Enabling a complete antibiotic resistance annotation using crowdsourcing

G. A. Arango-Argoty1, A. Pruden2, L. S. Heath1, P. Vikesland2, and L. Zhang1\*

1Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

2Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA

\* Corresponding author (lqzhang@vt.edu)

# **ABSTRACT**

# **INTRODUCTION**

What is antibiotic resistance?

What have been done for the annotation of ARGs?

What is crowdsourcing and why is useful for the annotation of ARGs?

What have been done using crowdsourcing or another area?

What is the potential of crowdsourcing?

None of the ARG databases cited in this paper make a distinction between chromosomal ARGs with those that are present in Mobile Genetic Elements (MGEs), such as plasmids and transposons; which can effectively mobilize ARGs to other bacteria via horizontal gene transfer (HGT) [1].

# **CROWDSOURCING**

What type of crowdsourcing is used?

Why type of model is employed for the annotation?

How spammers are avoided or how to avoid misclassifications?

Challenges of the crowdsourcing methods?

# **METHODS**

## ARG DATABASE

Antibiotic resistance genes were collected from different resources. First, the Comprehensive Antibiotic Resistance Database (CARD) [2] which to date is the most exhaustive resource for antibiotic resistance gene information. The ARDB [3] database that comprises a vast number of homology predicted ARGs. The Deep learning Antibiotic Resistance database (DeepARG-DB) [4] which incorporates antibiotic resistance genes from the UniProt [5] database with the CARD and ARDB databases. Lastly, the MEGARes [6] database that simplifies the nested structure of the CARD database. MEGARes incorporates genes from the ARG-ANNOT [7], RESFINDER [8] and the Lahey Clinic beta-lactamase archive [9] from the National Center for Biotechnology Information (NCBI). MEGARes also integrates the antibiotic resistance mechanisms.

To obtain a clean collection of ARGs, the deepARG database was updated with the latest version of the CARD and UniProt databases using their sequence identifiers. Several genes were then integrated or removed from the deepARG database depending on their evidence. For instance, deprecated UniProt sequences were removed from deepARG-DB, whereas, the newly ARGs from CARD were integrated. Also, genes that where erroneously included in CARD that confer resistance due mutations were removed. The resulting collection of ARGs was then aligned to the CARD, ARDB and MEGARes databases using DIAMOND and TBLASTN to subtract the best hit of each ARG along with its metadata. Therefore, each ARG is represented by its best hits through the databases that allows annotation consistency through the ARG resources. Because the deepARG-DB contains information about the origin of the ARGs, the metadata from the UniProt database is accessed via the UniProt API (Application Programming Interface) that allows the retrieval of updated information for each gene. The goal of this process is to link the ARGs sequences to their most up to date metadata. Therefore, each ARG is depicted in the user interface as a set of panels containing the ARG’s best hits, their metadata and the alignment quality (see **Figure X.1**). Scores are ranked in a color scale to enhance readability and human interpretation.

## MOBILE GENETIC ELEMENTS

The ACLAME (Classification of Mobile Genetic Elements) database [10] was used to identify the ARGs that have the potential of being mobilized by MGEs. DIAMOND [11] was used to perform the comparison of ARGs to MGEs via sequence alignment (parameters e-value < 1e-10). Alignment information along with MGEs metadata is present in the user guide for workers to make a decision whether the ARG has enough evidence of being carried by an MGE or not. This, evidence is scored from 0 to 5. Colors, depict the degrees of confidence of the information present in the MGE panel (see **Figure X.2**).

## PATHOGENS

98,758 bacterial genomes were downloaded from the PATRIC [12] database. This resource comprises information about bacterial pathogenicity, antimicrobial resistance phenotype, diseases and host organisms. This information is valuable to identify ARGs that are present in pathogens. For instance, the gene xxx was present in xxx bacterial genomes, out of those xxx are pathogens. On the other hand, the gene xxx was present in xxx genomes, but, none of those genomes corresponded to a pathogenic genome. The collection of ARGs were then screened against the genome sequences from PATRIC using DIAMOND. To ensure the quality of the assignments, all genes with an identity below 90% and a minimum coverage of 90% were discarded. Workers are asked to rate the pathogenicity of an ARG based on the evidence provided (frequency of pathogenic genomes, diseases, antimicrobial phenotype and hosts. See **Figure X.3**).

## ARG-MINER PLATFORM

REMOVING CROWDSOURCING NOISE

CROWDSOURCING CONSISTENCY

EXPERIMENT INTERFACE

ADMINISTRATION INTERFACE

## CASE STUDIES AND EVALUATION

GENE XXX

GENE XXX

EVALUATING THE CROWD

## ARG-MINER API

How is the method designed?

Insert a figure here showing the general process backend and front end

Which are the components of the platform annotation?

What is the database collection and information obtained?

How the data was compared to what ARGs?

How is the system being evaluated?

How each worker is scored?

How each gene is scored?

How is the evaluation made?

Which platform used for the crowdsourcing?

How the platform is dedicated to non-experts?

# **RESULTS**

Average annotation

Average time for annotation

Number of different workers

Invalid user responses

How the scammer-free strategy improved the performance of the annotations?

Quality of the workers and quality of the classifications.

# **FUTURE WORK**

# **REFERENCES**

1. Frost, L.S., et al., *Mobile genetic elements: the agents of open source evolution.* Nature reviews. Microbiology, 2005. **3**(9): p. 722.

2. McArthur, A.G., et al., *The comprehensive antibiotic resistance database.* Antimicrobial agents and chemotherapy, 2013. **57**(7): p. 3348-3357.

3. Liu, B. and M. Pop, *ARDB—antibiotic resistance genes database.* Nucleic acids research, 2008. **37**(suppl\_1): p. D443-D447.

4. Arango-Argoty, G.A., et al., *DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data.* bioRxiv, 2017: p. 149328.

5. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2014: p. gku989.

6. Lakin, S.M., et al., *MEGARes: an antimicrobial resistance database for high throughput sequencing.* Nucleic acids research, 2017. **45**(D1): p. D574-D580.

7. Gupta, S.K., et al., *ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.* Antimicrob Agents Chemother, 2014. **58**(1): p. 212-20.

8. Zankari, E., et al., *Identification of acquired antimicrobial resistance genes.* J Antimicrob Chemother, 2012. **67**(11): p. 2640-4.

9. Bush, K. and G.A. Jacoby, *Updated functional classification of β-lactamases.* Antimicrobial agents and chemotherapy, 2010. **54**(3): p. 969-976.

10. Leplae, R., et al., *ACLAME: a CLAssification of Mobile genetic Elements.* Nucleic acids research, 2004. **32**(suppl\_1): p. D45-D49.

11. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. **12**(1): p. 59-60.

12. Wattam, A.R., et al., *PATRIC, the bacterial bioinformatics database and analysis resource.* Nucleic Acids Res, 2014. **42**(Database issue): p. D581-91.