ARG-miner: A web platform for crowdsourcing-based curation of antibiotic resistance genes

G. A. Arango-Argoty[1], G. K. P. Guron[2,3], E. Garner[2], M.V. Riquelme[2], L. S. Heath[1], A. Pruden[2], P. Vikesland[2], and L. Zhang[1*]

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

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

[3]Department of Food Science and Technology, Virginia Tech, Blacksburg, VA, USA

* Corresponding author (lqzhang@vt.edu)

ABSTRACT

Curation of antibiotic resistance gene (ARG) databases is a labor-intensive process that requires expert knowledge to manually collect, correct, and/or annotate individual genes. Correspondingly, updates to existing databases tend to be infrequent, commonly requiring years for completion and often containing inconsistences. Further, because of limitations of manual curation, most existing ARG databases contain only a small proportion of known ARGs (~5k genes). A new approach is needed to achieve a truly comprehensive ARG database, while also maintaining a high level of accuracy. Here we propose a new web-based curation system, ARG-miner, which supports annotation of ARGs at multiple levels, including: gene name, antibiotic category, resistance

mechanism, and evidence for mobility and occurrence in clinically-important bacterial strains. To overcome limitations of manual curation, we employ crowdsourcing as a novel strategy for expanding curation capacity towards achieving a truly comprehensive, up-to-date database. We develop and validate the approach by comparing performance of multiple cohorts of curators with varying levels of expertise, demonstrating that ARG-miner is more cost effective and less time-consuming relative to traditional expert curation. We further demonstrate the reliability of a trust validation filter for rejecting confounding input generated by spammers. Crowdsourcing was found to be as accurate as expert annotation, with an accuracy >90% for the annotation of a diverse test set of ARGs. ARG-miner provides a public API and database available at http://bench.cs.vt.edu/argminer.

INTRODUCTION

Antimicrobial resistance (AMR) has been identified by the World Health Organization (WHO) as a major global health threat. It is projected that AMR will increase exponentially by 2050, leading to substantial human morbidity and mortality (1,2). Therefore, swift action is required to enable enhanced monitoring and help tackle the spread of AMR, including: understanding the mechanisms controlling dissemination of antibiotic resistance genes (ARGs) via environmental sources and pathways (3-5), discovering novel and newly evolved ARGs before they are found to be problematic in the clinic (6), developing new computational strategies for ARG annotation (7-11), and expansion of current ARG repositories (7,9).

Metagenomics-based approaches have proved to be a powerful means of accessing the diverse array of ARGs, or "resistomes," (12) characteristic of various environments (13-16) and has supported the discovery of novel ARGs and their interactions (17,18). Metagenomic data can be expressed in terms of "total ARGs" or can be further mined for ARGs corresponding to specific antibiotics or mechanisms of interest. However, existing metagenomics approaches are largely dependent upon predicting the antibiotic resistance attributes through sequence similarity computation, which is subject to major limitations. First, such similarity computations require a high quality and up-to-date ARG reference/annotation database to enable consistent and accurate ARG identification. Second, the scope of these kinds of analyses is limited to previously characterized ARGs, either due to the parameter cutoff stringency employed in the sequence alignment or to lack of a comprehensive target gene for alignment (10).

To improve the capacity of metagenomic-based approaches to broadly and accurately detect the full range of ARGs present in a given sample, it is necessary to continuously expand and improve curation of corresponding databases (7). However, risk of incorporation of false positives, i.e., "ARG-like" genes that do not necessarily induce an AMR phenotype, stands as a major impediment to expanded curation efforts. Therefore, inspection and validation of new ARG entries is a critical aspect of ensuring the validity of AMR databases and their application. Manual curation of ARGs is typically carried out by a few researchers associated with particular laboratories. This process can be complex, tedious, and time-consuming. For instance, the last update of the Antibiotic Resistance Database (ARDB) was in 2009 (19) and therefore it does not contain any

newly discovered ARGs, such as *bla*$_{\text{NDM-1}}$ or *mcr*-1. The MEGARes database (9), which was designed to simplify the organization of ARG annotation, has not been updated since December, 2016. The SARGs database (10), which integrated genes available from the Comprehensive Antibiotic Resistance Database (CARD) and ARDB, has not had any update since its first release on August, 2016. In addition, the SARGs database is closed to the public, which limits its use and prevents integration with additional databases and other applications. Similarly, the resqu database, which contains genes for which there is evidence of having been transferred via Mobile Genetic elements (MGEs), has not been updated since 2013 (20) and it is also closed to the public, limiting its use. Further, the ARG-ANNOT resource (21), released in 2014, was updated most recently in March 2017. In summary, only a handful of ARG databases have been maintained and updated, albeit infrequently.

We recently introduced the Deep learning Antibiotic Resistance Gene DataBase (DeepARG-DB) (7) (first released in July, 2017 and most recently updated August, 2017), which employs manual curation, literature review, and computational-based annotation. The CARD database (11), perhaps the most up-to-date ARG resource, was most recently updated in October 2017 and, since its first major upgrade in 2016, it has been updated a total of 21 times with changes to the ARG sequences and metadata (e.g., antibiotic class, gene name, and mechanism). The multiple updates of CARD illustrate how complex and time-consuming this task can be, even for domain experts. Another problem that encompasses all ARG resources is the lack of a standardized nomenclature. For instance, the gene aadA1 is also named ANT(3'')-I, the BacA gene is

also named UppP, and the *tet*(A) gene can be found as TetA and tetA, depending on the ARG resource. On the other hand, most ARG databases contain approximately 4,000 genes, with the exception of DeepARG, which contains approximately 15,000 genes. For DeepARG-DB database, even a simple manual curation is essentially an impossible task.

Another major concern in AMR research is the identification of mechanisms for the ARGs mobility within and among bacterial species (22). In this aspect, Mobile Genetic Elements (MGEs) have been recognized as a major player that facilitates the transmission of ARGs via horizontal gene transfer. For instance, it has been found that plasmids are responsible for the transmission of particular beta-lactamase resistance genes ($bla_{\text{TEM-52B}}$, $bla_{\text{TEM-52C}}$, bla_{KPC}) among different bacterial strains (23) (24). Thus, it is important to detect ARGs that have the potential of being transferred by MGEs.

To overcome the difficulties in curation and manual validation of an extensive number of ARGs, a novel approach that breaks down this complex task into simpler and smaller microtasks is proposed. The core of this methodology consists of a compendium of AMR resources and a crowdsourcing strategy, which simplifies the ARG information to allow nonexperts, the general public, and domain experts collectively to execute curation of the ARG database.

Application of crowdsourcing in biology, particularly for data curation, is not new and comprises a variety of areas including: name entity recognition (NER) for drug and

diseases (25-27), identification of medically-relevant terms from patient online posts (28), annotation of diseases described in PubMed (29), systematic examination of databases and other resources for drug indications, biomedical ontologies and gene-disease interactions (26,30-33), identification of the relationship between genes and mutations (34), and annotation of medical data for electronic health records (35). Interestingly, in most of the studies, crowdsourcing has proven to be as effective as expert curation (26,36).

Here we introduce an online platform for the manual curation of ARGs. The system, called ARG-miner, enables users to curate and retrieve all the information available from several ARG resources including the DeepARG-DB (7), CARD (11), ARDB (19), MEGARes (9), and UniProt (37) databases. Moreover, to provide information on whether an ARG might be present in a pathogen, the PATRIC (38) database is used. ARG-miner provides evidence of ARGs that are potentially carried by MGEs, particularly plasmids, viruses or prophages. This information is obtained by looking at the gene sequences and metadata from the Classification of Mobile Genetic Elements database (ACLAME) (37). The ARG-miner platform is designed, built, and implemented as an open-source project facilitating a collaborative and integrative approach for the standardization of ARG annotation by the broader community of scientists and citizens with a desire to contribute towards combatting the spread of AMR. All data associated with ARG miner, as well as the source code, is available under a public repository accessed freely online at http://bench.cs.vt.edu/argminer.

MATERIALS AND METHODS

ARG Database

ARGs were downloaded from the following resources: CARD (11), which contains ARG information; the ARDB (19) database; which comprises a vast number of homology-predicted ARGs; DeepARG-DB (7), which integrates ARGs from UniProt (39), CARD, and ARDB; and the MEGARes (9) database, which incorporates genes from the ARG-ANNOT (21), RESFINDER (40), and the Lahey Clinic beta-lactamase archive (41) available from the National Center for Biotechnology Information (NCBI).

To obtain a clean collection of ARGs, the DeepARG-DB database was updated with the latest version of the CARD (v 1.1.8) and UniProt databases using their corresponding sequence identifiers. Discontinued UniProt sequences were removed from DeepARG-DB, whereas the newly-added ARGs from CARD were incorporated. Also, genes from CARD known to confer resistance due to single point mutations were removed. The resulting collection of ARGs was then aligned to the CARD, ARDB, and MEGARes databases using DIAMOND (42) and TBLASTN (43) to extract the best hit of each ARG along with its corresponding metadata. In this manner, each ARG is represented by its best hit to each database, upholding consistency in annotation among the ARG resources. Because 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), which allows retrieval of up-to-date information for each gene. Therefore, each ARG is displayed in the user interface as a set of sections containing an ARG's best hits, its metadata, and the alignment quality. Scores are ranked

160 according to a color scale to enhance readability and human interpretation (see
161 **Supplementary Figure S3-A)**.
162
163 **Mobile Genetic Elements**
164 The ACLAME database (37) was used to identify ARGs that have potential of being
165 mobilized by MGEs (plasmids, viruses, and phages). DIAMOND (42) was used to
166 perform the comparison of ARGs to MGEs via sequence alignment (parameters e-value
167 < 1e-10). Alignment information along with MGE metadata is presented in the interface
168 for users to make a decision on whether an ARG has enough evidence of being carried
169 by an MGE or not. This evidence is scored from 0 to 5. Color depicts the degree of
170 confidence for the information presented in the MGE panel (see **Supplementary Figure**
171 **S3-B**).
172
173 **Pathogens**
174 A total of 98,758 bacterial genomes were downloaded from the PATRIC (38) database.
175 This database contains information about bacterial pathogenicity, antimicrobial
176 resistance phenotype, corresponding diseases, and host organisms. The information is
177 valuable for identifying ARGs that are present in pathogens. For instance, the gene
178 entry BAE06009.1 was present in 2,037 bacterial genomes, of which, 1,004 belong to
179 pathogenic bacteria, 40 are involved in cystic fibrosis disease in humans, and 706
180 exhibit intermediate and resistant phenotypes (see **Supplementary Figure S3-C**). The
181 collection of ARGs were then screened against the genome sequences from PATRIC
182 using DIAMOND. To ensure the quality of the assignments, all genes with an identity

below 90% and an alignment coverage below 90% were discarded. Users are asked to rate the pathogenicity of known bacterial hosts of ARGs based on the evidence provided by PATRIC (frequency of pathogenic genomes, diseases, antimicrobial phenotype, and hosts).

Annotation microtasks

An annotation task consists of labeling ARGs based on the evidence provided on the web site. Users are requested to classify an ARG in terms of gene name, antibiotic class, and antibiotic mechanism. In addition, users are asked to rank the evidence of ARGs being carried by MGEs and pathogen genomes. A user-friendly web interface makes it easy to follow the annotation process. By employing simple tasks and a crowdsourcing strategy, ARG-miner advocates mass collaboration from an open community that includes experts and the general public.

Expert gold standard data set

To assess the accuracy and quality of classifications generated by crowd-sourcing, three domain experts who are actively engaged in shotgun metagenomic-based ARG research applied to various environments were asked to annotate a gold standard data set of 35 ARGs. Experts were required to annotate the ARGs by their name, antibiotic class, and mechanism. In total, 34 out of the 35 ARG annotations were in agreement among at least two of the three experts in terms of antibiotic class and gene name. These 34 ARGs were further considered in downstream analysis.

Crowdsourcing microtasks

Annotations were obtained using Amazon Mechanical Turk (MTurk), an online platform that allows access to a broad crowdsourcing audience to perform Human Intelligent Tasks (HITs). ARG-miner requests were submitted to MTurk in batches; crowdsourcing workers were requested without prior knowledge of ARGs. Then, once a worker performed a microtask, the system prompted a token number that workers needed to submit to the MTurk web site to validate the annotation and to obtain a monetary reward. Because of the high diversity of MTurk workers, the ARG-miner HITs were opened to a broad audience including domain experts and nonexperts. In addition, users were allowed to perform multiple annotations (maximum 20). Finally, workers were asked to indicate their expertise and confidence for each annotation performed. This information was used to score the individual ARG classifications (explained under Section **Annotation Score**). On the other hand, a diverse domain-knowledge group of workers were also evaluated; this group was consisted of students enrolled in a graduate-level microbiology class. Not all of them had antimicrobial resistance knowledge, but they were at least familiar with microbiology in general. While they did not receive a monetary reward, they were directed to follow the same instructions as the MTurk workers.

User interface

The ARG-miner interface has three main components or sections:

1. **Current Annotation**: comprises the known information available for an ARG. It consists of the gene name, antibiotic class, the database from which the

sequence was extracted, and number of times the gene has been inspected by workers/users (see **Figure 1A**).

2. **Evidence**: corresponds to the metadata available for the ARG as well as the best hit from the CARD, ARDB, and MEGARes databases. It also provides evidence/information on whether the gene is likely carried by an MGE (the ACLAME database) and whether the gene can be found in a pathogen genome (the PATRIC database, see **Figure 1B**).
3. **Microtask**: refers to the section where a worker/user enters his/her annotation. The information in this panel has to be consistent with the observations from the evidence. It comprises three simple steps. First, workers must validate the gene name, antibiotic class, and mechanism by looking at the Evidence section. Second, workers must rank the MGE and pathogen evidence. Third, workers must rank their annotation overall by scoring their expertise (how familiar are they with ARGs) and confidence (how strong is the evidence, see **Figure 1C**).

The web interface provides a training step for new users that is mandatory for AMT workers (required for getting a monetary reward). The goal of this step is to familiarize the workers with the platform environment by performing two microtasks. ARG-miner also provides a list of problematic ARGs that have inconsistent annotation. These problematic ARGs are identified by comparing the annotation of the genes with their best hits from ARDB, CARD, and MEGARes. All tests performed during validation were completed using these problematic ARGs.

ARG-miner also provides an administrative interface to update the ARG database. This interface comprises a set of figures that show the distribution of different labels as well as the MGE and pathogenic evidence scores. In this interface, ARG-miner administrators are able to accept/reject the annotations made by the crowd and update the ARG database (see **Supplementary Figure S1**).

Annotation score

Each gene g_i is classified into three annotation fields ($F_{mechanism}$, $F_{antibiotic\ category}$ and $F_{gene\ name}$), were each field is composed by a list of labels $L \in \{mechanism, antibiotic\ category, gene\ name\}$. For example, the $F_{antibiotic\ category}$ field contains the set of labels $L_{antibiotic\ category}$ that corresponds to the name of the antibiotic categories (e.g., multidrug, beta-lactam, peptide, aminoglycoside). Thus, each gene g_i contains a set of annotations $A_{L_k}^{F_k}(g_i) = \{a_{l_1}^{F_k}, \ldots, a_{l_p}^{F_k}\}$, where p is the total number of labels, with each element $a_{l_j}^{F_k}$ corresponding to the number of workers that assigned the label l_j to the gene g_i.

The ARG-miner score uses the majority voting strategy described in (44), but it is weighted by the evidence, expertise, and the validation scores. Therefore, the annotation score of the gene g_i for the label L_p of the field F_k (Eq 2) is calculated as follows:

$$S_{F_k,L_p}^{g_i} = V_p * H_p * \frac{a_{L_p}^{F_k}}{\sum_p a_{L_p}^{F_k}}, \quad (1)$$

$$\mathrm{H}_p = \frac{\sum_p E_p * X_p}{25 * a_{L_p}^{F_k}}, \quad (2)$$

$$V_p = \frac{L_p \cap \Lambda}{L_p}, \quad (3)$$

where

H_p (Eq 2) describes the expertise (X_p) and confidence (E_p) scores normalized to the $[0,1]$ interval. V_p (Eq 3) computes the similarity between the label L_p and the gene evidence Λ.

Trust validation filter

Because of the unsupervised nature of crowdsourcing, users can provide erroneous feedback or just ignore the evidence and enter random inputs. Under an uncontrolled scenario, spammers can even get a monetary reward. More critically, too much random and/or erroneous feedback can increase the variance in ARG annotations and propagate annotation error. To circumvent the problem, ARG-miner implements a trust-validation filter and use the metric $\mathbf{V_p}$ (computed by Equation 3) to evaluate whether the input corresponds to real evidence or not. $\mathbf{V_p}$ is computed in real time, and, unless the user provides valid information, the system will not proceed to the next stage. **Figure 2** shows an example of a user providing erroneous input. The user entered *polyamine* as the antibiotic class, whereas the evidence shows that this gene belongs to the *polymyxin* antibiotic class. Despite the similarity between the two words, the system will not allow the submission until the answer has a minimum $\mathbf{V_p}$ score of 50 (50% of similarity) in the evidence section.

RESULTS AND DISCUSSION

To assess the effectiveness of the crowdsourcing approach for ARG annotation, three experiments were performed with the following contributes:

1. A set of crowdsourcing workers from MTurk, referred to as **AMT-Free**. In this scenario workers were paid $0.10 for each annotation, with the trust validation filter disabled to examine the reliability of the crowd. Therefore, workers could input anything as feedback without restriction. A total of 100 annotations were requested on MTurk for this test.
2. A second batch of crowd workers from MTurk, referred to as **AMT-Val**. In this case the trust validation filter was enabled. The main purpose of this experimental group was to measure the effectiveness of the trust validation filter. In this scenario, workers were paid $0.05 per annotation, with a total of 200 requested annotations.
3. A group of users with general microbiological knowledge, with varying levels of experience in ARG research, referred to as **LAB**. This group comprises Masters and Ph.D. students from a microbiology class at Virginia Tech. They completed this work as an assignment and did not receive any monetary reward. Here the annotations were performed with the validation filter on and each worker was requested to perform 15 annotations (540 microtasks in total). The goal of the LAB scenario was to compare its performance against the nonexpert community of MTurk (AMT-Val, AMT-Free).

Effectiveness of the trust validation filter

"Spammers" are workers that intend to obtain monetary reward by submitting invalid information, which is a major confound of crowdsourcing. In the present study, although the ARG-miner website provides workers with detailed instructions on how to handle the annotation process, many of the **AMT-Free** workers submitted misleading and/or unrelated feedback. Particularly, for the antibiotic category annotation task, workers must choose the antibiotic class that they believe the gene belongs to from a dropdown menu that contains a list of antibiotic classes. Results indicated that many **AMT-Free** workers simply picked up the first option on the dropdown menu, most likely without reading the evidence section of the web page. Therefore, most of the antibiotic class annotations under the AMT-Free group were labeled as aminoglycosides. This is a serious hurdle to accurate database curation and indicates the need for a real time control that guarantees correctness of the annotation (see Section Trust validation filter). In terms of performance, as expected, the AMT-Free group achieved very low scores for all annotations (**Figure 3**). However, not all workers were spammers. Interestingly it was observed that workers who performed more than ten microtasks responded correctly and consistently to their observations and evidence. In addition, this test was designed with the aim to evaluate the impact of controlling the validity of the worker's feedback and to check the performance of the proposed real time validation. Thus, after integration of the trust validation filter, MTurk workers were not allowed to input false annotations (see **Figure 3**). As a result, the performance of the AMT-Val workers improved significantly (p-value<1e-10) for all the fields (antibiotic class, ARG name, and ARG mechanism) over the AMT-Free group. Under the new

342 policy, MTurk workers were not allowed to continue with the microtask until their
343 annotation was valid (the input was related to the evidence), as **Figure 2** shows. In this
344 test, all nonsense input was completely removed, and all annotations from the AMT-Val
345 group corresponded to actual ARG evidence. These results demonstrate the
346 effectiveness of the trust validation filter for the control of spam annotations. In addition,
347 it was imperative to test the performance of the MTurk workers against domain
348 knowledge users. The main goal of this test case was to investigate whether a
349 nonexpert crowd community (AMT-Val) can perform a complex task in a similar fashion
350 to a group of workers with domain-knowledge (LAB). As expected, the LAB workers
351 achieved a much higher average score (0.146) than the AMT-Free workers (0.06), but,
352 surprisingly, a rather similar score to the AMT-Val workers (0.114). This shows that
353 crowdsourcing is indeed a powerful alternative to manual inspection and annotation of
354 ARGs. As expected, MTurk annotations (AMT-Val) had a higher variance compared to
355 the LAB group in all annotation fields, but the two distributions are not significantly
356 different (Kolmogrov-Smirnov test: p-value > 0.05).
357
358 **Effectiveness of the scoring strategy**
359 To evaluate the quality of the scoring strategy, four genes were selected among the
360 total set of curated genes and examined in greater detail, as illustrated in Figure 4. For
361 instance, the UniProt entry A0A0D0NPG2 is a bifunctional polymyxin resistance protein,
362 ArnA, that is involved in several biological processes including coenzyme binding, UDP-
363 glucuronic acid dehydrogenase activity, lipid A biosynthetic process, and response to
364 antibiotic. This protein builds the UDP-L-4-formamido-arabinose attached to lipid A,

which is required for conferring resistance to polymyxin and cationic antimicrobial peptides (45). From the crowdsourcing classification, both peptide and polymyxin antibiotic classes were identified, where polymyxin was characterized by a slightly higher score (Figure 4A). A closer look at the evidence from the antibiotic resistance databases (CARD, ARDB, and MEGARes) reveals a consensus of the gene towards the polymyxin antibiotic class. As another example, the gene entry A0A127SI91 was inspected 62 times and found to belong to the beta-lactamase antibiotic class. This ARG was tagged as a novel, identified by Pehrsson et al. (17) in a study that analyzes linkages in antibiotic resistance exchange among different human environments. The evidence from the antibiotic resistance databases strongly suggests this gene as a bl1-EC beta lactamase gene. **Figure 4C** shows different crowd classifications (including all evaluation scenarios). Note that beta-lactam is the class with the highest score. However, as a consequence of disabling the trust validation filter, several unrelated categories were accepted, such as aminoglycoside, MLS, multidrug, nitrofurantoin, polyamine, polymyxin, and even the word “yes”. Fortunately, the scoring strategy was able to positively weight and assign the correct classification. One particularly interesting observation is the close proximity between valid annotations. For instance, in **Figure 4D**, the gene AAC76733.1 was correctly assigned to multidrug as its best classification and to the “multi-drug resistance” category as its second best classification. These small semantic differences are not detected by the trust validation filter. Therefore, under the validation interface, the administrators of ARG-miner have the ability to validate or reject the annotations if needed. **Figure 4B** shows that most

387 workers assigned the gene A0A0Q9QYU5 to the beta lactamase category. Note that the
388 suggested name “beta_lactam” is the highest scored among all choices.
389
390 **Figure 5** shows the crowdsourced score for the ARG name classification. As seen for
391 the antibiotic category annotation, there are cases where the annotations are
392 semantically close. For instance, the gene A0A127SI91 was tagged as bl1_ec, bl1-ec,
393 or blaec, all corresponding to the bla1-EC gene name (**Figure 5C**). Note that all these
394 labels were ranked higher than the other gene names (macb, baca, ba1) and all the
395 unrelated tags such as “mm-58”, “15”, “yes”, and “middle”. Also, all unrelated
396 annotations were ranked low by the scoring strategy.
397
398 Although identification of the antibiotic category for the gene A0A0Q9QYU5 was
399 straightforward, the detection of its gene name is challenging. Primarily, because the
400 metadata of this entry does not include the gene name and because the identity of its
401 best hit alignments is below 30%. This indicates that the gene has a potential homology
402 to known ARGs. Two ARG databases (CARD and MEGARes) show a significant best
403 hit e-value (<1e-22) over the mecB gene. For this example, 50% of the workers
404 annotated the gene as mecB whereas the other 50% annotated it as ash00_000180.
405 Also, workers yielded a higher confidence for the mecB gene (2.6 average confidence
406 score) compared to the ash00_000180 (2.3 average confidence score). As a result,
407 mecB had a slightly higher scoring. To avoid uncertainty, ARG-miner recommends that
408 users retain the original label if the evidence is not convincing. For the other examples

409 (**Figure 5A** and **5D**), the crowd classified the gene names according to the observed
410 evidence.
411
412
413 **Annotation analysis**
414 To assess the accuracy of the crowdsourcing annotation, genes that were inspected by
415 at least 10 workers were removed from the total pool of classified genes. A total of 35
416 genes were identified and manually curated by three domain experts according to the
417 *antibiotic class* and *gene name* annotation. It was found that experts achieved an
418 annotation pairwise correlation of 0.96 ± 0.02, indicative of an almost perfect
419 classification agreement. Thus, genes that were classified to the same label by at least
420 two experts were used as the gold standard data set (see **Supplementary Table ST1**).
421 This benchmark was then used to measure the performance of the crowdsourcing
422 workers where labels were selected based on the greatest annotation score.
423
424 The crowdsourcing classification of the antibiotic classes was essentially just as
425 accurate as the expert annotation (94% Positive Predicted Value - PPV). In other words,
426 33 out of 35 genes labeled via crowdsourcing matched the expert classification (see
427 **Supplementary Material 1**). The genes for which the workers failed to identify the
428 correct antibiotic class were a quinolone ARG annotated as multidrug (YP_001693238)
429 and a multidrug gene annotated as quinolone (NP_358469.1). On the other hand, the
430 classification of the ARG names proved to be a challenging task. Indeed, experts did
431 not fully agree about the correct name for five ARGs (see **Supplementary Table ST1**).

However, only one of those conflicting genes had a different classification from all three experts. This gene corresponded to a macrolide gene (AFU35065.1), which was tagged as: Isa, Isa-A, and Isa-E by the three experts. Thus, this gene entry was removed for the gene name analysis comparison and the final control data set contained 34 genes. When comparing the gene name annotation from the crowdsourcing workers, their prediction had a 97% PPV (see **Supplementary Material 2**). This indicates that only one gene was not correctly classified by the crowd (J2LT98). By examining the details of this gene in ARG-miner, all three ARG databases agreed that the gene belongs to the SHV group, with markedly high scores. However, CARD labeled it as the SHV variant 1 (SHV-1), ARDB labeled it as variant 2 (SHV-2), and MEGARes labeled it as the group SHV, without specification of a variant. An interesting aspect with respect to this particular ARG is that variants are defined by specific amino acid modifications (46), so these genes have a high identity and identifying the correct variant by using sequence alignment is a particularly difficult task, as shown in **Supplementary Figure S2**. This aspect has the potential to confuse workers when classifying genes that are very similar. Interestingly, by looking at the crowd results, workers were able to discard the SHV variant 2 (99.3% identity), but they were not able to differentiate between the SHV variant 1 and the SHV group (both have the same score). These results suggest that crowdsourcing workers are able to follow the correct track, even in the face of particularly complex tasks. Because of the risk of propagation errors, the updating process is not fully automated and administrator approval is required to approve/reject new classifications that will be updated in new database releases generated by ARG-miner. Overall the crowd exhibited performance comparable to that of the expert panel,

455 but in much less time. These results suggest that crowdsourcing annotation is a strong
456 alternative to the classification and validation of ARGs.

457

458 **Expertise and confidence**

459 ARG-miner asks users to rate their own expertise in the analysis of ARGs on a scale of
460 0 to 5. **Figure 6A** shows the distribution of the expertise score against the right or wrong
461 annotations for the antibiotic category classification (including all scenarios: AMT-Free,
462 AMT-VAL and LAB). Surprisingly, it is clear that having expert knowledge does not
463 really make a difference in the quality of the classification. Indeed, because of the open
464 nature of AMT, most of the workers are not experts and have little knowledge about
465 ARGs. From Figure 6A, it is also evident that the proportion of correct annotations was
466 higher compared to the incorrect classifications (the size of the dot indicates the number
467 of annotations). This result suggests that accurate detection of the correct antibiotic
468 resistance category does not necessarily require domain experts. On the other hand,
469 workers were also required to rate their confidence in the annotation. Results show that
470 self-rated confidence is a strong predictor for the quality of the annotation (**Figure 6B**).
471 The distribution of the confidence score shows that higher confidence correlates with
472 more accurate results. For instance, from the workers that rated their confidence with 5
473 stars, 95% obtained a successful annotation and only 5% missed the correct
474 classification. This strongly suggests the confidence score is a superior indicator of
475 correct annotation than the expertise score.

476

CONCLUSIONS

Here we develop, launch, and validate a new web platform, ARG-miner, as a powerful system for advancing robust and comprehensive curation of publicly-available ARG database drawing from the power of crowdsourcing. ARG-miner enables access to key relevant information pertaining to ARGs, including up-to-date ARG metadata, evidence of ARGs being carried by pathogens, and the possibility of ARGs being mobilized by MGEs. Further, it enables a simple, but powerful, tool for the curation of ARGs designed to provide accurate information represented in a noncomplex way that can be validated by users without the requirement of domain knowledge. Results demonstrated that crowdsourcing workers are as accurate as experts in curating ARGs. However, it was found that implementation of a trust-validation filter was essential to overcome the potential for confounds introduced by “spammers” and other untrustworthy crowd workers. Incorporation of the trust-validation filter, which forces users to input data that is related to the evidence provided in the platform, was found to markedly improve accuracies of the annotations. Various scenarios, including comparison to individuals with specific ARG-expert domain and general microbiology-knowledge domain, along with a novel scoring strategy, were implemented to measure the efficiency of the crowd. The crowd workers were not only able to identify the correct ARG classifications and other relevant metadata, but were much more efficient than ARG-domain experts alone. Thus, ARG-miner opens the possibility of a truly comprehensive, accurate, and perpetually up-to-date publicly-available ARG database.

FUNDING

Funding was provided in part for this effort by the United States Department of Agriculture (USDA) – National Institute of Food and Agriculture (NIFA) Award # 2015-68003-2305 and # 2017-68003-26498 Effective Mitigation Strategies for Antimicrobial Resistance program, the National Science Foundation (NSF) Partnership in International Research and Education (PIRE) award #1545756, the Virginia Tech Institute for Critical Technology and Applied Science Center for the Science and Engineering of the Exposome (SEE) and the Virginia Tech Sustainable Nanotechnology Interdisciplinary Graduate Education Program (IGEP).

REFERENCES

1. O'Neill, J. (2016) Tackling drug-resistant infections globally: final report and recommendations. . *The review on antimicrobial resistance*.
2. Pires, D., de Kraker, M.E.A., Tartari, E., Abbas, M. and Pittet, D. (2017) 'Fight Antibiotic Resistance—It's in Your Hands': Call From the World Health Organization for 5th May 2017. *Clinical Infectious Diseases*, **64**, 1780-1783.
3. Pruden, A., Larsson, D.J., Amézquita, A., Collignon, P., Brandt, K.K., Graham, D.W., Lazorchak, J.M., Suzuki, S., Silley, P. and Snape, J.R. (2013) Management options for reducing the release of antibiotics and antibiotic resistance genes to the environment. *Environmental health perspectives*, **121**, 878.
4. Martínez, J.L. (2008) Antibiotics and antibiotic resistance genes in natural environments. *Science*, **321**, 365-367.
5. Bengtsson-Palme, J., Kristiansson, E. and Larsson, D. (2017) Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*.
6. Berglund, F., Marathe, N.P., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.-F., Larsson, D.J. and Kristiansson, E. (2017) Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, **5**, 134.
7. Arango-Argoty, G.A., Garner, E., Pruden, A., Heath, L.S., Vikesland, P. and Zhang, L. (2017) DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *bioRxiv*, 149328.

8. Gibson, M.K., Forsberg, K.J. and Dantas, G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*, **9**, 207-216.
9. Lakin, S.M., Dean, C., Noyes, N.R., Dettenwanger, A., Ross, A.S., Doster, E., Rovira, P., Abdo, Z., Jones, K.L. and Ruiz, J. (2017) MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research*, **45**, D574-D580.
10. Yang, Y., Jiang, X., Chai, B., Ma, L., Li, B., Zhang, A., Cole, J.R., Tiedje, J.M. and Zhang, T. (2016) ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics*, **32**, 2346-2351.
11. McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., De Pascale, G. and Ejim, L. (2013) The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, **57**, 3348-3357.
12. Sello, J.K. (2012) Mining the antibiotic resistome. *Chemistry & biology*, **19**, 1220-1221.
13. Li, B., Chen, H., Su, J.-Q., Gillings, M.R., Chen, Q.-L., Zhang, T., An, X.-L. and Zhu, Y.-G. (2017) Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome*, **5**, 84.
14. Garner, E., Wallace, J.S., Argoty, G.A., Wilkinson, C., Fahrenfeld, N., Heath, L.S., Zhang, L., Arabi, M., Aga, D.S. and Pruden, A. (2016) Metagenomic profiling of historic Colorado Front Range flood impact on distribution of riverine antibiotic resistance genes. *Scientific reports*, **6**, 38432.
15. Bengtsson-Palme, J., Hammarén, R., Pal, C., Östman, M., Björlenius, B., Flach, C.-F., Fick, J., Kristiansson, E., Tysklind, M. and Larsson, D.J. (2016) Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Science of the Total Environment*, **572**, 697-712.
16. Pal, C., Bengtsson-Palme, J., Kristiansson, E. and Larsson, D.G. (2016) The structure and diversity of human, animal and environmental resistomes. *Microbiome*, **4**, 54.
17. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W. and Bertoli, M.T. (2016) Interconnected microbiomes and resistomes in low-income human habitats. *Nature*, **533**, 212-216.
18. Forsberg, K.J., Patel, S., Gibson, M.K., Lauber, C.L., Knight, R., Fierer, N. and Dantas, G. (2014) Bacterial phylogeny structures soil resistomes across habitats. *Nature*, **509**, 612-616.
19. Liu, B. and Pop, M. (2008) ARDB—antibiotic resistance genes database. *Nucleic acids research*, **37**, D443-D447.
20. Bengtsson-Palme, J., Larsson, D.J. and Kristiansson, E. (2017) Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy*, **72**, 2690-2703.
21. Gupta, S.K., Padmanabhan, B.R., Diene, S.M., Lopez-Rojas, R., Kempf, M., Landraud, L. and Rolain, J.M. (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother*, **58**, 212-220.
22. Jitwasinkul, T., Suriyaphol, P., Tangphatsornruang, S., Hansen, M.A., Hansen, L.H., Sorensen, S.J., Permpikul, C., Rongrungruang, Y. and Tribuddharat, C. (2016) Plasmid

metagenomics reveals multiple antibiotic resistance gene classes among the gut microbiomes of hospitalised patients. *J Glob Antimicrob Resist*, **6**, 57-66.

23. Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K.L. and Threlfall, E.J. (2005) Identification of plasmids by PCR-based replicon typing. *Journal of microbiological methods*, **63**, 219-228.
24. Song, X., Sun, J., Mikalsen, T., Roberts, A.P. and Sundsfjord, A. (2013) Characterisation of the plasmidome within Enterococcus faecalis isolated from marginal periodontitis patients in Norway. *PLoS One*, **8**, e62248.
25. Islamaj Dogan, R., Murray, G.C., Névéol, A. and Lu, Z. (2009) Understanding PubMed® user search behavior through log analysis. *Database*, **2009**.
26. Khare, R., Good, B.M., Leaman, R., Su, A.I. and Lu, Z. (2015) Crowdsourcing in biomedicine: challenges and opportunities. *Briefings in bioinformatics*, **17**, 23-32.
27. Lu, Z., Kim, W. and Wilbur, W.J. (2009) Evaluation of query expansion using MeSH in PubMed. *Information retrieval*, **12**, 69-80.
28. MacLean, D.L. and Heer, J. (2013) Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, **20**, 1120-1127.
29. Good, B.M., Nanis, M., Wu, C. and Su, A.I. (2014), *Pacific Symposium on Biocomputing Co-Chairs*. World Scientific, pp. 282-293.
30. Wei, C.-H., Kao, H.-Y. and Lu, Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, **41**, W518-W522.
31. Wei, C.-H., Harris, B.R., Li, D., Berardini, T.Z., Huala, E., Kao, H.-Y. and Lu, Z. (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, **2012**.
32. Arighi, C.N., Carterette, B., Cohen, K.B., Krallinger, M., Wilbur, W.J., Fey, P., Dodson, R., Cooper, L., Van Slyke, C.E. and Dahdul, W. (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database*, **2013**.
33. Lu, Z. and Hirschman, L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database*, **2012**.
34. Burger, J.D., Doughty, E., Khare, R., Wei, C.-H., Mishra, R., Aberdeen, J., Tresner-Kirsch, D., Wellner, B., Kann, M.G. and Lu, Z. (2014) Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. *Database*, **2014**.
35. Cocos, A., Qian, T., Callison-Burch, C. and Masino, A.J. (2017) Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of Biomedical Informatics*, **69**, 86-92.
36. Good, B.M. and Su, A.I. (2013) Crowdsourcing for bioinformatics. *Bioinformatics*, **29**, 1925-1933.
37. Leplae, R., Hebrant, A., Wodak, S.J. and Toussaint, A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic acids research*, **32**, D45-D49.
38. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*, **42**, D581-591.
39. Consortium, U. (2014) UniProt: a hub for protein information. *Nucleic acids research*, gku989.

40. Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M. and Larsen, M.V. (2012) Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*, **67**, 2640-2644.
41. Bush, K. and Jacoby, G.A. (2010) Updated functional classification of β-lactamases. *Antimicrobial agents and chemotherapy*, **54**, 969-976.
42. Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature methods*, **12**, 59-60.
43. Gertz, E.M., Yu, Y.-K., Agarwala, R., Schäffer, A.A. and Altschul, S.F. (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC biology*, **4**, 41.
44. Prill, R.J., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K. and Stolovitzky, G. (2011) Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Science signaling*, **4**, mr7.
45. Huja, S., Oren, Y., Trost, E., Brzuszkiewicz, E., Biran, D., Blom, J., Goesmann, A., Gottschalk, G., Hacker, J. and Ron, E.Z. (2015) Genomic avenue to avian colisepticemia. *MBio*, **6**, e01681-01614.
46. Paterson, D.L., Hujer, K.M., Hujer, A.M., Yeiser, B., Bonomo, M.D., Rice, L.B. and Bonomo, R.A. (2003) Extended-spectrum β-lactamases in Klebsiella pneumoniae bloodstream isolates from seven countries: dominance and widespread prevalence of SHV-and CTX-M-type β-lactamases. *Antimicrobial agents and chemotherapy*, **47**, 3554-3560.

FIGURES

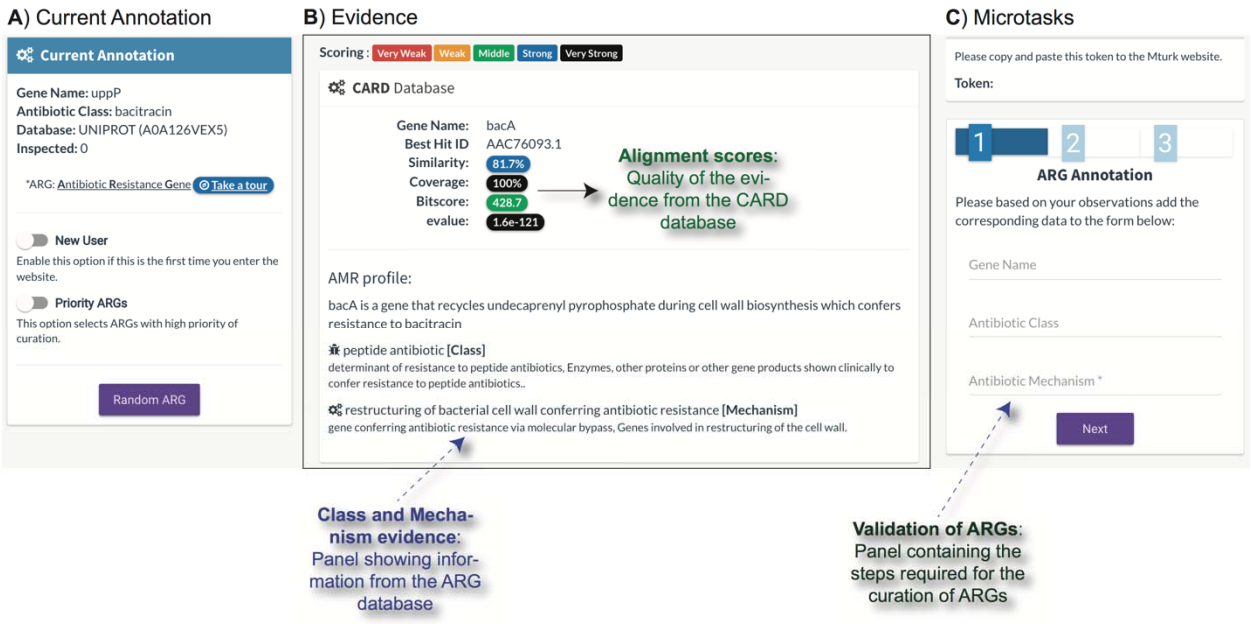

Figure 1: General overview of the ARG-miner platform. A) Current annotation. This panel contains the current information available for the ARG entry that requires validation. B) Evidence. This is the main panel and provides all of the metadata and information extracted from the different databases and resources. Note that in this panel there are colors that describe the relevance of each scoring metric. This is useful for users that are not familiar with alignment scores. C) Microtasks. This section contains the three microtasks needed for the ARG curation. It also contains real-time validation, which prompts error messages if the user inputs errors.

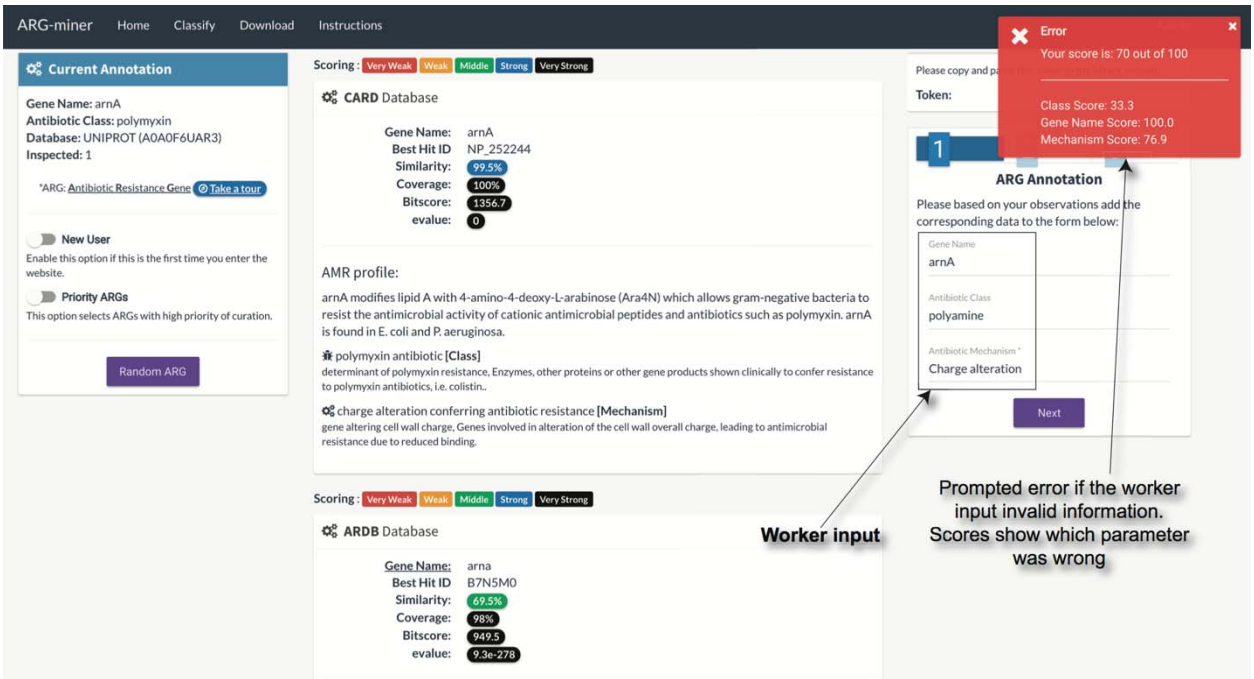

Figure 2: Trust validation strategy. The system rejects all inputs that are not consistent with the evidence. For instance, the antibiotic class *polyamine* does not correspond to the actual antibiotic class.

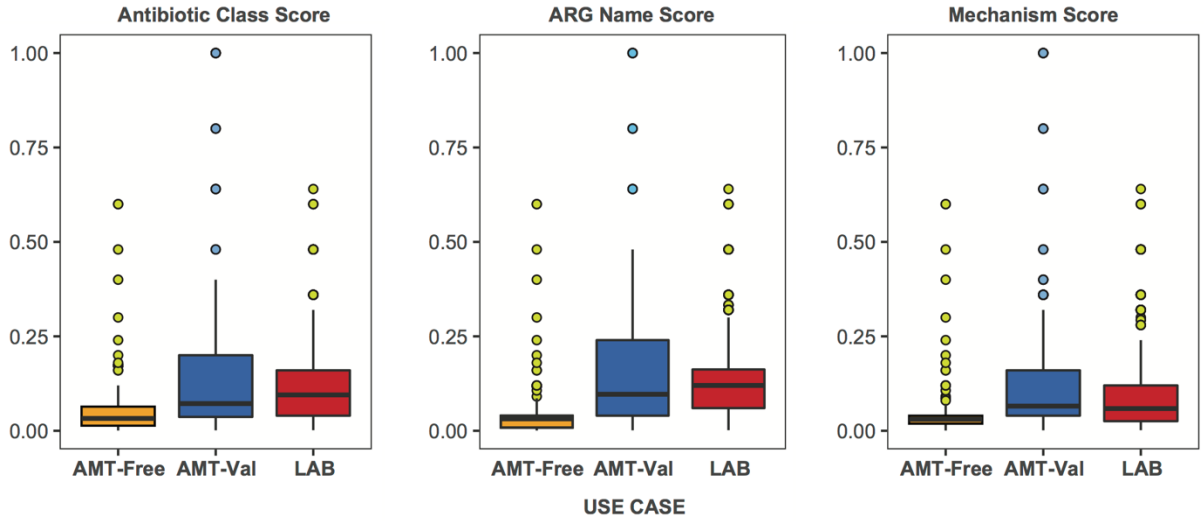

Figure 3: Annotation score of the three crowdsourced use cases (AMT-Free: Amazon MTurk workers without the true validation filter, AMT-Val: Amazon MTurk workers with the validation filter enabled and LAB: a group of workers with general microbiology domain knowledge and some antibiotic resistance knowledge. AMT-Val displayed the highest variance. However, this distribution was closer to that obtained by the workers with domain knowledge. Scores from the AMT-Free workers were the lowest among the three scenarios, indicating the ineffectiveness of the crowdsourcing annotation when the worker's input was not validated.

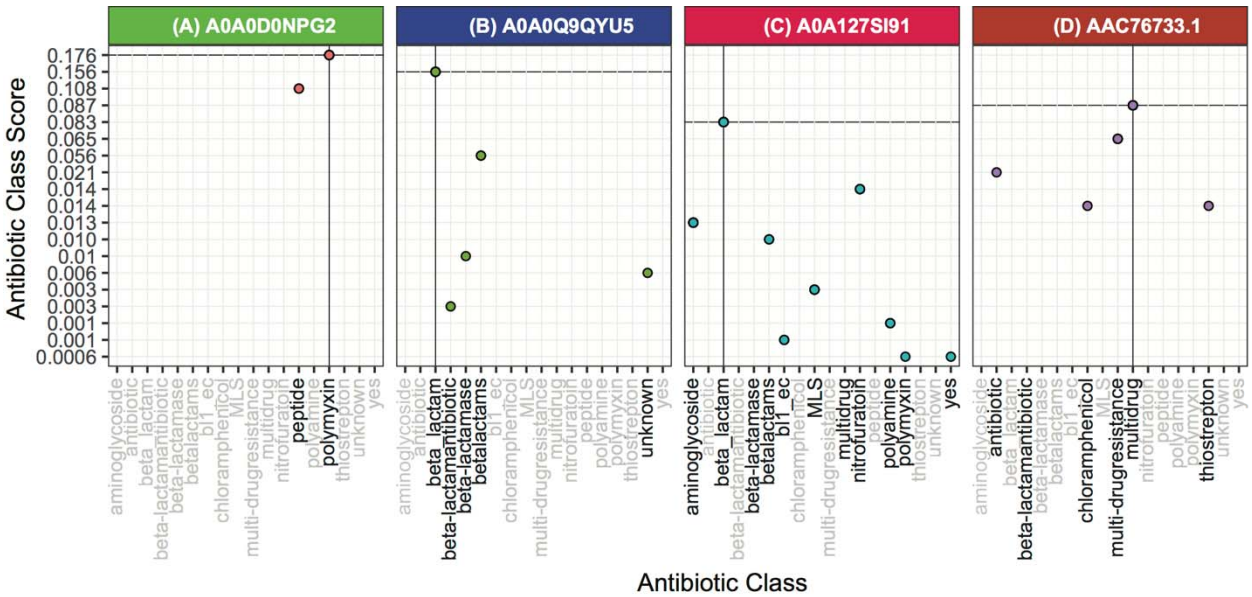

Figure 4: Distribution of the antibiotic class annotation by the crowdsourcing workers using the annotation score. X axis corresponds to the antibiotic resistance categories, where black labels indicate the categories reported by the workers and the top of each box corresponds to the ARG identifier.

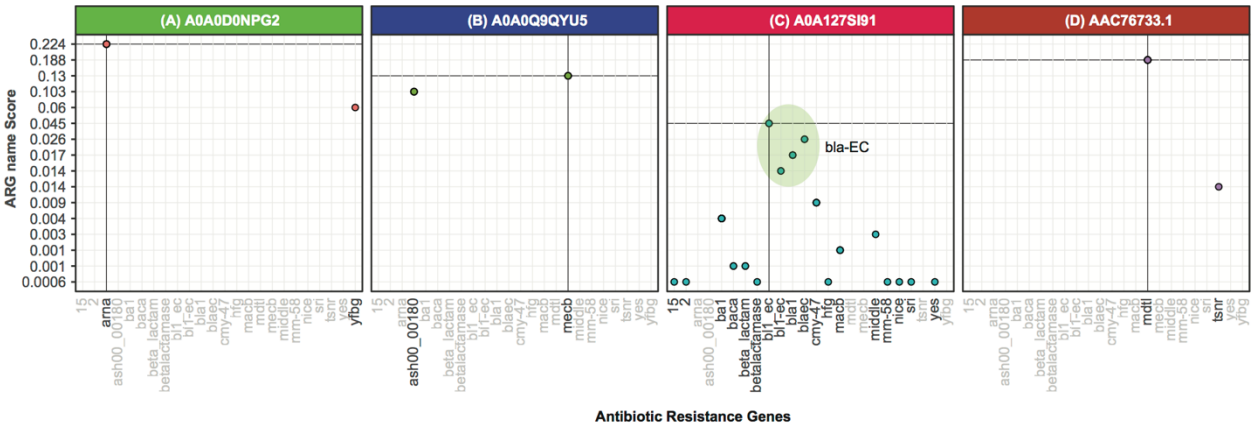

Figure 5: Distribution of the prediction of ARG names. ARG names are represented on the x axis and the y axis indicates the corresponding annotation score. The top of each box corresponds to the ARG identifier.

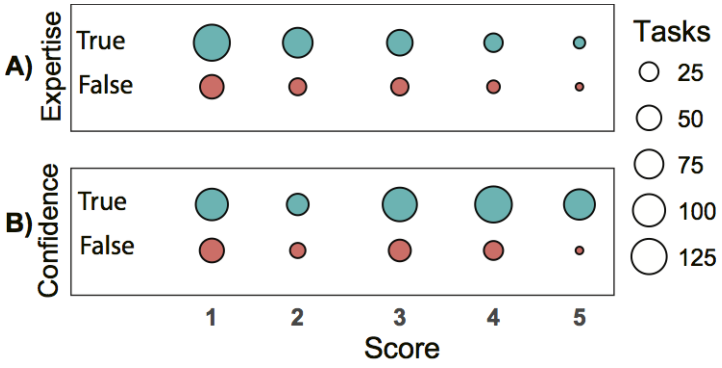

Figure 6: Expertise and confidence levels of the workers. The size of the points indicates the number of tasks; the x axis corresponds to the score level and the y label shows the expertise and confidence parameters. Color depicts correct and incorrect classifications.

Supplementary Figure S1: Administrator interface. This interface contains the elements to update a single ARG (accept/reject annotations from the crowd) as well as the main tools for releasing a new version of the database and to compute the ARGs that have conflicting annotations. This process is recommended to make ensure the annotations are valid.

Supplementary Figure S2: Case study of data provided by ARG-miner for the gene J2LT98. This gene represents a difficult ARG naming and annotation case. First, it has been identified in three major databases with an alignment coverage below 90% indicating genomic variability. Second, all three databases indicate a high bitscore and percentage of identity that can potentially confuse the user.

Supplementary Figure S3: ARG-miner evidence. A) Color scale describes the quality of the evidence from very weak (yellow) to very strong (black). B) Evidence of the ARG being carried by a MGE. This panel shows the alignment scores and the number of MGEs carrying the ARG. C) pathogen evidence, this panel illustrates the evidence of the ARG being carried by a pathogenic genome. More detailed information is also provided by ARG-miner, such as diseases and antimicrobial resistance phenotype.

Supplementary Material 1: Antibiotic classification of all gene entries from the expert validated dataset.

Supplementary Material 2: Antibiotic resistance names annotation of all gene entries from the expert validated dataset.

Supplementary Table ST1: Curated dataset from the three experts (A, B and C). This table shows the classification of the ARG name and Antibiotic category.