# Translating the Comprehensive Antimicrobial Resistance Database to act as a Stopgap for MEGARes

Marcel J. Jansen, Brian P. Alcock,

Amogelang R. Raphenya, Dr. Andrew G. McArthur

April 2020

**Abstract**

Antimicrobial resistance (AMR) is quickly rendering many vital drugs useless. AMR databases keep up-to-date records of bacterial resistomes, allowing clinicians and researchers quick access to AMR data. One of those databases, MEGARes, fell out of date for over two years, leaving the users of its AMR++ software without an accurate source of reference sequences. This project translated the CARD AMR database into a format which AMR++ could read, providing a stopgap measure in the event a gap in MEGARes curation appears again. The effectiveness of MEGARes and the translated database were compared. Significant variation in the results suggested that further analysis is required before translated CARD can confidently be used as a replacement for MEGARes in AMR++.

## 1 Introduction

As antimicrobial resistance (AMR) becomes increasingly prominent worldwide, the medical community struggles with infections that have not beset their patients in generations. Procedures that were once mundane have become risky due to the potential for infection by resistant microbes. Surveillance and prediction are critical to combating this rapidly-evolving threat.

## 1.1 Antimicrobial Resistance

As organisms, pathogenic bacteria evolve in response to environmental pressures. Some randomly develop mutations that make them resistant to various antibiotics, whether human-made or naturally-occurring. Should the environment suddenly change to contain large quantities of such a compound, any unresistant microbes are negatively affected or killed outright, allowing the resistant strain to gain a foothold. Thus, it makes up a larger portion of the microbial population within that environment [1]. This is known as antimicrobial resistance (AMR). Excessive antibiotic use, especially for broad-spectrum antibiotics, increases the risk of a pathogen developing antibiotic resistance [2]. Additionally, microbes can engage in horizontal evolution, transferring resistance genes to one another or collecting them from dead, previously resistant microbes[1]. The collective set of resistances available to a population of microbes is known as its resistome[3].

Powerful antibiotics like carbapenems and polymyxins are known as last-line agents [4, 5]. They are employed when infections have proven resistant to all other antimicrobial drugs. Bacterial populations are rapidly developing resistances to these drugs of last resort, with fewer companies willing to provide the necessary resources to develop new therapies [4–6].

The global health implications of this are clear. According to a report by the World Health Organization's Interagency Coordination Group (IACG) on Antimicrobial Resistance, deaths associated with resistant strains are already on the rise. Across the globe today, 700 000 individuals die of drug-resistant infections annually. By 2050, that number is expected to rise as high as 10 000 000 annual deaths, with 700 000 of those deaths being North Americans or Europeans [7]. Within the next thirty years, 2 400 000 deaths are expected in North America, Europe, and Australia alone [6, 8]. Although the specific numbers have been disputed [9], AMR is universally recognized as a considerable threat to global health.

According to the IACG, the most effective way to combat AMR is through surveillance. Access to up-to-date resistance data is a critical component of such surveillance [10]. AMR databases make this vital knowledge easily accessible to clinicians, researchers, and policymakers. For example, AMR databases can provide a clear understanding of a drug's targets, allowing clinicians to be more specific with their drug choices [3]. Broad-spectrum antibiotics increase the risk of resistance, so allowing clinicians to tailor drugs to their patients' infection can both improve clinical outcomes and reduce evolution rates

[1, 6, 10]. Some bacteria, such as *E. coli* O157:H7, produce dangerous toxins in response to antibiotic treatment[11]. In such cases, knowing the specific genotype of the infectious agent is vital to patient well-being. In addition to informing clinical decisions, surveillance is also critical to decision-making at governmental levels, allowing policymakers to make informed choices when implementing interventions [10].

## 1.2 Ontologies

Ontologies are a hierarchical structure for sorting and categorizing concepts and their relationships with one another [12]. The resultant web of terms and relationships can be used to display a holistic representation of an idea. Each idea is known as a ”term”, and each type of connection between terms is known as a ”relationship”. Each relationship connection has one subject (child term) and one object (parent term). The subject is a sub-term of the object. A simple example is the phrase ”I love you”. The ”I” term is the subject, ”you” is the object, and ”love” is the relationship (See Fig. 1). Conceivably, there could be multiple terms that ”is_loving” the ”you” term, such as your friends, family, or cat. Each of these could be a sub-term of ”you” and would be connected to it by an ”is_loving” relationship.

For a more complex example, take a vehicle ontology. One could break vehicles into three classes: land, sea, and air (See Fig. 2). Each of these can then be further subdivided. In this example, there are only two relationships: ”is_a” and ”transported_by”. Every term - excluding the root term - must have an ”is_a” relationship. A car is a land vehicle, which in turn is a vehicle. In an ontology, it is possible for a term to have multiple parents. Take the hovercraft, for example. It is both a land vehicle and a water vehicle.

The ”transported_by” relationship identifies that a sub-term can be loaded onto and moved by a parent object. For example, a car is small enough to be transported by a cargo ship, but a passenger plane is not. This relationship carries not only classificational information, but also functional information. It tells you what a term can do, not only what it is.

Ontologies are also known as Controlled Vocabularies, so-named because they restrict the number of words that can be used and predetermine their definitions. The goal of any controlled vocabulary is consistent language processing. In natural English, a word can have multiple meanings depending on context. This is not typically the case in a controlled vocabulary. This rigidity is vital because it reduces the risk of miscommunication and allows computers to easily comprehend the terms[12]. For example, if CARD is used to analyze a query gene, when it finds a match, it can now classify the new sequence using all of the ontological information connected to the reference sequence. Take, for instance, the following scenario: A user inputs a query sequence which is a new mutant of an aminocoumarin-resistant gyrB. Once compared, CARD's ontology is able to infer that the gyrase derives from an aminocoumarin-sensitive gyrB, or that this gyrB is an antibiotic-resistant DNA topoisomerase subunit, or any other predefined term which falls into the web of terms connected to ”aminocoumarin-resistant gyrB” [3]. This may seem simple to the language-focused human brain, but computers struggle immensely with such information, hence the need for curation. Other resistance databases lack this level of granularity, often storing this information in annotations, which are inaccessible to the database itself. Such in-depth classification benefits standardization as it keeps all information classified within strict terminology.

Ontologies can be used for numerous applications, including categorizing databases.

## 1.3 AMR Databases

Tracking the resistome is complex. The development of highly efficient whole-genome and community-based sequencing have contributed to an enormous quantity of data being rapidly generated [10, 13, 14]. The distributed nature of this data makes it extremely difficult to search for relevant information. For example, if a researcher performs a whole-genome sequencing of a bacterium and believes they have found a new resistance gene, it would be impractical for them to search the literature for every homologue. AMR databases make it far simpler for researchers, medical professionals, and regulatory bodies to track the resistome without needing to conduct a continuous and extensive review of the literature.

There are numerous ”boutique” databases, which track resistance data for specific species, such as Tuberculosis. However, there are three major comprehensive databases: CARD [15], ResFinder[16], and ARG-ANNOT[17]. The NCBI database feeds off of these to form a meta-database. As an offshoot of CARD, MEGARes is also considered to be a comprehensive database. Comprehensive databases seek to collect all available resistance data in a single, easily-searched location. Of the ”big three”, CARD is the only one which is ontology-focused, while ResFinder and ARG-ANNOT are sequence-focused. In the latter two, curators attach their sequence data to a term, but any ontological data is optional. CARD, on the other hand, requires curators to fit any new entries into a set of classification tags, detailed below. This takes additional time, but ensures that CARD is ontologically complete, with every term being linked back to the root term of the ontology. This maintains consistency and accessibility. This concept is detailed further below.

### 1.3.1 CARD and RGI

CARD uses four ontologies: The Antibiotic Resistance Ontology (ARO) to organize the definitions of its AMR terms, the Model Ontology (MO) to collect the models which are used to interpret AMR data, the Relations Ontology (RO) which contains the eight ARO relationships, and the NCBI Taxonomy Ontology for classifying pathogens[15, 18].

The ARO is the most critical ontology and has seven major branches (See Fig. 3) which then split off into numerous subcategories. As the as-yet-unpublished CARD 2020 manuscript details, The three most important branches are *Determinant of Antibiotic Resistance*, which contains the genetic and phenotypic causes of resistance, *Antibiotic Molecule*, which contains all classes of antibiotic, and *Mechanism of Antibiotic Resistance*, which describes how the determinants perform their resistance-producing function. The other four branches contain related topics in AMR: the original non-resistant antibiotic targets, the cellular processes which produce antibiotics, agents which modify resistance, and a catch-all branch for terms relating to AMR research. [3, 15, 18].

The ARO contains fifteen relationship types (See Fig. 4), the most important of which is *is_a*. Using a series of *is_a* relationships, any term can be traced back to the root term ”process or component of antibiotic biology or chemistry” (ARO:1000001). Thus, *is_a* forms the core of each branch of the ARO. The other nine relationships allow for connections between the seven main branches [3]. This is one of the key factors separating CARD from MEGARes. Based on Directed Acyclic Graphs instead of ontologies, MEGARes’ terms can have only one parent.

The names of the ARO’s relationships have changed slightly over time [3, 15, 18], however their definitions have remained functionally the same (See Fig. 4). Aside from *is_a*, the two most common relationships are *confers_resistance_to_drug_class* and *participates_in*. These connect gene families to the drug classes they resist and AMR genes to their mechanisms of resistance, respectively These are the up-to-date definitions of these relationships. Usage has changed since the 2013 paper. [15, 18].

In order to cope with the enormous quantities of data being published daily, as well as the ever-evolving nature of prokaryotic resistance, databases like CARD must be continuously updated in order to remain useful[12, 19]. Ontology designers need to try to predict what their vocabulary will be used for, but must update it if new ideas appear. Curators use vocabulary terms to classify sequences found in reference papers. Active curation involves reviewing literature and adding relevant genes to the database, along with relevant relationships. Critical information like gene or protein sequences, relevant resistomes, and detection models are added as annotations [3]. These annotated sequences are known as reference sequences.

There are four major classification tags: Gene family, Resistance Mechanism, Affected Drug Class, and Specific Antibiotic. All AMR genes must be classified using the first three tags and work is currently underway to add connections between all possible AMR genes to their specific antibiotic. Therefore, all properly-curated AMR determinants must be assigned a tag from each of those four classes [18]. Additionally, each sequence is assigned a gene name. This is used to tag the sequence in its database header.

Classification is not the sole purpose of these databases. External software is used to interpret and apply them. MEGARes uses AMR++ and Meta-MARC, which are detailed below. CARD uses the Resistance Gene Identifier (RGI). One of CARD's four ontologies, the Model Ontology, contains the models with which its genes can be analyzed. RGI uses these models to determine whether or not a gene is considered to be contributing to resistance. The protein homology model, for example, compares test sequences against the reference sequence stored in CARD. RGI has three levels of accuracy with which it can search: Perfect, Strict, and Loose. If a 100% match between the reference and query sequences is required, Perfect is used. If some variation is acceptable, for example if a query sequence is believed to be a minor variant of an existing gene, Strict is applied. Loose is useful for detecting potential resistance mutations which have yet to be discovered *in vivo*, but tends to produce more false positives [3, 18, 19]. These three algorithms can be applied to any model within the model ontology as long as that model supports them. For example, the protein variant model cannot have a perfect match [15].

### 1.3.2 MEGARes, AMR++, and Meta-MARC

MEGARes is a conglomeration of NCBI's Bacterial Antimicrobial Resistance Reference Gene Database, ResFinder, BacMet, CARD [20, 21]. As a result, it contains 7,868 unique reference sequences to CARD's 2,984, over two and a half times as many (as of April 2020) [15, 21].

MEGARes' classification system is far simpler than that of CARD, using a Directed Acyclic Graph, or DAG. In a DAG, each term can have only one parent, similar to a phylogenetic tree. Compared to the numerous hierarchical levels in CARD, MEGARes has only three: Drug Class, Mechanism, and Group. Drug Class contains the broad antibiotic classification to which resistance is provided. The next level down, Mechanism, indicates how that resistance is biochemically facilitated. The lowest level, Group, provides the genetic regions which contribute to that resistance mechanism and contains the list of specific sequences [20–22]. These groups correspond to CARD's gene families. Collectively, groups and families can be referred to as "bins". Drugs are not the only antimicrobial agents that are developing resistance. An additional level of classification, Type, was added in MEGARes 2.0 [21]. This allows differentiation between resistance to antibiotic drugs, metals, biocides like peroxides and alcohols, or any combination of the three. All non-drug resistance genes in MEGARes come from BacMet [21]. CARD exclusively focuses on drug resistance, so any comparison between the two databases must be done in terms of drug resistance only.

Different databases are better suited to different tasks. MEGARes' simple tree structure is far more computationally inexpensive, making it preferable for count-based analyses, for example. These determine the number of reads that align with predetermined reference genes. Because each child term has multiple parents in an ontology, count-based analyses can cause inflated results when applied to ontology-based databases [20]. However, accurate representation of biological classification can be lost, and MEGARes makes sure to warn users that it is not meant to replace databases that provide more detailed information [20].

To perform its alignments, MEGARes' uses two pieces of software: AMR++ and Meta-MARC [20]. These fill a similar role to RGI, allowing a user to input DNA sequences and receive information regarding that sequence's relevance to AMR. Unlike RGI, AMR++ is restricted to a single alignment method, the same Burrows-Wheeler-Aligner which the RGI Protein Homology model uses [15, 20]. Meta-MARC, on the other hand, uses Hidden Markov Models to extend the reach of its search beyond that of a Burrows-Wheeler transform, allowing it to identify potential AMR resistance genes which have not been previously identified [21, 23]. This is inherently prone to false-positives, as those genes have not been lab-tested. Given the same data, RGI's protein homology model and AMR++ produces the same output. RGI, however, can alternatively use any other model which is present in the Model Ontology [15]. Both are able to manage metagenomic information, however AMR++ has historically been more commonly used for such work.

Although originally adapted from data stored in CARD, much of MEGARes' data was manually transferred and reformatted from one to the other, with significant portions being done manually. As a result, it has been updated only once since its release. It launched in August of 2016. The final update for version 1.0 was on December 2016 [22]. As an AMR database, this resulted in it falling behind the most recent research. A totally revamped version, MEGARes 2.0, was released at the end of 2019 [21], a nearly three-year hiatus of updates. For reference, CARD is typically updated quarterly to keep pace with advancements in AMR detection [15].

## 1.4 MEGARes-CARD Comparability

The databases are each contained in a FASTA file where each entry is composed of a DNA sequence that provides resistance and a header that contains its classification information. CARD's headers are structured like so:

**sequence source|DNA Accession|Directionality(+/-)|Gene Start-Gene Stop|ARO:###|Gene [species of origin]**

MEGARes's headers are structured like so:

**MEG_###|Type|Class|Mechanism|Group|(optionally)RequiresSNPConfirmation**

The octothorpes (#) indicate numbers used to identify database entries. They are always unique in MEGARes's case, but not in CARD's case.

There are some key differences between the two. First, CARD uses a GenBank DNA accession where MEGARes assigns each entry a unique number. Second, CARD identifies each sequence with a specific gene where MEGARes merely provides the group. Finally, MEGARes includes a "type" identifier for differentiating between resistance to drugs, biocides, metals, or a combination of those three. CARD tracks only drugs, which becomes important when comparing count data from the two databases. CARD's header does not contain drug class information. Other distinctions like CARD's inclusion of directionality and gene location are not relevant to this project.

In addition to the database FASTA file, each database also has annotation files which contain the header information without sequence information. In CARD's case, it has an index file which contains many pieces of information about each sequence, including the DNA Accession, class, mechanism, gene, and gene family information of each determinant.

Because MEGARes and CARD vary significantly in their curation methodologies, identical sequences can be classified very differently. Families and groups could overlap to varying degrees and the classification systems of each database may represent some areas more accurately than others. A group and a family overlapping means that they share at least one sequence. If a group overlaps entirely with a family, that means that the family contains all of the sequences that the group contains, but it could also contain more sequences, which could in turn be shared with other groups. This leads to a complex web of possible relationships between bins. Prior to this project, both teams behind MEGARes and CARD were aware that significant differences existed between their classification systems, but the extent of those differences was not well defined.

MEGARes' focus on computational simplicity results in some differences in drug classes as well. For example, it has a single class for multidrug-resistant determinants; These are determinants that provide resistance to more than two drugs. Strangely, MLS, which is a combination of three classes, is treated as a separate class. CARD, on the other hand, tracks all its multidrug-resistant classes independently by combining them into a cluster of classes separated by semicolons. MLS stands for macrolide-lincosamide-streptogramin resistance, so MLS would be macrolide;lincosamide;streptogramin in CARD. macrolide-lincosamide-streptogramin-penam resistance, however, would be classed as multidrug resistant in MEGARes. In CARD it would be macrolide;lincosamide;streptogramin;penam.

Other, smaller differences exist. Certain CARD families and MEGARes groups are ”prime” or ”double-prime” versions of other groups or families. However, CARD uses the symbols ‘ and “ respectively to indicate these families. MEGARes uses -PRIME and -DPRIME. Additionally, AMR++ treats spaces as the end of a line while RGI does not. Thus, CARD annotations contain spaces while MEGARes annotations use underscores.

### 1.5 Goals

The goals of this project were to:

1. Write software to translate CARD's Protein Homolog Model headers into database and annotation files that AMR++ can read and use to generate resistome information from a sample
2. Determine the major differences between CARD and MEGARes' structures
3. Assess the usefulness of translated CARD as a replacement for MEGARes data in AMR++

## 2 Methods

The goal of translation was accomplished by writing a Python script to combine the data into a header identical in structure to that of MEGARes and then search through CARD's database, identify each entry, and assign it that new, AMR++-compliant header. GenBank DNA Accessions included in the index file are used in place of MEG's unique identifier number. These were used because they are almost always unique while making it possible to identify each entry and search through the database. Only protein homolog model entries were converted, as the PHM is the most similar in its application to MEGARes' structure.

The new headers were assembled from the DNA Accession, Class, Mechanism, and Family columns of the index file. A type column was added to conform to MEGARes standards, but all entries were assigned the ”Drugs” type, as CARD only contains drug resistance genes. The final translated headers have the following format: **DNA Accession|Type|Class|Mechanism|Family**

Because the CARD index file and database file are ordered differently, the database entries had to be found by DNA accession and gene family instead of by numerical order. CARD database headers contain only gene information and not gene family information, so the header's gene information was used to match each entry to its new annotation. If two sequences have different genes that come from the same gene family, however, it resulted in the creation of duplicate database headers. For AMR++ to function, every entry in an AMR database must have a unique header, so the software removed any entries with duplicate headers. They must also have the same resisted classes and mechanisms for this to be an issue. This conversion from gene to gene group resulted in a loss of 59 database entries, or 2.2% of the database. See supplementary item 8 to see the list of culled entries.

To ensure compatibility and improve comparability, the CARD prime symbols are converted to MEGARes format. The translator also replaces the spaces present in CARD's headers with underscores. Step-by-step details on the function of the translator in the README included with the translator (See supplementary item 1).

I evaluated the goal of assessing major differences by writing a comparison program that first identified how many groups overlap with each family and how many families overlap with each group. For example, multiple groups might be perfectly contained by family A, but another group might be spread between families A, B, and C. Ideally, there would be one group for every family and vice versa, suggesting perfect overlap. As will be seen later, this is not the case. Second, given that there are multiple families per group and multiple groups per family, the software determined which specific groups overlapped with which families and vice versa (See Fig. 5 and 6).

I wrote the translation and comparison software in Python with the packages csv, Pandas, numpy, datetime, Biopython, re, and argparse. Because AMR++ runs on a Linux server, the program dos2Unix was run on the translated files to make them AMR++-legible.

I evaluated CARD's usefulness as a MEGARes replacement by the similarity between AMR++'s output when using MEGARes versus translated CARD on the same wastewater genome, ERR1713335 [24]. If AMR++ produced a nearly identical number of counts when using either one as its source of reference sequences, then CARD would be identified as a reasonable replacement for MEGARes. Ideally, there should be similar counts of both related families/groups and the same drug classes. If each database found different levels of the same drug classes, then they would be deemed to have significant classification differences. Mechanism was not a point of comparison. CARD and MEGARes differ too much in how they define resistance mechanisms for that to be a useful measure of similarity. For the count information, I removed AMR++ results that had a non-Drug type because CARD tracks only drugs, not biocides or metals. Some MEGARes determinants operate on multiple types, so some determinants that act on drugs were lost in the process. MEGARes' structure did not contain the specific classes that these entries provide resistance to, so the only lost data which was comparable to CARD was the group information.

# 3 Results

## 3.1 Classification Comparison

When analyzing the differences between the two classification systems, I found that Mechanisms are incomparable between the two. CARD uses the physical mechanism of action, such as antibiotic efflux or inactivation. MEGARes, on the other hand, uses the biochemical mechanism of action, such as protein pumps. MEGARes groups and CARD families are similar in structure and purpose, allowing for comparison between them as long as care is taken to keep track of how each group and family relate to one another. Bins can overlap (or not) in very unpredictable ways, and may even fall into a grey area of describing mechanisms instead. For example, Rifampin Phosphotransferases are a CARD family, but a MEGARes mechanism. The effects that such fundamental structural differences would have on output when used with AMR++ is difficult to accurately determine. Drug classes are probably the most similar structural component between the two, although the two databases have their own ways of categorizing multidrug resistance.

The output of the group to family comparison software was dense and complex, and a great deal more analysis could still be done. MEGARes contains over 1300 gene groups, and those that relate to CARD can overlap totally or partially with CARD families [22]. Ideally, every one of these differences would be analyzed on a case-by-case basis, but the sheer volume of groups, families, and their overlaps requires a more general viewpoint based on a few cases. See supplementary items 2-5 for raw comparison data.

Many more MEGARes groups fit into a single CARD family than vice versa, suggesting that each CARD family encompasses a larger proportion of CARD's AMR space than each MEGARes group does. Only part of MEGARes overlaps with CARD. Of 554 groups that overlapped in any way with CARD families, 537, or 97%, overlapped with a single family (See Fig. 6). Meanwhile, only 163 families overlapped with a single MEGARes group. This suggests a many-to-one relationship between MEGARes and CARD, where many MEGARes bins fall under a single CARD bin. Many CARD families covers a broad collection of MEGARes groups. More CARD families contain sequences that are in multiple MEGARes groups and each family contains many groups. The most families that a MEGARes group overlapped with was seven, while the maximum number of groups in a single family was 72, associated with the *major facilitator superfamily (MFS) antibiotic efflux pump* CARD family (See Fig. 5).

There are many nuances, especially in MEGARes, to be investigated. MEGARes' *ERM32* group, for example, shared sequences with CARD's *Erm 23S ribosomal RNA methyltransferase* and *non-erm 23S ribosomal RNA methyltransferase G748*. That is to say that CARD believes that the one set of sequences do not code for an ERM protein, but MEGARes believes that at least one in that set does code for an ERM protein. The databases agree that the sequences code for 23s methyltransferases, so at least they are not disagreeing on the function of the gene product.

Strange overlaps like this were not uncommon. The VAN bins of each database were quite scrambled. Sequences that fell into *vanR* in CARD would fall into MEGARes' *VANXYC* and *VANXYE* groups.

I also identified peculiarities in MEGARes' VAN sequences themselves. Some sequences do not start with ATG, starting instead with TTG (MEG_7331, MEG_7333), or GTG (MEG_7332) [22]. The next ATG in MEG_7333 is also outside of the reading frame starting from the first nucleotide in the sequence. The next start codon in the same reading frame is at position 115 in the sequence [22]. It is unclear whether these pre-start segments are critical to the gene or some kind of error. For comparison, CARD contains only one *VanA* gene and it begins on an ATG (ARO:3000010) [15].

These sequences were all identified in a single MEGARes group, *VAN*, which was itself an anomaly. The description of the group says that ”The VAN group of genes confer resistance ... through the mechanism, VanA-type resistance protein.”. However, there is a separate group labelled *VANA*. Both groups are under the mechanism *VanA-type resistance protein* [22]. Why the sequences in the *VAN* group are not therefore under *VanA* is not explained.

## 3.2 Wastewater Counts Comparison

After running AMR++ on the same metagenomic sample, MEGARes found 547,790 determinants, Removing all MEGARes determinants that were not of the "Drug" type resulted in a total of 460,859 total counts, or 84.4% of total MEGARes counts. Meanwhile translated CARD only found 70,567, just under 15% of MEGARes' drug-only counts. See supplementary items 6 and 7

When sorted by drug class, CARD has three times as many multidrug-resistant hits as MEGARes does, but far fewer individual drug hits (See Fig. 7 and 8) in spite of the massive difference in total results. However, this appears to be due to the fact that MEGARes considers MLS to be a separate class from "multidrug resistance". MEGARes has nearly twenty times the MLS count as it does multidrug resistance. Combining the two, MEGARes' counts are roughly 41% multidrug resistance to CARD's 53%. Such differences persisted at the gene level. Not only did counts vary significantly, the relative proportions of different classes and bins varied as well. Aminoglycosides made up 12.7% of CARD's hits and were its most-represented class. They made up 21.3 % of MEGARes and were the second-highest after MLS (See Fig. 8).

One of the most bizarre results was that of *CBLA*. Both databases found twenty instances of *CBLA* genes. The odd part is that CARD only has one *CblA* sequence, while MEGARes has two, one of which is sequence-identical to the CARD entry. The counts were split eleven to nine between MEGARes' two entries for *CBLA*, while all 20 were associated with the single CARD entry. This effect persisted when the CARD data had the "RequiresSNPConfirmation" flag attached to it, meaning that the single CARD sequence must be a perfect match for the metagenomic data. This is strange because it is the same data being run through AMR++, meaning that the sequence in MEGARes that does not overlap with CARD is being assigned counts despite there being a better match for the data present in the MEGARes database.

# 4 Discussion

Using CARD as a replacement for MEGARes results in highly divergent results. Whether this is overcounting on MEGARes' part, under-counting on CARD's part, or simply a product of MEGARes having nearly three times as many sequences cannot be conclusively determined. CARD should not be considered a complete replacement until the cause for the difference in counts is found or in the event that MEGARes becomes so out of date that under-counting is preferable to inaccurate counting. As of MEGARes 2.0, 97% of CARD's current sequences are already present in MEGARes, so inaccuracy due to old data should be low for the time being.

With that being said, several potential errors were identified in MEGARes, suggesting that CARD may be useful as a verification tool. CARD's data, due to its active curation and strict QA, can be generally taken to be more accurate than that of MEGARes.

The VAN category scrambling was unusual. *VANR* genes being confused with core resistance genes could be a serious issue for van genes, as the *vanR* gene is a regulatory gene. These serve a function which is distinct from the core resistance genes and should not be lumped in with the rest [15, 22]. MEGARes has distinct categories for *VANR* genes, so it is unclear why sequences that CARD defines as *vanR* appear in MEGARes' *VANXYE*/*C*. Additionally, The splitting of sequences between *VAN* and *VANA* would make sense if *VAN* was a parent of *VANA, VANHD*, etc., but MEGARes' structure only allows for 5 specific levels, so this is not the case. The fact that VAN lacks a reference to any paper suggest that a curation error has occurred somewhere, splitting the groups for an unknown reason.

An unknown number of MEGARes entries do not even have a reference paper attached to them, leaving their accuracy unsupported. For example, none of the groups under the mechanism *Fosfomycin target mutation* or ”Fosfomycin MFS efflux pump” contained a reference paper [22]. This represented only twelve sequences in total, but it is impossible to know how many unverified sequences are present in MEGARes. Furthermore, a degree of curational inconsistency in MEGARes is implied by the inconsistencies in *erm*, *VANR* and *VANA* classification, and anomalies in sequence, multidrug classification, and *CBLA* counting. As a result, the use of CARD as a verification tool for MEGARes counts is reasonable. If a determinant is found by AMR++ using only MEGARes, it will be less likely to be a genuine hit. If it is found by both, the determinant is more likely to be a true AMR determinant.

Due to the incredible complexity and size of the AMR space, it is exceptionally difficult to classify whether such large differences between databases is a symptom of poor curation practices on the part of any database or simply a result of different approaches and goals. For example, there is debate between bioinformaticians about whether regulatory genes qualify as AMR genes or not, as well as the acceptable parameters with which to run bioinformatic software [25, 26]. The field is subject to a degree of subjectivity. With that said, very little independent research has been done to verify and compare the quality of antimicrobial databases. ResFinder, which makes up over a quarter of MEGARes, has been phenotypically verified to a certain extent [16, 21, 27]. However, no verification has been done independently of the original ResFinder authors and no extensive comparison has been made between it and other databases. Therefore, the source databases that feed MEGARes should be investigated for their accuracy.

This ambiguity, combined with the significant anomalies present in MEGARes, suggest that the discrepancy in wastewater determinant counts is most likely not a result of a CARD error. However, that ambiguity also makes it difficult to conclusively say whether CARD can be used as a replacement for MEGARes. Would reducing the number of counts be an improvement in accuracy, avoiding pitfalls unseen in MEGARes, or is some unknown information that leads to such high counts being lost? Further investigation is needed to determine the extent of CARD's usefulness as a MEGARes replacement.

## 4.1 Future Steps

Comparing the MEGARes results to those produced individually by ResFinder and NCBI's Bacterial Antimicrobial Resistance Reference Gene Database on the same data would help to track down the source of the discrepancy in counts between MEGARes and CARD. These two databases make up the remainder of MEGARes' drug-exclusive sequences. However, MEGARes has its own classification system, so care must be taken to account for differences in classification that might influence the output. Searching those databases for unreferenced sequences would also identify the source of the unreferenced sequences making their way into MEGARes.

The first improvement I would make to the translator would be to add a unique number to the end of each DNA Accession in the translated header. This was not implemented due to time constraints. The conversion to gene family causes 59 headers to be identical, prompting their removal. Adding a unique number would render this removal unnecessary, improving database size slightly. Translator longevity would also be improved because CARD's database will continue to grow in size, increasing the risk of such overlaps happening again if no measures are taken to prevent them. Following this adjustment, re-running the same sample would allow a more accurate comparison between MEGARes and CARD.

Retaining the DNA Accessions instead of replacing them entirely would maintain identifiability of each sequence, making it easier to identify bugs in future development.

I recommend running the same wastewater sample through RGI using the original CARD and comparing its results to those of the two AMR++ tests: Original MEGARes and translated CARD. This would provide a baseline for the expected output of CARD, allowing for analysis of how the translation affects the integrity of the CARD database. Ostensibly, they use the same BWA algorithm, but RGI can also use Bowtie2. Additionally, bioinformatics software is complex and comparing the outputs could help identify nuances in approach between the different software. If RGI's results are similar to those of AMR++ when using CARD, then the difference in output is not a result of translation and is wholly due to differences in the databases.

Some verification could be done on the drug classes used by both data sets. CARD and MEGARes use different wording for the same drug classes, so creating a list of all classes present in both databases would ensure that comparisons can be clearly made and that they are covering the same drug space. If one contains drug classes that the other does not, counts will differ.

Counts are an imperfect assessment. Different samples may be skewed towards the presence of certain genes over others. The count assessment was done on a single metagenomic sample. Some groups and families were not represented at all. Testing with other metagenomic data would make it possible to determine how widespread the disparity is. If different groups are represented differently, then sequence number is not the only factor. Some drugs, mechanisms, and genes may be over-represented in certain environments over others.

MEGARes contains over two and a half times as many entries as CARD does, likely accounting for the count difference. Comparing the number of gene sequences in each group and family would make it easier to determine the source of differences in count between otherwise overlapping bins. For example, the MEGARes group RPH contains four sequences. The associated CARD family, rphB, contains only one sequence. More sequences in a bin means more opportunities to match with the query data. 258 counts of RPH were found for MEGARes, but no instances of rphB were found for CARD. Currently, variation in sequences per bin is one of the most likely culprits of the variation in wastewater count results per bin.

## 4.2 Conclusion

AMR databases, like the biology that they track, are incredibly complex webs of information. This makes it difficult to assess the exact source of this disparity. However, CARD's database may be useful for improving the accuracy of MEGARes' outputs. I would recommend more investigation by the Microbial Ecology Group and the McArthur lab to identify and catalogue specific differences in curation approach.

# References

(1) Tenover, F. C. *The American Journal of Medicine* **2006**, *119*, S3–S10.

(2) WHO *AntiMicrobial Resistance - Global Report on Surveillance*; tech. rep.; WHO, 2014.

(3) McArthur, A. G. et al. *Antimicrobial agents and chemotherapy* **2013**, *57*, 3348–57.

(4) Velkov, T.; Roberts, K. D.; Nation, R. L.; Thompson, P. E.; Li, J. *Future Microbiology* **2013**, *8*, 711–724.

(5) Papp-Wallace, K. M.; Endimiani, A.; Taracila, M. A.; Bonomo, R. A. Carbapenems: Past, present, and future, 2011.

(6) IACG *NO TIME TO WAIT: SECURING THE FUTURE FROM DRUG-RESISTANT INFECTIONS REPORT TO THE SECRETARY-GENERAL OF THE UNITED NATIONS*; tech. rep.; WHO, 2019.

(7) O'neill, J. *Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations*; tech. rep.; London, 2014.

(8) OECD, *Stemming the Superbug Tide*; OECD Health Policy Studies, Vol. 2015; OECD: 2018, p 2.

(9) De Kraker, M. E.; Stewardson, A. J.; Harbarth, S. *PLoS Medicine* **2016**, *13*, DOI: `10.1371/journal.pmed.1002184`.

(10) IACG *Surveillance and monitoring for antimicrobial use and resistance*; tech. rep.; WHO, 2018.

(11) Walterspiel, J. N.; Morrow, A. L.; Cleary, T. G.; Ashkenazi, S. *Infection* **1992**, *20*, 25–29.

(12) Mungall, C. J.; Emmert, D. B. *Bioinformatics* **2007**, *23*, i337–i346.

(13) Wright, G. D.; Poinar, H. *Trends in Microbiology* **2012**, *20*, 157–159.

(14) McArthur, A. G.; Wright, G. D. *Current Opinion in Microbiology* **2015**, *27*, 45–50.

(15) The Comprehensive Antibiotic Resistance Database, 2013.

(16) Zankari, E.; Hasman, H.; Cosentino, S.; Vestergaard, M.; Rasmussen, S.; Lund, O.; Aarestrup, F. M.; Larsen, M. V. *Journal of Antimicrobial Chemotherapy* **2012**, *67*, 2640–2644.

(17) Gupta, S. K.; Padmanabhan, B. R.; Diene, S. M.; Lopez-Rojas, R.; Kempf, M.; Landraud, L.; Rolain, J.-M. *Antimicrobial agents and chemotherapy* **2014**, *58*, 212–20.

(18) Alcock, B. et al. **2020**.

(19) Jia, B. et al. *Nucleic Acids Research* **2017**, *45*, D566–D573.

(20) Lakin, S. M.; Dean, C.; Noyes, N. R.; Dettenwanger, A.; Ross, A. S.; Doster, E.; Rovira, P.; Abdo, Z.; Jones, K. L.; Ruiz, J.; Belk, K. E.; Morley, P. S.; Boucher, C. *Nucleic acids research* **2017**, *45*, D574–D580.

(21) Doster, E.; Lakin, S. M.; Dean, C. J.; Wolfe, C.; Young, J. G.; Boucher, C.; Belk, K. E.; Noyes, N. R.; Morley, P. S. *Nucleic Acids Research* **2019**, DOI: `10.1093/nar/gkz1010`.

(22) Microbial Ecology Group MEGARes.

(23) Lakin, S. M.; Kuhnle, A.; Alipanahi, B.; Noyes, N. R.; Dean, C.; Muggli, M.; Raymond, R.; Abdo, Z.; Prosperi, M.; Belk, K. E.; Morley, P. S.; Boucher, C. *Communications Biology* **2019**, *2*, DOI: `10.1038/s42003-019-0545-9`.

(24) DTU Illumina HiSeq 3000 paired end sequencing - SRA - NCBI.

(25) Gupta, S. K.; Rolain, J. M. Reply to ”Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in Bacteria”, 2014.

(26) Zankari, E. Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in Bacteria, 2014.

(27) Zankari, E.; Hasman, H.; Kaas, R. S.; Seyfarth, A. M.; Agersø, Y.; Lund, O.; Larsen, M. V.; Aarestrup, F. M. *Journal of Antimicrobial Chemotherapy* **2012**, DOI: `10.1093/jac/dks496`.

# 5 Appendix

## 5.1 Supplementary data

1. Readme file - usage and structure of translator
2. meg_num_bins Analyzed.xlsx - overlap of bins in CARD-in-MEGARes direction
3. card_num_bins Analyzed.xlsx - overlap of bins in MEGARes-in-CARD direction
4. across_card_spread.csv - MEGARes groups and the CARD families across which they are distributed
5. across_meg_spread.csv - CARD families and the MEGARes groups across which they are distributed
6. ERR1713335_MEGARes_AMR_analytic_matrix Analyzed.xlsx - Wastewater data AMR++ output using MEGARes
7. ERR1713335_CARD_AMR_analytic_matrix Analyzed.xlsx - Wastewater data AMR++ output using CARD
8. Overlap_culled_DB.csv - List of database entries culled for overlapping

## 5.2 Figures

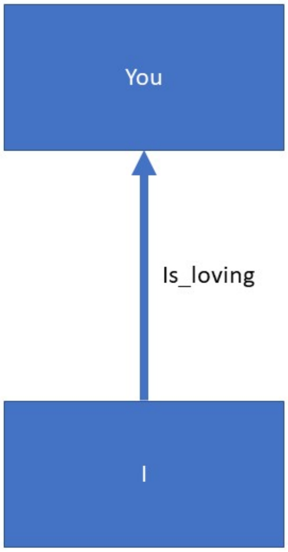


Figure 1: A simple example of an ontology.

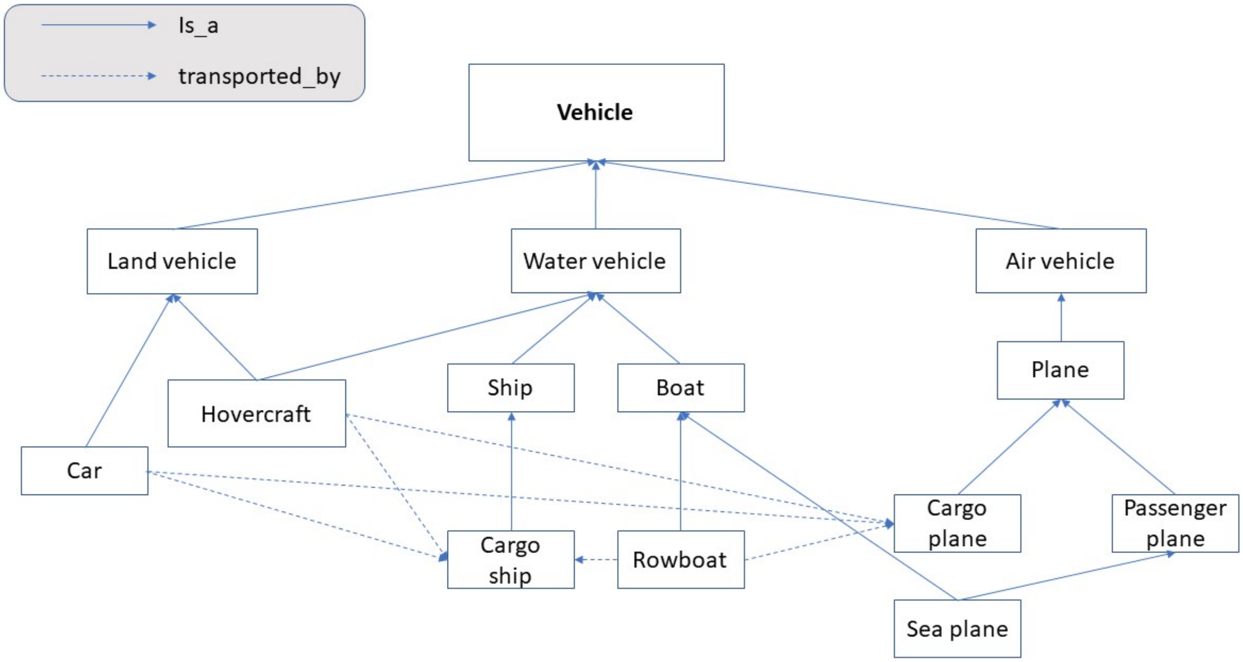


Figure 2: A more complex example of an ontology. Arrows the relationship between two terms. In this case, that the subject is a sub-term (*is_a*) or can be transported by the object (transported_by). Note that each term can have multiple parents.

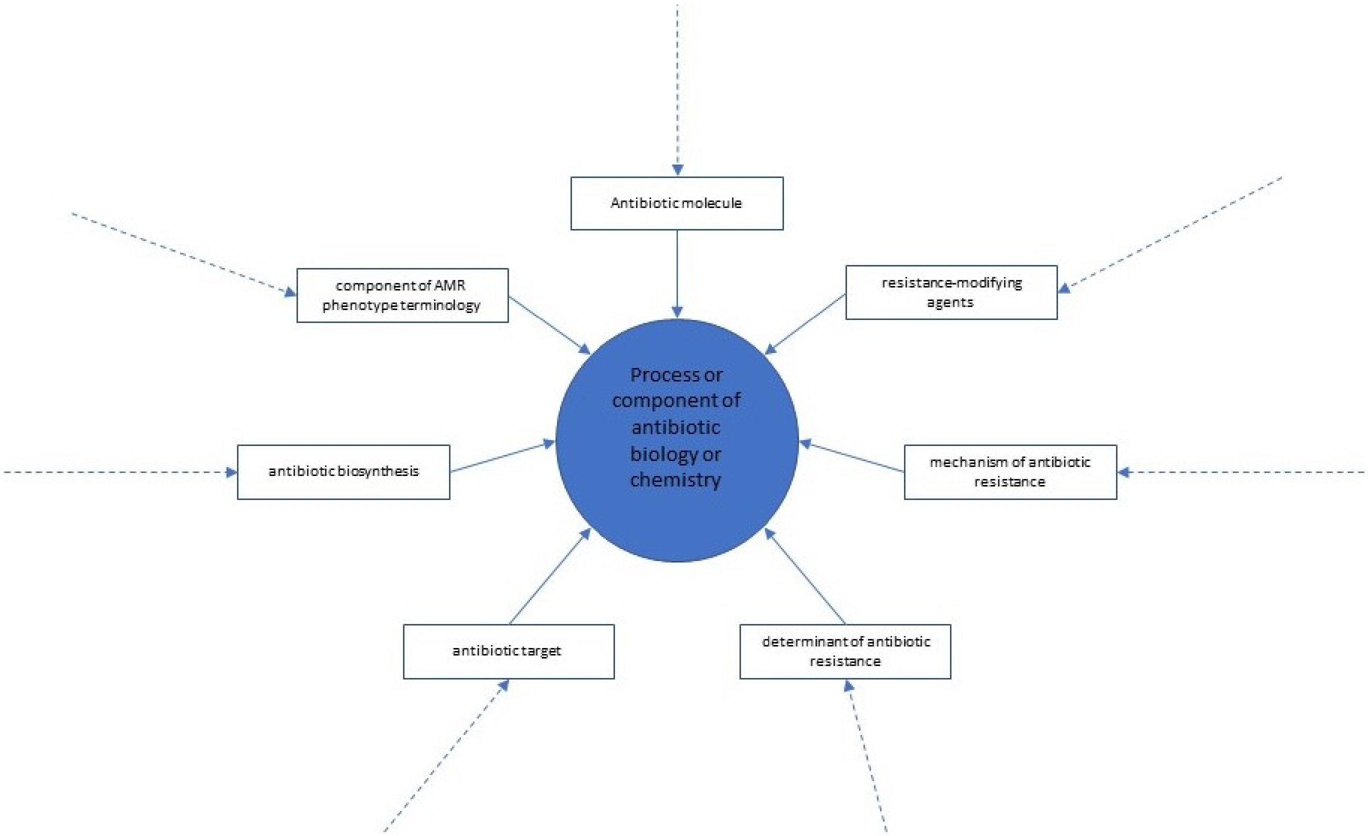
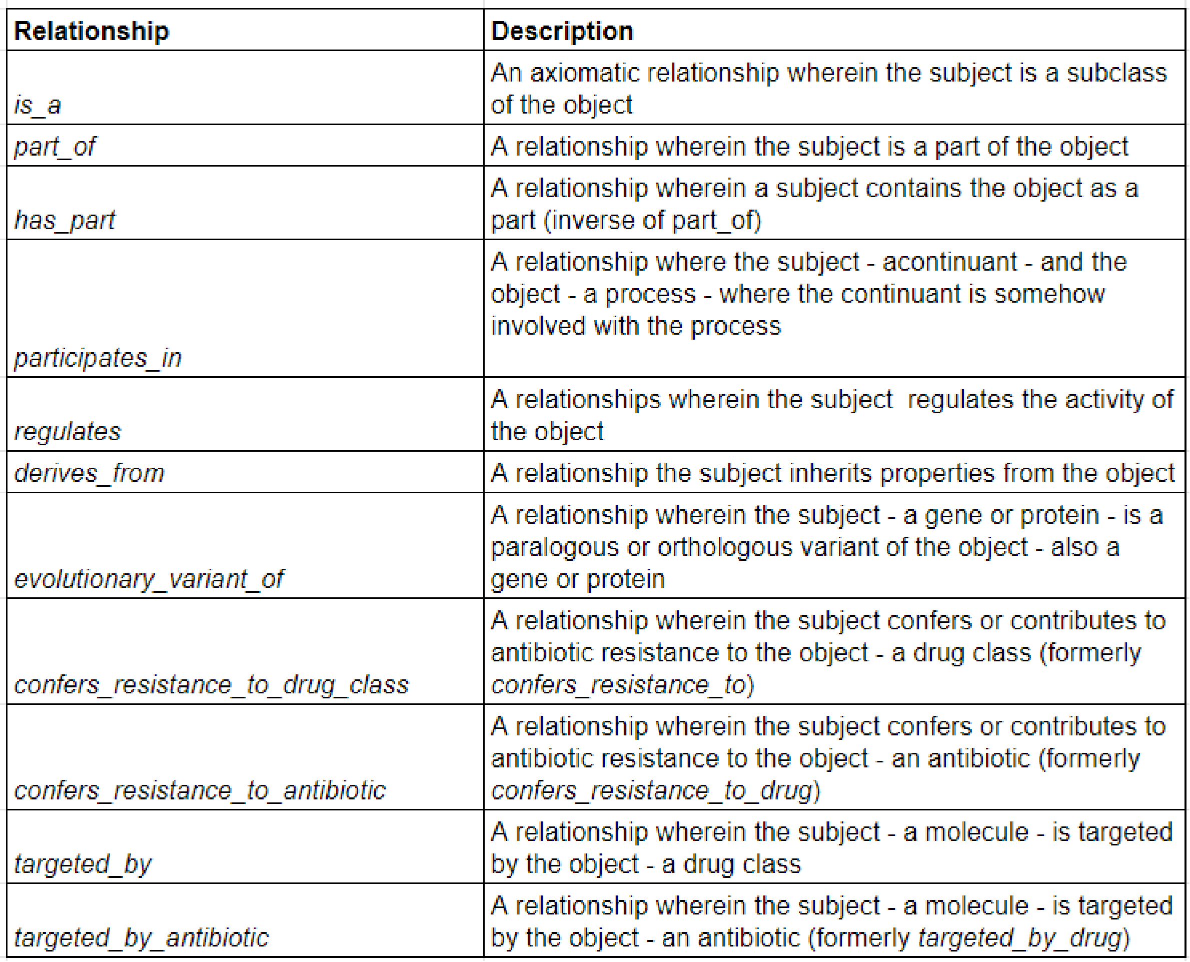

Antibiotic molecule
component of AMR phenotype terminology
resistance-modifying agents
Process or component of antibiotic biology or chemistry
antibiotic biosynthesis
mechanism of antibiotic resistance
antibiotic target
determinant of antibiotic resistance

Figure 3: The seven branches of the ARO. Each one describes a critical branch of antimicrobial resistance.



| **Relationship** | **Description** |
|---|---|
| *is_a* | An axiomatic relationship wherein the subject is a subclass of the object |
| *part_of* | A relationship wherein the subject is a part of the object |
| *has_part* | A relationship wherein a subject contains the object as a part (inverse of part_of) |
| *participates_in* | A relationship where the subject - acontinuant - and the object - a process - where the continuant is somehow involved with the process |
| *regulates* | A relationships wherein the subject  regulates the activity of the object |
| *derives_from* | A relationship the subject inherits properties from the object |
| *evolutionary_variant_of* | A relationship wherein the subject - a gene or protein - is a paralogous or orthologous variant of the object - also a gene or protein |
| *confers_resistance_to_drug_class* | A relationship wherein the subject confers or contributes to antibiotic resistance to the object - a drug class (formerly *confers_resistance_to*) |
| *confers_resistance_to_antibiotic* | A relationship wherein the subject confers or contributes to antibiotic resistance to the object - an antibiotic (formerly *confers_resistance_to_drug*) |
| *targeted_by* | A relationship wherein the subject - a molecule - is targeted by the object - a drug class |
| *targeted_by_antibiotic* | A relationship wherein the subject - a molecule - is targeted by the object - an antibiotic (formerly *targeted_by_drug*) |

Figure 4: All relationships currently present in CARD. Adapted from the CARD 2020 manuscript

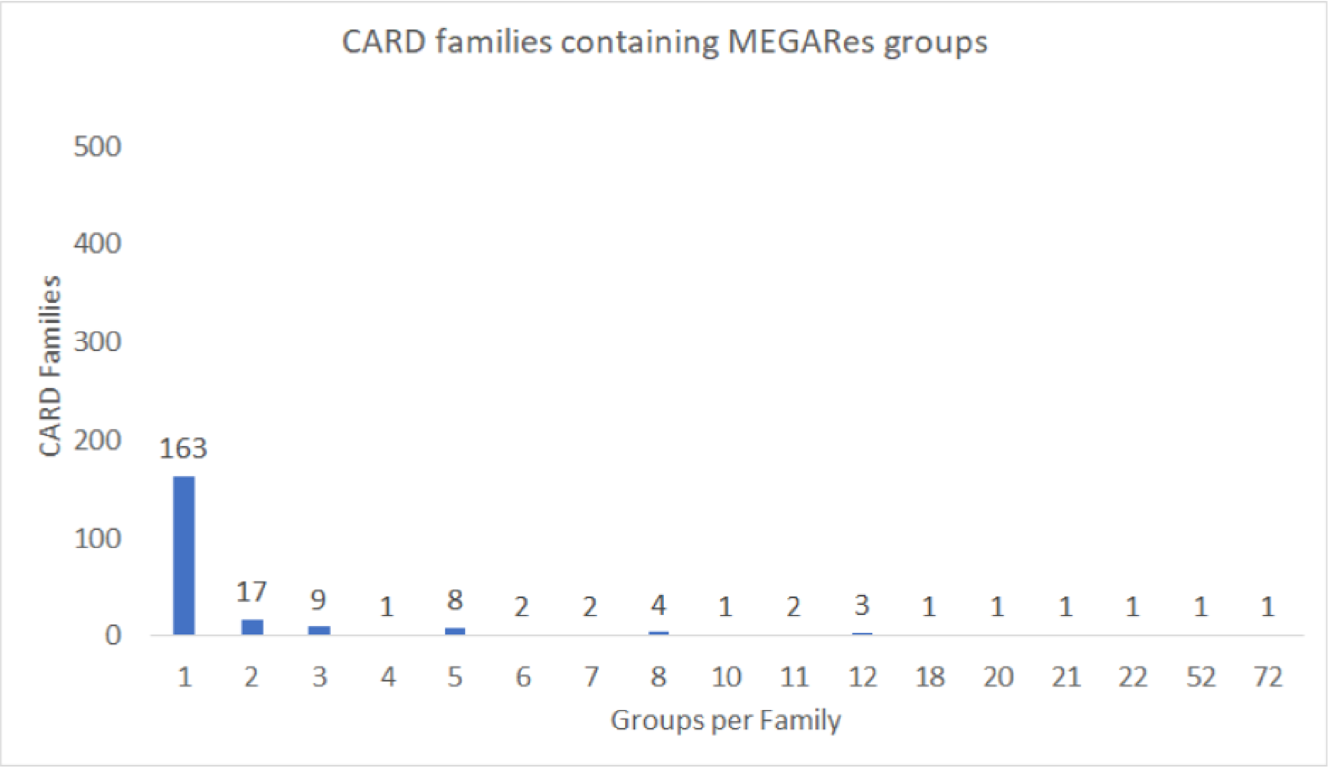


Figure 5: The number of CARD families that overlap with each quantity of MEGARes groups. For example, only one CARD family contains sequences that are also in 72 MEGARes groups, but 163 families contain sequences that are also in a single group. See supplementary item 2

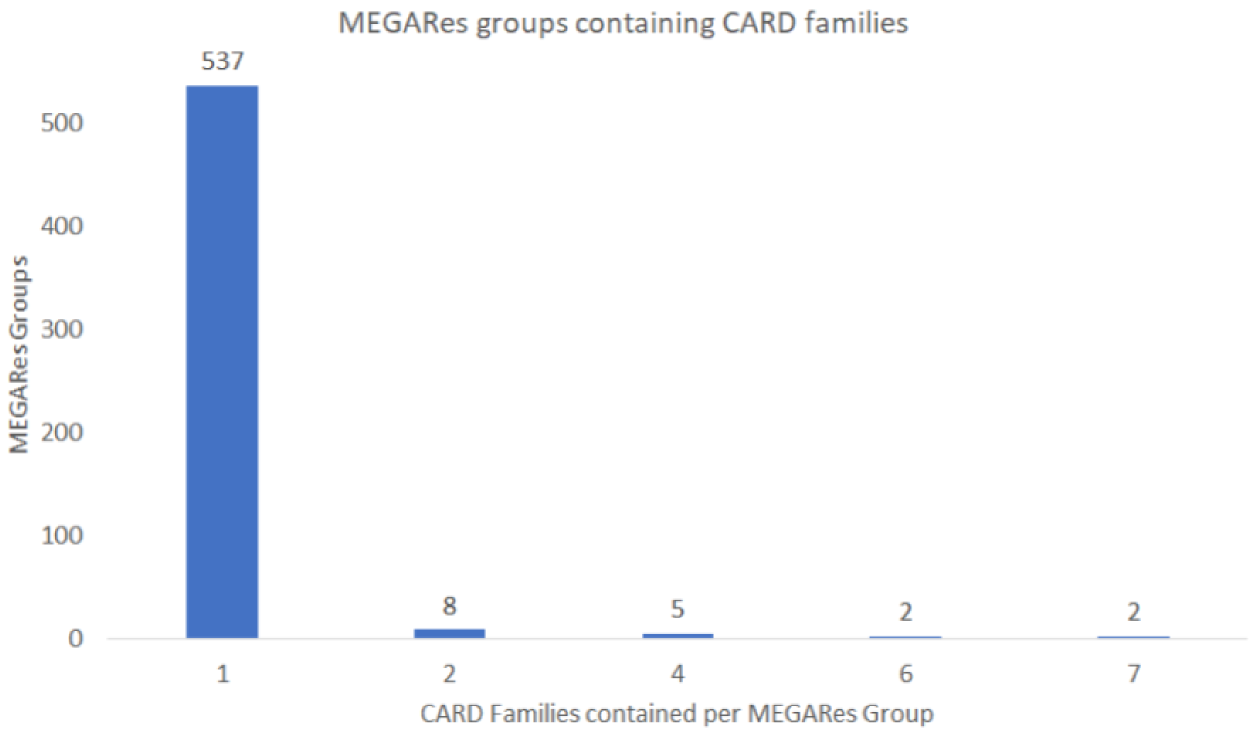
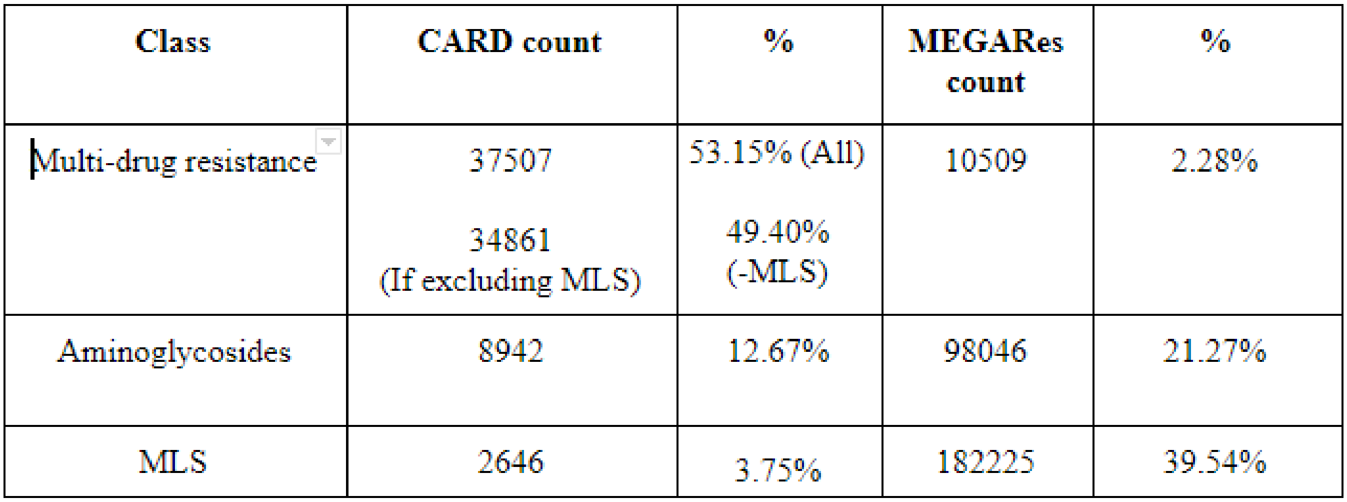
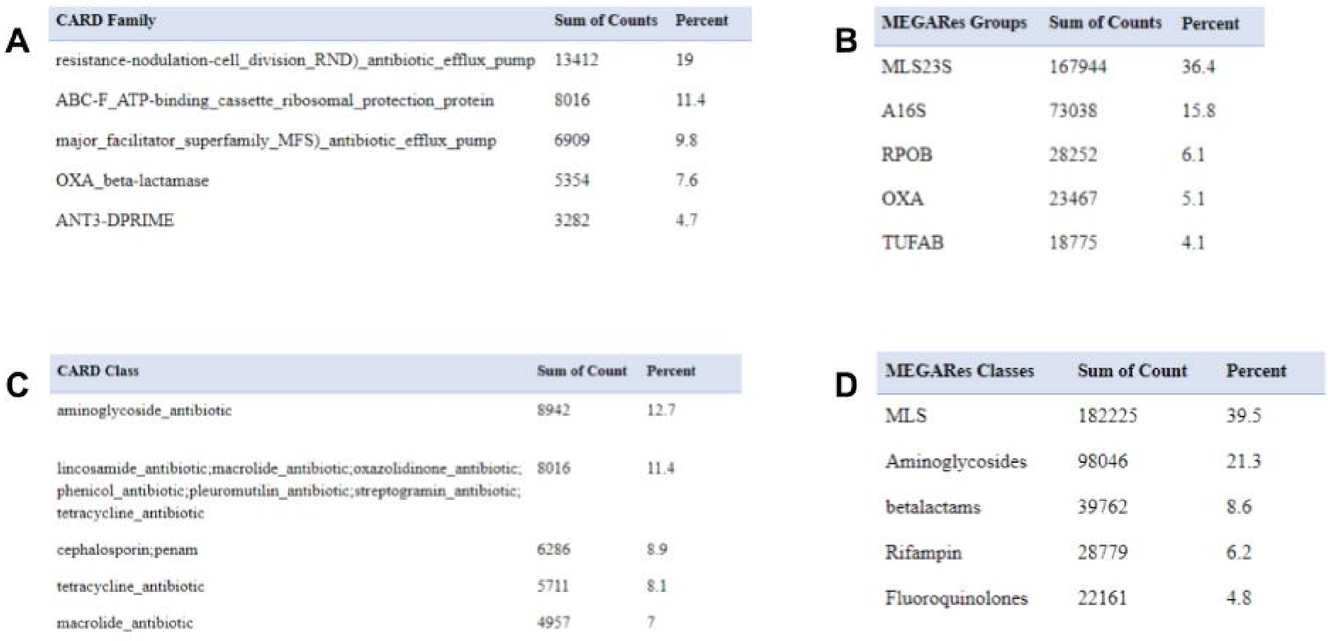


Figure 6: The number of MEGARes groups that contain sequences that are shared with each quantity of CARD families. For example, two MEGARes groups overlap with seven families each, while 537 groups overlap with one family each. See supplementary item 3



| Class | CARD count | % | MEGARes count | % |
|---|---|---|---|---|
| Multi-drug resistance | 37507<br>34861<br>(If excluding MLS) | 53.15% (All)<br>49.40%<br>(-MLS) | 10509 | 2.28% |
| Aminoglycosides | 8942 | 12.67% | 98046 | 21.27% |
| MLS | 2646 | 3.75% | 182225 | 39.54% |

Figure 7: A selection of classes and their relative presence in the CARD and MEGARes wastewater output. Percentage is relative to total counts per database.



| CARD Family | Sum of Counts | Percent |
|---|---|---|
| resistance-nodulation-cell_division_RND)_antibiotic_efflux_pump | 13412 | 19 |
| ABC-F_ATP-binding_cassette_ribosomal_protection_protein | 8016 | 11.4 |
| major_facilitator_superfamily_MFS)_antibiotic_efflux_pump | 6909 | 9.8 |
| OXA_beta-lactamase | 5354 | 7.6 |
| ANT3-DPRIME | 3282 | 4.7 |

| MEGARes Groups | Sum of Counts | Percent |
|---|---|---|
| MLS23S | 167944 | 36.4 |
| A16S | 73038 | 15.8 |
| RPOB | 28252 | 6.1 |
| OXA | 23467 | 5.1 |
| TUFAB | 18775 | 4.1 |

| CARD Class | Sum of Count | Percent |
|---|---|---|
| aminoglycoside_antibiotic | 8942 | 12.7 |
| lincosamide_antibiotic;macrolide_antibiotic;oxazolidinone_antibiotic; phenicol_antibiotic;pleuromutilin_antibiotic;streptogramin_antibiotic; tetracycline_antibiotic | 8016 | 11.4 |
| cephalosporin;penam | 6286 | 8.9 |
| tetracycline_antibiotic | 5711 | 8.1 |
| macrolide_antibiotic | 4957 | 7 |

| MEGARes Classes | Sum of Count | Percent |
|---|---|---|
| MLS | 182225 | 39.5 |
| Aminoglycosides | 98046 | 21.3 |
| betalactams | 39762 | 8.6 |
| Rifampin | 28779 | 6.2 |
| Fluoroquinolones | 22161 | 4.8 |

Figure 8: Top five wastewater AMR determinants by count, separated by bin (A,B) and class (C,D) as well as those detected using translated CARD (A,C) and those detected using MEGARes (B,D). MEGARes' multi-drug resistance group, with only 10,509 counts, did not make its top five. C demonstrates CARD's use of semicolons to indicate multidrug resistance.