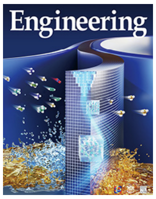
Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Antimicrobial Resistance—Article

ARGs-OAP v3.0: Antibiotic-Resistance Gene Database Curation and Analysis Pipeline Optimization

Xiaole Yin, Xiawan Zheng, Liguan Li, An-Ni Zhang, Xiao-Tao Jiang, Tong Zhang *

Environmental Microbiome Engineering and Biotechnology Laboratory, Center for Environmental Engineering Research, Department of Civil Engineering, the University of Hong Kong, Hong Kong 999077, China

ARTICLE INFO

Article history:
Received 4 May 2022
Revised 31 August 2022
Accepted 12 October 2022
Available online 27 December 2022

Keywords:
SARG database
ARGs-OAP
Antibiotic-resistance genes
Environmental metagenome
Quantification

ABSTRACT

Antibiotic resistance, which is encoded by antibiotic-resistance genes (ARGs), has proliferated to become a growing threat to public health around the world. With technical advances, especially in the popularization of metagenomic sequencing, scientists have gained the ability to decipher the profiles of ARGs in diverse samples with high accuracy at an accelerated speed. To analyze thousands of ARGs in a high-throughput way, standardized and integrated pipelines are needed. The new version (v3.0) of the widely used ARGs online analysis pipeline (ARGs-OAP) has made significant improvements to both the reference database—the structured ARGs (SARG) database—and the integrated analysis pipeline. SARG has been enhanced with sequence curation to improve annotation reliability, incorporate emerging resistance genotypes, and determine rigorous mechanism classification. The database has been further organized and visualized online in the format of a tree-like structure with a dictionary. It has also been divided into sub-databases for different application scenarios. In addition, the ARGs-OAP has been improved with adjusted quantification methods, simplified tool implementation, and multiple functions with user-defined reference databases. Moreover, the online platform now provides a diverse biostatistical analysis workflow with visualization packages for the efficient interpretation of ARG profiles. The ARGs-OAP v3.0 with an improved database and analysis pipeline will benefit academia, governmental management, and consultation regarding risk assessment of the environmental prevalence of ARGs.

1. Introduction

The discovery of antibiotics changed the world of clinical therapeutics and has saved hundreds of millions of lives threatened by infectious diseases. However, the misuse and abuse of antibiotics subsequently led to global concern about antimicrobial resistance (AMR) in the post-antibiotic era. Joint efforts from many countries have been made to advance antibiotic stewardship and enhance the surveillance of antibiotic-resistance genes (ARGs) [1,2]. Genomic sequencing with increasing applications in ARG surveillance is advantageous in high-throughput analysis and decoding the genomic context of ARGs. With the evolution of sequencing technology and the AMR crisis threatening human health, there is a growing need for reliable reference databases and bioinformatic tools for the fast and accurate annotation, classification, and quantification of ARGs using big data of DNA sequences [3].

The structured ARGs (SARG) database, first published in 2016, is one of the most popular ARG databases. It was constructed based on the comprehensive antibiotic-resistance database (CARD) [4] and the ARGs genes database (ARDB) [5] to generate a collection of 4 049 variants with a hierarchical structure of type-subtype-sequence and a clear classification of each sequence [6]. ARG types represent antibiotics against which the genes encoding proteins are resistant (like the antibiotic/drug classes used in some studies), while subtypes represent the genotypes of the genes (like the ARG families used in some studies). Further expansion was conducted on the SARG database to evolve it into v2.0 in 2018, which included more curated ARG reference sequences from the National Center for Biotechnology Information (NCBI) non-redundant (NR) database[†] through robust selection criteria such as sequence alignment and keyword matching. The ARGs genes online analysis pipeline (ARGs-OAP) can be used for ARG annotation, classification, and

* Corresponding author.
E-mail address: zhangt@hku.hk (T. Zhang).

[†] https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/.

https://doi.org/10.1016/j.eng.2022.10.011
2095-8099/

quantification with two-step analysis. Its first step incorporates the fast filtering of ARG sequences via Usearch [7], and the second step uses the basic local alignment search tool (BLAST) for accurate classification [8]. Global attention has been given to the ARGs-OAP, and the increasing number of users has motivated the continuous improvement of this tool, leading to an update to v2.0 with the deployment of SARGfam and the application of essential single-copy marker genes in cell number quantification [9].

Continuous improvement of this analytic tool, the SARG-based ARGs-OAP, is required to advance its performance and integration with other downstream analyses. Thus, this study describes recent updates to the ARGs-OAP v3.0, as shown in Fig. 1, which include ① a database extensively curated for reducing annotation bias with a revised hierarchical structure; ② upgraded annotation, classification, and quantification tools with increased coverage of ARGs for environmental samples, and a new method of calculating ARG abundance; and ③ improvement of the website for the integrative in-depth analysis of ARGs and statistical visualizations.

2. Methods

2.1. Database curation

Robust curation of all reference sequences was conducted using in-house scripts, followed by manual validation referring to the literature,[†] molecular experts, other relevant databases, and NCBI annotations. Through sequence alignment and keyword matching for individual sequences, an accurate classification was obtained if the alignment results matched the keyword searching. In detail, first, ARGs for specific antibiotic types were curated according to the most up-to-date knowledge, such as the terminology of tetracycline and macrolide-lincosamide-streptogramin (MLS) resistance genes [10,11]. Second, the name and classification of the ARG subtypes were supplemented by other databases, including the CARD (v3.2.4, downloaded on 27 July 2022) [12]. After manual filtration, 713 out of 4 641 sequences from the CARD were included into the SARG database, providing an updated collection of terms of ARG subtypes (Table 1). Furthermore, the rest of the reference sequences in SARG were reviewed individually according to their classification in the published papers. Those sequences without available classification of subtypes/types were removed to avoid potential misannotation and false positives in quantification. Lastly, SARG database alignment against the NCBI NR database (downloaded on 28 August 2022) was conducted to retrieve more reference sequences followed by rigorous selection criteria [9].

2.2. Simulated datasets for evaluation of the ARGs-OAP v3.0

To evaluate the performance of the ARGs-OAP v3.0, simulated datasets were generated from the Swiss-Prot database with customized scripts [13]. Sequences with the keywords “antibiotic resistance” in the Swiss-Prot database (downloaded on 20 April 2020) were treated as ARGs, while other sequences in the Swiss-Prot database were treated as non-ARGs. The whole collection was enumerated to produce k-mers of 50, 67, and 100 amino acid (aa) protein sequences to represent metagenomic datasets with read lengths of 150, 201, and 300 base pairs (bp). A gradient of cutoffs (i.e., E-value, identity, and hit length ratio) were assessed when applying the ARGs-OAP to the simulated datasets. The robustness of the pipeline with different cutoffs was evaluated based on the Matthews correlation coefficient (MCC), sensitivity, and precision using the calculation methods summarized in the Appendix A.

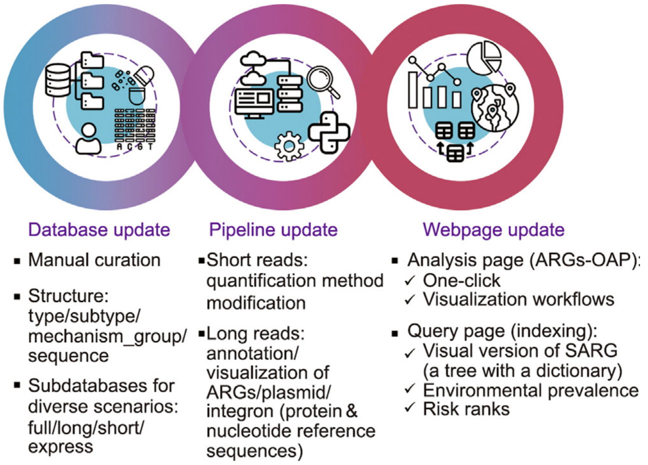

Database update
▪ Manual curation
▪ Structure: type/subtype/ mechanism_group/ sequence
▪ Subdatabases for diverse scenarios: full/long/short/ express
Pipeline update
▪Short reads: quantification method modification
▪Long reads: annotation/ visualization of ARGs/plasmid/ integron (protein & nucleotide reference sequences)
Webpage update
▪ Analysis page (ARGs-OAP):
✓ One-click
✓ Visualization workflows
▪ Query page (indexing):
✓ Visual version of SARG (a tree with a dictionary)
✓ Environmental prevalence
✓ Risk ranks

Fig. 1. The ARGs-OAP v3.0 has been updated to include a new database, a polished pipeline, and webpages with multiple functions.

2.3. Datasets for the evaluation of different versions of the ARGs-OAP

To evaluate the changes in the ARG abundance and diversity profiles introduced by the database update, quantification analyses were conducted on 36 samples from seven different environmental types, including four samples from river water, three samples from sediment, four samples of anaerobic digested sludge (ADS) from wastewater treatment plants (WWTPs), nine samples from WWTP activated sludge (AS), two samples from WWTP effluent, two samples from WWTP influent, and 12 samples from livestock feces or swine farm wastewater. For each environmental type, the abundances of ARG types quantified from different metagenomes were averaged to represent that environmental type.

2.4. Risk ranking of reference ARGs

The risk-ranking framework used for reference ARGs is based upon the work by Zhang et al. [14], which classifies reference sequences in the SARG database into four risk ranks (Ranks I, II, III, and IV) by a decision tree according to three criteria. First, all reference ARGs in the SARG v3.0 database were searched in global metagenome collections (n = 1 427, data obtained before 17 September 2022) and the Refseq genome collection (n = 256 788, downloaded on 26 August 2022) using the default cutoff values. Those reference sequences that could not be detected in any metagenome were ranked as “Unassessed.” Second, the prevalence of ARGs in human-associated environments (including human feces, cattle feces, swine feces, sewage, wastewater treatment facilities, the agricultural field, industrial wastewater treatment facilities, and mines) was compared against the ARGs abundance in non-impacted environments (including marine water, natural water, natural sediment, and natural soil) (Table S1 in Appendix A). Those ARGs that were not found to be enriched (a cutoff of < 100 times) in human-associated environments were categorized as “Rank IV.” Third, among the ARGs enriched in human-associated environments, nonmobile ones judged by searching the Mobile Genetic Elements (MGEs) database (Refseq plasmid database downloaded on 4 April 2022) were categorized as “Rank III.” Fourth, among the mobile and human-associated ARGs, those not carried by pathogens were categorized as “Rank II.” Lastly, those reference ARGs that met all three criteria, including being ① enriched in human-associated environments, ② mobile, and ③ carried by pathogens, were categorized as “Rank I,” which indicates the highest risk.

[†] https://smile.hku.hk/ARGs/Indexing.

Table 1
Counts of type/mechanism/subtype/sequence in the databases SARG v1.0/v2.2/v3.0-F and CARD v3.2.4.

Item	Database			
	SARG v1.0	SARG v2.2	SARG v3.0-F	CARD v3.2.4
Type	24	24	32	–
Mechanism	–	–	11	–
Subtype	1 208	1 244	2 842	–
Sequence	4 499	12 085	13 672	4 641

SARG v3.0-F: SARG v3.0 full version.

2.5. Information technology

ARGs-OAP v1.0 [6] and v2.0 [9] were deployed based on the Galaxy project [15]. In the updated version, the Galaxy project was developed *in situ* with a customized Python Flask framework, Vue.js framework, and Quasar framework, with supportive datasets generated by R-Studio. Database indexing with mass data was supported by MySQL storage and visualized by the markmap package.

3. Results and discussion

3.1. SARG database update

As in the previous two versions of SARG, the reference sequences of ARGs in SARG v3.0 are organized in a hierarchical structure (type-subtype-sequence), which is beneficial for the top–down interpretation of the resistome in an environmental sample, especially when applying the ARGs-OAP in the quantification of both the phenotypes (ARG types) and genotypes (ARG subtypes) of ARGs. In SARG v3.0, the resistance mechanisms are identified to form a new structure with four layers (type-mechanism-subtype-sequence). Six mechanism groups are included: antibiotic target alteration, antibiotic target protection, antibiotic target replacement, efflux pump, enzymatic inactivation, and reduced permeability [16–26]. For some groups, the mechanisms are further classified into subgroups. For example, the efflux pump is further classified into five subgroups: adenosine triphosphate (ATP)-binding cassette (ABC) transporter; major facilitator superfamily (MFS) transporter; multidrug and toxic compound extrusion (MATE) transporter; resistance-nodulation-cell division (RND) transporter; and small multidrug resistance (SMR) transporter [24,25].

Moreover, in SARG v3.0, special "two-component" and "three-component" tags are given to those ARG subtypes that have two-component systems or three-component systems encoding antibiotic resistance. For example, a pair of genes conferring an efflux pump (*tetA*(46) and *tetB*(46)) is required for tetracycline resistance [27]. AcrEF-TolC is another example of a three-component system from the subfamily of RND transporters, the function of which requires a membrane fusion protein (AcrE), inner membrane transporter (AcrF), and outer membrane factor (TolC) [28].

Moreover, the curation of the name list of ARG types and subtypes has resulted in 1 717 new ARG subtypes being added to the SARG database, including 157 aminoglycoside, 230 beta-lactam, 35 chloramphenicol, 96 MLS, 99 multidrug, 106 quinolone, 73 vancomycin, and other resistance subtypes (Table 1; Table S2 in Appendix A). Eleven synonyms have been identified. In particular, the SARG database now includes 127 more ARG subtypes in addition to those of CARD, including *mdtL*, *SHV-112*, *SHV-39*, and *tetX1*. Manual curation of these subtype names has been conducted for the SARG v3.0 database.

To summarize, 1 425 sequences were removed from SARG v2.2 due to inconsistent classification into specific types or subtypes, and 3 012 sequences were added, resulting in an updated database SARG v3.0 full version (SARG v3.0-F) with 32 types, 2 842 subtypes, and 13 672 sequences (Table 1 and Table S1).

For better annotation and classification of ARGs using DNA sequences of different lengths, SARG v3.0-L (a sub-database for long-read annotation, n = 13 439) was constructed from SARG v3.0-F by removing 233 sequences that evolved from mutations or functioned when overexpressed, which were not suitable for annotation based on a similarity search. Furthermore, SARG v3.0-S (for short read quantification, n = 12 746) was created as a sub-database of SARG v3.0-L that excludes the 693 sequences tagged with transcriptional regulators (including activators and repressors) that cannot be correctly annotated using short reads. The sub-database SARG 3.0-E (for express analysis, currently n = 10 538) only includes environmentally prevalent SARG sequences that have been detected at least once in a comprehensive survey across diverse environments. Both SARG v3.0-S and SARG v3.0-E are provided as reference databases in the ARGs-OAP for the full or fast analysis of environmental metagenomic short-read datasets using similarity search algorithms.

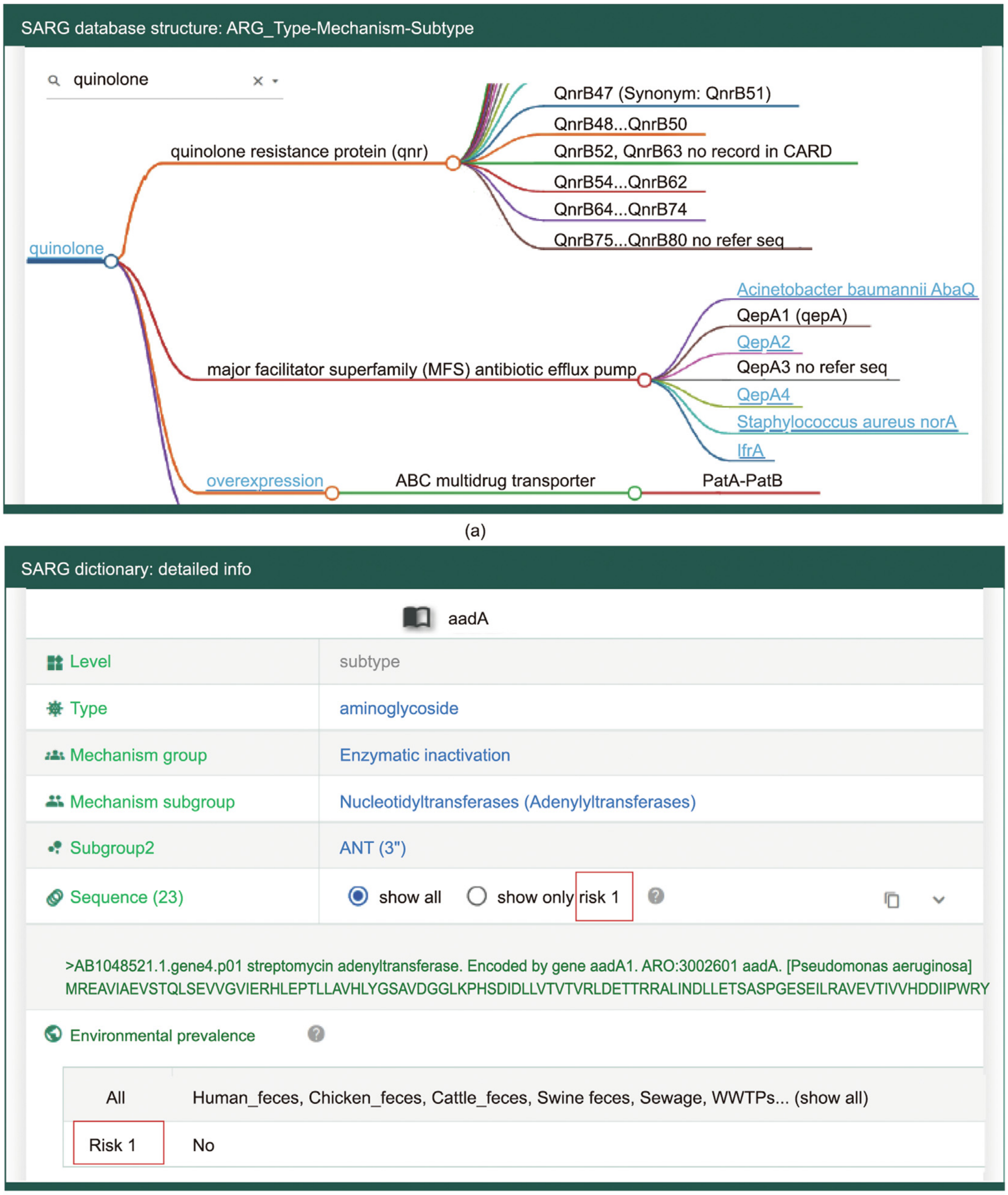
3.2. Database indexing platform

The structure of the SARG v3.0-F database is clearly displayed on the ARGs-OAP website for users to retrieve information about each gene and reference sequence and refer to any ARG of interest among the 13 672 reference sequences.

Two formats have been adopted for indexing the SARG database (Fig. 2). One is the hierarchical tree view, in which each tree is rooted in one ARG type and then grown into different resistance mechanisms, protein families, and ARG subtypes as branches. The tree-structure indexing is a user-friendly visualization of the hierarchical structure of the SARG database. The other display format is designed to archive each ontology in the SARG database with comprehensive descriptions, similar to other databases such as UniProt [29] and Pfam [30].

The most notable section of the ARGs-OAP indexing is the environmental prevalence information on each ARG subtype, which is summarized from the data mining results of over 1 000 metagenomic datasets from various environmental samples (Table S1). In addition, the 12 746 reference sequences in SARG v3.0-S are classified into risk ranks I, II, III, and IV.[†] Based on our recently published risk-ranking scheme [14], the risk Rank I ARGs claim the most attention due to their high mobility across phylogenetic boundaries, their wide dissemination under anthropogenic activities, and the pathogenicity of their hosts. Thus, risk Rank I ARGs are highlighted on the webpage in both the "environmental prevalence" and "sequence" sections in the dictionary. In general, ARG prevalence data and the risk-ranking scheme provide a valuable reference for both academia and government in understanding the dissemination of ARGs and developing control strategies.

[†] https://smile.hku.hk/ARGs/Indexing/riskranking.


SARG database structure: ARG_Type-Mechanism-Subtype
quinolone
quinolone
quinolone resistance protein (qnr)
QnrB47 (Synonym: QnrB51)
QnrB48...QnrB50
QnrB52, QnrB63 no record in CARD
QnrB54...QnrB62
QnrB64...QnrB74
QnrB75...QnrB80 no refer seq
major facilitator superfamily (MFS) antibiotic efflux pump
Acinetobacter baumannii AbaQ
QepA1 (qepA)
QepA2
QepA3 no refer seq
QepA4
Staphylococcus aureus norA
lfrA
overexpression
ABC multidrug transporter
PatA-PatB

(a)

SARG dictionary: detailed info

aadA

Level	subtype
Type	aminoglycoside
Mechanism group	Enzymatic inactivation
Mechanism subgroup	Nucleotidyltransferases (Adenylyltransferases)
Subgroup2	ANT (3")
Sequence (23)	● show all ○ show only risk 1

>AB1048521.1.gene4.p01 streptomycin adenyltransferase. Encoded by gene aadA1. ARO:3002601 aadA. [Pseudomonas aeruginosa]
MREAVIAEVSTQLSEVVGVIERHLEPTLLAVHLYGSAVDGGLKPHSDIDLLVTVTVRLDETTRRALINDLLETSASPGESEILRAVEVTIVVHDDIIPWRY

Environmental prevalence

All	Human_feces, Chicken_feces, Cattle_feces, Swine feces, Sewage, WWTPs... (show all)
Risk 1	No

(b)

Fig. 2. A visual version of the SARG v3.0-F (n = 13 672) in two formats. (a) Trees of ARG types with a supported searching function; (b) archived information of each ontology in the SARG database, including the relevant type, subtype, mechanism groups, subgroups, reference sequences, and environmental prevalence information. aadA: aminoglycoside 3' adenyltransferase.

3.3. Updates to the ARG quantification tool and visualization

The integrated tool for the annotation and quantification of ARGs by leveraging the SARG database is termed the ARGs-OAP[†]. It has been improved in version 3.0 to accurately quantify ARG abundance from metagenomic datasets, with the following modified equation:

$$\text{Abundance} = \sum_{i=1}^{n}\left(k \times \frac{Ni_{\text{ARG-like sequence}} \times Li_{\text{read}}/Li_{\text{ARGs reference sequence}}}{N_{\text{cell number}}}\right)$$

where $Ni_{\text{ARG-like sequence}}$ is the number of ARG-like reads annotated to one specific ARG reference sequence; Li_{read} is the read length; $Li_{\text{ARGs reference sequence}}$ is the nucleotide sequence length of the corresponding ARG reference sequence; and n is the number of mapped ARG reference sequences belonging to that ARG type or subtype. $N_{\text{cell number}}$ is the cell number estimated either by mapping against an essential single-copy marker gene database or by correction from copy numbers of 16S ribosomal RNA (rRNA) sequences [9]. The parameter k equals 0.5 if the specific ARG reference sequence is a two-component system, 0.33 if the specific ARG reference sequence is a three-component system, and 1.0 for all ARGs other than the above two categories.

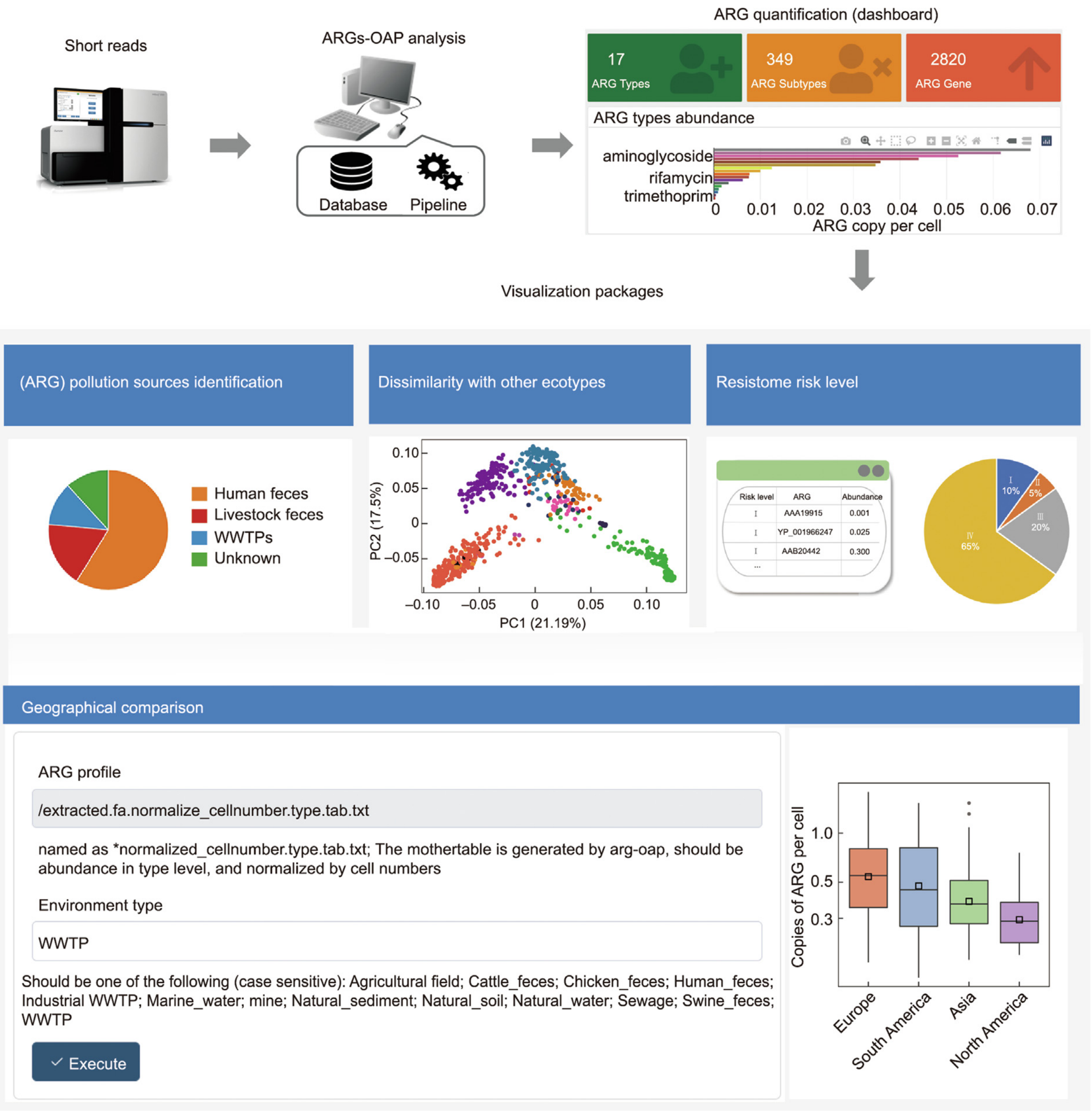
In SARG v3.0, different subtypes of ARGs are tagged with different k values, which are used to adjust the quantification. A total of

[†] https://smile.hku.hk/SARGs.

3 552 sequences are tagged as "three-component" systems with *k* values of 0.33, because the synchronous occurrence of three genes in this category as a group is required for genuine resistance and, without the adjustment parameter of 0.33, the old quantification method of counting every component as 1.0 will result in a threefold overestimation. Similarly, 65 sequences are tagged as "two-component" systems with *k* values of 0.5, because the occurrence of two genes in this category as a group leads to resistance; thus, a single occurrence has been adjusted by a parameter of 0.5. This modification will help reduce the bias in the quantification of a few ARG types, including multidrug ARGs, MLS resistance genes, and so forth. Those subtypes of ARGs whose *k* values are equal to 1.0 are not affected in the quantification process in the updated formula.

Moreover, the online analysis platform has been updated to a new file management system that facilitates much more user-friendly online analysis in many aspects (Fig. 3). First, the old versions of the ARGs-OAP required local sample pretreatment. With the update, users can choose to upload raw reads (via the webpage or FTP) and then go through the quantification steps of the ARGs-OAP with just one click. Second, the updated online pipelines provide multiple downstream analyses after the above quantification using the ARGs-OAP. Visualization packages have been integrated into the ARGs-OAP and the downstream analysis tools to display the results for better interpretation. In detail, short reads of environmental samples can be used as the query input for the ARGs-OAP analysis to classify ARGs and quantify ARG prevalence, generating abundance tables of ARG types, subtypes, and variants. A dashboard is also available with summarized counts of detected ARGs and a bar chart of ARG abundance (in the unit of copies of ARGs per cell).

The workflow of the downstream analyses, taking the geographical comparison analysis as an example, is demonstrated and illustrated in Figs. 3 and 4. The whole package of downstream analyses includes the following:



Risk level	ARG	Abundance
I	AAA19915	0.001
I	YP_001966247	0.025
I	AAB20442	0.300
...		

ARG profile

/extracted.fa.normalize_cellnumber.type.tab.txt

named as *normalized_cellnumber.type.tab.txt; The mothertable is generated by arg-oap, should be abundance in type level, and normalized by cell numbers

Environment type

WWTP

Should be one of the following (case sensitive): Agricultural field; Cattle_feces; Chicken_feces; Human_feces; Industrial WWTP; Marine_water; mine; Natural_sediment; Natural_soil; Natural_water; Sewage; Swine_feces; WWTP

Fig. 3. Workflows of the ARGs-OAP v3.0 platform for short reads. The query datasets can be analyzed to quantify ARGs in an efficient and accurate way, followed by the use of integrated tools for visualization and interpretation. One example is the "geographical comparison" package, whose interface requires selecting the environmental type of the query sample and then uploading an input file, which is the mother table of the ARG abundance after the analysis of the ARGs-OAP. The resulting profiles include a boxplot and a map that are generated based on the query sample and the archived database, which contains 1 427 samples from 13 types of habitats.

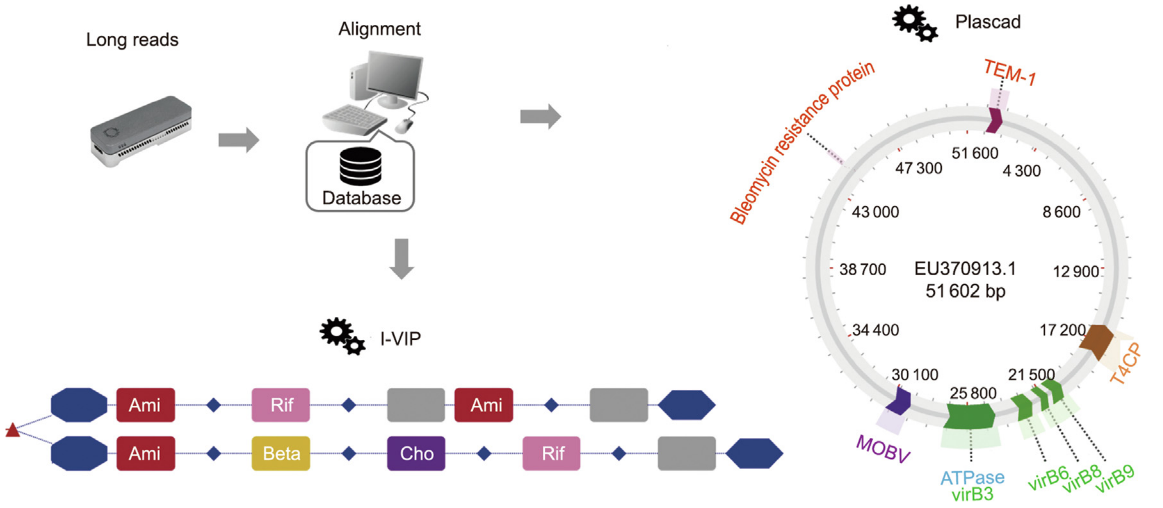


Fig. 4. Tools available on the platform of the ARGs-OAP v3.0 for long reads, including integron identification and plasmid classification. I-VIP: integron visualization and identification pipeline.

(1) A geographical comparison performed by importing an abundance table of ARG types of samples from different locations for benchmarking ARG pollution levels with a global collection of data from the same type of habitat;

(2) ARG pollution source identification based on microbial source tracking (MST) to identify the proportions of different sources (including sewage, human feces, livestock feces, WWTPs, the agricultural field, industrial WWTPs, mine, and natural sources) contributing to ARGs in a sample of interests;

(3) An ordination analysis of a sample performed by referring to the ARG profiles in a collection of various ecosystems to demonstrate similarity and dissimilarity;

(4) The profile of four ranks regarding ARG risks in a sample of interests.

In addition to a metagenomic analysis on short reads, there is a growing application of long-read-based ARG annotation, which is generated either from third-generation sequencing [31,32] or the *de novo* assembly of short reads into contigs [33]. By referring to the SARG database, long reads can easily be aligned to either protein or nucleotide reference sequences,[†] depending on the sequencing accuracy and the research scenarios, to annotate ARGs. The genetic context can be further deciphered by MGE analysis by the integron visualization and identification pipeline (I-VIP) for integrons [34] or pipeline for plasmid classification (Plascad) for plasmids [35]. Identification of the colocalization of ARGs and MGEs provides critical information for further exploration of potential horizontal gene transfer across bacterial communities.

3.4. Evaluation of the performance of the ARGs-OAP v3.0

We evaluated the performance of the updated pipeline based on the MCC, sensitivity, and precision by annotating ARGs in the simulated metagenomic datasets with the read lengths of 150, 201, and 300 bp (Fig. 5; Fig. S1 in Appendix A). The evaluation results revealed the excellent performance of the ARGs-OAP v3.0, which showed high precision and sensitivity for ARG identification in three sets of environmental metagenomes when the recommended cutoffs were applied (i.e., *E*-value: 1e−7; identity: 80%; hit length ratio: 75%). False positives are always a concern when annotating genes in complex samples. The ARGs-OAP has been demonstrated to be of high quality, with a false-positive rate of less than 2%.

To further evaluate the performance, all three versions (v1.0, v2.2, and v3.0) of the ARGs-OAP were applied to analyze 36 metagenome datasets of seven typical environmental sources representing diverse levels of anthropogenic impacts. The results showed clear improvement using the updated database—that is, increased ARG abundance (copies of ARGs per cell) and richness (number of detected ARG subtypes) were found in all the studied environmental samples with varying ARG levels.

As shown in Fig. 5(d), the application of SARG v3.0 changed ARG detection in environments with different ratios, from −12.6% in sewage to 28.8% in river water, compared with SARG v2.2. Samples from natural environments (river water and sediment) were found to have diverse detection improvement (4.7%–28.8%), while some samples from wastewater treatment facilities (WWTP ADS, AS, and effluent) were detected to have total abundances improved in similar ratios (12.4%–15.9%), and other samples showed a decrease of total abundance. Based on the Mann–Whitney test ($P < 0.05$), significant differences were found in the total abundance of ARGs in different environmental types, resulting in the stratification of four levels, as follows (from high to low abundance): livestock feces > sewage > wastewater treatment facilities > natural samples. The stratification of the abundance levels of ARGs remained the same, regardless of which version of the database was applied.

The increased detection of ARGs in nature samples and WWTP were furtherly confirmed by the extended detection richness (the number of detected ARG subtypes) in these samples, while the decrease of total abundances of ARGs in sewage and livestock feces were attributed to the removal of ambiguous reference sequences in the database, mainly multidrug resistance genes. Overall, SARG v2.2 detected 736 subtypes (1 244 were available in the database), whereas SARG v3.0 detected 1 019 subtypes. That is, the updated database retrieved 283 more ARG subtypes that were present in at least one sample from the seven environmental types and were not detected using the previous versions (Table S3 in Appendix A). The abundances of these newly retrieved subtypes, which included aminoglycoside, beta-lactam, MLS, multidrug, polymyxin, and other resistance types, ranged from 6.17×10^{-6} to 0.92 copies of ARGs per cell in the tested samples. The newly detected subtypes with abundances of more than 0.4 copies of ARGs per cell included *lnuC* and *optrA* from the resistant type MLS, and *lnuH* and *fexB* from

[†] https://smile.hku.hk/ARGs/Indexing/download.

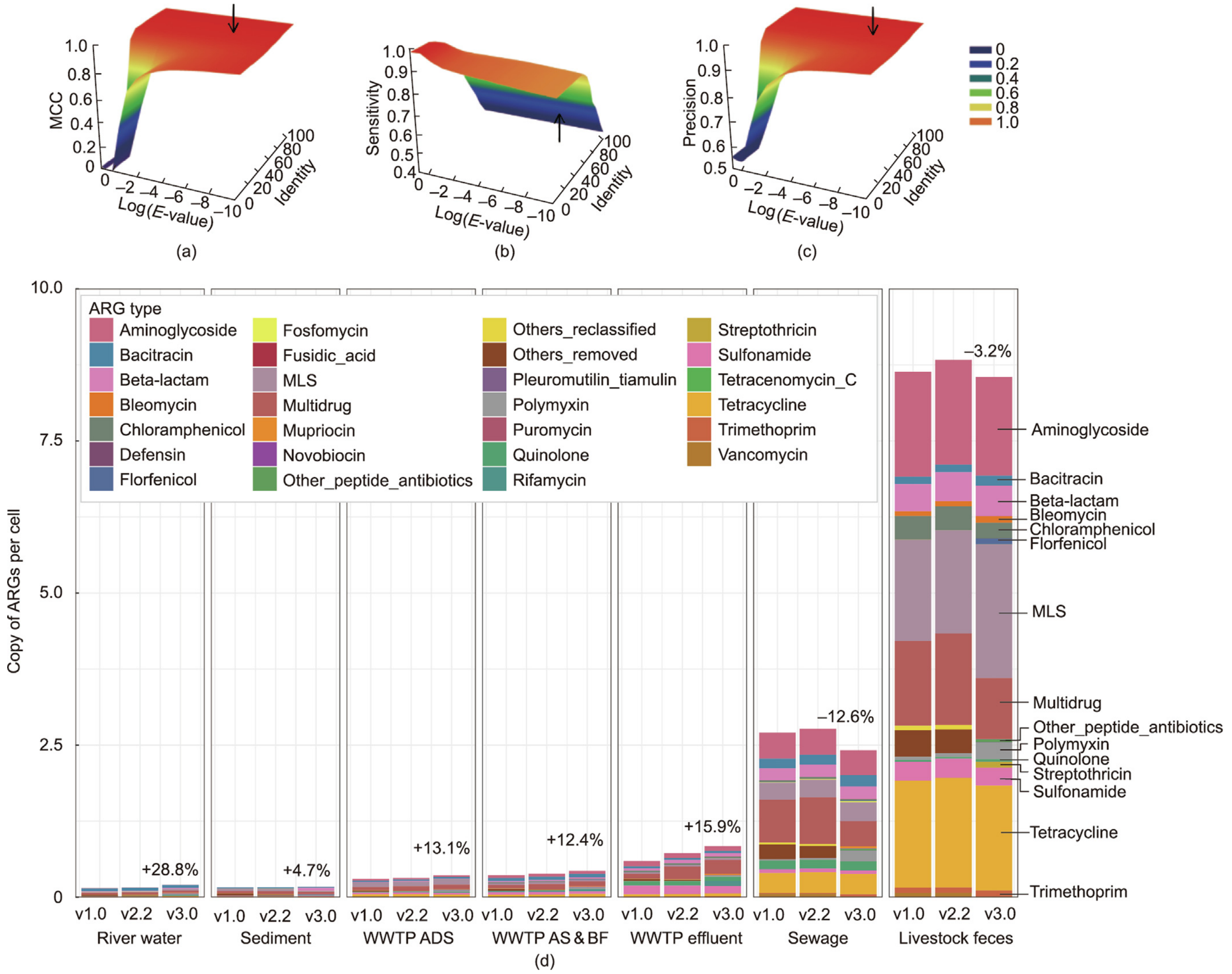


Fig. 5. Evaluation of the updated databases and the pipeline for ARG annotation and quantification. (a) MCC, (b) sensitivity, and (c) precision were assessed when applying the ARGs-OAP v3.0 with a gradient of cutoff values on simulated metagenomic datasets with 150 bp read length. The color gradients represent the values of (a) MCC, (b) sensitivity, and (c) precision within the range of 0–1.0. (d) A further evaluation using metagenomes from different environments was conducted by applying the three versions of ARGs-OAP. For each environment, the bars indicate the reference databases used: (left) SARG v1.0; (middle) SARG v2.2; (right) SARG v3.0. The ARGs were quantified in units of copies of ARGs per cell. The percentage labels in the figures are the increased numbers of ARGs detected using SARG v3.0 compared with v2.2. BF: biofilm.

the resistant type florfenicol, indicating that these newly detected subtypes would be covered after adding the novel reference genes to SARG v3.0. Therefore, the updated database will improve detection coverage in the surveillance of ARGs in diverse environmental samples. And at the meantime, the new database will facilitate accurate prediction of ARGs by reducing false positives.

4. Conclusions

The ARGs-OAP was first released in 2016 and then updated in 2018. As described in this study, continued development has been conducted on this analytical tool to achieve better performance in studies on the environmental dimension of antibiotic resistance. In the ARGs-OAP v3.0, improvements have been introduced in both the database updates and the integration of different analytical tools. First, the reference database SARG has been updated to v3.0 to remove/add sequences and adjust the names of types and subtypes according to the updated knowledge, add information on mechanism families and subfamilies, and expand the coverage through curation on the basis of other databases, such as CARD. SARG v3.0-S and SARG v3.0-E, which exclude genes related to mutation, repressors, and regulators, have been embedded in the ARGs-OAP v3.0 as reference databases, while SARG v3.0-F is publicly accessible for visualization through a tree structure and dictionary form. Second, user-friendly workflows have been developed with integrated tools starting from the ARGs-OAP with follow-up analysis, including a risk-ranking scheme, geographical comparison, MST, and similarity/dissimilarity analysis with other ecosystems. Visualization has been implemented in the analysis pipelines, which will facilitate data interpretation and effective communication.

Acknowledgments

The work was substantially supported by a Theme-based Research Scheme grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (T21-705/20-N). Dr. Xiaole Yin would like to thank the University of Hong Kong for a postdoctoral fellowship. Ms. Xiawan Zheng thanks the University of Hong Kong for a postgraduate studentship. The authors also thank the lab technician, Ms. Vicky Fung, for assisting in the whole experiment process. The computations were performed using research computing facilities offered by Information Technology Services, the University of Hong Kong.

Compliance with ethics guidelines

Xiaole Yin, Xiawan Zheng, Liguan Li, An-Ni Zhang, Xiao-Tao Jiang, and Tong Zhang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.eng.2022.10.011.

References

[1] Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell 2021;184(13):3376–93.

[2] Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 2019;10(1):1124.

[3] Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019;20(6):356–70.

[4] McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 2013;57(7):3348–57.

[5] Liu B, Pop M. ARDB-antibiotic resistance genes database. Nucleic Acids Res 2009;37(Suppl 1):D443–7.

[6] Yang Y, Jiang X, Chai B, Ma L, Li B, Zhang A, et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics 2016;32(15):2346–51.

[7] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–1.

[8] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.

[9] Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, et al. 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 2018;34(13):2263–70.

[10] Roberts MC, Schwarz S. Tetracycline and chloramphenicol resistance mechanisms. In: Mayers DL, Sobel JD, Ouellette M, Kaye KS, Marchaim D, editors. Antimicrobial drug resistance: mechanisms of drug resistance. New York City: Springer; 2017. p. 231–43.

[11] Chopra I, Roberts M. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. Microbiol Mol Biol Rev 2001;65(2):232–60.

[12] Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017;45(D1):D566–73.

[13] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28(1):45–8.

[14] Zhang AN, Gaston JM, Dai CL, Zhao S, Poyet M, Groussin M, et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. Nat Commun 2021;12(1):4765.

[15] Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 2018;46(W1):W537–44.

[16] Roberts MC. Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. FEMS Microbiol Lett 2008;282(2):147–59.

[17] De Oliveira DMP, Forde BM, Kidd TJ, Harris PNA, Schembri MA, Beatson SA, et al. Antimicrobial resistance in ESKAPE pathogens. Clin Microbiol Rev 2020;33(3):e00181–219.

[18] Munita JM, Arias CA. Mechanisms of antibiotic resistance. Microbiol Spectr 2016;4(2):VMBF-0016-2015.

[19] Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol 2015;13(1):42–51.

[20] Bush K. The ABCD's of β-lactamase nomenclature. J Infect Chemother 2013;19(4):549–59.

[21] Rodríguez-Martínez JM, Velasco C, Álvaro P, Cano ME, Luis MM. Plasmid-mediated quinolone resistance: an update. J Infect Chemother 2011;17(2):149–82.

[22] Wright GD. Q&A: antibiotic resistance: where does it come from and what can we do about it? BMC Biol 2010;8(1):123.

[23] Ramirez MS, Tolmasky ME. Aminoglycoside modifying enzymes. Drug Resist Updat 2010;13(6):151–71.

[24] Piddock LJV. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. Clin Microbiol Rev 2006;19(2):382–402.

[25] Poole K. Efflux-mediated antimicrobial resistance. J Antimicrob Chemother 2005;56(1):20–51.

[26] Connell SR, Tracz DM, Nierhaus KH, Taylor DE. Ribosomal protection proteins and their mechanism of tetracycline resistance. Antimicrob Agents Chemother 2003;47(12):3675–81.

[27] Warburton PJ, Ciric L, Lerner A, Seville LA, Roberts AP, Mullany P, et al. TetAB46, a predicted heterodimeric ABC transporter conferring tetracycline resistance in *Streptococcus australis* isolated from the oral cavity. J Antimicrob Chemother 2013;68(1):17–22.

[28] Nishino K, Yamada J, Hirakawa H, Hirata T, Yamaguchi A. Roles of TolC-dependent multidrug transporters of *Escherichia coli* in resistance to β-lactams. Antimicrob Agents Chemother 2003;47(9):3030–3.

[29] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–9.

[30] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49(D1):D412–9.

[31] Yang Y, Zhang AN, Che Y, Liu L, Deng Y, Zhang T. Underrepresented high diversity of class 1 integrons in the environment uncovered by PacBio sequencing using a new primer. Sci Total Environ 2021;787:147611.

[32] Che Y, Xia Y, Liu L, Li AD, Yang Y, Zhang T. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. Microbiome 2019;7(1):44.

[33] Ma L, Xia Y, Li B, Yang Y, Li LG, Tiedje JM, et al. Metagenomic assembly reveals hosts of antibiotic resistance genes and the shared resistome in pig, chicken and human feces. Environ Sci Technol 2016;50(1):420–7.

[34] Zhang AN, Li LG, Ma L, Gillings MR, Tiedje JM, Zhang T. Conserved phylogenetic distribution and limited antibiotic resistance of class 1 integrons revealed by assessing the bacterial genome and plasmid collection. Microbiome 2018;6(1):130.

[35] Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, et al. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. Proc Natl Acad Sci USA 2021;118(6):e2008731118.