Ontology-Aware Deep Learning Enables Novel Antibiotic Resistance Gene Discovery Towards Comprehensive Profiling of ARGs

Yuguo Zha[1], Cheng Chen[2], Qihong Jiao[2], Xiaomei Zeng[1, *], Xuefeng Cui[2, *], Kang Ning[1, *]

[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

[2]School of Computer Science and Technology, Shandong University, Qingdao 266237, Shandong, China

[*]Correspondence should be addressed to K.N (Email: ningkang@hust.edu.cn), X.C (Email: xfcui@email.sdu.edu.cn) and X.Z (Email: xmzeng@hust.edu.cn)

Abstract

Antibiotic resistance genes (ARGs) have emerged in pathogens and spread faster than expected, arousing a worldwide concern. Current methods are suitable mainly for the discovery of close homologous ARGs and have limited utility for discovery of novel ARGs, thus rendering the profiling of ARGs incomprehensive. Here, an ontology-aware deep learning model, ONN4ARG (http://onn4arg.xfcui.com/), is proposed for the discovery of novel ARGs based on multi-level annotations. Experiments based on billions of candidate microbial genes collected from various environments show the superiority of ONN4ARG in comprehensive ARG profiling. Enrichment analyses show that ARGs are both environment-specific and host-specific. For example, resistance genes for rifamycin, which is an important antibacterial agent active against gram-positive bacteria, are enriched in Actinobacteria and in soil environment. Case studies verified ONN4ARG's ability for novel ARG discovery. For example, a novel streptomycin resistance gene was discovered from oral microbiome samples and validated through wet-lab experiments. ONN4ARG provides a complete picture of the prevalence of ARGs in microbial communities as well as guidance for detection and reduction of the spread of resistance genes.

Keywords: antibiotic resistance gene, ontology-aware, deep learning, novel ARG, microbiome

Introduction

With the development of metagenomics and next-generation sequencing, many new microbial taxa and genes have been discovered, but different kinds of "unknowns" remain. For instance, the microbes found in the human gut microbiome involve 25 phyla, more than 2,000 genera, and 5,000 species[1]. However, the functional diversity of microbiomes has not been fully explored, and about 40% of microbial gene functions remain to be discovered[2]. A typical example is the antibiotic resistance gene (ARG), which is an urgent and growing threat to public health[3]. In the past few decades, problems caused by antibiotic resistance have drawn the public's attention[4]. Antibiotic resistance in pathogens has been an increasing threat to human health over the past decade, and it is widely accepted that antibiotic resistance development and spread in microbes can be largely attributed to the abuse and misuse of antibiotics. A direct correlation between antimicrobial use and the extent of antimicrobial resistance has been reported[4]. Antimicrobial resistance genomic data is an ever-expanding data source, with many new ARG families discovered in recent years[5,6]. The discovery of resistance genes in diverse environments offers possibilities for early surveillance, actions to reduce transmission, gene-based diagnostics, and, ultimately, improved treatment[7].

Existing annotated ARGs have been curated manually or automatically for decades.

Presently, there are 2,979 annotated ARGs in the reference database CARD[5,6] (v3.1.2, released in April 2021), 3,159 in the ResFinder database[8] (as of May 2021), and 2,675 in SwissProt[9] (as of May 2021). These annotated ARGs are categorized into antibiotic resistance types, which are organized in an ontology structure (see **Methods**, **Supplementary Figure 1**), in which higher-level ARG types cover lower-level ARG types. For example, AHE40557.1 is annotated in the CARD database as a streptomycin resistance gene, which belongs to a lower-level ARG type aminoglycoside and a higher-level ARG type non-beta-lactam. Current ARG databases are far from complete: though no ARG database contains more than 4,000 well-annotated ARGs, NCBI non-redundant database searches yielded more than 7,000 putative genes annotated with “antibiotic resistance” as of May 2021. Therefore, we deemed that there is a large gap between the genes annotated in ARG databases and the possible ARGs that already exist in general databases, not to mention ARGs that are not yet annotated.

Many ARG prediction tools have been proposed in the past few years[8,10-20]. These tools can generally be divided into two approaches. One approach is sequence-alignment, such as BLAST[21], USEARCH[22], and Diamond[23], which uses homologous genes to annotate unclassified genes. A confident prediction requires a homolog with sequence identity greater than 80% in many programs, such as ResFinder[8,11]. The other approach is deep learning, such as DeepARG[12] and HMD-ARG[16], which uses neural network models to predict and annotate ARGs. The input of deep learning approach can be bit-score (for DeepARG) or one-hot encoding vector of protein sequence (for HMD-ARG).

Several limitations still preclude comprehensive profiling of antibiotic resistance genes. A more comprehensive set of ARGs could be roughly defined as having more ARGs in type and number with less false-positive entries, regardless of the homology with known ARGs, and many of these ARGs could be experimentally validated. Based on this definition, existing tools fall short in comprehensive profiling of ARGs.

First, existing tools are limited to a few types of ARGs due to the fact that the datasets used for building models are specialized and therefore cannot reconstruct the comprehensive profile of ARGs across various environments. For example, HMD-ARG[16] identifies only 15 types of resistance genes, and PATRIC[13] is limited to identifying ARGs encoding resistance to carbapenem, methicillin, and beta-lactam antibiotics. Second, existing tools fall short in discovering novel ARGs, which usually lack homology to known sequences in the reference databases. For instance, the gene POCOZ1 (VraR) that confers resistance to vancomycin has a sequence identity of only 24% to the homolog from the CARD[12]. Recognizing such remote homologs requires the ability to perceive the correlation between the internal features of genes, which is challenging for existing tools. Therefore, there is an urgent need for a new approach to address these limitations.

Here, we propose an ontology-aware deep learning approach, ONN4ARG, which allows comprehensive identification of ARGs. ONN4ARG is an ontology-aware neural network model that employs a novel ontology-aware layer and generates multi-level annotations of antibiotic resistance types (**Figure 1**). Systematic evaluations show that the ONN4ARG model has a profound performance improvement over state-of-the-art models such as DeepARG, especially for the detection of remotely homologous ARGs. The application of ONN4ARG has uncovered a total of 120,726 ARGs from the microbiome, which has greatly expanded the existing ARG repositories. Enrichment analyses have confirmed the enrichment patterns of ARG types across multiple environments, showing that ARGs are both environment-specific and host-specific. For example, resistance genes for rifamycin, which is an important antibacterial agent active against gram-positive bacteria, are enriched in Actinobacteria and in soil environment. Case studies have also verified the ability of ONN4ARG for novel ARG discovery. For example, a recently experimentally validated ARG gene GAR[7], which is not in the CARD database, could not be identified by DeepARG or HMD-ARG but was predicted by ONN4ARG. A novel streptomycin resistance gene was also discovered by ONN4ARG from oral microbiome data and validated through wet-lab experiments.

In summary, ONN4ARG is a comprehensive deep learning method for ARG discovery, which provides a complete picture of the prevalence of ARGs in microbial communities as well as guidance for detection and reduction of the spread of resistance genes.

Results

ONN4ARG model employs an ontology-aware neural network for ARG identification and classification

To address the large gap between the genes annotated in ARG databases and the possible ARGs that already exist in general databases along with the ARGs that are not yet annotated, we propose ONN4ARG, which is an ontology-aware neural network model (**Figure 1a**), that could be used to predict ARGs in a comprehensive manner. ONN4ARG takes similarities (e.g., identity, e-value, bit-score) between the query gene sequence and ARG gene sequences and profiles (i.e., PSSM) as inputs and predicts ARG annotations for the query gene. These sequence-alignment similarities and profile-alignment similarities are pre-processed by calling Diamond[23] and HHsearch[24]. ONN4ARG generates multi-level annotations of antibiotic resistance types, which are compatible with the antibiotic resistance ontology structure. One advantage of ONN4ARG over state-of-the-art models is that ONN4ARG employs a novel ontology-aware layer that incorporates ancestor and descendent annotations to enhance annotation accuracies. ONN4ARG outperforms existing models, including DeepARG, with higher average accuracies and better generalization ability for unseen data. To train and evaluate our ONN4ARG model and for rapid deployment of ARG discovery in multiple contexts, we also built an ARG database (**Figure 1b**), namely, ONN4ARG-DB, which comprises ARGs from CARD and UniProt (see **Methods**).

Systematic evaluation and comparison

ONN4ARG has high efficiency, high accuracy, and comprehensiveness for ARG identification based on our systematic evaluation of ONN4ARG and comparison with other models. The evaluation and comparison were based on ONN4ARG-DB, with 28,396 positive ARGs and 17,937 negatives, out of which 75% of the dataset was randomly selected for training and the remaining 25% of the dataset was selected for testing (see **Methods**).

We evaluated ONN4ARG's efficiency, accuracy, and comprehensiveness. As an ontology-aware deep learning model, ONN4ARG is fast: it could complete ARG identification for all genes in the testing dataset within four hours, which is equivalent to one second per gene identification. As shown in **Figure 2a**, ONN4ARG was more accurate for ARG identification (overall accuracy of 97.70%) compared to sequence alignment (overall accuracy of 69.11%) and DeepARG (overall accuracy of 96.39%). Moreover, ONN4ARG achieved an overall precision of 75.59% and an overall recall of 89.93%, which were higher than DeepARG's overall precision of 68.30% and overall recall of 77.84% (**Figure 2b**). It is natural that ONN4ARG could not outperform DeepARG in all resistance types and this is exemplified by results on pleuromutilin due to the small number of sequences for pleuromutilin in the ONN4ARG-DB. In addition, for most of the resistance types that have adequate number of sequences, ONN4ARG's results could achieve higher precision and recall. Thus, with the accumulation of annotated ARG sequences, greater advantages of both ONN4ARG-DB and ONN4ARG could be expected. Furthermore, ONN4ARG was more comprehensive for ARG identification: there were 4,916 ARGs in the testing set (with the masking threshold of testing equal to 0.4, see **Methods**), out of which 4,913 were identified by the ONN4ARG model, whereas DeepARG identified 4,906 (**Supplementary Table 1**).

ONN4ARG demonstrates an advantage over other methods in identification of remotely homologous ARGs whose sequences are not similar to existing ARG sequences (**Supplementary Tables 2 and 3**). In this context, when testing with only remote homologs (i.e., the masking threshold of testing set equal to 0.4), ONN4ARG achieves an accuracy of 94.26%, which is significantly improved from 89.85% of DeepARG. When testing with all close and remote homologs (i.e., the masking threshold of testing set equal to 1.0), both ONN4ARG and DeepARG achieved high accuracies. These results validate ONN4ARG's significantly better generalization abilities than sequence-alignment and DeepARG, which makes ONN4ARG especially suitable for identification of remotely homologous ARGs and indicates ONN4ARG's ability for novel ARG discovery (**Supplementary Tables 1–3**).

In summary, ONN4ARG has high efficiency, accuracy, and comprehensiveness for ARG identification, and it possesses the ability for identification of remotely homologous ARGs.

Applications of ONN4ARG on metagenomic data

We collected metagenomic samples from several published studies[25,26]. These samples were mainly from "marine," "soil," and "human" environments. Human-associated samples consisted of two gut groups (one group from Madagascar, i.e., GutM; the other group from Denmark, i.e., GutD), one oral group, and one skin group (both oral and skin groups were from the HMP project). Details about these samples are provided in **Supplementary Table 4**. Then, genes were obtained by calling Prodigal[27]. The ONN4ARG model was used to predict whether these unclassified genes were ARGs and their corresponding resistance types. In total, 120,726 ARGs were identified from microbiome samples, many of which are novel, which greatly expands the existing ARG repositories.

Broad-spectrum profile of predicted ARGs among diverse environments

We first investigated the proportion of predicted ARGs for different sequence lengths. The distribution shows that about half of the predicted ARGs have a length of 128–256 amino acid residues (**Figure 3a**). We found that human-associated microbiome samples carry a higher abundance of ARGs, especially for the oral group, in which more than one resistance gene could be observed out of a hundred genes on average (**Figure 3b**, **Supplementary Table 5**).

For ARGs detected in samples from all environments, we found that about a third of them (42,848 out of all 120,726 ARGs) had sequence identity of less than 40% to their homologs in the ONN4ARG-DB (**Figure 3c**). We define these ARGs as novel ARGs, which have low sequence identities when aligned to their homologs in the reference database (i.e., ONN4ARG-DB). For example, we found 45% of predicted ARGs in the marine group belonged to novel ARGs (**Figure 3c**).

In total, 31 ARG types were detected in these various environments (**Figure 3d**, **Supplementary Figure 2**). The number of predicted ARG sequences for different types varied greatly (**Figure 3d**), from a few (i.e., nitrofuran) to thousands (i.e., fluoroquinolone). In general, fluoroquinolone and tetracycline resistance genes were more abundant than other types (**Figure 3d**). As expected, these abundant ARGs were usually associated with the antibiotics used extensively in human medicine or veterinary medicine, including growth promotion[28]. Novel ARG detection indicates the unique ability of ONN4ARG in novel ARG discovery and ARG abundance profiling in various environments, which would help researchers to better understand the prevalence of antibiotic resistance genes.

Enrichment of predicted ARGs among diverse hosts and environments

Rapid deciphering of potential antimicrobial-resistant pathogens is necessary for effective public health monitoring. The host-tracking of ARGs allows for accurate identification of pathogens. Therefore, we conducted Kraken2[29] analysis to track the hosts of these predicted ARGs. Results showed that there are 949 genera, each genus carries at least one type of ARG (**Supplementary Table 6**). The host composition and distribution of all classified ARGs for the most abundant 20 genera are displayed in **Supplementary Figure 3**. The host distribution shows that these predicted ARGs are primarily affiliated with Proteobacteria (38.2%), including *Candidatus Pelagibacter*, *Pseudomonas*, *Bradyrhizobium*, and *Escherichia* (**Supplementary Figure 3**). The most abundant ARGs carried by the 20 genera were resistance types of fluoroquinolone, macrolide, peptide, penam, and tetracycline, accounting for about half of the total detected ARGs (**Supplementary Figure 3**). We used network inference based on strong (Pearson's correlation $\rho > 0.8$) and significant (P-value < 0.01) correlations to investigate the co-occurrence patterns among ARG types and microbial taxa (**Supplementary Figure 4**, **Supplementary Note**). The co-occurrence network indicated the co-occurrence patterns between ARGs and microbial taxa. For example, ARGs that belong to beta-lactam resistance type (e.g., cephamycin, penam, penem, and monobactam) were observed to appear together in Proteobacteria.

Enrichment analyses showed that ARGs are both environment-specific and host-specific (**Figure 4**). We found that the proportion of certain types of ARGs was significantly higher in certain environments than in others. For example, rifamycin resistance genes were found enriched in Actinobacteria (with proportion of 0.1%) and enriched in the soil environment (with proportion of 4.7%) (**Figure 4**). Rifamycin is an important antibacterial agent active against gram-positive bacteria, and it has a wide range of applications[30,31]. The enrichment results were not surprising because *Actinomycetes* is a representative genus widely distributed in various soil environments, and its rifamycin resistance is compatible with its ability for rifamycin production[32-35].

Evaluation of the ability for novel ARG identification using a recently annotated ARG

We further evaluated ONN4ARG's ability for novel ARG identification based on the assessment of a newly annotated aminoglycoside resistance gene, GAR[7]. GAR is a recently reported aminoglycoside resistance gene (e.g., gentamicin, micronomicin) that belongs to non-beta-lactam, which is not present in CARD (v3.0.3), UniProt (as of May 2021), DEEPARG-DB (v1.0.2), HMD-ARG-DB (as of May 2021), and ONN4ARG-DB. We searched the sequence of GAR with both DeepARG and

HMD-ARG models, and the results showed that both of these models indicated it as non-ARG. We searched the sequence of GAR against all the sequences in ONN4ARG-DB using Diamond and did not find any homologous gene as well. However, the prediction by ONN4ARG identified GAR as an ARG resistant to non-beta-lactam with high confidence (probability score = 100%). We should emphasize that though ONN4ARG could only predict GAR as non-beta-lactam and not as sub-type of aminoglycoside, it was the only method used in this study that could predict GAR as an ARG gene, which again confirms ONN4ARG's better generalization ability for novel ARG discovery.

Functional verification of candidate novel resistance genes

To identify promising putative novel resistance genes, we used four criteria: (i) remote homologs to reference ARGs, (ii) prediction with high confidence, (iii) predicted to be single-type resistance, and (iv) the host is known. Despite the large number of candidate genes discovered by the ONN4ARG model (**Supplementary Table 5**), only 4,365 ARGs fulfilled all mentioned criteria (**Supplementary Table 7**).

We selected one candidate ARG (Candi_60363_1) for further experimental validation (**Supplementary Tables 8 and 9**). Candi_60363_1, detected in *Streptococcus* in the oral environment, was predicted to be streptomycin (belonging to aminoglycoside) resistant with high confidence by the ONN4ARG model, and the closest homolog of Candi_60363_1 in ONN4ARG-DB is P12055 (sequence identity of 37.2%). One positive control from CARD (AHE40557.1, streptomycin resistance gene) was used in our experiments for verification of the experimental system. All these genes were heterologously expressed in the *E. coli* BL21 (DE3) host by the induction of Isopropyl β-D-1-thiogalactopyranoside (IPTG) and tested for minimal inhibitory concentration (MIC) (**Figure 5a**). The result showed that the mRNA level of the genes increased with the addition of 1 mM IPTG compared with that without IPTG (**Figure 5b**), which verified the expression of the genes induced by IPTG. Furthermore, the MIC of the strain containing the positive control gene AHE40557.1 was more than 1,024 μg/ml (**Supplementary Figure 5**), which is consistent with previous reports[36,37]. This verified that our MIC measuring experimental system works well. Our results showed that the MIC of the strain containing Candi_60363_1 was significantly higher than the negative control containing no insert (**Figure 5c**, **Supplementary Figure 5**), which demonstrated the increased resistance to streptomycin of the novel candidate gene Candi_60363_1 and verified the good performance of our model.

Phylogeny and structure of Candi_60363_1

There are remote similarities between Candi_60363_1 and all known ARGs in the reference database, including aminoglycoside resistance genes (closest homolog is P12055, sequence identity of 37.2%). The function annotation of P12055 shows that it has the catalytic activity of reaction between streptomycin and ATP, and it is required for streptomycin resistance (https://www.uniprot.org/citations/3357770). Additionally, the search result of Candi_60363_1 using InterPro (the Integrated Resource of Protein Domains and Functional Sites) shows the protein family matching to Candi_60363_1 is IPR007530, which is also known as aminoglycoside 6-adenylyltransferase that confers resistance to aminoglycoside antibiotics. Then, we used BLAST to search homologs of Candi_60363_1 from the NCBI non-redundant protein database. The BLAST result showed that there are 44 homologs with sequence identity greater than 80%, and they are from various organisms (**Supplementary Table 10**), such as *Streptococcus oralis*, *Peptoniphilus lacrimalis DNF00528*, and *Mycobacteroides abscessus subsp. Abscessus*. Considering that Candi_60363_1 is harbored by distantly related species, it obviously has mobility. Notably, the most similar protein of Candi_60363_1 from the NCBI non-redundant protein database (87.5% identity, SHZ78752.1) is also annotated as aminoglycoside adenylyltransferase (**Supplementary Table 10**). The result of BLAST search against the NCBI non-redundant protein database and other databases showed that Candi_60363_1, which is absent in all the existing ARG databases, is highly likely to be an ARG that confers resistance to aminoglycoside antibiotics.

Aminoglycoside modifying enzymes are the most clinically important resistance mechanism against aminoglycosides[38]. Aminoglycoside modifying enzymes are divided into three enzymatic classes, namely, aminoglycoside N-acetyltransferase (AAC), O-nucleotidyltransferase (ANT), and O-phosphotransferase (APH). We investigated the phylogenetic relationship between Candi_60363_1 and the known aminoglycoside modifying enzymes. The phylogenetic tree of Candi_60363_1 and related proteins (**Figure 6a**) shows that Candi_60363_1 is clearly separated from the known aminoglycoside modifying enzymes and is located among proteins mostly annotated as aminoglycoside adenylyltransferase. Phylogenetic analysis indicated its evolutionarily close relationships with known aminoglycoside adenylyltransferase.

Protein structure prediction results confirmed the anti-microbial functionality of Candi_60363_1. The optimal Candi_60363_1-streptomycin complex structure and the corresponding interaction details are described in **Figure 6b**. The optimal binding affinity between the Candi_60363_1 and streptomycin is -7.7 kcal/mol (**Supplementary Table 11**), which is 1.6 kcal/mol lower than the negative control. As shown in **Figure 6b**, the Streptomycin ligand can fit the ARG protein structure well and generate a geometric and energetic docking complex.

From wet-lab experiments, phylogenetic analysis, and protein structure docking, we consider that Candi_60363_1 predicted by ONN4ARG is highly likely a real ARG gene.

Discussion

In this study, we proposed an ontology-aware deep learning method, ONN4ARG, for the detection and understanding of antibiotic resistance genes. The ONN4ARG model is capable of accurately identifying ARGs from coarse to fine levels and discovering novel ARGs that lack homology to known sequences in the reference databases. To complement ONN4ARG for ARG mining applications, we have also created a custom

ARG database, ONN4ARG-DB, that contains 28,396 well-curated ARGs. The application of ONN4ARG uncovered 120,726 ARGs from microbiome samples, out of which 42,848 are novel, which substantially expands the existing ARGs repositories.

The novelty of this work is in three contexts. First, ONN4ARG has the potential for detection of remotely homologous ARGs and thus generates a more comprehensive set of ARGs. The advantage of our ONN4ARG model over state-of-the-art models is that ONN4ARG employs a novel ontology-aware layer that incorporates ancestor and descendant annotations to enhance annotation accuracies. The comprehensive antibiotic resistance ontology used in the ONN4ARG model consists of four levels and more than 100 resistance types (**Supplementary Table 12**), which includes hierarchical antibiotic resistance annotations from the most popular ARG database, CARD. Thus, the classification range of the ONN4ARG model is substantially larger than current tools (e.g., 30 types supported for DeepARG and 15 types supported for HMD-ARG). The ability of ONN4ARG to identify remote homologs (i.e., sequence identity between 30% and 40%) allows more accurate prediction for those misclassified by sequence-alignment based tools as false negatives. Therefore, ONN4ARG greatly reduces false negatives and offers a powerful approach for comprehensive and accurate profiling of ARGs.

Second, it enabled the comprehensive enrichment analysis of ARGs, species-wise and environment-wise. For the actual application of the ONN4ARG model, we investigated the presence of ARGs in a variety of environments, including water, soil, and the human gut, and the results showed that ARGs are environment-specific and host-specific (**Figure 4**). The environment-specific and host-specific phenomenon of ARGs may be caused by specific bacteria evolving to possess specific types of ARGs in response to specific environments, and horizontal gene transfer may be one of the mediating pathways of this process. For example, one published study has reported that *Amycolatopsis* in the soil environment produces rifamycin and thus gains ecological advantages over other bacteria[32].

Third, the novel ARGs predicted by ONN4ARG could be functionally validated. Functional verification of a novel streptomycin resistance gene (i.e., Candi_60363_1) with wet-lab experiments demonstrated the ability of the ONN4ARG model for novel ARG discovery. Although the MIC test value of Candi_60363_1 was only two times higher than that of the control (**Figure 5**), this increase was still sufficient to indicate the presence of resistance. Moreover, phylogenetic analysis and protein structure docking further confirmed that Candi_60363_1 is highly likely to be an ARG that confers resistance to aminoglycoside antibiotics. Another validation of novel ARG identification based on the assessment of a recently annotated ARG (i.e., GAR) also indicated the ability of the ONN4ARG model for novel ARG discovery. GAR is a novel ARG that is resistant to a variety of aminoglycosides (e.g., gentamicin and micronomicin). We searched the sequence of GAR using other tools (i.e., DeepARG and HMD-ARG), and the results showed that both of those models indicated it as non-ARG. We emphasize that the ONN4ARG model only identified GAR as non-beta-lactam. This shows that the multi-level annotations of ONN4ARG allow low resolution recognition, which can greatly decrease the false negative rate.

In summary, ONN4ARG is a deep learning approach for ARG identification. It allows in-depth gene mining on large-scale metagenomic data and helps researchers discover novel ARGs. ONN4ARG provides a complete picture of the prevalence of ARGs in the microbial communities and guidance for detection and reduction of the spread of resistance genes in such scenarios, including clinical research, environmental monitoring, and agricultural management.

ONN4ARG could be improved in a few ways. For more comprehensive ARG prediction, continuous improvement of curating ARG nomenclature and annotation databases is required. For novel ARG prediction, especially those belonging to entirely new ARG families, deep learning models might need to consider more information other than sequence alone. We believe these efforts could lead to a holistic view about ARGs in diverse environments around the globe.

Methods

Dataset

The ARGs we used in this study for model training and testing were from the Comprehensive Antibiotic Resistance Database (CARD[5,6], v3.0.3). We also used protein sequences from the UniProt (SwissProt and TrEMBL) database to expand our training dataset. First, genes with ARG annotations were collected from CARD (2,587 ARGs) and SwissProt (2,261 ARGs). Then, their close homologs (with sequence identities greater than 90%) were collected from TrEMBL (23,728 homologous genes). These annotated and homologous ARGs made up our positive dataset. The negative dataset was made from non-ARG genes that had relatively weak sequence similarities to ARG genes (with sequence identities smaller than 90% and bit-scores smaller than alignment lengths) but not annotated as ARG genes in SwissProt (17,937 genes). Finally, redundant genes with identical sequences were filtered out. As a result, our ARG gene dataset, namely, ONN4ARG-DB, contained 28,396 positive and 17,937 negative genes. For evaluation and comparison of ONN4ARG, 75% of the dataset was randomly selected for training, and the remaining 25% of the dataset was selected for testing.

Antibiotic resistance ontology

The antibiotic resistance ontology was organized into an ontology structure, which contains four levels. The root (first level) is a single node, namely, “arg” (**Supplementary Table 12**). There are 1, 2, 34, and 277 nodes from the first level to the fourth level, respectively. For instance, there are “beta-lactam” and “non-beta-lactam” in the second level, “acridine dye” and “aminocoumarin” in the third level, and “acriflavine” and “clorobiocin” in the fourth level. For example, AHE40557.1 is annotated in the CARD database as a streptomycin resistance gene, which belongs to a lower-level ARG type aminoglycoside and a higher-level ARG type non-beta-lactam (**Supplementary Figure 1**).

Protein annotations

The protein sequences for training and testing were annotated according to the antibiotic resistance ontology. For example, AHE40557.1 is annotated in the CARD database as a streptomycin resistance gene, which belongs to a lower-level ARG type aminoglycoside and a higher-level ARG type non-beta-lactam. Accordingly, this protein will be annotated as “arg” at the first level, “non-beta-lactam” at the second level, “aminoglycoside” at the third level, and “streptomycin” at the fourth level.

Sequence-alignment

We used Diamond[23] as the sequence-alignment tool for comparison. For queries in the testing set, we searched them against the training set. The target with the highest identity was defined as the closest homologous gene for each query. Then, we compared whether the actual annotation of the query was consistent with the annotation of its closest homologous gene to evaluate the accuracy.

DeepARG

DeepARG[12] is a newly developed tool that applies a neural network to identify antibiotic resistance genes. For queries in the testing set, we used the DeepARG[12] model to predict their annotations. Then, we compared whether the actual annotation of the query was consistent with the predicted annotation to evaluate the accuracy.

Evaluation and comparison

In this study, the performance of ONN4ARG was evaluated and compared to state-of-the-art models, including sequence-alignment and DeepARG. For these three models, the training dataset was used to train the model parameters, and the testing dataset was used to calculate the prediction accuracies. Both DeepARG and ONN4ARG are deep learning models that use millions of parameters. Unlike deep learning models, sequence-alignment (i.e., Diamond) has only one parameter (i.e., the identity cutoff to distinguish ARG and non-ARG genes).

Masking threshold

To simulate remotely homologous ARG genes in our experiments, similarities between the query protein and its close homologs with sequence identities greater than a threshold were masked as zeros (i.e., no signals). For instance, when the masking threshold of testing set equaled 0.4, similarities between the query protein (in the testing set) and its close homologs (in the training set) with sequence identities greater than 40% were masked as zeros. Occasionally, all homologs were masked for a query protein, and such query proteins were removed during training and testing. For example, if query *X* had two homologs, *M* and *N*, and assuming the identity of *M* is 0.35 and the identity of *N* is 0.85, when the masking threshold of the testing set equaled 0.4, similarities between query *X* and homolog *M* were masked as zeros. When the masking threshold of the testing set equaled 0.9, query *X* was removed during testing.

Benchmark method

In this study, a prediction was defined to be correct if and only if all ARG annotations (including ancestor annotations from ARG ontologies) were correctly predicted. The accuracy of the tested model was defined as the number of correct predictions over the total number of predictions. The precision of the tested model was defined as the number of true positive predictions over the total number of positive predictions, and the recall was defined as the number of true positive predictions over the total number of true positive plus false negative predictions.

ARG mining on metagenomic data

We collected microbiome sequencing data from several published studies (**Supplementary Table 4**), including samples from soil, water, and human body. The gene contigs were processed by Prodigal[27]. Protein sequences were also obtained by the Prodigal program. Then, the ARG annotations of these protein sequences were predicted by using ONN4ARG.

Taxonomy annotation

Kraken2[29] program was used to identify the host of gene contigs. Then, each ARG predicted by ONN4ARG was annotated according to the host of its gene contigs.

Phylogenetic tree

Sequences of the 44 proteins most closely related to Candi_60363_1 were collected using BLASTP with default parameters on the NCBI non-redundant protein database. The retrieved proteins, Candi_60363_1 and all aminoglycoside resistance proteins from ResFinder[8] (https://bitbucket.org/genomicepidemiology/resfinder_db/src/master, last update March 2021), were aligned with ClustalW. The phylogenetic tree was calculated by MEGA[39] (v10) using the maximum likelihood algorithm with default parameters. The Interactive Tree of Life (iTOL v6) online tool[40] was used to prepare the phylogenetic tree for display.

Protein model and docking

Rosetta[41] was utilized to predict the protein structure using ab initio protein folding (http://robetta.bakerlab.org/). The top five protein pockets were generated for docking calculation with Surface Topography of proteins[42] (CASTp). We used the Cambridge Structure Database[43] to generate streptomycin conformers. The 3D protein-ligand complexes were obtained from AutoDock Vina[44].

ARG candidate gene expression plasmids construction and expression verification

The candidate resistance gene Candi_60363_1 and a positive control resistance gene AHE40557.1 were synthesized and subcloned into pUC19 vector, replacing *lacZ'* gene. The recombinant plasmids were then transformed into *E. coli* BL21 (DE3). The expression of resistance genes was induced by Isopropyl β-D-1-thiogalactopyranoside

(IPTG) and verified by quantitive Real-time PCR (qRT-PCR) assay. Briefly, bacteria were grown in LB supplemented with ampicillin (100 μg/ml) to OD600 of 0.5-0.6 by incubation at 37 °C with 220 rpm agitation, and the bacterial cultures were continued to grow until OD600 reached to 1.0 by adding or without adding 1 mM IPTG. The cells were harvested and total RNAs were purified using Bacterial RNA Extraction Kit (Vazyme Biotech). RNA reverse transcription was performed by using HiScript® II Q Select RT SuperMix for qPCR kit (Vazyme Biotech). qRT-PCR was performed by using SYBR Green Master Mix-High ROX Premixed (Vazyme Biotech) in a Stepone Plus system (Applied Biosystems). The *ldh* gene was used as internal control in all reactions. The relative fold changes were determined using the $2^{-\Delta\Delta Ct}$ method, in which *ldh* was used for normalization. The protein sequences of the synthesized genes were presented in **Supplementary Table 8** and the primer sequences for qRT-PCR were listed in **Supplementary Table 9**.

MIC determination

Minimal inhibitory concentrations (MICs) of the antibiotic for the strains containing resistance genes were determined using E-tests. Single colonies of the strains were incubated in 3 ml Mueller-Hinton (MH) medium with the addition of 100 μg/ml ampicillin at 35 °C for 4 hours, and the cells equal to $1.5X10^8$ cells/ml were spread on MH agar plates with the addition of 100 μg/ml ampicillin and 1 mM IPTG, and streptomycin MIC Test Strips (Liofilchem®) were put in the middle of the plates. The plates were incubated at 35 °C for 18-24 hours, and the MICs were read. The strain containing empty vector was used as a negative control.

Data availability

We collected metagenomic samples from several published studies[25,26], and these samples are mainly from "marine", "soil" and "human" associated environments. For human associated samples, including two gut groups (one group from Madagascar, i.e., GutM, the other group from Denmark, i.e., GutD), one oral group and one skin group (both oral and skin groups are from HMP project). Details and links about these samples are shown in **Supplementary Table 4**. The ONN4ARG-DB dataset could be accesses at: http://onn4arg.xfcui.com/.

Code availability

All source codes can be accessed at: https://github.com/xfcui/onn4arg, and http://onn4arg.xfcui.com/.

Acknowledgments

We are grateful to Mingyue Cheng and Hui Chong for insightful discussions. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. This work was partially supported by National Science Foundation of China grant 81774008, 81573702, 31871334 and 31671374, and the Ministry of Science and Technology’s national key research and development program grant (No. 2018YFC0910502).

Author contributions

KN, XC conceived and proposed the idea, and designed the study. YZ, CC, QJ, XZ, XC performed the experiments and analyzed the data. YZ, CC, XZ, KN and XC contributed to editing and proof-reading the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable

References

1 Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *bmc biology* **17**, doi:10.1186/S12915-019-0667-Z (2019).

2 Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *nature biotechnology* **32**, 834-841, doi:10.1038/NBT.2942 (2014).

3 Brogan, D. M. & Mossialos, E. A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *globalization and health* **12**, 8-8, doi:10.1186/S12992-016-0147-Y (2016).

4 Goossens, H., Ferech, M., Stichele, R. V. & Elseviers, M. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *the lancet* **365**, 579-587, doi:10.1016/S0140-6736(05)17907-0 (2005).

5 Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *nucleic acids research* **45**, doi:10.1093/NAR/GKW1004 (2017).

6 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *nucleic acids research* **48**, doi:10.1093/NAR/GKZ935 (2019).

7 Böhm, M.-E., Razavi, M., Marathe, N. P., Flach, C.-F. & Larsson, D. G. J. Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities. *microbiome* **8**, 1-11, doi:10.1186/S40168-020-00814-Z (2020).

8 Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *journal of antimicrobial chemotherapy* **75**, 3491-3500, doi:10.1093/JAC/DKAA345 (2020).

9 Bateman, A. *et al.* UniProt: A hub for protein information. *nucleic acids research* **43**, doi:10.1093/NAR/GKU989 (2015).

10 Rowe, W. *et al.* Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. *plos one* **10**, doi:10.1371/JOURNAL.PONE.0133492 (2015).

11 Kleinheinz, K. A., Joensen, K. G. & Larsen, M. V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage* **4**, e27943-e27943, doi:10.4161/bact.27943 (2014).

12 Arango-Argoty, G. *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *microbiome* **6**, 23-23, doi:10.1186/S40168-018-0401-Z (2018).

13 Davis, J. J. *et al.* Antimicrobial Resistance Prediction in PATRIC and RAST. *scientific reports* **6**, 27930-27930, doi:10.1038/SREP27930 (2016).

14 Lakin, S. M. *et al.* Hierarchical Hidden Markov models enable accurate and diverse detection of antimicrobial resistance sequences. *Commun Biol* **2**,

294-294, doi:10.1038/s42003-019-0545-9 (2019).

15 Doster, E. *et al.* MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic acids research* **48**, D561-D569, doi:10.1093/nar/gkz1010 (2020).

16 Li, Y. *et al.* HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *microbiome* **9**, 40-40, doi:10.1186/S40168-021-01002-3 (2021).

17 Gupta, S. K. *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy* **58**, 212-220, doi:10.1128/AAC.01310-13 (2014).

18 Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial agents and chemotherapy* **63**, e00483-00419, doi:10.1128/AAC.00483-19 (2019).

19 Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *genome medicine* **6**, 90-90, doi:10.1186/S13073-014-0090-6 (2014).

20 Rowe, W. P. M. & Winn, M. D. Indexed variation graphs for efficient and accurate resistome profiling. *bioinformatics* **34**, 3601-3608, doi:10.1093/BIOINFORMATICS/BTY387 (2018).

21 Altschul, S. F., Gish, W., Miller, W. C., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

22 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

23 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *nature methods* **12**, 59-60, doi:10.1038/NMETH.3176 (2015).

24 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *bmc bioinformatics* **20**, 1-15, doi:10.1186/S12859-019-3019-7 (2019).

25 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *science* **348**, 1261359-1261359, doi:10.1126/SCIENCE.1261359 (2015).

26 Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic acids research* **46**, D726-D735, doi:10.1093/nar/gkx967 (2018).

27 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *bmc bioinformatics* **11**, 119-119, doi:10.1186/1471-2105-11-119 (2010).

28 Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* **9**, 2490-2502, doi:10.1038/ismej.2015.59 (2015).

29 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *genome biology* **20**, 1-13, doi:10.1186/S13059-019-1891-0 (2019).

30 Qi, F. *et al.* Deciphering the late steps of rifamycin biosynthesis. *Nature Communications* **9**, 2342, doi:10.1038/s41467-018-04772-x (2018).

31 Floss, H. G. & Yu, T.-W. RifamycinMode of Action, Resistance, and Biosynthesis. *Chemical Reviews* **105**, 621-632, doi:10.1021/cr030112j (2005).

32 Yao, Y., Zhang, W., Jiao, R., Zhao, G. & Jiang, W. Efficient isolation of total RNA from antibiotic-producing bacterium Amycolatopsis mediterranei. *Journal of Microbiological Methods* **51**, 191-195, doi:doi.org/10.1016/S0167-7012(02)00078-7 (2002).

33 Wilson, M. C., Gulder, T. A. M., Mahmud, T. & Moore, B. S. Shared biosynthesis of the saliniketals and rifamycins in Salinispora arenicola is controlled by the sare1259-encoded cytochrome P450. *J Am Chem Soc* **132**, 12757-12765, doi:10.1021/ja105891a (2010).

34 Saxena, A., Kumari, R., Mukherjee, U., Singh, P. & Lal, R. Draft Genome Sequence of the Rifamycin Producer Amycolatopsis rifamycinica DSM 46095. *Genome Announc* **2**, e00662-00614, doi:10.1128/genomeA.00662-14 (2014).

35 Huang, H. *et al.* Micromonospora rifamycinica sp. nov., a novel actinomycete from mangrove sediment. **58**, 17-20, doi:doi.org/10.1099/ijs.0.64484-0 (2008).

36 Pinto-Alphandary, H., Mabilat, C. & Courvalin, P. Emergence of aminoglycoside resistance genes aadA and aadE in the genus Campylobacter. *antimicrobial agents and chemotherapy* **34**, 1294-1296, doi:10.1128/AAC.34.6.1294 (1990).

37 Holden, M. T. G. *et al.* Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen Streptococcus suis. *plos one* **4**, doi:10.1371/JOURNAL.PONE.0006072 (2009).

38 Ramirez, M. S., Nikolaidis, N. & Tolmasky, M. Rise and dissemination of aminoglycoside resistance: the aac(6′)-Ib paradigm. *frontiers in microbiology* **4**, 121-121, doi:10.3389/FMICB.2013.00121 (2013).

39 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

40 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

41 Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *methods in enzymology* **383**, 66-93, doi:10.1016/S0076-6879(04)83004-0 (2004).

42 Tian, W., Chen, C. & Liang, J. CASTp 3.0: Computed Atlas of Surface Topography of Proteins and Beyond. *biophysical journal* **114**, doi:10.1016/J.BPJ.2017.11.325 (2018).

43 Cole, J. C., Korb, O., McCabe, P., Read, M. G. & Taylor, R. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *journal of chemical information and modeling* **58**, 615-629, doi:10.1021/ACS.JCIM.7B00697 (2018).

44 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455-461, doi:10.1002/jcc.21334 (2010).

Figure 1

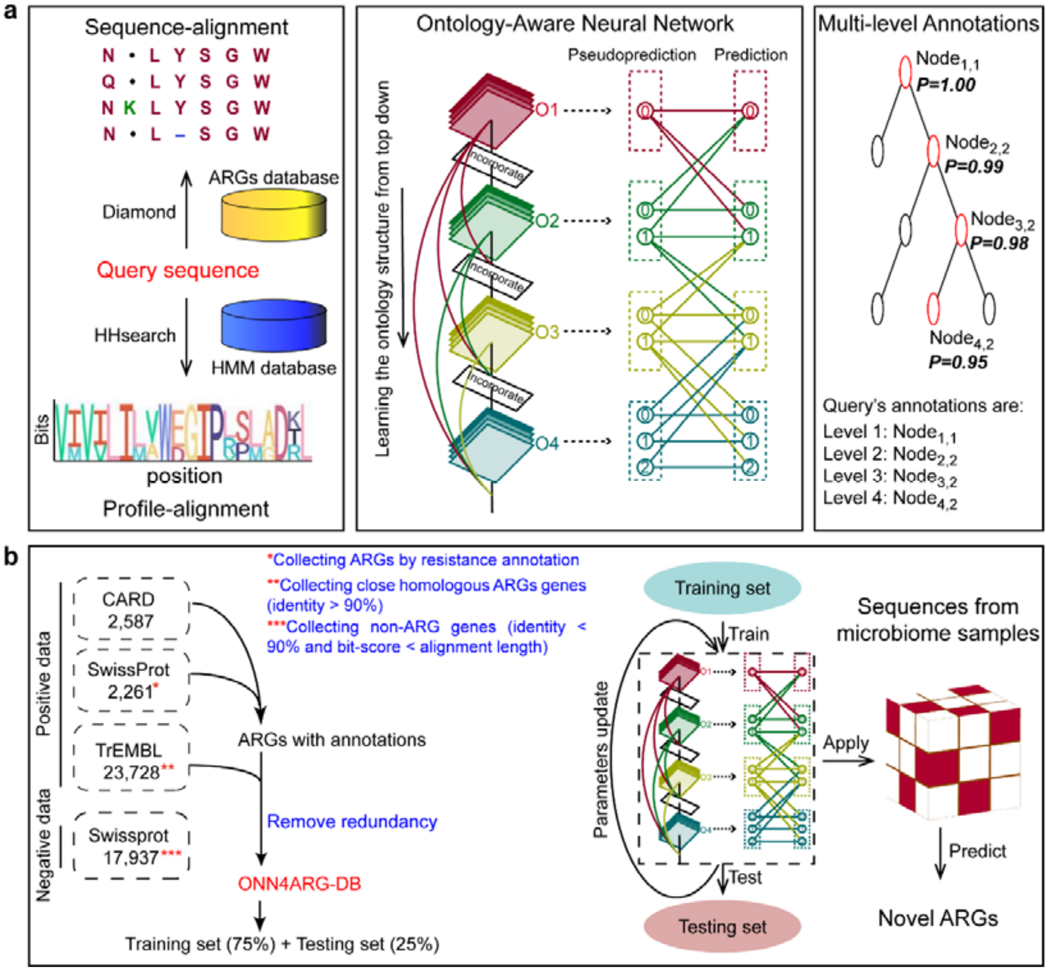


Figure 1. Overview of the ONN4ARG model and its use for novel ARG discovery. a. The input (left), architecture (middle), and output (right) of the ONN4ARG model. ONN4ARG takes similarities between the query gene sequence and ARG gene sequences and profiles as inputs. Then, ontology-aware layers (i.e., O1, O2, O3, and O4) are employed to incorporate ancestor and descendant annotations to enhance annotation accuracy. ONN4ARG outputs multi-level annotations of antibiotic resistance types, which are compatible with the antibiotic resistance ontology structure. **b.** Building the dataset for training and testing and applying it on microbiome sequencing data to discover novel ARGs.

Figure 2

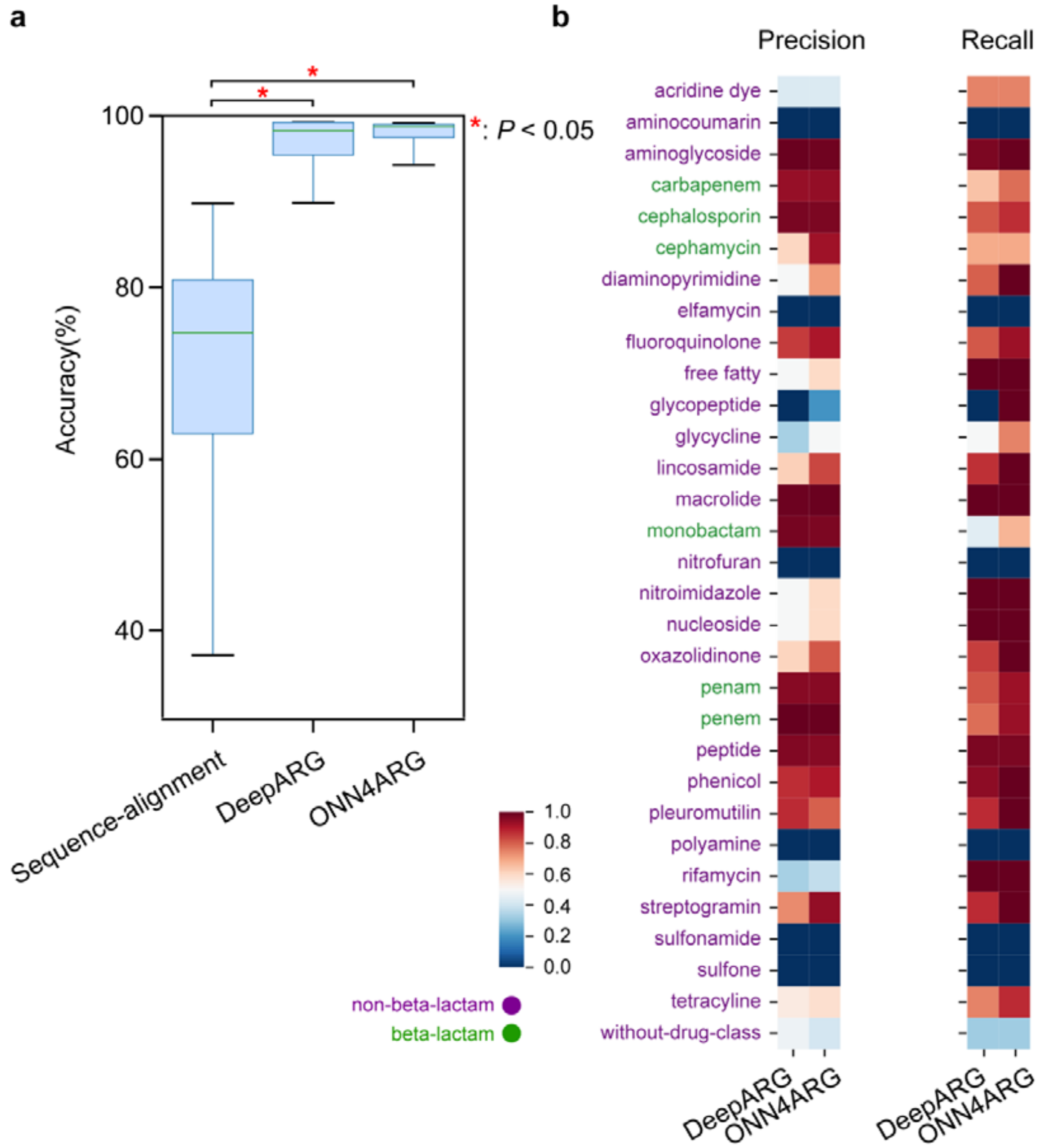


Figure 2. Systematic evaluation and comparison between sequence-alignment, DeepARG, and ONN4ARG. a. The accuracy of three models on ARG classification was assessed using a box plot. Diamond was used for sequence-alignment; significance test was based on the *t*-test. **b.** The precision and recall of DeepARG and ONN4ARG on ARG classification for each antibiotic resistance type. The masking threshold of testing set equaled 0.4 (details of masking threshold are provided in **Methods**).

Figure 3

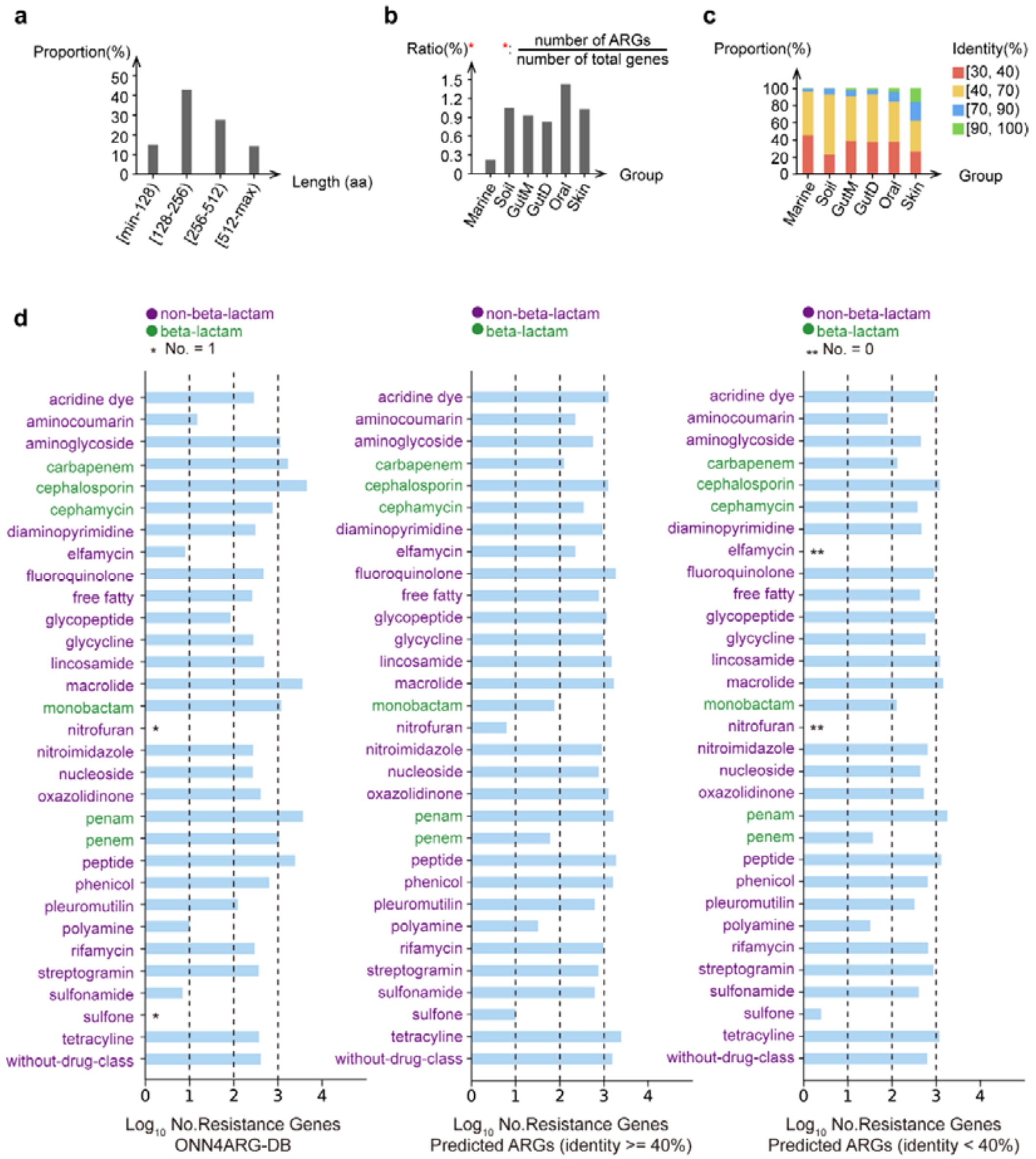


Figure 3. Broad-spectrum profile of predicted ARGs among diverse environments. a. The proportion of predicted ARGs for different protein sequence lengths. **b.** The abundance ratio of predicted ARGs among diverse environments. Abundance ratio was defined as the number of ARGs divided by the number of total genes. **c.** The proportion of predicted ARGs for different sequence identities among diverse environments. **d.** Number of genes in ONN4ARG-DB (left), predicted homologous ARGs (middle), and predicted novel ARGs (right) for various resistance types. The horizontal axis indicates the logarithmic number of genes, and the vertical axis indicates different antibiotic resistance types. We collected metagenomic samples from several

published studies; these samples were mainly from “marine,” “soil,” and “human” environments. Human-associated samples consisted of two gut groups (one group from Madagascar, i.e., GutM; the other group from Denmark, i.e., GutD), one oral group, and one skin group (both oral and skin groups were from the HMP project).

Figure 4

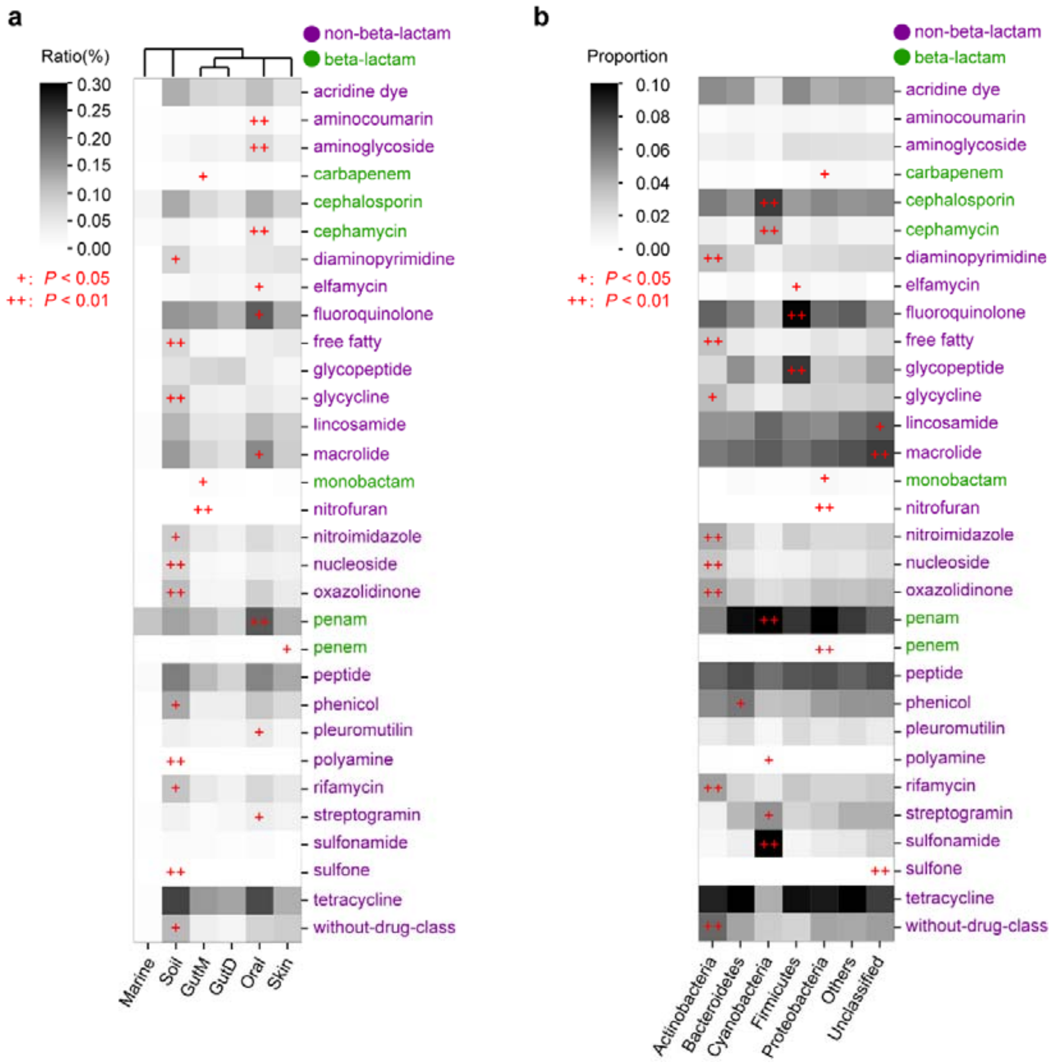


Figure 4. Enrichment of predicted ARGs among diverse environments and hosts. a. Relative abundance and enrichment of ARGs among diverse environments. Abundance ratio was defined as the number of ARGs divided by the number of total genes. **b.** Proportion and enrichment of ARGs among diverse hosts. Colors indicate the proportion of ARGs for each phylum and resistance type. Results for the most abundant five phyla that carry ARGs are shown. “+”: P-value < 0.05 (*t*-test); “++”: P-value < 0.01 (*t*-test).

Figure 5

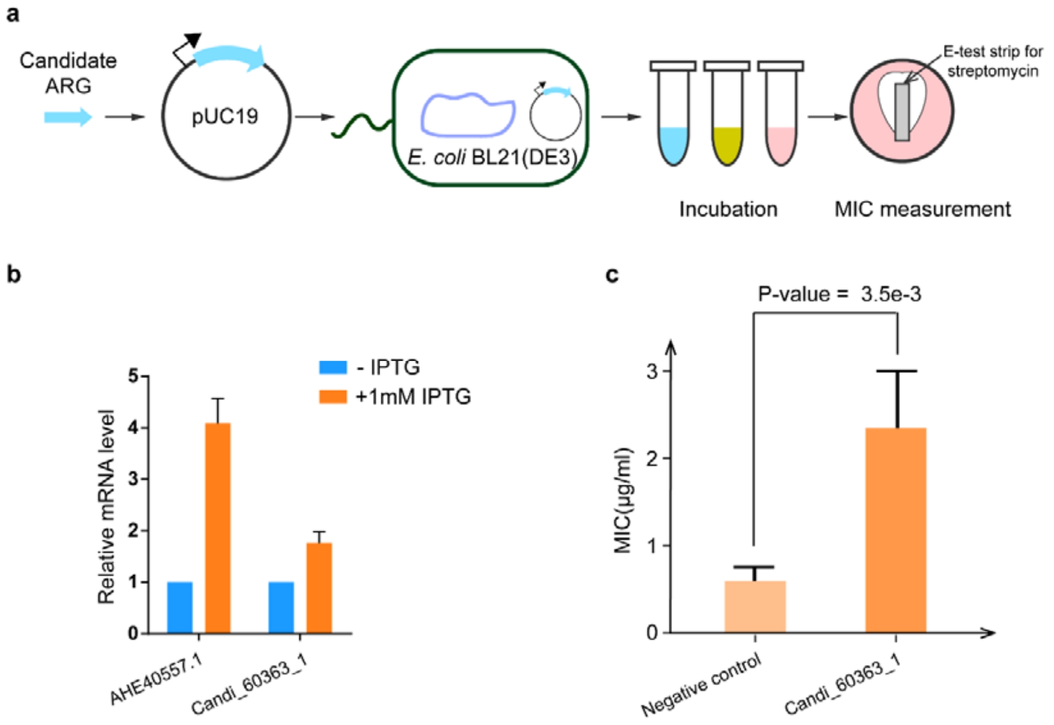


Figure 5. Functional validation of a predicted candidate novel ARG. a. A diagram showing the procedure of heterologous expression and functional analysis of the predicted candidate ARG in the *E. coli* BL21 (DE3) host. **b.** Gene expression validation of the predicted candidate ARG. The vertical axis indicates the relative mRNA level. **c.** The MIC of the predicted candidate ARG and negative control. The vertical axis indicates the MIC value. The MIC of the predicted candidate novel ARG is significantly higher than the negative control (*t*-test, P-value = 3.5e-3).

Figure 6

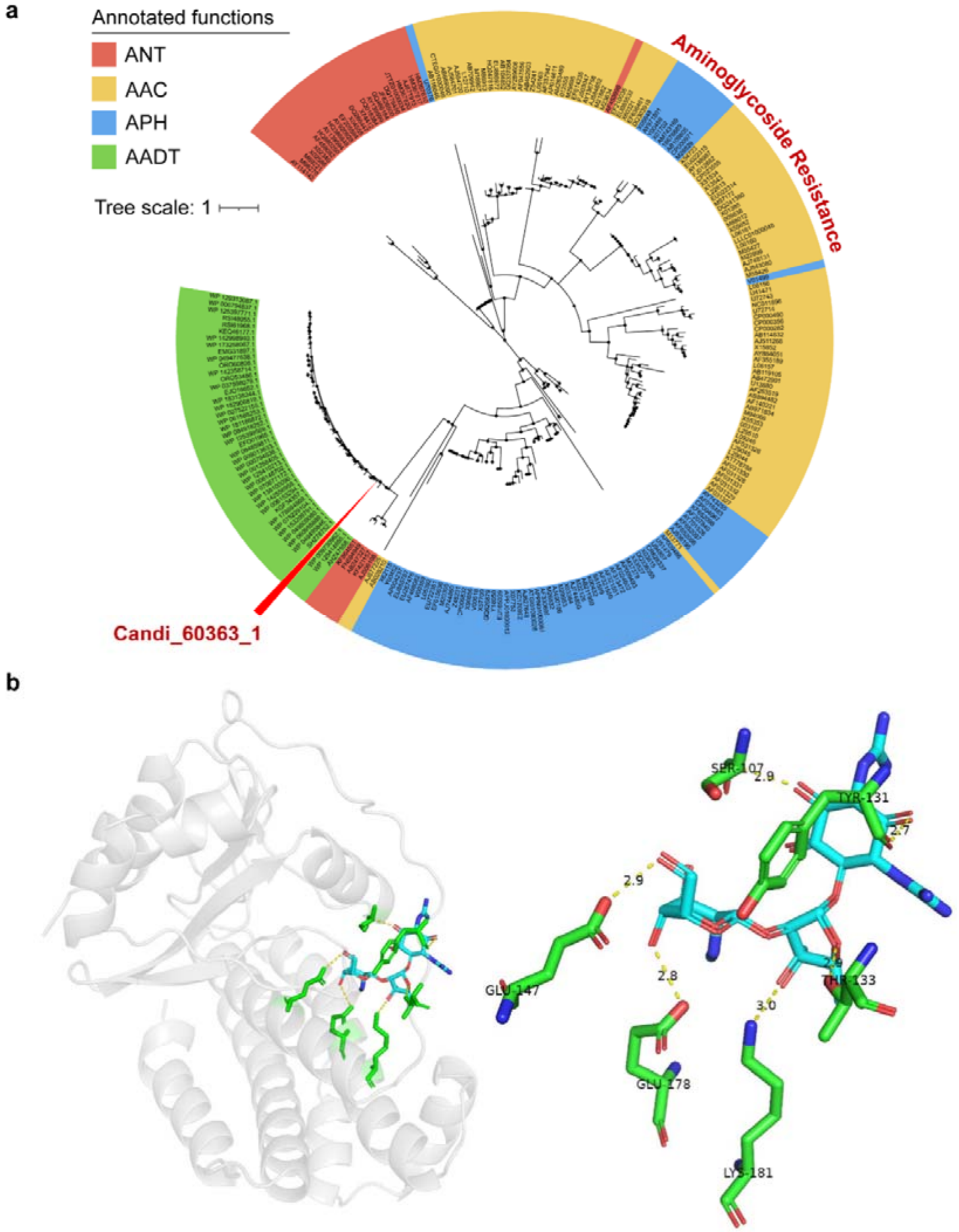
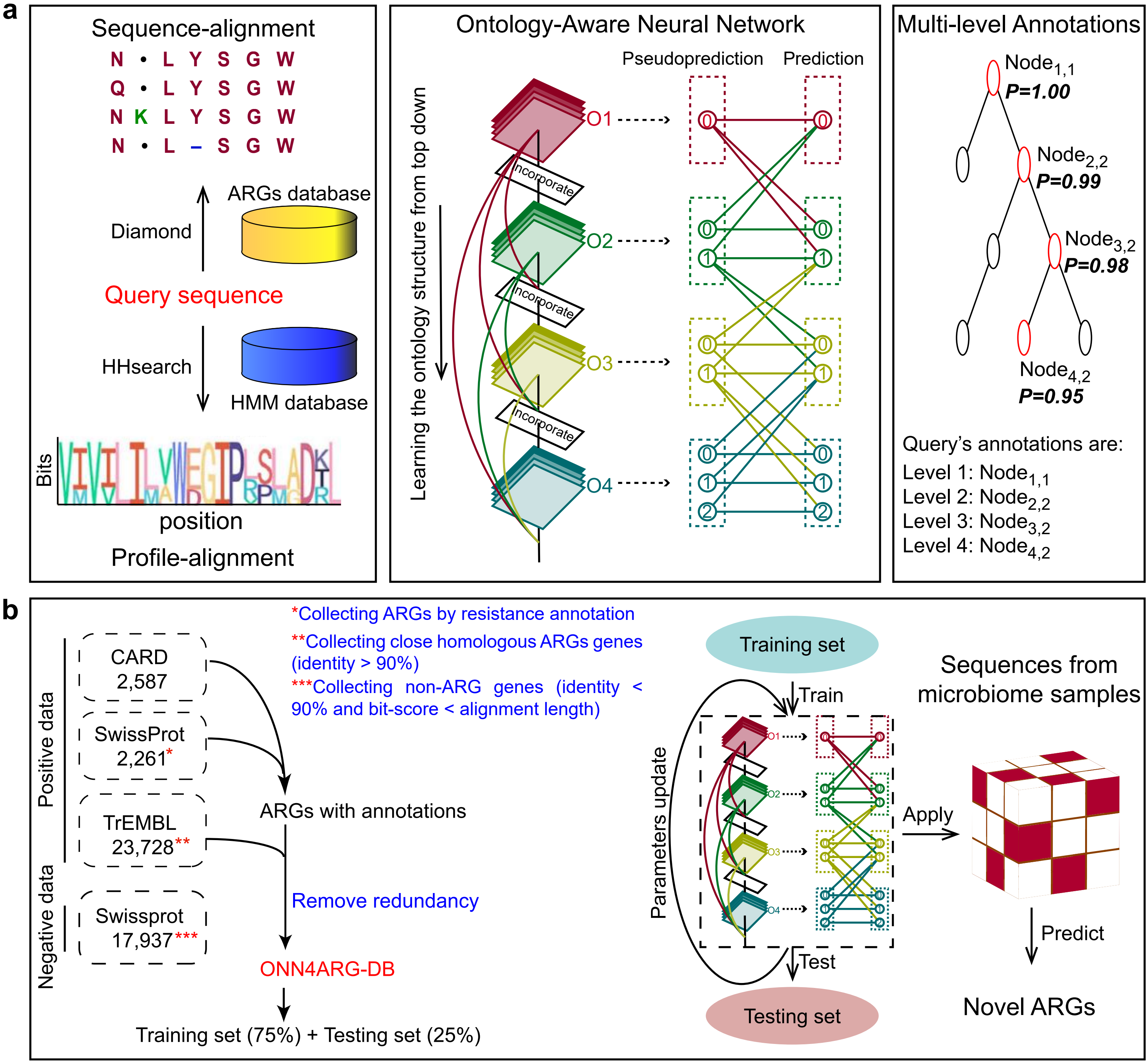
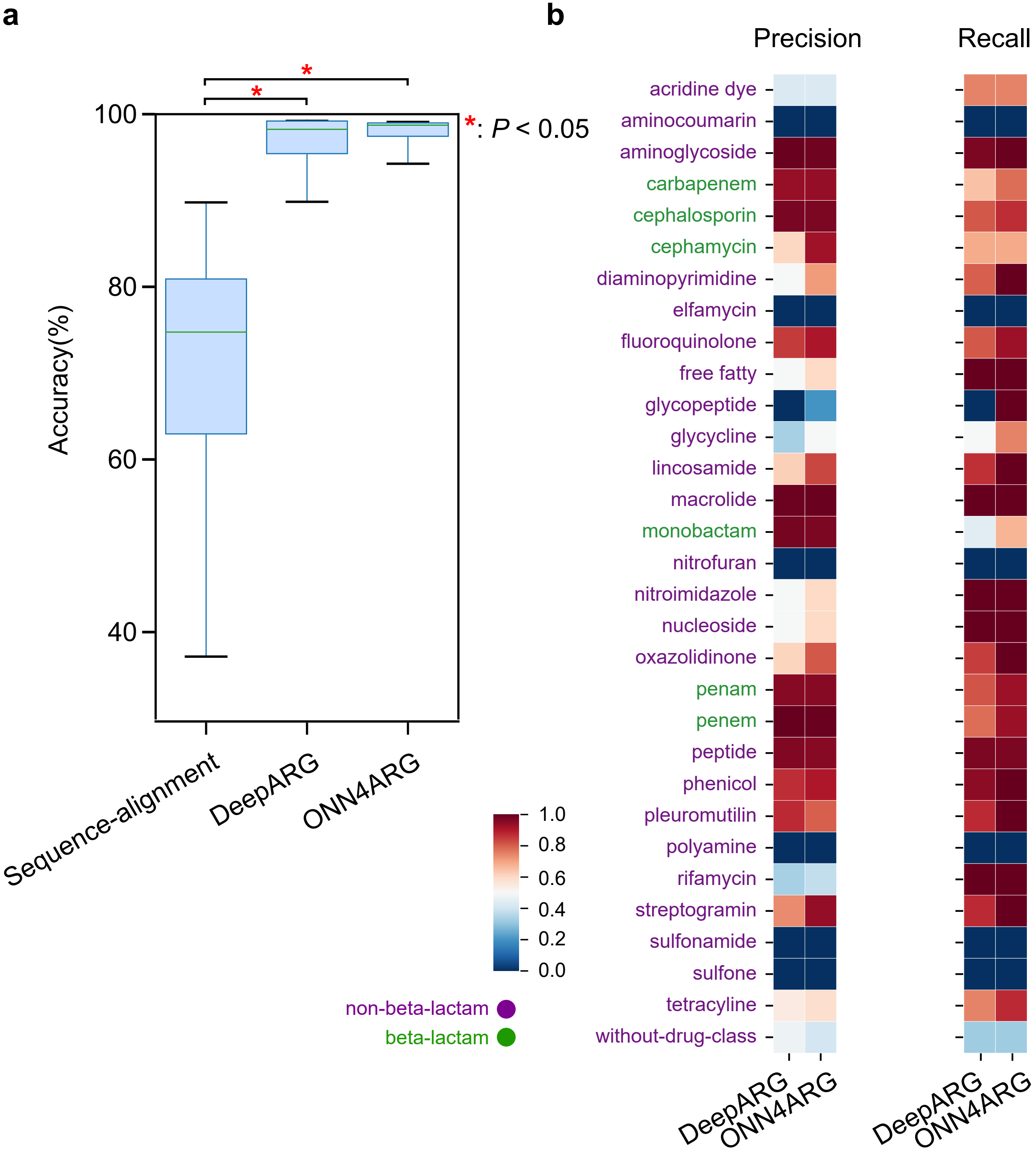
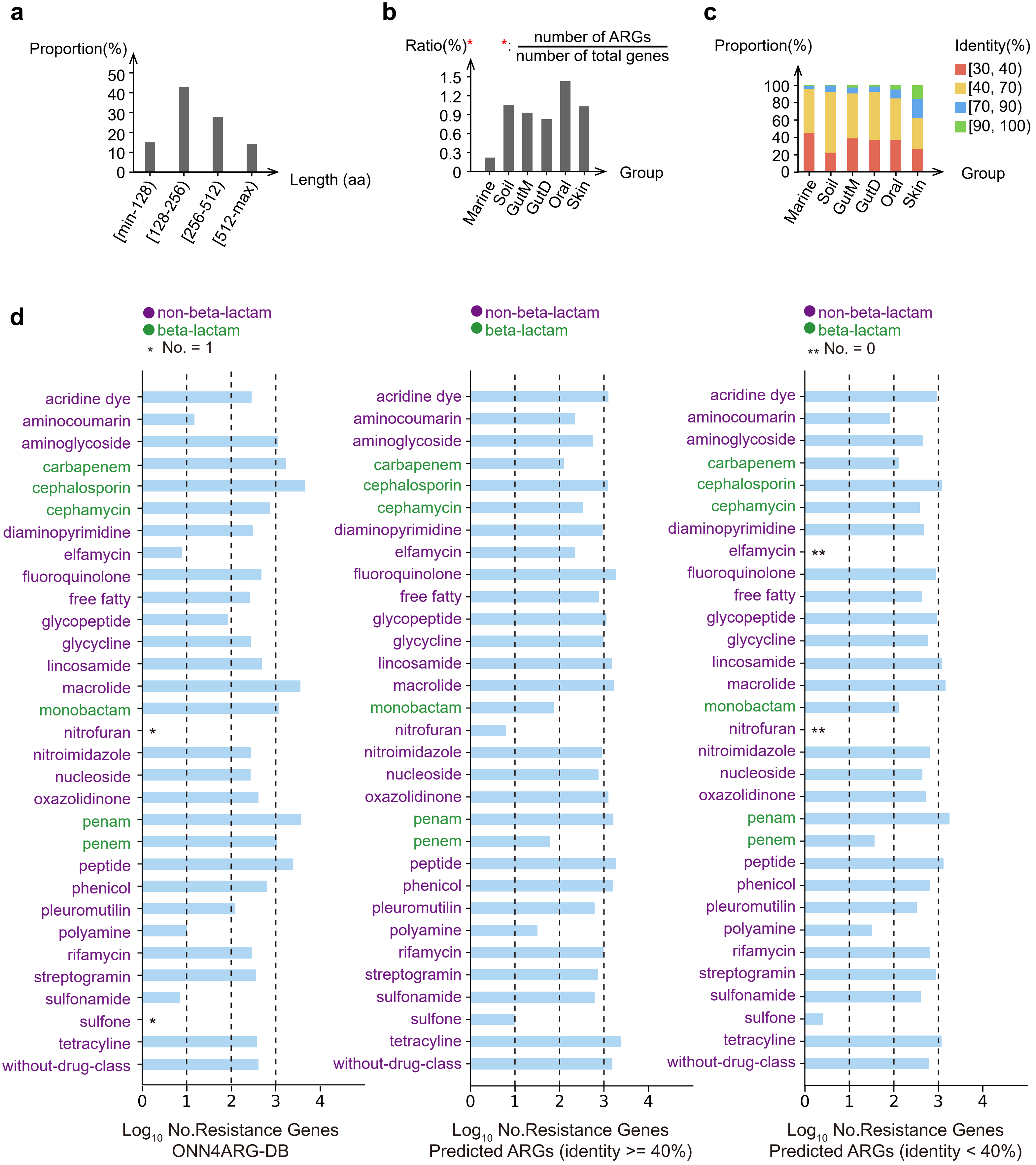


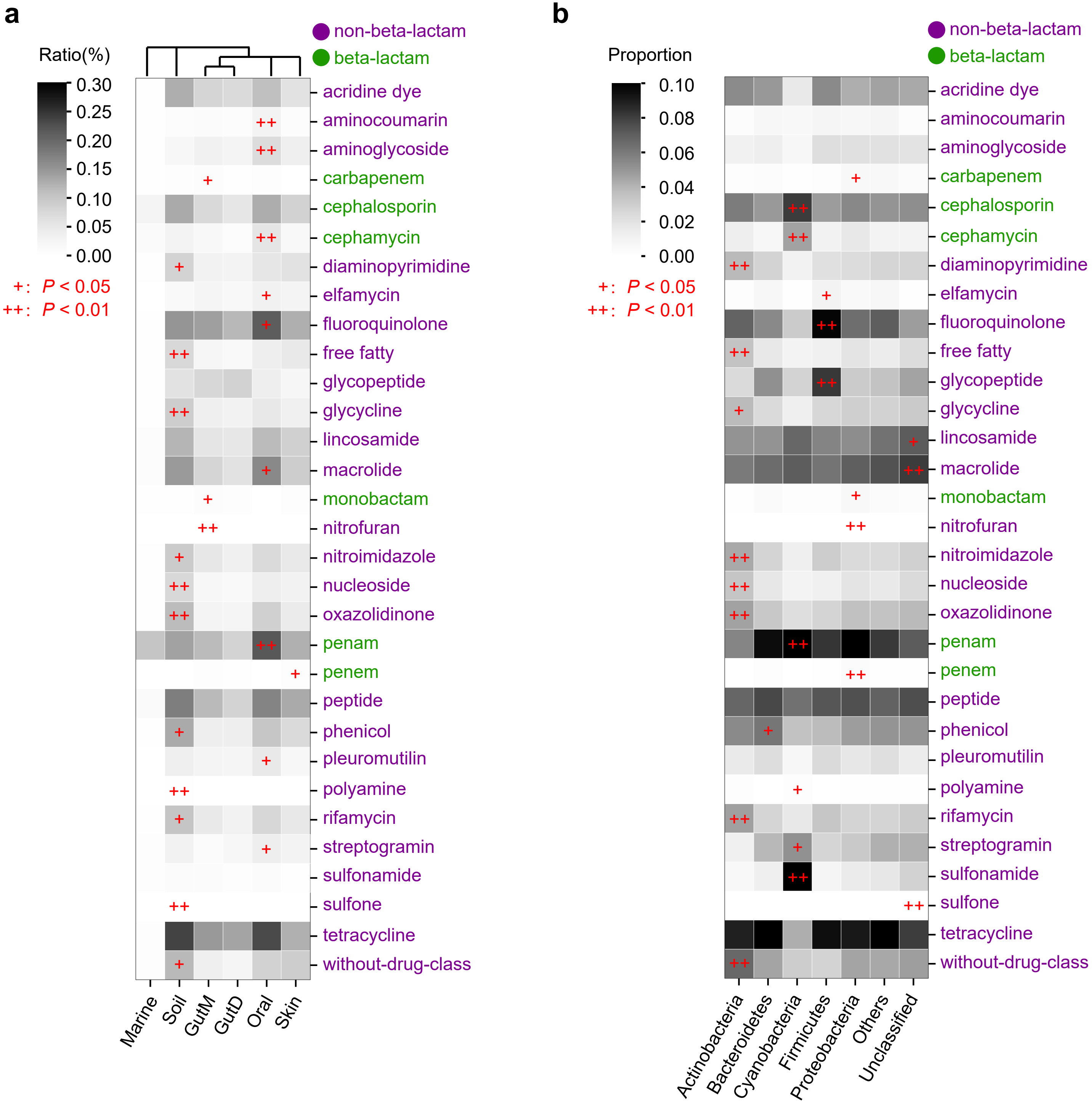
Figure 6. Phylogenetic analysis and structure investigation of Candi_60363_1. a. Phylogenetic tree of aminoglycoside resistance enzymes, Candi_60363_1, and its homologs from the NCBI non-redundant protein database. ANT: O-nucleotidyltransferase, AAC: N-acetyltransferase, APH: O-phosphotransferase, AADT: aminoglycoside adenylyltransferase. **b.** The optimal Candi_60363_1-streptomycin complex structure (left), and the local interactions

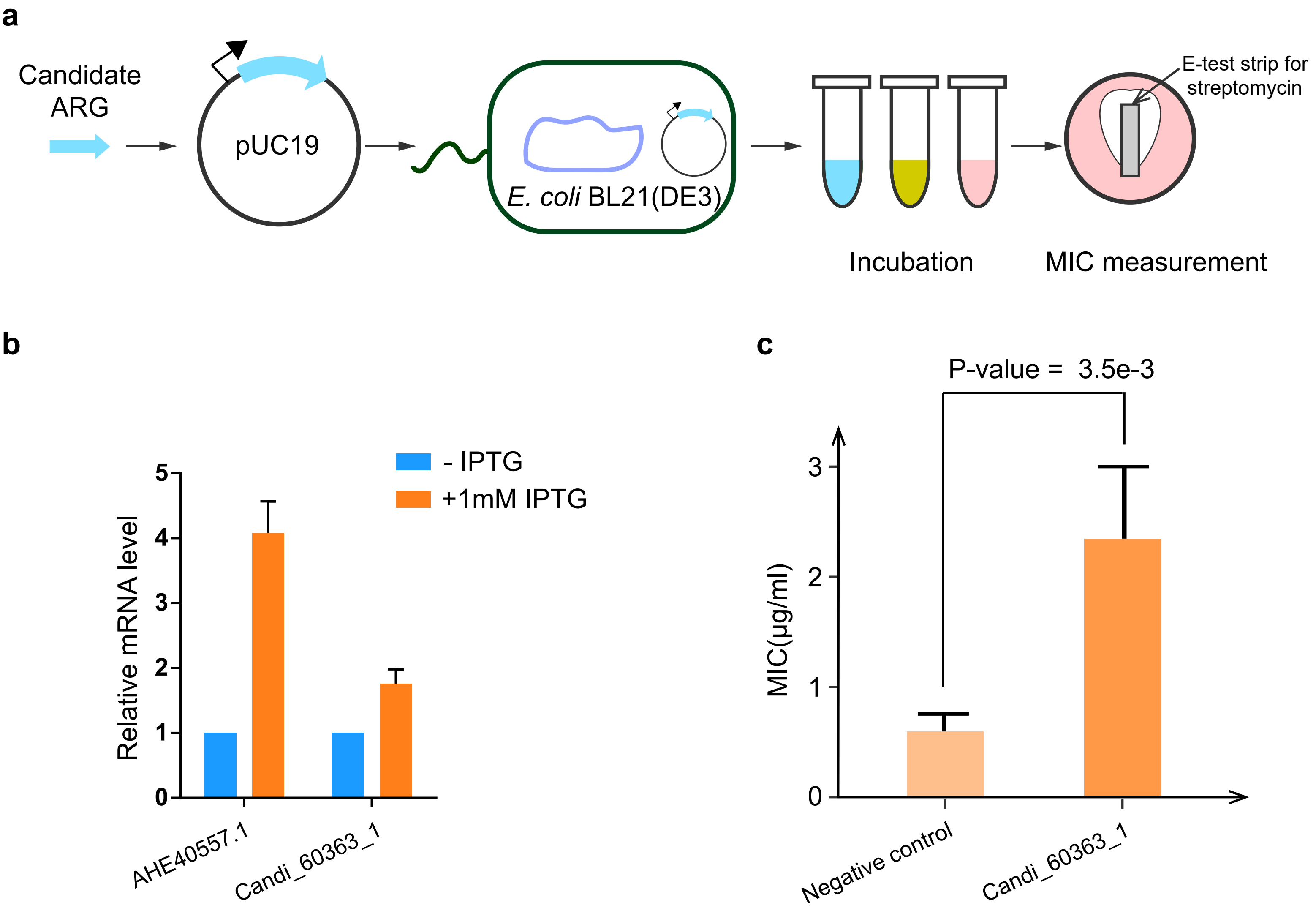
between ligand and neighboring residues (right). The docking experiment indicates there are six neighboring residues whose distances are less than three angstroms.

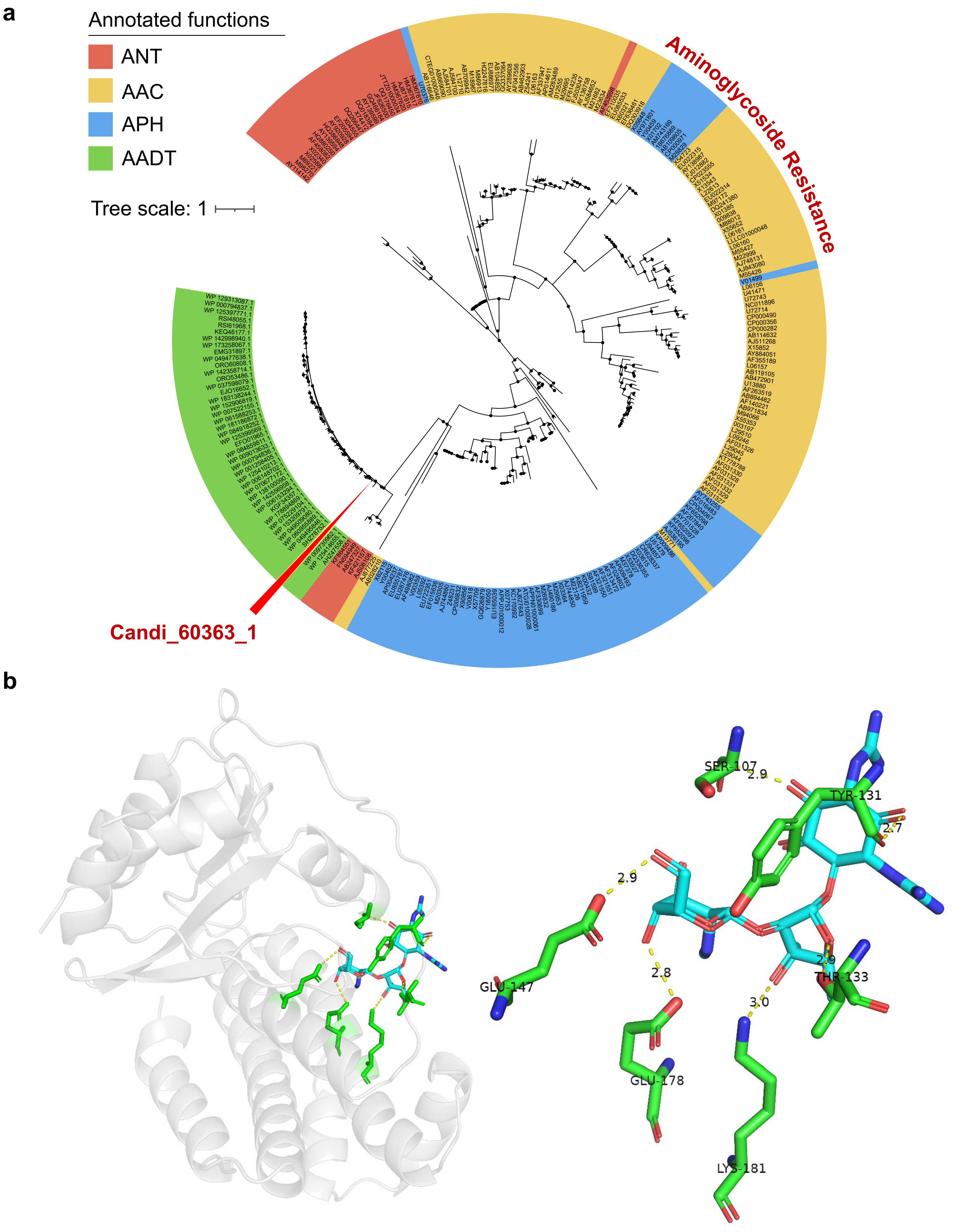

a
Sequence-alignment
N • L Y S G W
Q • L Y S G W
N K L Y S G W
N • L – S G W
ARGs database
Diamond
Query sequence
HHsearch
HMM database
Bits
VIVLILVWEGIPLSLADKT
position
Profile-alignment
Ontology-Aware Neural Network
Pseudoprediction
Prediction
Learning the ontology structure from top down
O1
O2
O3
O4
Incorporate
Incorporate
Incorporate
0
0 1
0 1
0 1 2
Multi-level Annotations
Node$_{1,1}$
P=1.00
Node$_{2,2}$
P=0.99
Node$_{3,2}$
P=0.98
Node$_{4,2}$
P=0.95
Query's annotations are:
Level 1: Node$_{1,1}$
Level 2: Node$_{2,2}$
Level 3: Node$_{3,2}$
Level 4: Node$_{4,2}$
b
*Collecting ARGs by resistance annotation
**Collecting close homologous ARGs genes (identity > 90%)
***Collecting non-ARG genes (identity < 90% and bit-score < alignment length)
Positive data
Negative data
CARD 2,587
SwissProt 2,261*
TrEMBL 23,728**
Swissprot 17,937***
ARGs with annotations
Remove redundancy
ONN4ARG-DB
Training set (75%) + Testing set (25%)
Training set
Train
Parameters update
Test
Testing set
Sequences from microbiome samples
Apply
Predict
Novel ARGs


a
b
*
*
*: $P < 0.05$
100
80
60
40
Accuracy(%)
Sequence-alignment
DeepARG
ONN4ARG
Precision
Recall
acridine dye
aminocoumarin
aminoglycoside
carbapenem
cephalosporin
cephamycin
diaminopyrimidine
elfamycin
fluoroquinolone
free fatty
glycopeptide
glycycline
lincosamide
macrolide
monobactam
nitrofuran
nitroimidazole
nucleoside
oxazolidinone
penam
penem
peptide
phenicol
pleuromutilin
polyamine
rifamycin
streptogramin
sulfonamide
sulfone
tetracyline
without-drug-class
1.0
0.8
0.6
0.4
0.2
0.0
non-beta-lactam ●
beta-lactam ●
DeepARG
ONN4ARG
DeepARG
ONN4ARG


a
Proportion(%)
50
40
30
20
10
0
[min-128)
[128-256)
[256-512)
[512-max)
Length (aa)
b
Ratio(%)*
*: number of ARGs / number of total genes
1.5
1.2
0.9
0.6
0.3
0
Marine
Soil
GutM
GutD
Oral
Skin
Group
c
Proportion(%)
100
80
60
40
20
0
Identity(%)
[30, 40)
[40, 70)
[70, 90)
[90, 100)
Marine
Soil
GutM
GutD
Oral
Skin
Group
d
non-beta-lactam
beta-lactam
* No. = 1
** No. = 0
acridine dye
aminocoumarin
aminoglycoside
carbapenem
cephalosporin
cephamycin
diaminopyrimidine
elfamycin
fluoroquinolone
free fatty
glycopeptide
glycycline
lincosamide
macrolide
monobactam
nitrofuran
nitroimidazole
nucleoside
oxazolidinone
penam
penem
peptide
phenicol
pleuromutilin
polyamine
rifamycin
streptogramin
sulfonamide
sulfone
tetracyline
without-drug-class
0
1
2
3
4
Log$_{10}$ No.Resistance Genes
ONN4ARG-DB
Predicted ARGs (identity >= 40%)
Predicted ARGs (identity < 40%)


a
Ratio(%)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
+: P < 0.05
++: P < 0.01
● non-beta-lactam
● beta-lactam
acridine dye
aminocoumarin
aminoglycoside
carbapenem
cephalosporin
cephamycin
diaminopyrimidine
elfamycin
fluoroquinolone
free fatty
glycopeptide
glycycline
lincosamide
macrolide
monobactam
nitrofuran
nitroimidazole
nucleoside
oxazolidinone
penam
penem
peptide
phenicol
pleuromutilin
polyamine
rifamycin
streptogramin
sulfonamide
sulfone
tetracycline
without-drug-class
Marine
Soil
GutM
GutD
Oral
Skin
b
Proportion
0.10
0.08
0.06
0.04
0.02
0.00
+: P < 0.05
++: P < 0.01
● non-beta-lactam
● beta-lactam
acridine dye
aminocoumarin
aminoglycoside
carbapenem
cephalosporin
cephamycin
diaminopyrimidine
elfamycin
fluoroquinolone
free fatty
glycopeptide
glycycline
lincosamide
macrolide
monobactam
nitrofuran
nitroimidazole
nucleoside
oxazolidinone
penam
penem
peptide
phenicol
pleuromutilin
polyamine
rifamycin
streptogramin
sulfonamide
sulfone
tetracycline
without-drug-class
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Proteobacteria
Others
Unclassified


a
Candidate
ARG
pUC19
E. coli BL21(DE3)
E-test strip for
streptomycin
Incubation
MIC measurement
b
- IPTG
+1mM IPTG
Relative mRNA level
0
1
2
3
4
5
AHE40557.1
Candi_60363_1
c
P-value = 3.5e-3
MIC(μg/ml)
0
1
2
3
Negative control
Candi_60363_1


a
Annotated functions
ANT
AAC
APH
AADT
Tree scale: 1
Aminoglycoside Resistance
Candi_60363_1
b
SER-107
2.9
TYR-131
2.7
2.9
GLU-147
2.8
2.9
THR-133
3.0
GLU-178
LYS-181