include a Protein Data Bank structure (if available), private Curator Notes, or data for our Minimum Inhibitory Concentration module under development. If the ARO term instead describes an antibiotic molecule, we often curate a PubChem structure. We then create relationships between this term and the main branches of the ARO. This may include creating terms for new resistance mechanisms, antibiotics, or drug classes as well as adding ARO Classification Tags. With all the ontological and related information curated, the remaining major curation effort is defining the AMR gene detection model parameters for the gene, including reference sequences, mapped SNPs, bitscore cut-offs, definition of operons, or identification of all components of protein complexes.

One of CARD's strengths is a dedication to expert, human curation of data. Yet, mistakes happen and there is a learning curve to accurately curating CARD for new staff. As such, CARD maintains a suite of Quality Control software tools (CARD-QC), which check a multitude of rules based on the Broad Street schema plus the ARO and other ontologies. These include checking text for use of whitespace, punctuation, and special characters; checking accuracy of citations from PubMED (including checking for updated citations), validating relations among ontology terms, including ARO's curation paradigms, use of RO terms, and validation of synonyms; assessing direct and transitive (but not reflexive) closure for ontologies; check for missing curation of PDB or PubChem structures; ensure CARD Short Names are properly applied to all ARGs; validate ARO Classification Tags; ensuring the NCBITaxon ontology is taxonomically accurate and reflects any updated nomenclature; check reading frames, codons, and translation of reference sequences, including use of valid nucleotide or amino acid characters; check for updated sequences at NCBI; and checking completeness and values of model parameters based on detection model type.

DATA AVAILABILITY, BAIT CAPTURE PROTOCOLS, AND COMMUNITY AMR CURATION

The CARD website (<https://card.mcmaster.ca>) provides tools for browsing and searching the ARO and other ontologies, names and descriptions of ARGs, CARD Re-

sistomes & Variants & Prevalence data, the CARD:Live collection, tracking of changes for each release, plus laboratory protocols and bait sequences for performing ARG enrichment metagenomic sequencing based on CARD's curated sequences (28). Online search tools include BLAST for comparing sequences to CARD reference sequences and the RGI for resistome annotation. CARD data can be downloaded online in a number of formats (TSV, OBO, OWL, JSON, FASTA). RGI software and documentation is available at GitHub (<https://www.github.com/arpard/rgi>) and the ARO is additionally cross-posted to the Open Biomedical Ontologies' OBO Foundry (<http://purl.obolibrary.org/obo/aro>). Updates to CARD are announced on Twitter (@arpard) and via the CARD-L mailing list (see <http://arpard.mcmaster.ca/about>). CARD curators are available to provide assistance via email to card@mcmaster.ca or posting of issues at CARD's amr_curation GitHub repo (https://github.com/arpard/amr_curation).

CONCLUSIONS

In our 2020 update, we focused heavily on changes and improvements to the ontological structure underpinning CARD. Comparatively, only minor changes to the overall database schema and formatting behind the ARO are provided in this update. Instead, along with several quality improvements to general user experience and data access, as CARD continues to expand we have directed our efforts to more specialized and previously under-curated aspects of the ontology. This approach has led to large ontological and curation improvements for specific AMR gene families, drug classes, or individual branches of the ARO, as seen with the inclusion of over one-thousand additional β -lactamase sequences, the harmonization of MCR-based colistin resistance variants and nomenclature, the integration of likelihood-associated mutation parameters for *M. tuberculosis*, the addition of resistance to antiseptics, and the ongoing ontological overhaul of resistance-modifying agents. Additionally, greater effort has been expended towards integrating CARD with analytical software, including the significant expansion of CARD Resistomes & Variants, the introduction of CARD:Live, and development of the CARD Short Names standardized nomenclature. While CARD will continue to provide high-quality curated reference data, future improvements will expand from CARD's focus upon individual ARGs to a pathogen-based holistic view of AMR prediction by combining the ARO with CARD-R as well as developing pathogen-specific curation and annotation rules.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Drs. Michael Surette and Hendrik Poinar (McMaster University) for sharing their curated list of sepsis-associated pathogens as aid to expansion of CARD Resistomes, as well as Drs. Marilyn Roberts (University of Wash-

ington Seattle), Torsten Seemann (The Peter Doherty Institute for Infection and Immunity), Daniel Haft & Michael Feldgarden (National Center for Biotechnology Information), and Marco Schito & Daniel Hartley (ReSeqTB) for input into CARD curation. Hafsa Omer provided a review of AMR mutations in agricultural *Salmonella*. We also thank the community of antimicrobial resistance researchers making contributions to CARD's growth and accuracy via mailing lists, curation repositories, and email. Dr. Gerard Wright (McMaster University) provides numerous insightful comments on all aspects of CARD.

FUNDING

This research was funded by the Canadian Institutes of Health Research (CIHR) and Genome Canada (to A.G.M., F.B., R.G.B.); Cisco Research Chair in Bioinformatics (to A.G.M.); David Braley Chair in Computational Biology (to A.G.M.). B.J. and K.G. were supported by scholarships from NSERC-CREATE, CIHR, Simon Fraser University; B.J. and L.A.C. were additionally supported by a CIHR Canada Graduate Scholarship (CGS-D); K.K.T. was supported by an Ontario Graduate Scholarship, McMaster University's MacDATA Institute Graduate Fellowship, Michael G. DeGroote Institute for Infectious Disease Research (IIDR) Michael Kamin Hart Memorial Scholarship; J.A.N. was supported by a Fred and Helen Knight Enrichment Award; A.E. by an Ontario Graduate Scholarship, MacData Institute Graduate Fellowship, Ashbaugh Graduate Scholarship; F.M. was supported by a Donald Hill Family Fellowship in Computer Science and NSERC; T.B. was supported by a Michael G. DeGroote Centre for Medicinal Cannabis Research (MGD CMCR) postdoctoral fellowship; E.G. was supported by a CIHR Project Grant [PJT-159456 to W.W.L.H.]; K.R. was supported by an Ontario Graduate Scholarship; A.P. was supported by the Visual and Automated Disease Analytics graduate training program; E.D.B. was supported by a Tier 1 Canada Research Chair award, a Foundation grant from the Canadian Institutes of Health Research (CIHR) [FRN 143215], and the Ontario Research Fund [RE09-047]. Computer resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants from the Canada Foundation for Innovation and a donation of hardware by Cisco Systems Canada, Inc. Funding for open access charge: Canadian Institutes for Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. McArthur,A.G., Waglechner,N., Nizam,F., Yan,A., Azad,M.A., Baylay,A.J., Bhullar,K., Canova,M.J., De Pascale,G., Ejim,L. *et al.* (2013) The Comprehensive Antibiotic Resistance Database. *Antimicrob. Agents Chemother.*, **57**, 3348–3357.
2. Tacconelli,E., Carrara,E., Savoldi,A., Harbarth,S., Mendelson,M., Monnet,D.L., Pulcini,C., Kahlmeter,G., Kluytmans,J., Carmeli,Y. *et al.* (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.*, **18**, 318–327.
3. CDC (2019) *Antibiotic Resistance Threats in the United States, 2019*. U.S. Department of Health and Human Services, CDC, Atlanta, GA.
4. The Expert Panel on the Potential Socio-Economic Impacts of Antimicrobial Resistance in Canada (2019) *When Antibiotics Fail*. Council of Canadian Academies, Ottawa, ON.
5. O'Neill,J. (2016) Tackling drug-resistant infections globally: final report and recommendations. *The Review on Antimicrobial Resistance*. London.
6. Antimicrobial Resistance Collaborators (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, **399**, 629–655.
7. Liu,Y.-Y., Wang,Y., Walsh,T.R., Yi,L.-X., Zhang,R., Spencer,J., Doi,Y., Tian,G., Dong,B., Huang,X. *et al.* (2016) Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect. Dis.*, **16**, 161–168.
8. Surette,M.D., Waglechner,N., Kotova,K. and Wright,G.D. (2022) HelR is a helicase-like protein that protects RNA polymerase from rifamycin antibiotics. *Mol. Cell*, **82**, 3151–3165.
9. Zhang,A.-N., Gaston,J.M., Dai,C.L., Zhao,S., Poyet,M., Groussin,M., Yin,X., Li,L.-G., van Loosdrecht,M.C.M., Topp,E. *et al.* (2021) An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat. Commun.*, **12**, 4765.
10. Zhang,Z., Zhang,Q., Wang,T., Xu,N., Lu,T., Hong,W., Penuelas,J., Gillings,M., Wang,M., Gao,W. *et al.* (2022) Assessment of global health risk of antibiotic resistance genes. *Nat. Commun.*, **13**, 1553.
11. Ghosh,S., Bornman,C. and Zafer,M.M. (2021) Antimicrobial resistance threats in the emerging COVID-19 pandemic: where do we stand? *J. Infect. Public Health*, **14**, 555–560.
12. Knight,G.M., Glover,R.E., McQuaid,C.F., Olaru,I.D., Gallandat,K., Leclerc,Q.J., Fuller,N.M., Willcocks,S.J., Hasan,R., van Kleef,E. *et al.* (2021) Antimicrobial resistance and COVID-19: intersections and implications. *Elife*, **10**, e64139.
13. Alcock,B.P., Raphenya,A.R., Lau,T.T.Y., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.-L.V., Cheng,A.A., Liu,S. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.*, **48**, D517–D525.
14. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S., Sharma,A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.*, **45**, D566–D573.
15. Feldgarden,M., Brover,V., Gonzalez-Escalona,N., Frye,J.G., Haendiges,J., Haft,D.H., Hoffmann,M., Pettengill,J.B., Prasad,A.B., Tillman,G.E. *et al.* (2021) AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.*, **11**, 12728.
16. Bradford,P.A., Bonomo,R.A., Bush,K., Carattoli,A., Feldgarden,M., Haft,D.H., Ishii,Y., Jacoby,G.A., Klimke,W., Palzkill,T. *et al.* (2022) Consensus on β-Lactamase nomenclature. *Antimicrob. Agents Chemother.*, **66**, e00333–e22.
17. Jeon,J.H., Hong,M.K., Lee,J.H., Lee,J.J., Park,K.S., Karim,A.M., Jo,J.Y., Kim,J.H., Ko,K.S., Kang,L.W. *et al.* (2014) Structure of ADC-68, a novel carbapenem-hydrolyzing class c extended-spectrum β-lactamase isolated from *Acinetobacter baumannii*. *Acta Crystallogr. D Biol. Crystallogr.*, **70**, 2924–2936.
18. Berglund,F., Marathe,N.P., Österlund,T., Bengtsson-Palme,J., Kotsakis,S., Flach,C.-F., Larsson,D.G.J. and Kristiansson,E. (2017) Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, **5**, 134.
19. Wang,R., van Dorp,L., Shaw,L.P., Bradley,P., Wang,Q., Wang,X., Jin,L., Zhang,Q., Liu,Y., Rieux,A. *et al.* (2018) The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.*, **9**, 1179.
20. Partridge,S.R., Di Pilato,V., Doi,Y., Feldgarden,M., Haft,D.H., Klimke,W., Kumar-Singh,S., Liu,J.-H., Malhotra-Kumar,S., Prasad,A. *et al.* (2018) Proposal for assignment of allele numbers for mobile colistin resistance (mcr) genes. *J. Antimicrob. Chemother.*, **73**, 2625–2630.
21. Satapoomin,N., Dulyayangkul,P. and Avison,M.B. (2022) *Klebsiellapneumoniae* mutants resistant to ceftazidime-avibactam plus aztreonam, imipenem-relebactam, meropenem-vaborbactam, and

- cefepime-taniborbactam. *Antimicrob. Agents Chemother.*, **66**, e0217921.
22. Ezewudo,M., Borens,A., Chiner-Oms,Á., Miotto,P., Chindelevitch,L., Starks,A.M., Hanna,D., Liwski,R., Zignol,M., Gilpin,C. *et al.* (2018) Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci. Rep.*, **8**, 15382.
 23. The CRyPTIC Consortium (2022) A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13 antibiotics. *PLoS Biol.*, **20**, e3001721.
 24. The CRyPTIC Consortium (2022) Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biol.*, **20**, e3001755.
 25. Taccanello,E., Magrini,N., Kahlmeter,G. and Singh,N. (2017) Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *World Health Organ.*, **27**, 318–327.
 26. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
 27. Bertelli,C., Laird,M.R., Williams,K.P. and Simon Fraser University Research Computing GroupSimon Fraser University Research Computing Group, Lau,B.Y., Hoad,G., Winsor,G.L. and Brinkman,F.S.L. (2017) IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.*, **45**, W30–W35.
 28. Guitor,A.K., Raphenya,A.R., Klunk,J., Kuch,M., Alcock,B., Surette,M.G., McArthur,A.G., Poinar,H.N. and Wright,G.D. (2019) Capturing the resistome: a targeted capture method to reveal antibiotic resistance determinants in metagenomes. *Antimicrob. Agents Chemother.*, **64**, e01324-19.