TITLE

Accurate Pathogen-of-origin Prediction of Antibiotic Resistance Genes With CARD k-mers

AUTHORS

Mateusz A. Wlodarski, Tammy, Amogelang, etc author list

AFFILITATION

McArthur Lab, Institute for Infectious Disease Research, McMaster University – 1280 Main Street West, Hamilton, ON, CANADA, L8S 4L8

KEYWORDS

Metagenomics, Species classification, Antibiotic Resistance Genes, K-mers

ABSTRACT

Antimicrobial resistance (AMR) is a global health crisis, with Canadian cases exceeding 20 000 per year [1]. One challenge during outbreaks of pathogens exhibiting AMR is quickly identifying the pathogen-of-origin (PO) responsible for the onset of symptoms. This diagnosis is crucial for doctors to prescribe correct antibiotics and begin treatment. Traditional culture-based methods designed for this task remain the gold standard, however there is a need for faster results and the capability to sequence the entire microbial environment (i.e., metagenome). [2, 3]. Identifying the PO from metagenomes which may encompass millions of sequencing reads has proven difficult: alignment programs like BLAST are computationally burdensome [4, 5]. K-mer classifiers like Kraken2 circumvent these shortcomings by using exact word matching between a reference set of pre-computed, pathogen-specific k-mers and a query sequence [5] but are often designed for general use and perform poorly on niche sequence spaces [6]. Here we describe CARD k-mers, a novel computational tool that assigns taxonomic labels to sequences containing antibiotic resistance genes (ARGs). By synergizing with the Comprehensive Antibiotic Resistance Database (CARD), CARD k-mers predicts the PO of input sequence data in minutes. CARD k-mers accurately classified 95.71% of 103 456 *in silico* uni-pathogen alleles and was 91.8% accurate in predicting a species 'hit' for 9 722 *in silico* multi-pathogen alleles, compared to Kraken2 at 65.89% and 33.71%. This is also the first instance of tailoring a k-mer classifier to a desired sequence space, since using a ARG specific k-mer library for Kraken2 improved classification accuracy by 13.05% over the default distribution and reduced the rate of unclassified alleles by 1.496% and 9.618% in the classification tests, making this tool of interest to those working with ARG data and bioinformaticians alike.

DATA SUMMARY

CARD k-mers is written in Python and can be used once RGI has been downloaded, along with CARD-r to make the k-mer reference library. Full software documentation is available at from https://github.com/arpcard and all validation datasets and results from https://devcard.mcmaster.ca:8888/mwlodarski/mateuszmsc/-/tree/main?ref_type=heads

IMPACT

Antimicrobial resistance is a global health crisis that is motivating the need for accurate and faster pathogen-of-origin classification of metagenomic sequencing. We compare CARD kmers with the leading k-mer classifier, Kraken2 in a head-to-head classification test involving antibiotic resistance genes. This marks the pioneering adaptation of a k-mer classifier tailored for sequences containing antibiotic resistance genes. Overall our work illustrates that pathogen-of-origin classification using k-mer classifiers can be improved by defining a narrowed sequence space.

INTRODUCTION

Antibiotic misuse and a lack of novel drugs are driving the emergence of multidrug-resistant pathogens across agricultural, wastewater, and clinical networks [2, 3, 7]. Without serious mitigation efforts we can expect roughly 10 million yearly deaths worldwide by 2050 [8]. Bacterial pathogens are the principal exhibiters of AMR. These pathogens are equipped with ARGs that allow them to avoid the effects of certain antibiotics through several known resistance mechanisms. Bacterial ARGs can be in the bacterial chromosome or plasmids, with the latter being highly mobile between organisms. These lateral gene transfer (LGT) events further exacerbate the complexity of AMR and can lead to the acquisition of multiple ARGs by the same pathogen, potentially granting it resistance against several drug classes.

Thanks to recent improvements in sequencing technologies, it is now possible to perform mass sequencing and track the spread of antibiotic resistance genes (ARGs) and priority pathogens. Databases such as CARD serve as invaluable surveillance tools with the ability to detect known ARGs and discover new ones using CARD's Resistance Gene Identifier (RGI). While such global genomic surveillance has led to improved public health outcomes, one important surveillance problem remains the identification of the PO in question, as doctors need this diagnosis to prescribe an appropriate antibiotic in the face of AMR [9]. Traditionally, this has not been a challenge as genomic sequencing has been based on cultured isolates, where media and molecular diagnostics include pathogen identification [10]. However, we are seeing increasing need for culture-free PO classification methods for complex clinical (e.g., swabs) and wastewater samples, due to faster turnaround time and the ability to sequence the entire microbial environment (i.e., the metagenome) [7, 9].

For decades alignment-based paradigms have dominated the sequence search space to predict the PO from a metagenome. Programs like Burrows-Wheeler-Transform (BWT), Hidden-Markov-Model search (HMM), and Basic Local Alignment Search Tool (BLAST) all rely on DNA reads being aligned back to a reference genome before assigning a taxonomic label. These alignment tools massively burden computing performance and depend on subjective scoring matrices; they are primarily designed for comparing a limited set of queries to a collection of sequences, not PO. Thus, alignment-based methods are unfeasible for determining the PO from large metagenomes, which may encompass millions of sequencing reads. Furthermore, alignment programs assume that "homologous sequences comprise a series of linearly arranged and more or less conserved sequence stretches" [5]. This assumption is violated by LGT events and fast mutation rates that lead to massive genomic diversity in microbe populations. Alignment-based methods aren't suitable for two highly distant sequences either, as any two unrelated sequences will overlap to a certain degree [5]. Lastly, most alignment programs operate under a set of biological assumptions that dictate how the alignment should be scored. For the same pair of sequences, there likely exists a program or a parameter template that could be tweaked to enhance the score of the alignment to make it appear more credible. It has been shown that older BLOSUM matrices used in alignment scoring were miscalculated, like BLOSUM62, and consequently offer higher alignment scores than their updated versions [11].

To alleviate these qualms, bioinformaticians have begun to resort to alignment-free sequence classification. Word-based search methods circumvent performance shortcomings by using exact word matching between a reference set of pre-computed, pathogen-specific k-mers (substrings of a sequence of length 'k') and a query sequence [5]. As a result, several k-mer

'classifier' tools have become popular in metagenomic pipelines for PO prediction. Kraken2 has positioned itself as a top classifier tool, capable of classifying 1.5 million reads per minute [6]. Kraken2 achieves this by using ultra-fast k-mer lookups against their precomputed database of k-mers of microbial genomes. However, Kraken2, along with most other k-mer classifiers, have been designed with a generalist user base in mind and can be undermined by LGT, which is common for ARGs. Saliva samples collected for the Human Microbiome Project were chosen to test Kraken's ability to detect and classify emerging and unique microbe species. The results showed that 68.2% of all reads were not classified, and the reads that were classified were mostly only at the genus, not species level [6]. This begs the need for a k-mer classifier that is optimized for datasets with ARGs.

CARD currently supports a pilot k-mer classifier algorithm optimized for ARG data: CARD k-mers. By synergizing with CARD's Resistome and Variants data (CARD-r: predicted ARGs from many pathogens), CARD k-mers identifies the PO and type of sequence data (plasmid, chromosome, etc.) of input sequences containing any ARGs and can be used for the rapid taxonomic classification of resistant pathogens for metagenomic datasets (Figure 1).

First, CARD k-mers builds a reference k-mer set from the alleles stored in CARD-r. In CARD-r v.4.0.0 this equates to roughly 32 million k-mers of length = 61 (61-mers). CARD k-mers stores two different types of k-mers for classification: taxonomic and genomic k-mers. Taxonomic k-mers include two k-mer subtypes: single species, which are unique to one species, and genus k-mers, unique to one genus. These are used to predict the PO. Genomic k-mers include three k-mer subtypes: those unique to chromosomal sequences, those unique to plasmid sequences, or k-mers found in both plasmids and chromosomes (Table 1). Genomic k-mers can

be used to evaluate potential mobility risks of the associated ARG(s) on top of PO classification, a feature unique to this k-mer classifier. Next, to be considered for a taxonomic prediction, individual sequences (e.g. FASTA, RGI predicted ORF, metagenomic read) must pass the --*minimum* coverage value (default of 10, i.e., the number of k-mers in a sequence that need to match a single category, for both taxonomic and genomic classifications, in order for a classification to be made for that sequence). Subsequent classification is based on a following logic tree (Appendix 1). The algorithm is currently available as a command line tool encapsulated within the RGI software called RGI kmer_query and accepts: fasta, RGI json and RGI bam files as input. CARD k-mers outputs a summarizing json and text file containing the pathogen-of-origin prediction and k-mer count of chromosomes and plasmid results pertaining to the individual sequences contained in the input file (Figure 2). Currently, the tool can classify 377 pathogens and 2 data origin types: chromosome and plasmid [12], but it remains entirely unvalidated and performs slowly. Full software documentation on CARD k-mers is available at: https://github.com/arpcard/rgi#using-rgi-kmer-query-k-mer-taxonomic-classification

METHODS

Dataset construction:

Three datasets were constructed for three separate validation tests. First, 320 614 *in silico* predicted ARG alleles were sourced from CARD-r v.4.0.0. 9 722 of these alleles were excluded because they were observed in multiple species, and therefore were not suitable to test CARD k-mers' classification accuracy. For example, prevalence sequence "690" from CARD-r was excluded since it had entries from 48 different pathogens, whereas prevalence sequence

"150 182" is unique to *Salmonella Enterica* and was therefore included. A visual depiction of the number of alleles based on the number of pathogens they associate with is demonstrated in (Figure 3). The excluded multi-pathogen alleles were classified in a separate test to validate multi-pathogen allele classification accuracy. Next, two-thirds of the alleles (rounded down) belonging to each pathogen were dedicated to a 'training set' used to build the reference library of pathogen-specific k-mers for a custom version of CARD k-mers. The other third was dedicated to a 'testing set'. Therefore, for a pathogen to contribute to the dataset, it had to have at least 3 unique alleles in CARD-r, the minimum amount required for a one-third holdout validation strategy. As a result, 62 'rare' pathogens were excluded from the dataset due to a lack of data, for a total of 315 pathogens. Stratifying the data by pathogen maintained relative allele distribution rates of the starting CARD-r dataset. Altogether, the allele counts from 315 pathogens amalgamated to 207 589 in the training set, and 103 456 in the testing set to validate uni-pathogen classification accuracy. Lastly, 8 143 uni-type alleles (alleles found only in chromosomes or plasmids) from the testing set were used to validate the data type prediction accuracy of CARD kmers. 62 alleles found only in genomic islands were excluded from the data type prediction test because CARD kmers does not currently support a possible genomic island prediction. A visual depiction of the three validation datasets is shown in Figure 4.

CARD k-mers (RGI version 6.0.2) and two Kraken2 classifiers (version 2.08) were created: Kraken2-default used the standard 70GB library and Kraken2-custom used the training set alleles to build reference k-mer sets. CARD k-mers used the training set to build its k-mer library.

Classification rules:

Following classification of the 103 456 test set alleles, output files containing PO prediction data were generated from each of the three classifiers. Since Kraken2 assigned highly specialized taxonomy labels, going as far as the exact strain number in some cases, an intermediate taxonomy map was created using ENTREZ and NCBI's taxonomy database to scrape taxonomy identifiers and their respective scientific names. This step was omitted with CARD k-mers since its output already includes the scientific names of predicted POs. Classification accuracy was recorded by comparing the predicted PO to the labelled PO in CARD-r for each allele. For the two Kraken2 classifiers, a correct species prediction included an exact scientific name match. Since alleles in CARD-r are mapped to a base species, any Kraken2 PO prediction with an exact strain name was considered a correct species level prediction so long as the base species was an exact match. For example, if Kraken2 predicted the PO of prevalence sequence "93 594" from CARD-r as *Escherichia coli s8* 0145:H28 str. RM12581, this was considered a correct PO prediction since the CARD-r label for this allele was *Escherichia coli.* Genus level classification accuracy was recorded following similar rules.

For the multi-pathogen alleles, a correct call was granted if the predicted species or genus was a 'hit' part of the list of pathogens mapped to that corresponding allele, since there was no sole correct species call.

For the data type prediction test, a correct call was granted if CARD k-mers predicted the data type of the allele as either plasmid or chromosome. If the data type prediction was "chromosome or plasmid" or "no genomic info", this was considered uninformative.

RESULTS

For the uni-pathogen classification test, four metrics were considered: correct species accuracy, erroneous predictions, genus sensitivity (alleles with a correct genus, but false species prediction), and the number of unclassified alleles from each classifier. CARD k-mers returned the best species accuracy of 95.71% with a genus sensitivity of 2%, followed by the custom version of Kraken2 at (78.94%, 2.31%), then Kraken2 default at (65.89%, 1.84%). The Kraken2 classifiers yielded lower unclassified rates, the default and custom Kraken2 version rates were 1.603% and 0.107% compared to CARD kmers at 17.51% (Figure 5).

For the multi-pathogen test, species hit accuracy was recorded instead of species accuracy, since multi-pathogen alleles do not have a sole correct species call. CARD k-mers returned the best species hit accuracy at 91.8% with a genus sensitivity of 3.74%, followed by the custom version of Kraken2 at (47.91%, 4.81%), and Kraken2 default at (33.71%, 14.73%). Once again, the Kraken2 classifiers had lower unclassified rates of 10.05% (default) and 0.432% (custom), compared to CARD k-mers at 51.81% (Figure 6).

For the data type prediction test, CARD kmers correctly predicted the data type of origin of 99.67% of chromosome and 93.61% of plasmid ARGs, with uninformative rates (no given data type) of 36.61% and 78.61% (Figure 7).

DISCUSSION

CARD k-mers out-classified both Kraken2 classifiers in species accuracy for uni-pathogen alleles and in species hit accuracy for multi-pathogen alleles, supporting its utility as a viable alignment-free PO prediction tool of sequence data containing ARGs. At first glance, CARD k-mers' unclassified rate of 17.5% in the uni-pathogen test and 51.81% in the multi-pathogen test

may seem problematic. However, users must keep in mind that CARD k-mers default parameters dictate that for a sequence to be even considered for classification, it must return a minimum of ten informative k-mers in the reference library, compared to Kraken2's lone required informative k-mer [6]. CARD k-mers is conservative with low sensitivity to minimize false positive resistant pathogen predictions. Users can tweak these parameters to their desired sensitivity, however there is little empirical evidence to support modifying the required informative k-mer value from its default value of ten. This parameter should be explored further to maximize accuracy and minimize the number of unclassifiable or uninformative input sequences.

For both classification tests, Kraken2-custom significantly outperformed the Kraken2-default version by 13.05% in species accuracy prediction and 14.02% in species hit accuracy, while also having lower unclassified rates, despite the default Kraken2 database being 70GB larger than CARD-r. This suggests that k-mer classifier accuracy can be greatly improved by defining a precise sequence space to build the required reference k-mer sets.

The data type prediction test demonstrated that when an informative prediction is made, CARD kmers is highly accurate in predicting the correct ARG data type for chromosomes and plasmids. However, the tool currently suffers from a high uninformative rate. Since most CARD-r alleles are from contig data (unknown source) and with the recent incorporation of genomic islands into CARD-r, CARD kmers' data type predictive utility needs to be further validated and updated for genomic island support.

One important parameter not explored here is optimal k-mer length. K-mer length affects classification precision [17]. Tools such as KITSUNE or KAT should be run on the ARG sequence space provided by CARD-r to determine optimal k-mer length for CARD k-mers [17, 18].

Currently, CARD k-mers uses a default length of k=61 while Kraken2 a length of k=35, but these lengths need to be further validated.

While not explicitly measured in this standalone comparison, CARD k-mers has a speed constraint that will be addressed in future versions. To avoid longer wait times, users can pre-process their input data to contain only ARGs before classifying them.

Overall, CARD k-mers is a novel k-mer classifier that can be used to predict the PO of sequences with ARGs and secondarily used to detect the data type of those ARGs in mere minutes, thanks to the use of its highly specialized ARG reference library, CARD-r. Such specialization can likely be extended to other sequence spaces, such as classifiers for viral or microbiome genomes.

CONFLICTS OF INTERESTS

All authors have no competing financial or non-financial interests to report.

FUNDING INFORMATION

ETHICAL APPROVAL

Not required as all sequences were generated *in-silico.*

AUTHOR CONTRIBUTIONS

ACKNOWLEDGEMENTS

REFERENCES

1. Yunger, S., & Roche Canada, H.-L. Antimicrobial-Resistance Drugs: From Drug Discovery to Access Is Canada Prepared for Their Entry? (2018). https://www.cadth.ca/sites/default/files/symp-2018/presentations/april17-2018/Concurrent-Session-E6-Antimicrobial-Resistance-Drugs.pdf

2. Larsson, D. G. J., & Flach, C. F. Antibiotic resistance in the environment. Nature Reviews. Microbiology. (2022). https://doi.org/10.1038/S41579-021-00649-X

3. Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. (2019). https://doi.org/10.1016/S0140-6736(21)02724-0

4. Maguire, F. et al. A systematic evaluation of metagenomic read-based antimicrobial resistance gene detection. Nucleic Acids Research. (2020). doi:10.1093/nar/gkn000

5. Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. (2017). https://doi.org/10.1186/S13059-017-1319-7

6. Wood, D. E., Lu, J., & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biology. (2019). https://doi.org/10.1186/s13059-019-1891-0

7. Pazda, M. et al. Antibiotic resistance genes identified in wastewater treatment plant systems - A review. The Science of the Total Environment. (2019). https://doi.org/10.1016/J.SCITOTENV.2019.134023

8. O'Neill J. Review on Antimicrobial Resistance Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations. (2014). https://amr-review.org/sites/default/files/AMR%20Review%20Paper%20-%20Tackling%20a%20crisis%20for%20the%20health%20and%20wealth%20of%20nations_1.pdf

9. Ko, K. K. K., Chng, K. R., & Nagarajan, N. Metagenomics-enabled microbial surveillance. Nature Microbiology. (2022). https://doi.org/10.1038/s41564-022-01089-w

10. Abayasekara, L. M. et al. Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: A comparative study. BMC Infectious Diseases. (2017). https://doi.org/10.1186/S12879-017-2727-8/TABLES/5

11. Styczynski P. M. et al. BLOSUM62 miscalculations improve search performance. Nature Biotechnology. (2008). https://doi.org/10.1038/nbt0308-274

12. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research. (2022). https://doi.org/10.1093/nar/gkac920

13. Pornputtapong, N. et al. KITSUNE: A Tool for Identifying Empirically Optimal K-mer Length for Alignment-Free Phylogenomic Analysis. Frontiers in Bioengineering and Biotechnology. (2020). https://doi.org/10.3389/FBIOE.2020.556413/BIBTEX

14. Mapleson, D. et al. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. (2017). https://doi.org/10.1093/BIOINFORMATICS/BTW663

FIGURES

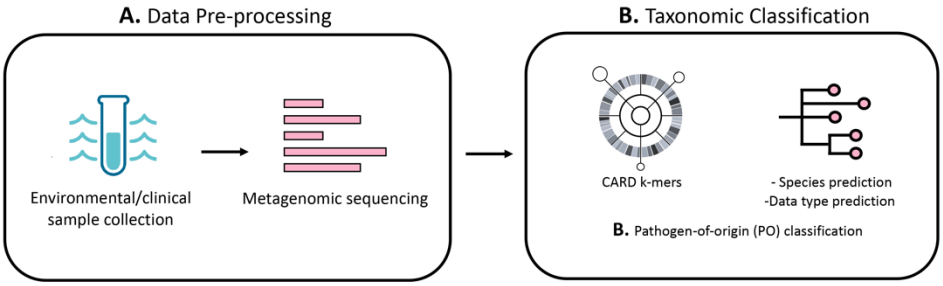


Figure 1: Top-level overview of a possible metagenomic workflow utilizing CARD k-mers for Pathogen-of-origin (PO) classification. Data must first be collected and preprocessed. **(A):** Environmental samples are captured and sequenced through culture-independent metagenomic sequencing. **(B):** CARD k-mers then queries input sequences against a precomputed database of 32 million k-mers to make an informed PO prediction. Note: only sequences with predicted ARG k-mers are suitable for classification with CARD k-mers.

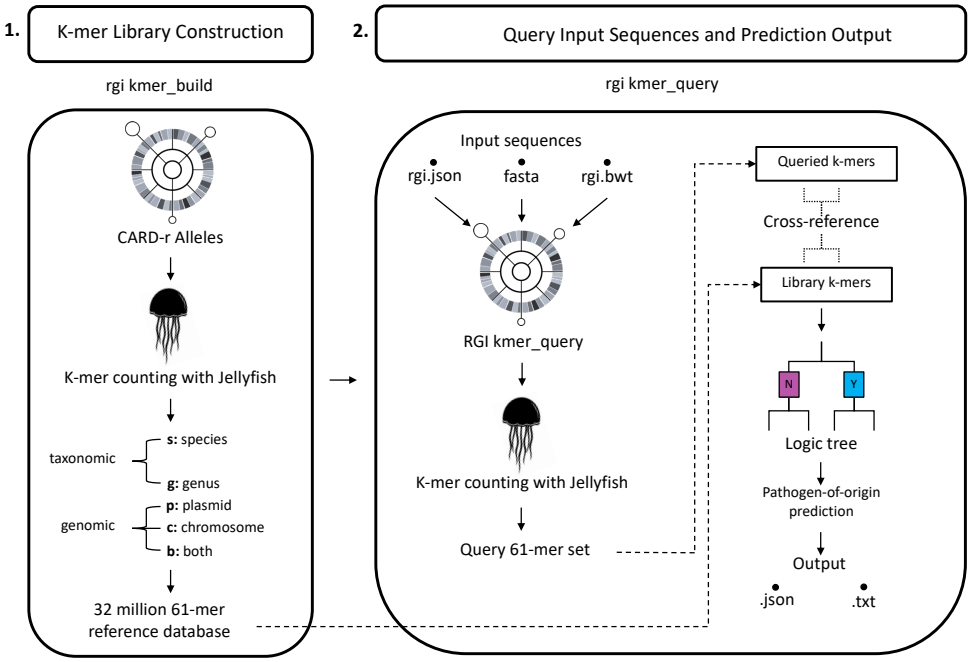


Figure 2: CARD k-mers classification algorithm. Users can classify their ARG containing sequences via a two-step process. First users must construct the reference k-mer library from CARD-r alleles using the rgi kmer_build command. As of CARD-r v.4.0.0, CARD k-mers stores over 32 million taxonomic and genomic 61-mers (1). Input sequences are then queried using rgi kmer_query (2): queried sequences are cross-referenced against the reference k-mer set using internal parameters and an independent logic tree with classification rules. The output results in pathogen-of-origin prediction of any input sequences containing ARGs. Different parameters exist to change k-mer length and the number of informative k-mers a query requires to be considered for classification.

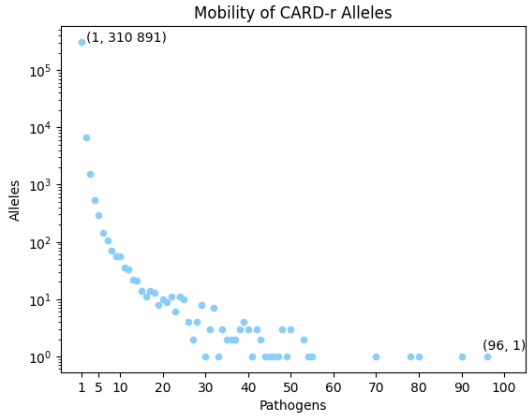


Figure 3: CARD-r v4.0.0 allele pathogen counts. CARD-r is primarily comprised of ARG alleles that have only been identified in one species. 9 722 alleles were identified in > one species. These mobile alleles can be classified with CARD k-mers to gain insight of potential species hits, and for users to gain valuable insight regarding the mobility of their ARG data if CARD k-mers predicts a plasmid genomic prediction, as these are mobile elements.

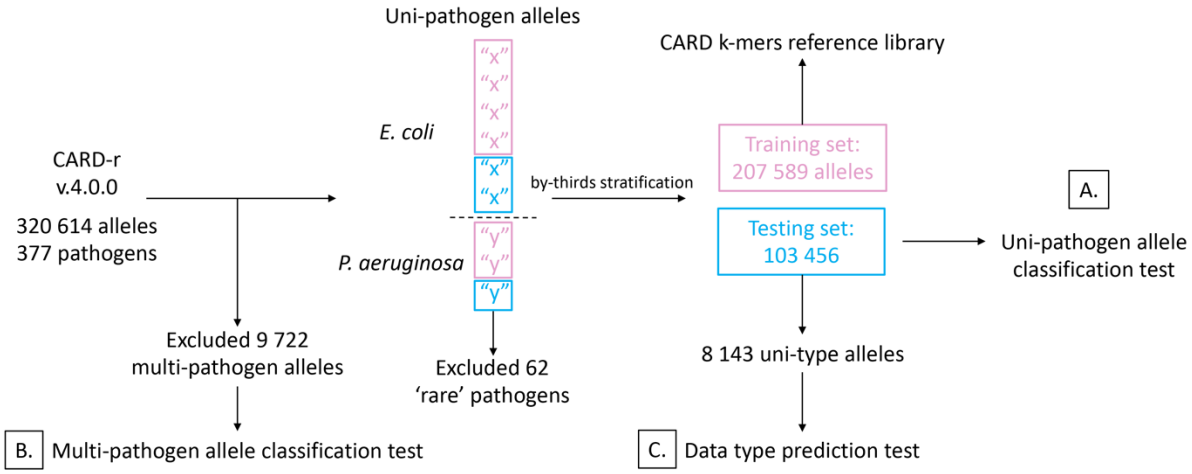


Figure 4: Validation dataset construction. Following exclusion of 9 722 multi-pathogen alleles from CARD-r v.4.0.0, ARGs from were stratified via a 'by thirds' holdout validation strategy by pathogen, to maintain the original allele distribution. The end results were a custom training and testing sets to validate CARD k-mers classification uni-pathogen classification accuracy (A). The excluded multi-pathogen alleles were classified in a separate test (B). Lastly, 8 143 uni-type alleles (only found in plasmids or chromosomes) were sourced from the testing set to validate CARD kmers' data type prediction accuracy (C).

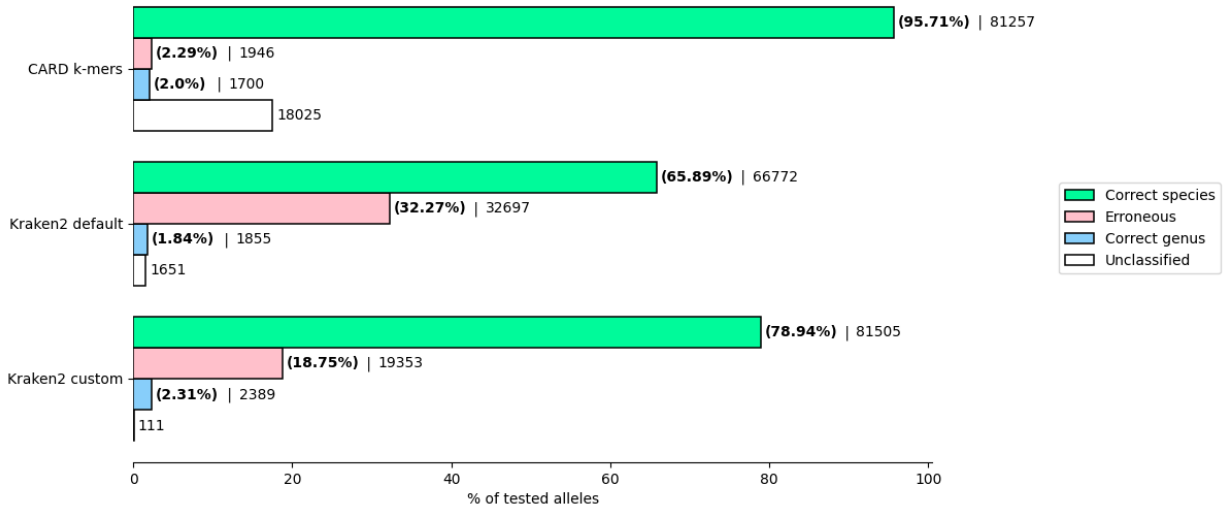


Figure 5: Classification accuracy comparison between CARD k-mers, default and custom Kraken2 versions for uni-pathogen alleles. Testing set classification results are shown for CARD k-mers (k = 61) and two Kraken2 classifiers (k = 35) each with four recorded metrics. Beside each bar is the percentage of alleles in that category, followed by the total number of alleles.

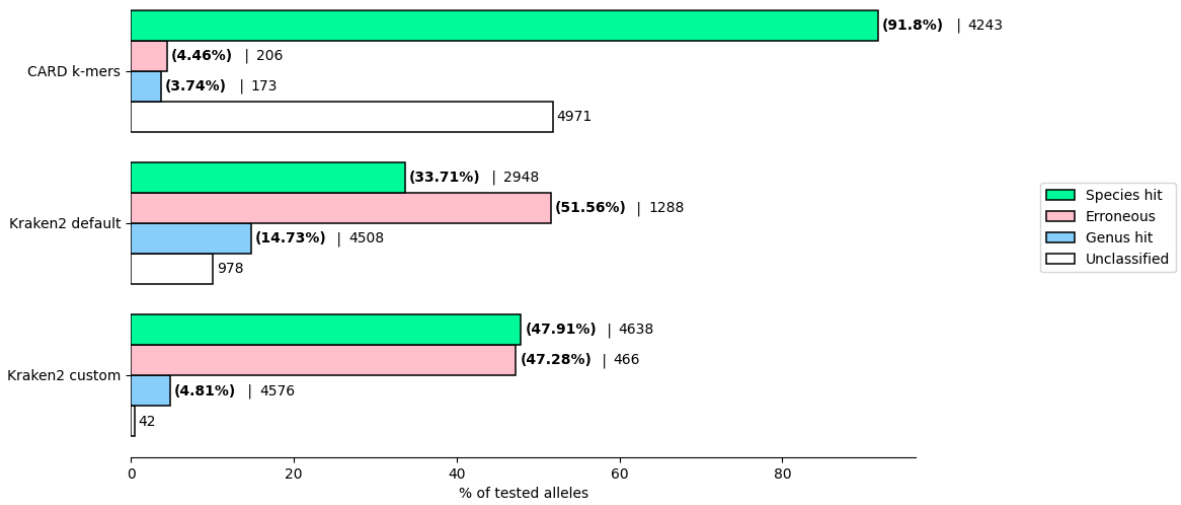


Figure 6: Multi-pathogen allele classification accuracy comparison between CARD k-mers, default and custom Kraken2 versions. Multi-pathogen allele classification results are shown for CARD k-mers (k = 61) and two Kraken2 classifiers (k = 35) each with four recorded metrics. Beside each bar is the percentage of alleles for that category, followed by the total number of alleles.

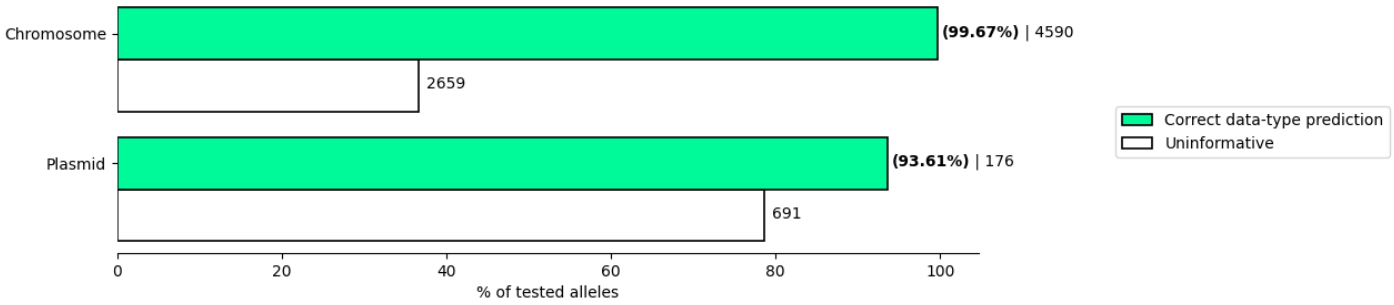
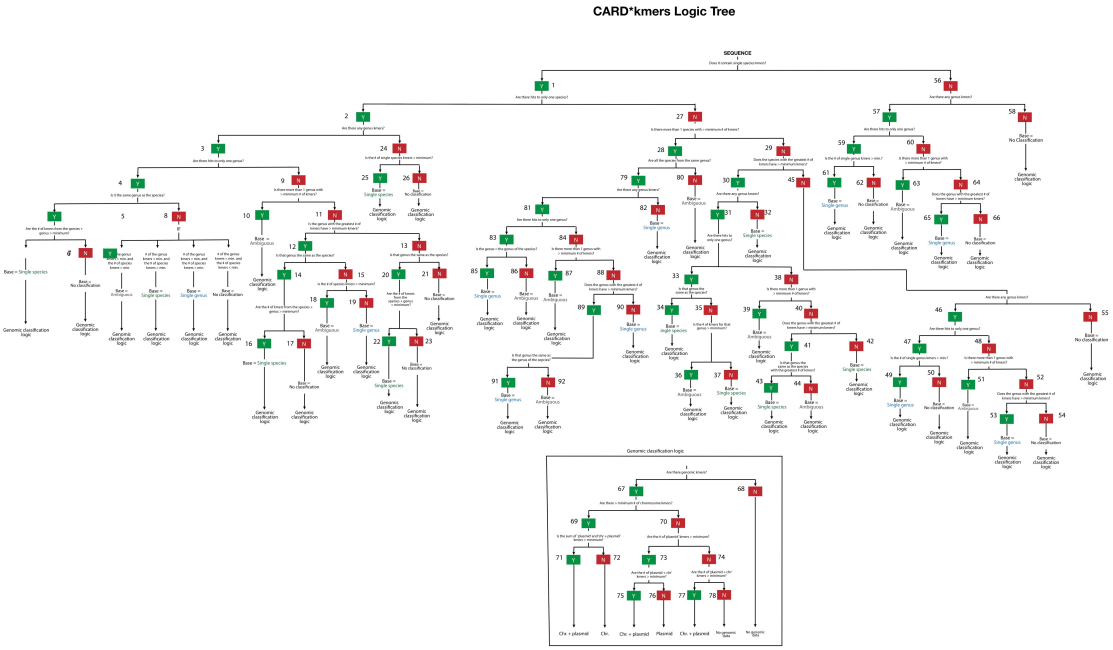


Figure 7: CARD k-mers data type prediction accuracy. Correct data type predictions are shown for ARGs originating from chromosomes and plasmids in green and the uninformative rate in white. Beside each bar is the percentage of predictions in that category, followed by the total number of alleles called.

Code	K-mer Type	Count
P	Plasmid	295 432
C	Chromosome	12 225 582
B	Plasmid + Chromosome	603 319
S	Species	27 401 594
G	Genus	1 259 938
		40 525927

Table 1: CARD k-mers reference library k-mer count. Over 40 million k-mers of length k = 61 are built and stored for taxonomic and genomic predictions. P, C, B class k-mers are used to predict the genomic location of in input ARG sequence data, while S and G class k-mers are used to predict the PO.



Appendix 1: CARD k-mers' classification logic tree. To receive a PO classification and data type prediction, input sequences are passed through this decision tree. The first goal is to provide the query sequence with a "base" value. Four possible base values exist: single species, single genus, ambiguous and no classification. Once a base value has been assigned, CARD k-mers predicts the ARG's genomic location. Four possible genomic predictions exist: chromosome, plasmid, chromosome and plasmid, or no genomic data. A more high-definition version can be visualized at https://github.com/arpcard