MICROBIAL GENOMICS

BIORESOURCE

Feldgarden *et al.*, *Microbial Genomics* 2022;8:000832

DOI 10.1099/mgen.0.000832

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Curation of the AMRFinderPlus databases: applications, functionality and impact

Michael Feldgarden*, Vyacheslav Brover, Boris Fedorov, Daniel H. Haft, Arjun B. Prasad and William Klimke

Abstract

Antimicrobial resistance (AMR) is a significant public health threat. Low-cost whole-genome sequencing, which is often used in surveillance programmes, provides an opportunity to assess AMR gene content in these genomes using *in silico* approaches. A variety of bioinformatic tools have been developed to identify these genomic elements. Most of those tools rely on reference databases of nucleotide or protein sequences and collections of models and rules for analysis. While the tools are critical for the identification of AMR genes, the databases themselves also provide significant utility for researchers, for applications ranging from sequence analysis to information about AMR phenotypes. Additionally, these databases can be evaluated by domain experts and others to ensure their accuracy. Here we describe how we curate the genes, point mutations and blast rules, and hidden Markov models used in NCBI's AMRFinderPlus, along with the quality-control steps we take to ensure database quality. We also describe the web interfaces that display the full structure of the database and their newly developed cross-browser relationships. Then, using the Reference Gene Catalog as an example, we detail how the databases, rules and models are made publicly available, as well as how to access the software. In addition, as part of the Pathogen Detection system, we have analysed over 1 million publicly available genomes using AMRFinderPlus and its databases. We discuss how the computed analyses generated by those tools can be accessed through a web interface. Finally, we conclude with NCBI's plans to make these databases accessible over the long-term.

DATA SUMMARY

(1) Both databases used by AMRFinderPlus can also be downloaded as part of the AMRFinderPlus installation process. AMRFinderPlus is available through GitHub (https://github.com/ncbi/amr), with a wiki that provides additional information on installation and programme use (https://www.github.com/ncbi/amr/wiki). AMRFinderPlus and its databases can be easily installed with Bioconda as described on the AMRFinderPlus wiki.

(2) Detailed documentation of database formats and contents is publicly available at https://www.github.com/ncbi/amr/wiki. All database historical releases are retained indefinitely, contain detailed change logs, and are backed up in multiple locations.

(3) The databases also are available through the following GUIs:

   a. The Pathogen Detection Reference Gene Catalog (https://www.ncbi.nlm.nih.gov/pathogens/refgene/) provides a visualization of acquired genes and point mutations used by AMRFinderPlus, where each row represents an acquired protein sequence or point mutation (see Fig. 1). Core genes are available in BioProject PRJNA313047.

   b. The Reference Gene Hierarchy (https://www.ncbi.nlm.nih.gov/pathogens/genehierarchy/) is a web-based view into the hierarchy of genes, families and upstream nodes that NCBI curators use to organize and relate the genes and hidden Markov models (HMMs) in the Reference Gene Catalog and Pathogen Detection Reference HMM Catalog.

   c. The Pathogen Detection Reference HMM Catalog (https://www.ncbi.nlm.nih.gov/pathogens/hmm/) is a web-based portal to our curated database of reference HMMs used by AMRFinderPlus in concert with gene sequences in the

Received 17 December 2021; Accepted 22 April 2022; Published 08 June 2022

Author affiliations: [1]National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894, USA.
***Correspondence:** Michael Feldgarden, michael.feldgarden@nih.gov

Keywords: curation; antimicrobial resistance; genomics.

Abbreviations: AMR, antimicrobial resistance; GUI, graphical user interface; HMM, hidden Markov model; MCR, moblie colistin resistance; NCBI, National Center for Biotechnology Information; Qnr, quinolone resistance.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. A supplementary table is available with the online version of this article.

000832

Significance as a BioResource to the community

AMR gene-detection tools are critical for using bacterial genomics to understand the spread of antibiotic resistance. Most AMR gene-detection tools rely on reference databases of nucleotide or protein sequences and collections of models and rules for analysis, which are derived from the literature and often designed to meet specific scientific needs, such as food-borne surveillance and other applied uses. These databases can have significant effects on the success of analytical procedures [30] and also can be reutilized for other research purposes. Here, we describe the curation process for the databases, models and rules used by NCBI's AMRFinderPlus. We also describe our procedure for making these databases publicly available, as well as user interfaces to interrogate these databases. We also describe how users can access computed AMRFinderPlus results for over 1000000 bacterial isolates in NCBI's Pathogen Detection system. To maintain relevance and accuracy, we describe possible areas of database improvement and how users can assist and guide our database curation.

Pathogen Detection Reference Gene Catalog to identify antimicrobial resistance (AMR) genes as well as some stress resistance and virulence genes.

(4) Computed analyses by AMRFinderPlus on the over 1,000,000 isolates in NCBI's Pathogen Detection system can be found in two different GUIs:
   a. The Isolates Browser (https://www.ncbi.nlm.nih.gov/pathogens/isolates/) provides a summary of AMR, stress response and virulence genes for each isolate.
   b. The Microbial Browser for Identification of Genetic and Genomic Elements (MicroBIGG-E; https://www.ncbi.nlm.nih.gov/pathogens/microbigge/), displays AMRFinderPlus results for those isolates, which have genomic data deposited in GenBank.

INTRODUCTION

Antimicrobial resistance (AMR) is a significant public health threat, and has been estimated to cause over one million deaths globally [1]. Low-cost whole-genome sequencing, which is often used in surveillance programmes, provides an opportunity to assess AMR gene content in these genomes using *in silico* approaches. These *in silico* approaches can lead to the discovery of novel resistance mechanisms [2], and also can be used to predict resistance phenotypes [3, 4]. A variety of bioinformatic tools have been developed to identify these genomic elements, for purposes ranging from basic research to applied uses such as surveillance and clinical use [4–6]. Most of those tools rely on reference databases of nucleotide or protein sequences and collections of models and rules for analysis.

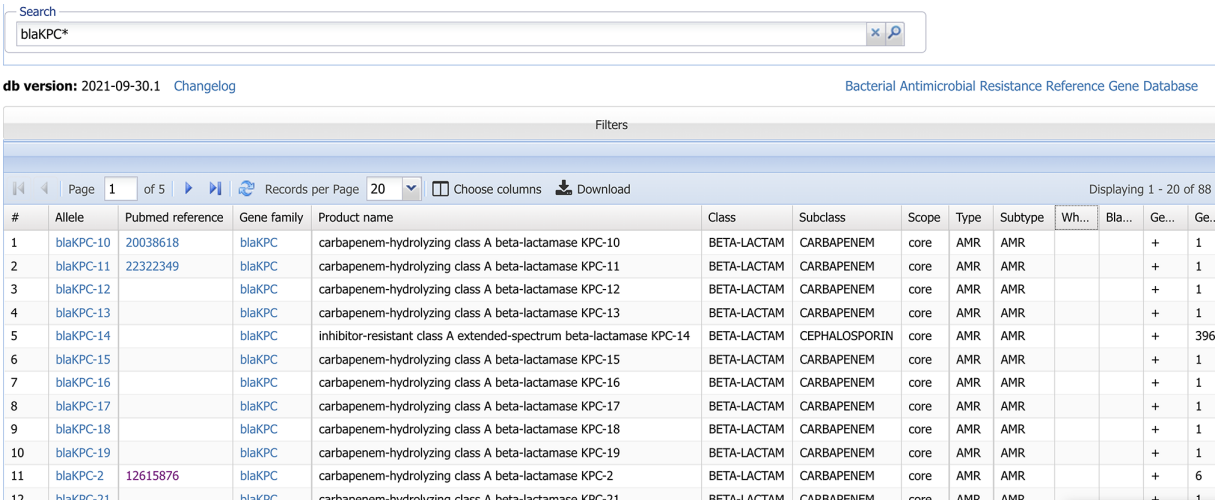
While the tools are critical for the identification of AMR genes, their underlying databases are a critical component of these tools' effectiveness; for example, if a gene is missing from a database, it is less likely to be found. The databases themselves also provide significant utility for researchers. For example, researchers interested in drug development can use existing sequence variation to design better drugs, and accessible databases make that process easier. Well-curated databases also contain biological information about those genetic elements they contain, which can be used to understand those elements' function. Users might use a curated database as the backbone for their own specialized analyses, such as using existing databases for large-scale metagenomic analyses [7]. Also, domain experts and others can evaluate public databases, improving those databases' accuracy. Thus, it is critical that other researchers are able to access these tools' databases [8].

Another key element is ongoing curation. Not only should databases be available to others, but ongoing curation, as novel genetic elements related to antimicrobial susceptibility and pathogenesis are routinely discovered, is essential for keeping databases functional and useful. To do this, database curation should involve:

- Gathering novel information.
- Ensuring the reliability of the data.
- Adding value through analysis.
- Making the data easily accessible to others (e.g. downloads, GUIs).

Most databases and their associated tools are developed for research and public health purposes. For these reasons, curators should make downstream analyses derived from the tools and curated databases available. In a sense, this is an additional layer of data that itself requires curation.

Here we describe how we acquire and curate the genes, point mutations, blast rules and HMM models used in NCBI's AMRFinderPlus [9, 10]. We then describe the steps we take to ensure database quality. Using the Reference Gene Catalog as an example, we detail how the databases, rules and models are made publicly available, as well as how to access the software. We discuss how the computed analysis of those tools can be accessed. Finally, we conclude with NCBI's plans to make these databases accessible over the long-term and plans for future improvements in our databases.


Search
blaKPC*

db version: 2021-09-30.1 Changelog

Bacterial Antimicrobial Resistance Reference Gene Database

Filters

Page 1 of 5 Records per Page 20 Choose columns Download Displaying 1 - 20 of 88

#	Allele	Pubmed reference	Gene family	Product name	Class	Subclass	Scope	Type	Subtype	Wh...	Bla...	Ge...	Ge..
1	blaKPC-10	20038618	blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-10	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
2	blaKPC-11	22322349	blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-11	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
3	blaKPC-12		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-12	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
4	blaKPC-13		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-13	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
5	blaKPC-14		blaKPC	inhibitor-resistant class A extended-spectrum beta-lactamase KPC-14	BETA-LACTAM	CEPHALOSPORIN	core	AMR	AMR			+	396
6	blaKPC-15		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-15	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
7	blaKPC-16		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-16	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
8	blaKPC-17		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-17	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
9	blaKPC-18		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-18	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
10	blaKPC-19		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-19	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1
11	blaKPC-2	12615876	blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-2	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	6
12	blaKPC-21		blaKPC	carbapenem-hydrolyzing class A beta-lactamase KPC-21	BETA-LACTAM	CARBAPENEM	core	AMR	AMR			+	1

Fig. 1. The Reference Gene Catalog. For acquired genes, each row contains the gene symbol, the allele symbol, GenBank and RefSeq nucleotide and protein accessions, phenotype information, and a PubMed citation. For point mutations, each row contains an allele symbol, which is a concatenation of the point mutation and gene symbol of the reference gene, the gene symbol of the reference gene, GenBank and RefSeq nucleotide or protein accessions of the reference sequence, phenotype information and a PubMed citation.

AMRFinderPlus and its associated databases as a model

Before we describe how we curate the databases used by AMRFinderPlus, we wish to review several key features of AMRFinderPlus (further details are available in [9, 10]). AMRFinderPlus searches either nucleotide or protein sequence, or both jointly, for acquired stress response, virulence and antimicrobial resistance (AMR) genes, as well as point mutations. Taxon-specific point mutations are identified by BLAST [11, 12] against a taxon-specific set of reference protein and nucleotide sequences. AMRFinderPlus output provides the position of each element (gene or point mutation), the method used for identification (BLAST or HMM), and possible phenotypes. Acquired genes are identified either through BLAST against a reference database, with each gene possessing a manually curated BLAST cutoff, or, if protein sequence is available, through a combination of BLAST and hidden Markov models (HMMs), both of which have manually curated cutoffs.

A novel feature of AMRFinderPlus is that genes, which can have one or more proteins, are assigned to a node in a hierarchy (Fig. 2a). For example, conceptually we can view how the classification of a beta-lactamase would work. A protein that is 100% identical to $bla_{\text{KPC-2}}$ is clearly $bla_{\text{KPC-2}}$. A novel, but only slightly divergent protein would be called bla_{KPC}, and would likely, though not necessarily confer resistance to carbapenems. A somewhat more divergent beta-lactamase would be assigned to a node composed of related class A beta-lactamases (with gene symbol *bla*), while even more divergent proteins could be identified as class A beta-lactamases or even as beta-lactamases of unknown class. This allows AMRFinderPlus to report the most accurate gene name, reflecting possible ambiguity in its functional annotation, as opposed to the name of the nearest gene as defined by sequence identity [13, 14].

Thus, the AMRFinderPlus database has four essential components:

(1) An acquired gene database of AMR, stress response and virulence genes. This is a collection of genes (and gene symbols), each of which contains one or more proteins, often with an associated BlastRule (protein identity threshold) or HMM, along with phenotypic data and other descriptive metadata.
(2) A collection of point mutations and reference sequences that contains the site(s) of the mutations, the reference sequence and the target organism, along with phenotypic data and other descriptive metadata.
(3) A collection of HMMs constructed in HMMER3 [15] with manually curated cutoffs. The raw HMMER3 file is stored, and these HMMs also are integrated into two separate NCBI sources.
(4) A gene family hierarchy, which enables the accurate naming and identification of both novel and known protein sequences.

To reflect the constant change in the literature, curators continuously update these databases with new releases made approximately every 2 months.

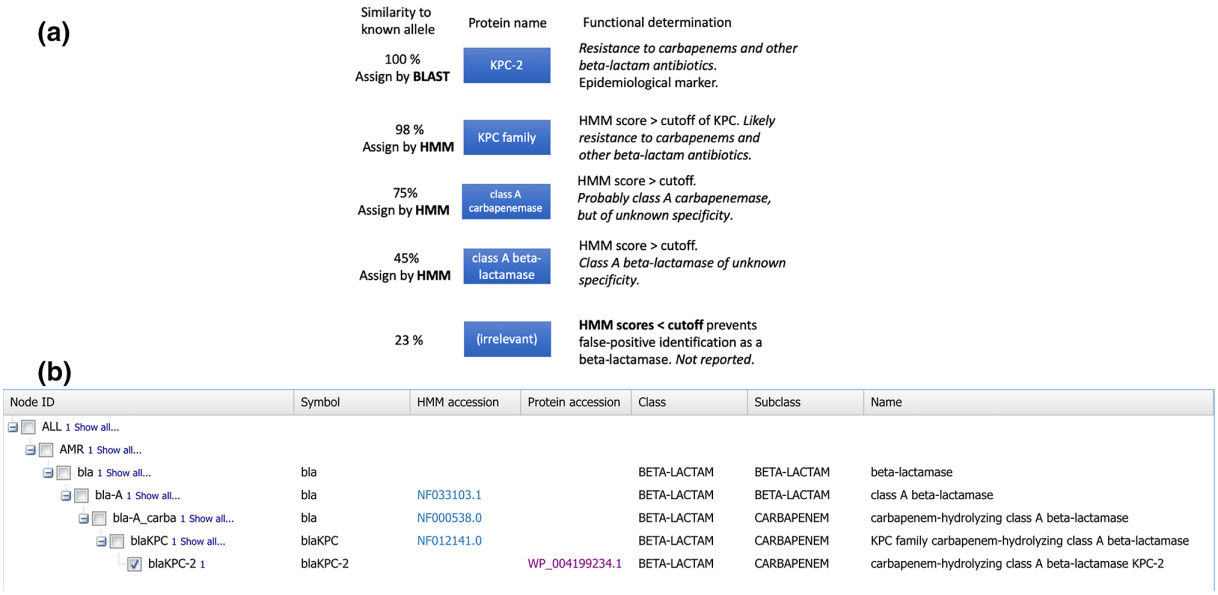


Fig. 2. . (a) Example of AMRFinderPlus' hierarchical structure, starting with *bla*$_{KPC-2}$ at the top and moving to less specific proteins. b) A screenshot showing how *bla*$_{KPC-2}$ is displayed in the Reference Hierarchy Viewers (https://www.ncbi.nlm.nih.gov/pathogens/genehierarchy/#blaKPC-2). Note that the uppermost 'bla' row is an organizational node, and lacks a HMM, so it is not represented in (a).

METHODS

Curation

Curation starts with the reporting of resistance and virulence mechanisms in the primary scientific literature. As described previously [9, 10], NCBI gathers novel genes and point mutations through a variety of mechanisms, including inter-organizational data exchanges, surveys of the literature, requests from collaborators and domain experts, pre-existing data sources, as well as requests for allele assignment of over 40 families of beta-lactamases, quinolone resistance genes (Qnr), and mobile colistin resistance genes (MCR).

We typically require experimental evidence or an extremely close identity hit to an experimentally verified protein for inclusion of acquired resistance alleles, genes and point mutations that are considered 'core' [10]. Curators do not require supporting evidence in the literature for allele assignment as a goal of resistance allele assignment is to encourage characterization and further study of the phenotypes and effects of these individual proteins (alleles are gene symbols that map to only one specific protein sequence) [16]. Here we describe specific features of the curation process for genes, alleles and point mutations.

Gene curation

For genes, curators will assign product names and a gene symbol (e.g. *bla*$_{TEM}$) and also ensure that the reference sequence is correct based on surveys of the literature and multiple sequence alignments. Then curators will edit the product name of the corresponding RefSeq protein record and identify a suitable RefSeq nucleotide sequence, which is subsequently assigned the curator-determined product name, gene symbol and allele designation. For genes that fall into the plus category – those genes related to biocide and stress resistance, general efflux, virulence, or antigenicity, or AMR genes whose phenotype is uncertain – curators assign product names and gene symbols, but no RefSeq nucleotide records are created.

Allele curation

When alleles are submitted to NCBI for allele assignment, we work with the submitter to rectify possible sequence problems. Specifically for alleles, curators compare submitted alleles to existing alleles in related families by sequence alignment [11, 12], as well as run AMRFinderPlus on these alleles to determine if the protein is full-length, functional, and if it belongs to an existing family. When a submission meets these requirements, then an automated process confirms that the allele is novel, and sequentially assigns it a new designation, preventing duplicate assignments. On occasion, researchers have assigned themselves alleles, creating conflicts, and NCBI curators attempt to resolve these conflicts.

Point-mutation curation

For point mutations, curators identify an appropriate reference protein or nucleotide sequence in GenBank, assign a name to be used by AMRFinderPlus, a mutation symbol that contains the location of the point mutation, and a gene symbol for the reference sequence. Point mutations are defined as *individual* differences from the reference. Our inclusion criteria for point mutations requires either experimental validation or strong phenotypic correlations that could not be accounted for by other resistance mechanisms.

Hierarchy node curation

Following sequence addition to the database, curators will develop either BlastRules or HMMs for a given node in the gene hierarchy. BlastRules are used when a small number of proteins exist for a gene, and novel proteins above a certain manually curated threshold are deemed members of that gene or gene family. If novel BlastRules are created or existing ones are updated due to addition of new sequences, these changes are added to NCBI's Protein Family Model database [17]. HMMs are built when there are multiple proteins with a gene or gene family, and they provide more evidence and resolution for inclusion within a gene or family. These HMMs are stored in NCBI's Protein Family Model database for use in PGAP [17]. We have described the details of HMM creation elsewhere [10]. Curators are alerted if there are conflicts between HMM results and the gene hierarchy; for example, curators are notified when a protein is hit by a HMM, but that HMM's gene family is not considered to be an ancestor of the node to which a protein is assigned.

Structural annotation

Along with functional annotation to determine if a genetic element should be included, curators assess structural annotation to determine that the sequence of the genetic element is appropriately characterized. For all alleles and genes, NCBI has a series of 54 automated quality-control checks (Table S1, available in the online version of this article) that are run during the release process to ensure the quality of each protein sequence added to the database. These checks serve as an automated backstop for the manual curation process. Some checks examine basic sequence problems, such as partial proteins that are misannotated, proteins with ambiguous bases, stop codons, frame shifts and other errors. Checks may flag proteins for quality reasons, such as partial proteins, determined either by the protein product name containing the word 'partial' or by comparison to similar proteins, and proteins with ambiguous residues. These issues often can be resolved by finding alternative sequences or contacting the original sequence submitters. Automatic and manually applied quality-control checks reduce the incidence of these errors. Alleles are automatically checked for possible errors, including ambiguous bases, frameshifts or stop codons.

Resolving nomenclature conflicts

Nomenclature conflicts and problems in the source literature also can affect gene databases. Multiple authors can use the same gene symbol for different sequences or genes, requiring curators to identify and resolve these conflicts. For example, when mobile colistin resistance genes were first identified, researchers unfortunately submitted different and diverse sequences with colliding gene names. As part of a large consortium, NCBI helped resolve these conflicts [18]. We have multiple checks that can identify such conflicts, summarized here and described in detail in Table S1. In addition, gene symbols are checked to ensure consistency and accuracy, and curators are alerted if gene symbols are non-standard or if they are linked to multiple product names.

Integration into other NCBI resources

These databases inform and are integrated into NCBI's annotation tools, such as PGAP [17] and RAPT (https://www.ncbi.nlm.nih.gov/rapt). An automatic system identifies any differences between the implementation of BlastRules and HMMs in AMRFinderPlus and their implementation in NCBI's Protein Family Model database so curators can harmonize these databases. All of these automated checks ensure that any obvious discrepancies will be reviewed by curators.

Testing of AMRFinderPlus

In addition, before release, developers initiate a series of over 200 tests of AMRFinderPlus using the release candidate database that can detect unwanted errors and highlight potential improvements curators might make (Fig. 3). These tests include processing all sequences from our internal database, all point mutations, a set of real public assemblies, varying software options, and known edge cases. All changes in AMRFinderPlus results from the previous database release are manually reviewed. New release candidates are generated, and all of the above QC tests are performed again until all tests pass and any automated test failures are manually reviewed by curators.

RESULTS

We have previously validated and published the AMRFinderPlus method, but we summarize briefly here. In one study, we compared the identification of AMR genes and point mutations to known phenotypes of 6242 isolates, and found that 98.4%

of phenotypes were consistent with the genotypic predictions [9]. In a separate study, we assessed the accuracy of the some of the plus genes, specifically metal resistance against a series of mercury-resistant isolates, and found a high correlation between genotypic predictions and observed phenotypes [10].

In addition, database curation is informed by the release of NCBI's publicly available AMRFinderPlus results. Where possible NCBI participates in nomenclature related collaborations; as described in Methods, NCBI was part of a multi-group effort to standardize mobile colistin resistance gene (MCR) nomenclature [18]. As a result, these updates were incorporated in the database and used in AMRFinderPlus analyses made publicly available through the Isolates Browser and MicroBIGG-E. Another example is a collaboration with the National Antimicrobial Resistance Monitoring System (NARMS) in the U.S. where NCBI Pathogen Detection found that *mcr-9* is the most common MCR variant in the US. NARMS initiated a phenotype study that found *mcr-9* did not confer resistance to colistin in over 100 natural *mcr-9* +isolates [19], despite laboratory evidence that *mcr-9* is capable of conferring resistance to colistin [20, 21]. Based on this finding, curators made the decision to move *mcr-9* to the plus category. This feedback loop between a publicly available database and analysis results and experimental verification is critical for improving our understanding of genotype-phenotype relationships and public health.

The publicly released data also enable analyses to better understand AMR trends and patterns. For example, a recent study used the Isolates Browser, which links AMRFinderPlus output with isolate metadata, to assess the global resistome and evolutionary epidemiology of multiple Enterobacterales species [22].

Database access and exploration

After new data is incorporated into internal curation tools, the data are released to the public, both as data files and in graphical user interfaces, as well as incorporated into NCBI's Pathogen Detection system (Fig. 3). AMRFinderPlus is run on over 1 million bacterial isolates and the results of these analyses are released in browsers and other formats. Here, we describe how these databases are made accessible to the public (Fig. 3).

One way for users to access the databases is through data files on our FTP site. These files can be of use for those who want to interrogate the data in detail or who want to use them as part their own data analysis process. From there, users can download the following as tab-delimited text files: the Reference Gene Catalog, the Reference Gene Hierarchy used by AMRFinderPlus, the allele counts of the AMR genes curated by NCBI by year. Users also can download the HMMs used by AMRFinderPlus and release notes describing the most recent curatorial changes. In addition, there are FASTA formatted files containing the protein and nucleotide reference sequences used by AMRFinderPlus.

NCBI has built multiple browsers to enable viewing and dynamic browsing of these various databases. A key new feature in all of the browsers is the ability to use hyperlinks that let users explore subsets of data identified in one browser in other browsers. Although we have described each of these databases as stand-alone entities, they have been designed to interact with each other, enabling links between the database views and the analytical views (Fig. 4, Table 1). We have described previously how both the Isolate Browser and MicroBIGG-E allow cross-browser selection, whereby sets of isolates or genes selected in one resource or the other allows selections in the other resource, either the set of genes encoded by the isolates, or the set of isolates that encode the genes, respectively [10], and so these will not be discussed here.

The Pathogen Detection Reference Gene Catalog (https://www.ncbi.nlm.nih.gov/pathogens/refgene/) provides a visualization of acquired genes and point mutations used by AMRFinderPlus. Each row represents an acquired protein sequence or point mutation (see Fig. 1). These fields have hyperlinks that connect to other NCBI databases such as GenBank and PubMed. Users also can use links to display every isolate with that genetic element in the Isolates Browser or to view all instances of that element in MicroBIGG-E (Fig. 3). For example, users in the Reference Gene Catalog can click a hyperlinked gene family symbol (e.g. *aac(3)-I*, and they will be automatically directed to MicroBIGG-E with every instance of *aac(3)-I* displayed. The Reference Gene Catalog also describes additional information about each sequence or point mutations such as the taxa for which detection of that sequence is excluded, information about its phenotype, location on the GenBank contig, a PubMed citation (if available), and other metadata (Fig. 4, Table 1). Users also can download these data in a table format.

The previously undescribed Reference Gene Hierarchy is a web-based view into the hierarchy of genes, families and upstream nodes that NCBI curators use to organize and relate the genes and HMMs in the Reference Gene Catalog and Pathogen Detection Reference HMM Catalog. This hierarchy drives the gene identification and naming algorithm of AMRFinderPlus. The Pathogen Detection Reference Gene Hierarchy provides the link between proteins, HMMs and protein names for those proteins that do not have an exact match in the Reference Gene Catalog. In the Reference Gene Hierarchy Viewer, this can be used (Fig. 2b) to better interpret AMRFinderPlus output and understand the organization of the data. For instance, in the Reference Gene Hierarchy, after searching for 'bla_{KPC}', which returns the entire family of KPC-family carbapenemases (https://www.ncbi.nlm.nih.gov/pathogens/genehierarchy/#blaKPC-2), users can see this family displayed in the Reference Gene Catalog, see every instance of a KPC-family carbapenemase displayed in MicroBIGG-E, or find which isolates in the Isolates Browser have a KPC-family carbapenemase.

Table 1. Field-specific links within browsers. 'Browser' describes the browser. 'Field in browser' describes the specific column with the hyperlink. 'Function' describes what is displayed upon selecting the hyperlink

Browser	Field in Browser	Function
Isolates Browser	Assembly	Links to Assembly record for that page
	BioSample	Links to BioSample page for that BioSample
	BioProject	Links to BioProject record for that BioProject
MicroBIGG-E	Assembly	Links to Assembly record for that assembly
	BioSample	Links to BioSample record for that BioSample
	BioProject	Links to BioProject record for that BioProject
	Closest Reference Accession	Links to Protein record of closest reference protein
	Contig	Links to Nucleotide record of contig containing the element
	HMM Accession	Links to HMM record in Protein Family Model database
	Protein	Links to Protein record
	PubMed ID	Links to related PubMed record(s) for that genetic element
	Start/Stop	Links to Nucleotide record of contig containing the element but displays only the element itself
Reference Gene Catalog	Allele	Links to all isolates in the Isolates Browser containing that allele
	Gene Family	Links to all isolates in the Isolates Browser containing that gene family
	GenBank Nucleotide Accession	Links to the GenBank Nucleotide Record for that element
	GenBank Protein Accession	Links to the GenBank Protein Record for that element
	Hierarchy Node ID	Displays that node in the Reference Gene Hierarchy
	PubMed ID	Links to related PubMed record(s) for that genetic element
	RefSeq Nucleotide Accession	Links to the RefSeq Nucleotide Record for that element
	RefSeq Protein Accession	Links to the RefSeq Protein Record for that element
Reference Gene Hierarchy	HMM Accession	Links to HMM record in Reference HMM Catalog
	Protein	Links to Protein record
Reference HMM Catalog	Accession	Links to HMM record in Protein Family Model database
	MicroBIGG-E	Displays all genetic elements identified by the HMM in MicroBIGG-E

The Pathogen Detection Reference HMM Catalog is a new web-based portal to our curated database of reference HMMs used by AMRFinderPlus in concert with gene sequences in the Pathogen Detection Reference Gene Catalog to identify AMR genes as well as some stress resistance and virulence genes. This is a highly curated subset of the HMMs included in NCBI's Protein Family Models database [17]. Every row in the Pathogen Detection Reference HMM Catalog is an individual HMM. Details including the seed alignment and HMM profile are available by clicking on the HMM accession in the table.

Computed analyses

In addition to the AMRFinderPlus databases, NCBI also runs AMRFinderPlus on genomes belonging to 47 bacterial taxonomic groups as part of NCBI's Pathogen Detection project [23]. The Isolates Browser (https://www.ncbi.nlm.nih.gov/pathogens/isolates/) displays a summary of AMR, stress response and virulence genes for each isolate of interest, where each row describes an isolate assembly. When submitted to NCBI antibiotic susceptibility testing data can be linked to each isolate. These data can be downloaded for further analysis. The Microbial Browser for Identification of Genetic and Genomic Elements (MicroBIGG-E; https://www.ncbi.nlm.nih.gov/pathogens/microbigge/), displays AMRFinderPlus results for those isolates that have genomic data deposited in GenBank, displayed in a more comprehensive format resembling the AMRFinderPlus output. In this case each row is a detected gene or point mutation. Information about how it was identified, links to the sequence, phenotype and reference data are included as well as isolate metadata such as BioSample and strain names, and isolate source that are also shown in the Isolates Browser. Sequences and table data can also be downloaded in bulk here. Importantly, both of these views display and

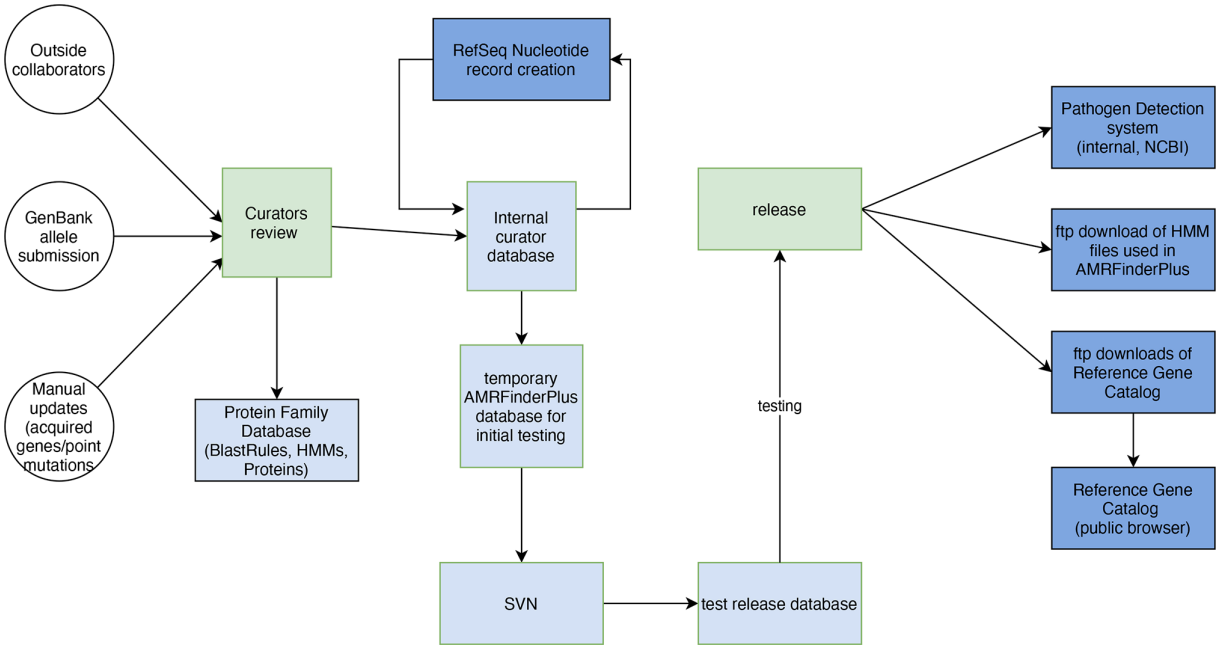


Fig. 3. How curators update and release AMR database products. Circles are data sources, squares are processes, and diamonds are tests. Orange is for curator supervised steps, light blue is for internal-only, light green is other NCBI resources and databases, dark blue is for public access to complete data files, and dark green is for public access through web interfaces. 'SAUTE guided assembler' refers to a set of non-redundant nucleotide sequences derived from the Refseq nucleotide sequences of acquired AMR genes and some virulence genes deemed of critical importance. These sequences are used by the SAUTE guided assembler in the Pathogen Detection assembly process to ensure assembly of these critical genes.

store the versions of the software and database that was used to analyse each isolate so users can determine if analyses were run with the database version containing their genes or mutations of interest since new genes and point mutations are added with each release and could affect the computed results. At each release, release notes indicate changes so users can determine if they deem it necessary to rerun AMRFinderPlus themselves with newer database or software version.

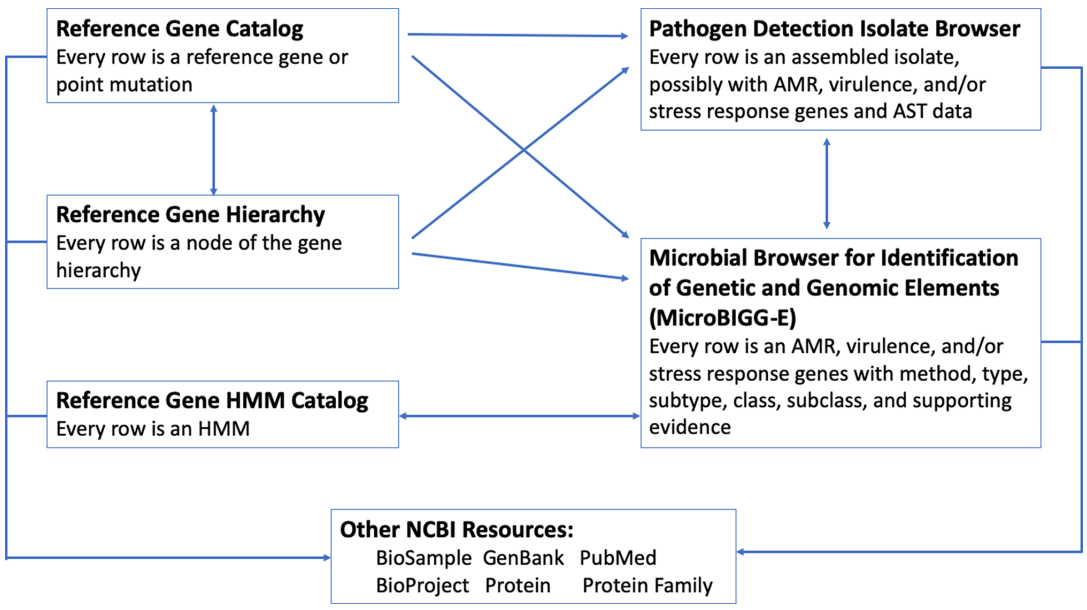


Fig. 4. Interactions among database viewers and existing NCBI resources. Arrows represent links resulting from selection of individual field values or cross-browser selection.

These computed analyses have been used successfully by other researchers. For example, one group used the output from Micro-BIGG-E to assessing the health risk of antimicrobial resistance genes [24]. These tools also have been used to survey *Acinetobacter* sp. for resistance mechanisms [25]. MicroBIGG-E enables researchers to collect a diverse set of AMR-related sequences, as was done in a study examining the function of erythromycin resistance methyltransferases [26].

DISCUSSION

Databases constructed for analytical tools are not only part of those tools, but are useful to researchers themselves. Here, we have described four primary components of the AMRFinderPlus database curated by NCBI. Inclusion requires either experimental evidence or strong correlations between genotypes and phenotypes. The database undergoes multiple quality-control checks that ensure data reliability and accuracy. We have also developed multiple online tools that make it easier for researchers to access the databases and examine their structure and data, as well as subdivide the data for further use. These databases are accessible in multiple formats and locations.

One advantage AMRFinderPlus has is that many of its components are integrated into other NCBI-maintained systems. The Protein Family Model database contains both the Blast Rules and HMMs used by AMRFinderPlus, and is the backbone of NCBI's PGAP tool, which is used for annotation at NCBI [17, 27]. Since all sequences in the Reference Gene Catalog are accessioned in GenBank, these will remain archived. The tight integration of the AMRFinderPlus database components with other NCBI systems along with being an integral component of NCBI's Pathogen Detection project and a part of the U.S. National Action Plan for Combating Antibiotic-Resistant Bacteria [28] further ensures that the underlying databases and their component data will continue to be available.

While we have an expansive database covering 5940 AMR genes, 914 point mutations, 233 stress response genes and 716 virulence genes, we recognize that there are still significant gaps. One way NCBI has attempted to fill the gaps is through collaborations with outside groups. For example, NCBI worked with multiple U.S. government agencies to add stress response and virulence genes found in food-borne pathogens that are critical for understanding the links among AMR, stress response (e.g. biocides) and virulence [10]. Not only does our database rely on these collaborators as well as the published literature, but we also use publicly available forums such as CARD's amr_curation discussion group (https://github.com/arpcard/amr_curation) and requests from outside investigators via the Pathogen Detection project (pd-help@ncbi.nlm.nih.gov) to improve and expand the databases. NCBI invites domain experts to contact us if they have requests to support organisms of interest, and we have multiple collaborations to expand and improve databases [18]. We also engage in data harmonization with CARD [29] and other sources.

In collaboration with our interagency partners, we have prioritized food-borne pathogens and high-priority AMR pathogens, so many of the 47 bacterial taxa in the Pathogen Detection system do not have taxon-specific virulence factors. In addition, only 14/47 taxa currently have point mutations, and AMRFinderPlus currently does not cover *Mycobacterium tuberculosis* point mutations at all. These are all directions for future work to improve the AMRFinderPlus database. Given the importance of bioinformatic approaches to combatting the problem of AMR, we hope that these databases will assist other researchers in their own efforts to combat this problem.

Funding information
This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

Acknowledgements
The authors would like to thank DeAnne Cravartis and Martin Shumway for comments on earlier drafts of the manuscript.

Author contributions
Conceptualization: V.B., M.F., D.H.H., W.K. and A.B.P. Data curation: V.B., M.F., D.H.H. and A.B.P. Methodology: V.B., M.F., D.H.H. and A.B.P. Software: V.B., B.F. and A.B.P. Project administration: W.K. Supervision: W.K. Visualization: M.F., A.B.P. and W.K. Writing – original draft: M.F. and A.B.P. Writing – review and editing: D.H.H., W.K. and A.B.P.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;399:629–655.
2. Rehman MA, Yin X, Persaud-Lachhman MG, Diarra MS. First Detection of a Fosfomycin Resistance Gene, *fosA7*, in *Salmonella enterica* Serovar Heidelberg Isolated from Broiler Chickens. *Antimicrob Agents Chemother* 2017;61:e00410-17.
3. Mellmann A, Bletz S, Böking T, Kipp F, Becker K, *et al*. Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting. *J Clin Microbiol* 2016;54:2874–2881.
4. Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, *et al*. Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in *Campylobacter* spp. *Appl Environ Microbiol* 2016;82:459–466.

5. Allard MW, Bell R, Ferreira CM, Gonzalez-Escalona N, Hoffmann M, *et al*. Genomics of foodborne pathogens for microbial food safety. *Curr Opin Biotechnol* 2018;49:224–229.
6. Moran RA, Anantham S, Holt KE, Hall RM. Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal *Escherichia coli* using PCR or WGS. *J Antimicrob Chemother* 2017;72:700–704.
7. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, *et al*. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 2021;6:960–970.
8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
9. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, *et al*. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 2019;63:11.
10. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, *et al*. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* 2021;11:12728.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
13. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, *et al*. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 2020;75:3491–3500.
14. Seemann T. tseemann/abricate; 2021. https://github.com/tseemann/abricate
15. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol* 2011;7:10.
16. Bush K, Jacoby GA. Updated functional classification of beta-lactamases. *Antimicrob Agents Chemother* 2010;54:969–976.
17. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, *et al*. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;49:D1020–D1028.
18. Partridge SR, Di Pilato V, Doi Y, Feldgarden M, Haft DH, *et al*. Proposal for assignment of allele numbers for mobile colistin resistance (mcr) genes. *J Antimicrob Chemother* 2018;73:2625–2630.
19. Tyson GH, Li C, Hsu CH, Ayers S, Borenstein S, *et al*. The *mcr-9* gene of *Salmonella* and *Escherichia coli* is not associated with colistin resistance in the United States. *Antimicrob Agents Chemother* 2020;64.
20. Carroll LM, Gaballa A, Guldimann C, Sullivan G, Henderson LO, *et al*. Identification of novel mobilized colistin resistance gene *mcr-9* in a multidrug-resistant, colistin-susceptible *Salmonella enterica* Serotype Typhimurium Isolate. *mBio* 2019;10:e00853-19.
21. Kieffer N, Royer G, Decousser J-W, Bourrel A-S, Palmieri M, *et al*. *mcr-9*, an inducible gene encoding an acquired phosphoethanolamine transferase in *Escherichia coli*, and its origin. *Antimicrob Agents Chemother* 2019;63:e00965-19.
22. Osei Sekyere J, Reta MA. Global evolutionary epidemiology and resistome dynamics of *Citrobacter* species, *Enterobacter hormaechei, Klebsiella variicola*, and Proteeae clones. *Environ Microbiol* 2021;23:7412–7431.
23. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2016;44:D7-19.
24. Zhang A-N, Gaston JM, Dai CL, Zhao S, Poyet M, *et al*. An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat Commun* 2021;12:4765.
25. Kyriakidis I, Vasileiou E, Pana ZD, Tragiannidis A. *Acinetobacter baumannii* antibiotic resistance mechanisms. *Pathogens* 2021;10:373.
26. Sharkey RE, Herbert JB, McGaha DA, Nguyen V, Schoeffler AJ, *et al*. Three critical regions of the erythromycin resistance methyltransferase, ErmE, are required for function supporting a model for the interaction of Erm family enzymes with substrate rRNA. *RNA* 2022;28:210–226.
27. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, *et al*. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res* 2018;46:D851–D860.
28. The Federal Task Force on Combating Antibiotic-Resistant Bacteria. National action plan for combating antibiotic-resistant bacteria 2020-2025. 2020.
29. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2020;48:D517–D525.
30. Mahfouz N, Ferreira I, Beisken S, von Haeseler A, Posch AE. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J Antimicrob Chemother* 2020;75:3099–3108.