1 **Ontology-Aware Deep Learning Enables Novel Antibiotic Resistance**

2 **Gene Discovery Towards Comprehensive Profiling of ARGs**

3

4 Yuguo Zha¹, Cheng Chen², Qihong Jiao², Xiaomei Zeng^{1,*}, Xuefeng Cui^{2,*}, Kang
5 Ning^{1,*}

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

11 ²School of Computer Science and Technology, Shandong University, Qingdao 266237,
12 Shandong, China

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

15

16 **Abstract**

17 Antibiotic resistance genes (ARGs) have emerged in pathogens and spread faster than
18 expected, arousing a worldwide concern. Current methods are suitable mainly for the
19 discovery of close homologous ARGs and have limited utility for discovery of novel
20 ARGs, thus rendering the profiling of ARGs incomprehensive. Here, an
21 ontology-aware deep learning model, ONN4ARG (<http://onn4arg.xfcui.com/>), is
22 proposed for the discovery of novel ARGs based on multi-level annotations.
23 Experiments based on billions of candidate microbial genes collected from various
24 environments show the superiority of ONN4ARG in comprehensive ARG profiling.
25 Enrichment analyses show that ARGs are both environment-specific and host-specific.
26 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 species¹. However, the functional diversity of microbiomes has not been fully explored, and about 40% of microbial gene functions remain to be discovered². A typical example is the antibiotic resistance gene (ARG), which is an urgent and growing threat to public health³. In the past few decades, problems caused by antibiotic resistance have drawn the public's attention⁴. 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 reported⁴. 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 treatment⁷.

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

56 Presently, there are 2,979 annotated ARGs in the reference database CARD^{5,6} (v3.1.2,
57 released in April 2021), 3,159 in the ResFinder database⁸ (as of May 2021), and 2,675
58 in SwissProt⁹ (as of May 2021). These annotated ARGs are categorized into antibiotic
59 resistance types, which are organized in an ontology structure (see **Methods**,
60 **Supplementary Figure 1**), in which higher-level ARG types cover lower-level ARG
61 types. For example, AHE40557.1 is annotated in the CARD database as a
62 streptomycin resistance gene, which belongs to a lower-level ARG type
63 aminoglycoside and a higher-level ARG type non-beta-lactam. Current ARG
64 databases are far from complete: though no ARG database contains more than 4,000
65 well-annotated ARGs, NCBI non-redundant database searches yielded more than
66 7,000 putative genes annotated with “antibiotic resistance” as of May 2021. Therefore,
67 we deemed that there is a large gap between the genes annotated in ARG databases
68 and the possible ARGs that already exist in general databases, not to mention ARGs
69 that are not yet annotated.

70

71 Many ARG prediction tools have been proposed in the past few years^{8,10-20}. These
72 tools can generally be divided into two approaches. One approach is
73 sequence-alignment, such as BLAST²¹, USEARCH²², and Diamond²³, which uses
74 homologous genes to annotate unclassified genes. A confident prediction requires a
75 homolog with sequence identity greater than 80% in many programs, such as
76 ResFinder^{8,11}. The other approach is deep learning, such as DeepARG¹² and
77 HMD-ARG¹⁶, which uses neural network models to predict and annotate ARGs. The
78 input of deep learning approach can be bit-score (for DeepARG) or one-hot encoding
79 vector of protein sequence (for HMD-ARG).

80

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

86 First, existing tools are limited to a few types of ARGs due to the fact that the datasets
87 used for building models are specialized and therefore cannot reconstruct the
88 comprehensive profile of ARGs across various environments. For example,
89 HMD-ARG¹⁶ identifies only 15 types of resistance genes, and PATRIC¹³ is limited to
90 identifying ARGs encoding resistance to carbapenem, methicillin, and beta-lactam
91 antibiotics. Second, existing tools fall short in discovering novel ARGs, which usually
92 lack homology to known sequences in the reference databases. For instance, the gene
93 POCOZ1 (VraR) that confers resistance to vancomycin has a sequence identity of
94 only 24% to the homolog from the CARD¹². Recognizing such remote homologs
95 requires the ability to perceive the correlation between the internal features of genes,
96 which is challenging for existing tools. Therefore, there is an urgent need for a new
97 approach to address these limitations.

98

99 Here, we propose an ontology-aware deep learning approach, ONN4ARG, which
100 allows comprehensive identification of ARGs. ONN4ARG is an ontology-aware
101 neural network model that employs a novel ontology-aware layer and generates
102 multi-level annotations of antibiotic resistance types (**Figure 1**). Systematic
103 evaluations show that the ONN4ARG model has a profound performance
104 improvement over state-of-the-art models such as DeepARG, especially for the
105 detection of remotely homologous ARGs. The application of ONN4ARG has
106 uncovered a total of 120,726 ARGs from the microbiome, which has greatly expanded
107 the existing ARG repositories. Enrichment analyses have confirmed the enrichment
108 patterns of ARG types across multiple environments, showing that ARGs are both
109 environment-specific and host-specific. For example, resistance genes for rifamycin,
110 which is an important antibacterial agent active against gram-positive bacteria, are
111 enriched in Actinobacteria and in soil environment. Case studies have also verified
112 the ability of ONN4ARG for novel ARG discovery. For example, a recently
113 experimentally validated ARG gene GAR⁷, which is not in the CARD database, could
114 not be identified by DeepARG or HMD-ARG but was predicted by ONN4ARG. A
115 novel streptomycin resistance gene was also discovered by ONN4ARG from oral

116 microbiome data and validated through wet-lab experiments.

117

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

122

123 **Results**

124 **ONN4ARG model employs an ontology-aware neural network for ARG** 125 **identification and classification**

126 To address the large gap between the genes annotated in ARG databases and the
127 possible ARGs that already exist in general databases along with the ARGs that are
128 not yet annotated, we propose ONN4ARG, which is an ontology-aware neural
129 network model (**Figure 1a**), that could be used to predict ARGs in a comprehensive
130 manner. ONN4ARG takes similarities (e.g., identity, e-value, bit-score) between the
131 query gene sequence and ARG gene sequences and profiles (i.e., PSSM) as inputs and
132 predicts ARG annotations for the query gene. These sequence-alignment similarities
133 and profile-alignment similarities are pre-processed by calling Diamond²³ and
134 HHsearch²⁴. ONN4ARG generates multi-level annotations of antibiotic resistance
135 types, which are compatible with the antibiotic resistance ontology structure. One
136 advantage of ONN4ARG over state-of-the-art models is that ONN4ARG employs a
137 novel ontology-aware layer that incorporates ancestor and descendent annotations to
138 enhance annotation accuracies. ONN4ARG outperforms existing models, including
139 DeepARG, with higher average accuracies and better generalization ability for unseen
140 data. To train and evaluate our ONN4ARG model and for rapid deployment of ARG
141 discovery in multiple contexts, we also built an ARG database (**Figure 1b**), namely,
142 ONN4ARG-DB, which comprises ARGs from CARD and UniProt (see **Methods**).

143

144 **Systematic evaluation and comparison**

145 ONN4ARG has high efficiency, high accuracy, and comprehensiveness for ARG
146 identification based on our systematic evaluation of ONN4ARG and comparison with
147 other models. The evaluation and comparison were based on ONN4ARG-DB, with
148 28,396 positive ARGs and 17,937 negatives, out of which 75% of the dataset was
149 randomly selected for training and the remaining 25% of the dataset was selected for
150 testing (see **Methods**).

151

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

171

172 ONN4ARG demonstrates an advantage over other methods in identification of
173 remotely homologous ARGs whose sequences are not similar to existing ARG
174 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**).

183

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

187

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 Prodigal²⁷. 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.

199

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,

205 in which more than one resistance gene could be observed out of a hundred genes on
206 average (**Figure 3b, Supplementary Table 5**).

207

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

214

215 In total, 31 ARG types were detected in these various environments (**Figure 3d**,
216 **Supplementary Figure 2**). The number of predicted ARG sequences for different
217 types varied greatly (**Figure 3d**), from a few (i.e., nitrofurantoin) to thousands (i.e.,
218 fluoroquinolone). In general, fluoroquinolone and tetracycline resistance genes were
219 more abundant than other types (**Figure 3d**). As expected, these abundant ARGs were
220 usually associated with the antibiotics used extensively in human medicine or
221 veterinary medicine, including growth promotion²⁸. Novel ARG detection indicates
222 the unique ability of ONN4ARG in novel ARG discovery and ARG abundance
223 profiling in various environments, which would help researchers to better understand
224 the prevalence of antibiotic resistance genes.

225

226 **Enrichment of predicted ARGs among diverse hosts and environments**

227 Rapid deciphering of potential antimicrobial-resistant pathogens is necessary for
228 effective public health monitoring. The host-tracking of ARGs allows for accurate
229 identification of pathogens. Therefore, we conducted Kraken2²⁹ analysis to track the
230 hosts of these predicted ARGs. Results showed that there are 949 genera, each genus
231 carries at least one type of ARG (**Supplementary Table 6**). The host composition and
232 distribution of all classified ARGs for the most abundant 20 genera are displayed in
233 **Supplementary Figure 3**. The host distribution shows that these predicted ARGs are
234 primarily affiliated with Proteobacteria (38.2%), including *Candidatus Pelagibacter*,

235 *Pseudomonas*, *Bradyrhizobium*, and *Escherichia* (**Supplementary Figure 3**). The
 236 most abundant ARGs carried by the 20 genera were resistance types of
 237 fluoroquinolone, macrolide, peptide, penam, and tetracycline, accounting for about
 238 half of the total detected ARGs (**Supplementary Figure 3**). We used network
 239 inference based on strong (Pearson's correlation $\rho > 0.8$) and significant (P-value $<$
 240 0.01) correlations to investigate the co-occurrence patterns among ARG types and
 241 microbial taxa (**Supplementary Figure 4, Supplementary Note**). The co-occurrence
 242 network indicated the co-occurrence patterns between ARGs and microbial taxa. For
 243 example, ARGs that belong to beta-lactam resistance type (e.g., cephamycin, penam,
 244 penem, and monobactam) were observed to appear together in Proteobacteria.

245
 246 Enrichment analyses showed that ARGs are both environment-specific and
 247 host-specific (**Figure 4**). We found that the proportion of certain types of ARGs was
 248 significantly higher in certain environments than in others. For example, rifamycin
 249 resistance genes were found enriched in Actinobacteria (with proportion of 0.1%) and
 250 enriched in the soil environment (with proportion of 4.7%) (**Figure 4**). Rifamycin is
 251 an important antibacterial agent active against gram-positive bacteria, and it has a
 252 wide range of applications^{30,31}. The enrichment results were not surprising because
 253 *Actinomycetes* is a representative genus widely distributed in various soil
 254 environments, and its rifamycin resistance is compatible with its ability for rifamycin
 255 production³²⁻³⁵.

256 257 **Evaluation of the ability for novel ARG identification using a recently annotated** 258 **ARG**

259 We further evaluated ONN4ARG's ability for novel ARG identification based on the
 260 assessment of a newly annotated aminoglycoside resistance gene, GAR⁷. GAR is a
 261 recently reported aminoglycoside resistance gene (e.g., gentamicin, micromycin)
 262 that belongs to non-beta-lactam, which is not present in CARD (v3.0.3), UniProt (as
 263 of May 2021), DEEPARG-DB (v1.0.2), HMD-ARG-DB (as of May 2021), and
 264 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

295 $\mu\text{g/ml}$ (**Supplementary Figure 5**), which is consistent with previous reports^{36,37}. This
 296 verified that our MIC measuring experimental system works well. Our results showed
 297 that the MIC of the strain containing Candi_60363_1 was significantly higher than the
 298 negative control containing no insert (**Figure 5c, Supplementary Figure 5**), which
 299 demonstrated the increased resistance to streptomycin of the novel candidate gene
 300 Candi_60363_1 and verified the good performance of our model.

301

302 **Phylogeny and structure of Candi_60363_1**

303 There are remote similarities between Candi_60363_1 and all known ARGs in the
 304 reference database, including aminoglycoside resistance genes (closest homolog is
 305 P12055, sequence identity of 37.2%). The function annotation of P12055 shows that it
 306 has the catalytic activity of reaction between streptomycin and ATP, and it is required
 307 for streptomycin resistance (<https://www.uniprot.org/citations/3357770>). Additionally,
 308 the search result of Candi_60363_1 using InterPro (the Integrated Resource of Protein
 309 Domains and Functional Sites) shows the protein family matching to Candi_60363_1
 310 is IPR007530, which is also known as aminoglycoside 6-adenylyltransferase that
 311 confers resistance to aminoglycoside antibiotics. Then, we used BLAST to search
 312 homologs of Candi_60363_1 from the NCBI non-redundant protein database. The
 313 BLAST result showed that there are 44 homologs with sequence identity greater than
 314 80%, and they are from various organisms (**Supplementary Table 10**), such as
 315 *Streptococcus oralis*, *Peptoniphilus lacrimalis* DNF00528, and *Mycobacteroides*
 316 *abscessus subsp. Abscessus*. Considering that Candi_60363_1 is harbored by distantly
 317 related species, it obviously has mobility. Notably, the most similar protein of
 318 Candi_60363_1 from the NCBI non-redundant protein database (87.5% identity,
 319 SHZ78752.1) is also annotated as aminoglycoside adenylyltransferase
 320 (**Supplementary Table 10**). The result of BLAST search against the NCBI
 321 non-redundant protein database and other databases showed that Candi_60363_1,
 322 which is absent in all the existing ARG databases, is highly likely to be an ARG that
 323 confers resistance to aminoglycoside antibiotics.

324

Aminoglycoside modifying enzymes are the most clinically important resistance mechanism against aminoglycosides³⁸. 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

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

358

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

374

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

ecological advantages over other bacteria³².

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 micromycin). 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,

443 which belongs to a lower-level ARG type aminoglycoside and a higher-level ARG
444 type non-beta-lactam (**Supplementary Figure 1**).

445

Protein annotations

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

453

Sequence-alignment

455 We used Diamond²³ as the sequence-alignment tool for comparison. For queries in the
456 testing set, we searched them against the training set. The target with the highest
457 identity was defined as the closest homologous gene for each query. Then, we
458 compared whether the actual annotation of the query was consistent with the
459 annotation of its closest homologous gene to evaluate the accuracy.

460

DeepARG

462 DeepARG¹² is a newly developed tool that applies a neural network to identify
463 antibiotic resistance genes. For queries in the testing set, we used the DeepARG¹²
464 model to predict their annotations. Then, we compared whether the actual annotation
465 of the query was consistent with the predicted annotation to evaluate the accuracy.

466

Evaluation and comparison

468 In this study, the performance of ONN4ARG was evaluated and compared to
469 state-of-the-art models, including sequence-alignment and DeepARG. For these three
470 models, the training dataset was used to train the model parameters, and the testing
471 dataset was used to calculate the prediction accuracies. Both DeepARG and
472 ONN4ARG are deep learning models that use millions of parameters. Unlike deep

473 learning models, sequence-alignment (i.e., Diamond) has only one parameter (i.e., the
474 identity cutoff to distinguish ARG and non-ARG genes).

475

476 **Masking threshold**

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

489

490 **Benchmark method**

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

498

499 **ARG mining on metagenomic data**

500 We collected microbiome sequencing data from several published studies
501 (**Supplementary Table 4**), including samples from soil, water, and human body. The
502 gene contigs were processed by Prodigal²⁷. Protein sequences were also obtained by

the Prodigal program. Then, the ARG annotations of these protein sequences were predicted by using ONN4ARG.

Taxonomy annotation

Kraken2²⁹ 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 ResFinder⁸ (https://bitbucket.org/genomicepidemiology/resfinder_db/src/master, last update March 2021), were aligned with ClustalW. The phylogenetic tree was calculated by MEGA³⁹ (v10) using the maximum likelihood algorithm with default parameters. The Interactive Tree of Life (iTOL v6) online tool⁴⁰ was used to prepare the phylogenetic tree for display.

Protein model and docking

Rosetta⁴¹ was utilized to predict the protein structure using ab initio protein folding (<http://rosetta.bakerlab.org/>). The top five protein pockets were generated for docking calculation with Surface Topography of proteins⁴² (CASTp). We used the Cambridge Structure Database⁴³ to generate streptomycin conformers. The 3D protein-ligand complexes were obtained from AutoDock Vina⁴⁴.

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 quantitative 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 $2^{-\Delta\Delta C_t}$ 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 1.5×10^8 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

591 **References**

- 592 1 Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome
593 research. *bmc biology* **17**, doi:10.1186/S12915-019-0667-Z (2019).
- 594 2 Li, J. *et al.* An integrated catalog of reference genes in the human gut
595 microbiome. *nature biotechnology* **32**, 834-841, doi:10.1038/NBT.2942
596 (2014).
- 597 3 Brogan, D. M. & Mossialos, E. A critical analysis of the review on
598 antimicrobial resistance report and the infectious disease financing facility.
599 *globalization and health* **12**, 8-8, doi:10.1186/S12992-016-0147-Y (2016).
- 600 4 Goossens, H., Ferech, M., Stichele, R. V. & Elseviers, M. Outpatient antibiotic
601 use in Europe and association with resistance: a cross-national database study.
602 *the lancet* **365**, 579-587, doi:10.1016/S0140-6736(05)17907-0 (2005).
- 603 5 Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the
604 comprehensive antibiotic resistance database. *nucleic acids research* **45**,
605 doi:10.1093/NAR/GKW1004 (2017).
- 606 6 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the
607 comprehensive antibiotic resistance database. *nucleic acids research* **48**,
608 doi:10.1093/NAR/GKZ935 (2019).
- 609 7 Böhm, M.-E., Razavi, M., Marathe, N. P., Flach, C.-F. & Larsson, D. G. J.
610 Discovery of a novel integron-borne aminoglycoside resistance gene present
611 in clinical pathogens by screening environmental bacterial communities.
612 *microbiome* **8**, 1-11, doi:10.1186/S40168-020-00814-Z (2020).
- 613 8 Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from
614 genotypes. *journal of antimicrobial chemotherapy* **75**, 3491-3500,
615 doi:10.1093/JAC/DKAA345 (2020).
- 616 9 Bateman, A. *et al.* UniProt: A hub for protein information. *nucleic acids*
617 *research* **43**, doi:10.1093/NAR/GKU989 (2015).
- 618 10 Rowe, W. *et al.* Search Engine for Antimicrobial Resistance: A Cloud
619 Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial
620 Resistance Genes Directly from Sequence Data. *plos one* **10**,
621 doi:10.1371/JOURNAL.PONE.0133492 (2015).
- 622 11 Kleinheinz, K. A., Joensen, K. G. & Larsen, M. V. Applying the ResFinder and
623 VirulenceFinder web-services for easy identification of acquired antibiotic
624 resistance and E. coli virulence genes in bacteriophage and prophage
625 nucleotide sequences. *Bacteriophage* **4**, e27943-e27943,
626 doi:10.4161/bact.27943 (2014).
- 627 12 Arango-Argoty, G. *et al.* DeepARG: a deep learning approach for predicting
628 antibiotic resistance genes from metagenomic data. *microbiome* **6**, 23-23,
629 doi:10.1186/S40168-018-0401-Z (2018).
- 630 13 Davis, J. J. *et al.* Antimicrobial Resistance Prediction in PATRIC and RAST.
631 *scientific reports* **6**, 27930-27930, doi:10.1038/SREP27930 (2016).
- 632 14 Lakin, S. M. *et al.* Hierarchical Hidden Markov models enable accurate and
633 diverse detection of antimicrobial resistance sequences. *Commun Biol* **2**,

634 294-294, doi:10.1038/s42003-019-0545-9 (2019).

635 15 Doster, E. *et al.* MEGARes 2.0: a database for classification of antimicrobial
636 drug, biocide and metal resistance determinants in metagenomic sequence data.
637 *Nucleic acids research* **48**, D561-D569, doi:10.1093/nar/gkz1010 (2020).

638 16 Li, Y. *et al.* HMD-ARG: hierarchical multi-task deep learning for annotating
639 antibiotic resistance genes. *microbiome* **9**, 40-40,
640 doi:10.1186/S40168-021-01002-3 (2021).

641 17 Gupta, S. K. *et al.* ARG-ANNOT, a new bioinformatic tool to discover
642 antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and*
643 *chemotherapy* **58**, 212-220, doi:10.1128/AAC.01310-13 (2014).

644 18 Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene
645 Database by Using Antimicrobial Resistance Genotype-Phenotype
646 Correlations in a Collection of Isolates. *Antimicrobial agents and*
647 *chemotherapy* **63**, e00483-00419, doi:10.1128/AAC.00483-19 (2019).

648 19 Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and
649 hospital microbiology labs. *genome medicine* **6**, 90-90,
650 doi:10.1186/S13073-014-0090-6 (2014).

651 20 Rowe, W. P. M. & Winn, M. D. Indexed variation graphs for efficient and
652 accurate resistome profiling. *bioinformatics* **34**, 3601-3608,
653 doi:10.1093/BIOINFORMATICS/BTY387 (2018).

654 21 Altschul, S. F., Gish, W., Miller, W. C., Myers, E. W. & Lipman, D. J. Basic
655 Local Alignment Search Tool. *journal of molecular biology* **215**, 403-410,
656 doi:10.1016/S0022-2836(05)80360-2 (1990).

657 22 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST.
658 *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

659 23 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment
660 using DIAMOND. *nature methods* **12**, 59-60, doi:10.1038/NMETH.3176
661 (2015).

662 24 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep
663 protein annotation. *bmc bioinformatics* **20**, 1-15,
664 doi:10.1186/S12859-019-3019-7 (2019).

665 25 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome.
666 *science* **348**, 1261359-1261359, doi:10.1126/SCIENCE.1261359 (2015).

667 26 Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of
668 microbial communities, from sequence reads to assemblies. *Nucleic acids*
669 *research* **46**, D726-D735, doi:10.1093/nar/gkx967 (2018).

670 27 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation
671 initiation site identification. *bmc bioinformatics* **11**, 119-119,
672 doi:10.1186/1471-2105-11-119 (2010).

673 28 Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and
674 co-occurrence of environmental antibiotic resistance genes. *ISME J* **9**,
675 2490-2502, doi:10.1038/ismej.2015.59 (2015).

676 29 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with
677 Kraken 2. *genome biology* **20**, 1-13, doi:10.1186/S13059-019-1891-0 (2019).

678 30 Qi, F. *et al.* Deciphering the late steps of rifamycin biosynthesis. *Nature*
679 *Communications* **9**, 2342, doi:10.1038/s41467-018-04772-x (2018).

680 31 Floss, H. G. & Yu, T.-W. Rifamycin Mode of Action, Resistance, and
681 Biosynthesis. *Chemical Reviews* **105**, 621-632, doi:10.1021/cr030112j (2005).

682 32 Yao, Y., Zhang, W., Jiao, R., Zhao, G. & Jiang, W. Efficient isolation of total
683 RNA from antibiotic-producing bacterium *Amiclatopsis mediterranei*.
684 *Journal of Microbiological Methods* **51**, 191-195,
685 doi:doi.org/10.1016/S0167-7012(02)00078-7 (2002).

686 33 Wilson, M. C., Gulder, T. A. M., Mahmud, T. & Moore, B. S. Shared
687 biosynthesis of the saliniketals and rifamycins in *Salinispora arenicola* is
688 controlled by the sare1259-encoded cytochrome P450. *J Am Chem Soc* **132**,
689 12757-12765, doi:10.1021/ja105891a (2010).

690 34 Saxena, A., Kumari, R., Mukherjee, U., Singh, P. & Lal, R. Draft Genome
691 Sequence of the Rifamycin Producer *Amiclatopsis rifamycinica* DSM 46095.
692 *Genome Announc* **2**, e00662-00614, doi:10.1128/genomeA.00662-14 (2014).

693 35 Huang, H. *et al.* *Micromonospora rifamycinica* sp. nov., a novel actinomycete
694 from mangrove sediment. **58**, 17-20, doi:doi.org/10.1099/ij.s.0.64484-0
695 (2008).

696 36 Pinto-Alphandary, H., Mabilat, C. & Courvalin, P. Emergence of
697 aminoglycoside resistance genes aadA and aadE in the genus *Campylobacter*.
698 *antimicrobial agents and chemotherapy* **34**, 1294-1296,
699 doi:10.1128/AAC.34.6.1294 (1990).

700 37 Holden, M. T. G. *et al.* Rapid evolution of virulence and drug resistance in the
701 emerging zoonotic pathogen *Streptococcus suis*. *plos one* **4**,
702 doi:10.1371/JOURNAL.PONE.0006072 (2009).

703 38 Ramirez, M. S., Nikolaidis, N. & Tolmasky, M. Rise and dissemination of
704 aminoglycoside resistance: the aac(6')-Ib paradigm. *frontiers in microbiology*
705 **4**, 121-121, doi:10.3389/FMICB.2013.00121 (2013).

706 39 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular
707 Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*
708 **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

709 40 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and
710 new developments. *Nucleic Acids Research* **47**, W256-W259,
711 doi:10.1093/nar/gkz239 (2019).

712 41 Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure
713 prediction using Rosetta. *methods in enzymology* **383**, 66-93,
714 doi:10.1016/S0076-6879(04)83004-0 (2004).

715 42 Tian, W., Chen, C. & Liang, J. CASTp 3.0: Computed Atlas of Surface
716 Topography of Proteins and Beyond. *biophysical journal* **114**,
717 doi:10.1016/J.BPJ.2017.11.325 (2018).

718 43 Cole, J. C., Korb, O., McCabe, P., Read, M. G. & Taylor, R. Knowledge-Based
719 Conformer Generation Using the Cambridge Structural Database. *journal of*
720 *chemical information and modeling* **58**, 615-629,
721 doi:10.1021/ACS.JCIM.7B00697 (2018).

722 44 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of
723 docking with a new scoring function, efficient optimization, and
724 multithreading. *J Comput Chem* **31**, 455-461, doi:10.1002/jcc.21334 (2010).
725

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.

737 **Figure 2**

738
739 **Figure 2. Systematic evaluation and comparison between sequence-alignment, DeepARG,**
740 **and ONN4ARG. a.** The accuracy of three models on ARG classification was assessed using a box
741 plot. Diamond was used for sequence-alignment; significance test was based on the t -test. **b.** The
742 precision and recall of DeepARG and ONN4ARG on ARG classification for each antibiotic
743 resistance type. The masking threshold of testing set equaled 0.4 (details of masking threshold are
744 provided in **Methods**).
745

746 **Figure 3**

747 **Figure 3. Broad-spectrum profile of predicted ARGs among diverse environments.** **a.** The
748 proportion of predicted ARGs for different protein sequence lengths. **b.** The abundance ratio of
749 predicted ARGs among diverse environments. Abundance ratio was defined as the number of
750 ARGs divided by the number of total genes. **c.** The proportion of predicted ARGs for different
751 sequence identities among diverse environments. **d.** Number of genes in ONN4ARG-DB (left),
752 predicted homologous ARGs (middle), and predicted novel ARGs (right) for various resistance
753 types. The horizontal axis indicates the logarithmic number of genes, and the vertical axis
754 indicates different antibiotic resistance types. We collected metagenomic samples from several
755

756 published studies; these samples were mainly from “marine,” “soil,” and “human” environments.
757 Human-associated samples consisted of two gut groups (one group from Madagascar, i.e., GutM;
758 the other group from Denmark, i.e., GutD), one oral group, and one skin group (both oral and skin
759 groups were from the HMP project).
760

761 **Figure 4**

762

763 **Figure 4. Enrichment of predicted ARGs among diverse environments and hosts. a.** Relative

764 abundance and enrichment of ARGs among diverse environments. Abundance ratio was defined as

765 the number of ARGs divided by the number of total genes. **b.** Proportion and enrichment of ARGs

766 among diverse hosts. Colors indicate the proportion of ARGs for each phylum and resistance type.

767 Results for the most abundant five phyla that carry ARGs are shown. “+”: P-value < 0.05 (*t*-test);

768 “++”: P-value < 0.01 (*t*-test).

769

770 **Figure 5**

771

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

778

779 **Figure 6**

780
781 **Figure 6. Phylogenetic analysis and structure investigation of Candi_60363_1. a.**
782 Phylogenetic tree of aminoglycoside resistance enzymes, Candi_60363_1, and its homologs from
783 the NCBI non-redundant protein database. ANT: O-nucleotidyltransferase, AAC:
784 N-acetyltransferase, APH: O-phosphotransferase, AADT: aminoglycoside adenyltransferase. **b.**
785 The optimal Candi_60363_1-streptomycin complex structure (left), and the local interactions

786 between ligand and neighboring residues (right). The docking experiment indicates there are six
787 neighboring residues whose distances are less than three angstroms.
788

a **b**

a **b**

a **b** **c**

a

Annotated functions

- ANT
- AAC
- APH
- AADT

Tree scale: 1

b