|  |
| --- |
| Metagenomic analysis  A Strategy and Pipeline for Visualizing Summarized Metagenomic Data on Antimicrobial Resistance Genes  John Barker1  1Johns Hopkins University  Received on 19 Dec 2022 |

[[1]](#footnote-2)\*abstract

**Motivation:** The spread of antimicrobial resistance is a growing concern worldwide. The ability to process large environmental datasets through metagenomic analysis is essential to understanding the impact of human activity on the spread of resistance genes throughout bacterial populations. An approach is proposed herein for quickly aligning known plasmids to microbiome datasets, then condensing the resultant analysis into a format for easy visualization and trend mapping to establish correlations with other known environmental factors. This solution promises generalizability to all datasets that include latitude and longitude of initial sampling

# introduction

The World Health Organization characterizes antimicrobial resistance (AMR) as a “global health and development threat,” that is, “one of the top 10 global public health threats facing humanity.” The growing use of antibiotics and other antimicrobial agents at scale – both in human medicine and to support agricultural concerns – are the most significant contributors to the reduced effectiveness of these agents (WHO, 2021). Different studies assign different significance to the contributions of agricultural antibiotic use (Manyi-Loh et al., 2018) or the general quality of water infrastructure (Nikolaos et al., 2022), but the existence of a broad and interconnected resistome allowing for the exchange and persistence of AMR genes within and between bacterial species is a known vector for the worsening of this public health challenge. (Su et al., 2017).

A wide array of tools exist for the prediction and characterization of particular genes and their homologs within a given sample from a microbiome. Unfortunately, it is a time consuming process to sift through the considerable volume of unstandardized data that is produced from these methods, and any insights gleaned from them are often specific to a particular generated gene assembly from a sequencing run. This project aims to select a very small number of features captured from metagenomic data and collect them into a condensed form that can leverage a small processed file size and the registered geographical coordinates of sampling projects to allow for location-based mapping of many different datasets, creating scatter plot maps that can provide a quick dashboard-like overview of regional variation in known or predicted antimicrobial resistance genes. The application of this approach to scan geo-tagged datasets for other genes of interest from a pre-established list of monitoring targets is beyond the scope of this project, however it may be that a similar approach to creating a small and easily maintained tool for processing and mapping insights from large data sets would be useful in addressing other public health questions as well.

# methods

This pipeline approach relies upon the existing work of Alcock et al. and the Resistance Gene Identifier tool linked to McMaster University’s Comprehensive Antimicrobial Resistance Database, however the strategy is designed to be generalizable to other environmental metagenomics questions. The RGI tool, in turn, relies upon the Prodigal tool for open reading frame (ORF) prediction and homolog detection using the DIAMOND rapid alignment tool. The CARD RGI web interface can be used to process smaller files and its command line version can be acquired and installed to process larger datasets – its predictions of antimicrobial resistance genes from metagenomic datasets taken from the Joint Genome Institute’s IMG/M database are here used to produce proof of concept visualizations using a standardized file format for easy later reprocessing.

The analysis of this data takes place in three steps: gene prediction, output parsing, and visualization, with options for user defined parameters at each step.

Initially, selected datasets were acquired from the IMG/M database and run through the RGI interface using default settings and the presets for low quality/coverage assemblies to allow for the prediction of partial genes. The RGI’s “nudge” setting was enabled to bump >95% identity hits to always clear the threshold for “strict” hit cutoffs, but the tool was also set to include Perfect and Loose hits to maximize available data for processing.

For output parsing, the rgiparser tool was implemented to quickly collect outputs from the RGI pipeline. After user definition of a manifest file associating manually entered latitude and longitude coordinates with the file paths of data to be processed, the tool will create a summarized list of identified resistances resulting from an RGI-predicted AMR gene and collect them together into a .csv file that can then be passed to the third visualization step of the process. This step is kept isolated from the others to improve modularity such that other scripts might be developed using the same summarized output format, and so that once these large files have been processed once, they need not be re-scanned every time a new visualization is needed.

The final step of this process is the creation of the end product – an interactive summary visualization of the processed data using the Plotly library to place scaled dots representing different incidences of antimicrobial resistance at the latitude and longitude coordinates of data sampling. The creation of this output is shaped by the settings in the user defined configuration file which allow for the application of a cutoff for sequence identity, a custom title for the visualization project, and a visualization of the total number of hits within an area. The program will also generate a histogram capturing the distribution of classified drug resistances within the provided data set, to provide further detail and insight.

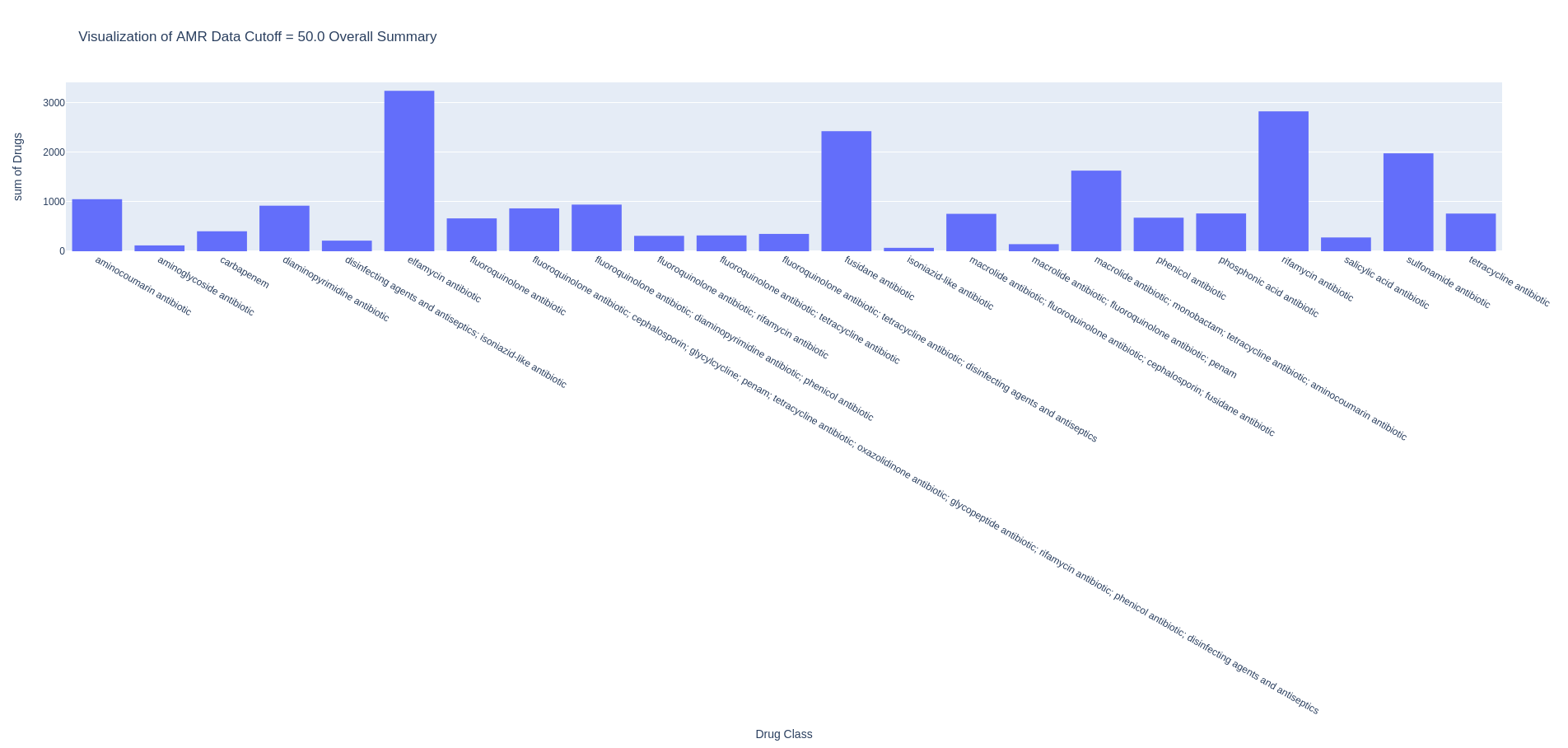
# results

For this study, three metagenomic datasets of human-connected microbiomes from a city subway, a groundwater contamination site, and a wastewater treatment plant were assessed using the RGI tool. These files were then loaded into the script and used to generate summarized data sets, which were later grouped according to the RGI model’s characterization of their antimicrobial resistance type. The resultant figures are captured below. Noteably, RGI did not identify antimicrobial resistance genes, even using the most permissive settings, for prokaryotes sequenced from the New York City subway, and so no hits are plotted for this data. A dummy entry is also included for reference to document a small number of hits, set with the coordinates of JHU in the manifest file.

**Figure 1. A color coded frequency plot of number of AMR gene hits above the cutoff threshold (50% sequence identity)**

****

**Figure 2. A histogram summarizing the data and demonstrating the distribution of types of AMR resistance genes**



This tool can be extended to process far larger data sets spanning far greater geographic areas, and the approach captured here should be considered when attempting to make sense of a number of large metagenomic datasets. While fragment recruitment plots and other tools are essential to understanding individual metagenomic reads, the bigger picture requires summarization and visualization in a more immediately human-readable manner. Encouragingly, the Comprehensive Antimicrobial Resistance Database and the SARG database are both attempting to do just that, creating live readouts mapping time and place of their datasets onto a global monitoring tool, although further work is needed to ensure that this is ultimately completed.

acknowledgements

Professor Joshua Orvis assisted with the planning of and impetus for this project.

References

Alcock et al. 2023. CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, in press.

Benjamin Buchfink, Chao Xie & Daniel H. Huson, Fast and Sensitive Protein Alignment using DIAMOND, Nature Methods, 12, 59–60 (2015) doi:10.1038/nmeth.3176.

Manyi-Loh C, Mamphweli S, Meyer E, Okoh A. Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications. Molecules. 2018 Mar 30;23(4):795. doi: 10.3390/molecules23040795. PMID: 29601469; PMCID: PMC6017557.

Nikolaos I. Rousis, Zhe Li, Richard Bade, Michael S. McLachlan, Jochen F. Mueller, Jake W. O'Brien, Saer Samanipour, Benjamin J. Tscharke, Nikolaos S. Thomaidis, Kevin V. Thomas, Socioeconomic status and public health in Australia: A wastewater-based study, Environment International, Volume 167, 2022, 107436, ISSN 0160-4120, <https://doi.org/10.1016/j.envint.2022.107436>.

Pillay, Manoj. Sewage treatment plant microbial communities from Vermont, USA – HR2\_B, JGI Accession #198747/IMG3300030034

Sorted cell/s from water treatment plant in SURF, South Dakota, USA - Parcubacteria bacterium SCGC\_AG-578-G05. Accession: PRJNA444638 ID: 444638

Sorted cells from water treatment plant in Sanford Underground Research Facility (SURF), South Dakota to study Microbial Dark Matter phase II

Stegen, James. Coupling Microbial Communities to Carbon and Contaminant Biogeochemistry in the Groundwater-Surface Water Interaction Zone: GW-RW metaG T4\_4-Nov-14. JGI IMG 1066008. https://www.osti.gov/award-doi-service/biblio/10.46936/jejc.proj.2014.48473/60005497

Su, JQ., An, XL., Li, B. et al. Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. Microbiome 5, 84 (2017). https://doi.org/10.1186/s40168-017-0298-y

Urban prokaryotic and eukaryotic communities from the subway and surrounding areas in New York, USA. Study ID: Gs0118444  Project ID: Gp0134944  Analysis ID: Ga0120709

World Health Organization. (2021) Antimicrobial resistance. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance (Accessed: December 8, 2022).

Yin, X, X-T Jiang, B. Chai, L. Li, Y. Yang, J. R. Cole, J. M. Tiedje\*, and T. Zhang\*.(2018) "ARGs-OAP v2. 0 with an Expanded SARG Database and Hidden Markov Models for Enhancement Characterization and Quantification of Antibiotic Resistance Genes in Environmental Metagenomes." Bioinformatics 1:8. DOI:10.1093/bioinformatics/bty053.

1. \*\*To whom correspondence should be addressed. [↑](#footnote-ref-2)