**Intro**

Over the past few decades, the massive use of antibiotics has resulted in the development of widespread resistance genes in microbial populations. Procedures that were once mundane have become risky due to the potential for infection by resistant microbes. Surveillance and prediction are critical to combating the rapidly-evolving threat of antimicrobial resistance, or AMR. The collective set of resistances available to a population of microbes is known as its resistome. Keeping up-to-date records of the bacterial resistome allows clinicians and researchers to effectively target their efforts. Without knowledge of what your opponent is doing, one cannot make clear decisions on how to combat them. AMR databases make this vital knowledge easily accessible to clinicians, researchers, and policymakers. 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. The distributed nature of this data makes it extremely difficult to search for relevant information. 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.

[Slide]

There are numerous ”boutique” databases, which track resistance data for specific species. There are a few major comprehensive databases, however. One of these is CARD, the Comprehensive Antibiotic Resistance Database.

[Slide]

CARD is structured as an ontology, a hierarchical structure for sorting and categorizing concepts and their relationships with one another. 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. Equivalently, the object is a parent term to the subject.

~~Take a vehicle ontology. One could break vehicles into three classes: land, sea, and air. 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. The ”transported by” relationship identifies that a sub-term can be loaded onto and moved by a parent object. This relationship tells you what a term can do, not only what it is.~~

[Slide]

MEGARes is an offshoot of CARD which has since grown to encompass several domains of AMR. 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.

[Slide]

CARD and MEGARes differ not only structurally, but also in their approach to classification. 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. Collectively, groups and families can be referred to as “bins”. As will be seen later, bins can overlap (or not) in very unpredictable ways, and may even fall into a grey area of describing mechanisms instead. Drug classes are probably the most similar structural component between the two. However, MEGARes has some unusual classifications such as a single class for multidrug-resistant determinants which contains any determinants which provide resistance to more than two drugs. Strangely, MLS, which is a combination of three classes, is treated as a separate class. MLS stands for macrolide-lincosamide-streptogramin.CARD, on the other hand, tracks all its multidrug-resistant classes by combining them into a cluster of classes separated by semicolons.

[Slide]

The databases are structured in a fasta file where each entry is the DNA sequence that provides resistance and a header that contains its classification information. 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. The other distinctions are not relevant to this project.

[Slide]

MEGARes and CARD are used to run AMR++ and RGI, respectively. These software compare reference sequences from their source databases to a query sequence and indicate the presence of AMR genes. AMR++ and RGI both match each sequence in the provided reference database and match that sequence to any copies in the query data. Both RGI and AMR++ are based on the same algorithm, the Burrows-Wheeler Transform, which has multiple implementations. For the sake of this project, both AMR++ and RGI use the BWA implementation. CARD contains multiple models which can contain different reference sequences and interpret them in different ways. I focused on the protein homologue model, which can be used for the metagenomic reads that AMR++ is built for. AMR++ runs verifies its outputs through RGI. 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. If MEGARes’ entries have the flag “RequiresSNPConfirmation”, anything not in the “Perfect” group is removed. SNP-reliant determinants can vary significantly in their expression if their sequence is not identical. Different databases are better suited to different tasks. MEGARes' DAGs are 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, the BWT approach can also cause inflated counts when applied to ontology-based databases.

[Slide]

Because of the constant growth of the AMR space, databases must be consistently updated in order to remain relevant. MEGARes went without an update for over three years, between December 2016 and January 2020. CARD, on the other hand, aims to update quarterly. This project assessed the feasibility of using CARD as a stopgap measure to fill in any future curation gaps.

The goals of this project were to:

1. Write software to translate CARD's Protein Homologue 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, especially in the relationship between CARD families and MEGARes groups
3. Assess the usefulness of translated CARD as a replacement for MEGARes data in AMR++

**Methods**

CARD has an index file which contains the DNA Accession, class, mechanism, and family information of each determinant in each CARD model.

[Slide]

The goal of writing translation software was accomplished by writing a nearly 500-line 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 a new, AMR++-compliant header. GenBank DNA Accessions included in the index file are used in place of MEG’s unique identifier number. Only protein homolog model entries were converted, as the PHM is the most similar in its application to MEGARes’ structure. Because the index file and database file are ordered differently, the database entries had to be found by DNA accession and gene family. CARD headers do not contain gene family information, so the gene family of each entry must be identified by another search in the index file before the database entry can be matched to a header. Every entry in an AMR database must have a unique header, so the software removes any entries with duplicate headers. This conversion from gene to gene group resulted in a loss of 59 database entries, or 2.2% of the database.

[Slide]

The goal of assessing major differences was evaluated by writing a 200-line 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, suggesting perfect overlap. As will be seen later, this is not the case. Second, the software, given that there are multiple families per group and multiple groups per family, determined which specific groups overlapped with which families and vice versa.

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. If CARD can function as an accurate replacement of MEGARes, AMR++ should produce a nearly identical number of counts. Additionally, there should be similar counts of both related families/groups and the same drug classes. If each database finds different levels of the same drug classes, then they are most likely not well aligned. 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. 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 would be 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.

[Slide]

**Results**

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, partially, or not at all. A few illustrative examples have been selected to provide a sense of the data.

Many more MEGARes groups fit into a single CARD family than vice versa, suggesting that each CARD group encompasses a larger area of the AMR space than each MEGARes group does. Of 554 groups, 537, or 97%, overlapped with a single CARD family.

[Slide]

Meanwhile, only 163 families overlapped with a single MEGARes group and seventy-two MEGARes groups overlapped with the “major facilitator superfamily (MFS) antibiotic efflux pump” CARD family.

[Slide]

The VAN bins of each database were quite scrambled. Sequences that fell into vanR in CARD would fall into MEGARes’ VANXYC and VANXYE categories. Strange overlaps like this were not uncommon. MEGARes’ ERM32 group, for example, shared sequences with CARD’s **Erm 23S** ribosomal RNA methyltransferase *and* **non-erm 23S** ribosomal RNA methyltransferase G748. CARD says that the one set of sequences do not code for an ERM protein, while MEGARes says that at least one in that set does. They are at least both 23s methyltransferases, so the two databases are not disagreeing on the function of the gene product.

Additionally, I discovered that MEGARes and CARD treat the same phenomena as totally different elements of their sorting hierarchy. For example, Rifampin Phosphotransferases are a CARD family, but a MEGARes mechanism. The effects that such fundamental structural differences would have on output is difficult to accurately determine, and there *were* significant differences.

[Slide]

When running AMR++ on the same metagenomic sample, MEGARes found 547,790 determinants, while translated CARD only found 70,567, just under thirteen percent of MEGARes’ counts. 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.

Looking at the counts sorted by drug class, CARD has 3 times as many multidrug-resistant hits as MEGARes does, but far fewer individual drug hits. All of this is 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 forty-one percent multidrug resistance when compared to CARD’s fifty-three percent. Such differences persisted at every level of classification.

[Slide]

**Discussion**

Using CARD as a replacement for MEGARes results in highly divergent results, and an alternative stopgap measure should be found. There are many factors that might contribute to the disparity in counts, but more investigation is needed. The information identified here opens up many avenues for analysis.

There are many nuances to the software to be investigated, such as the case of the non-erm ERM protein, or the MEGARes mechanisms classes as CARD families. 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 CARD sequence is a perfect match for the metagenomic data. This is strange because it’s 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.

The first step will be to run the same wastewater sample through RGI using the original CARD and compare its results to those of the two AMR++ tests. 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. Furthermore, it may be possible to adjust the implementation of the translator such that each of the 59 removed database entries could be added back in by replacing the DNA accession with a unique number, similar to MEGARes. Determining how many counts they would contribute would make it possible to separate any difference in results caused by their removal from differences caused by the conversion from gene to gene family during the translation.

Identifying 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 the most likely culprit of the variation in results.

~~Some verification could be done on the drug classes used by both datasets. 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.~~

~~The count assessment was done on a single metagenomic sample. Some groups 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.~~

**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. I would recommend more investigation by the two organizations to identify and catalogue specific differences in curation approach.