Title:

ARGs-OSP: online searching platform for antibiotic resistance genes distribution in metagenomic database and bacterial whole genome database

An Ni Zhang[1], Chen-Ju Hou[1], Li-Guan Li[1], Tong Zhang[1,2]*

[1]Environmental Biotechnology Laboratory, The University of Hong Kong, Hong Kong;

[2]International Center for Antibiotics and Resistance in Environments, Southern University of Science and Technology, China.

* To whom correspondence should be addressed. Tel: +852-2857 8551; Fax: +852-2859 8987; Email: zhangt@hku.hk and/or zhangt3@sustc.edu.cn

Abstract

Background

The antibiotic resistant genes (ARGs) have been emerging as one of the top global issue s in both medical and environmental fields. The metagenomic analysis has been widely adopted in ARG-related studies, revealing a universal presence of ARGs in diverse environments from medical settings to natural habitats, even in drinking water and ancient permafrost. With the tremendous resources of accessible metagenomic datasets, it would be feasible and beneficial to construct a global profile of antibiotic resistome as a guidance of its phylogenetic and ecological distribution. And such information should be shared by an open webpage to avoid the unnecessary repeat of data processing and the bias caused by incompatible search method.

Results

Two dataset collections, the Whole Genome Database (WGD, 54,718 complete and draft bacterial genomes) and the Metagenomic Database (MGD, 854 metagenomic datasets of 7 eco-types), were downloaded and analyzed using a standard method of ARG online analysis platform (ARGs-OAP v1.0). The representativeness of WGD and MGD was evaluated to have a comprehensive coverage of ARGs in bacterial genomes and metagenomes. Besides, an ARGs online searching platform (ARGs-OSP, http://args-osp.herokuapp.com/) was developed in this study to make the data

accessible to other researchers via the search and download functionality. Finally, flexible usage of the ARGs-OAP was demonstrated by evaluating the co-occurrence of class 1 integrases and total ARGs across different environments.

Conclusions

The ARGs-OSP is presented in this study as the valuable sources and references for future studies with versatile research interests, meanwhile avoiding unnecessary re-computations and re-analysis.

Keywords

Antibiotic resistant genes; whole genome analysis; metagenomic analysis; global antimicrobial resistant profile; class 1 integrases.

Background

Due to the intensive usage of the human and veterinary antibiotics, the antibiotic resistant genes (ARGs) are emerging in almost all environments as one of the top global concerns. Recently, through high-throughput sequencing and metagenomic analysis, the ARG profiles have been investigated in diverse habitats, especially the anthropogenic environments of human microbiome [1-6], animal microbiome [7-10] and WWTPs [11-13]. ARGs, especially clinically relevant ones [14-18], have been spreading from the anthropogenic habitats to the natural eco-systems [19], mainly contributed by WWTPs, pharmaceutical manufacturing plants, hospitals, and husbandry facilities [13, 20-23]. Universal presence of various ARGs has been revealed in all kinds of natural ecosystems by many metagenomic studies investigating the samples of sediment [23-25], soil [26-28], surface water [29-31], marine water [27] and even ancient permafrost [32-34]. The identification of ARGs in drinking water [35] and human food [36, 37] further reveals the potential of their direct exposure to human health. With the growing attention to the environmental issue of antibiotic resistance dissemination all over the world, ARG-related studies have gained momentum, and are covering all kinds of habitats. All these metagenomic

datasets are valuable and accessible sources to construct a global profile of antibiotic resistome.

Despite the recent increase in aforementioned metagenomic datasets, the approaches of the identification and annotation of ARGs varied in different researches regarding the searching methods, searching criteria, the reference databases and the units in the quantification. This makes direct inter-sample comparison infeasible. For example, the ARG profiles were evaluated to have significant difference of up to 5-20 fold [38] when using the domain-based searching method of Hidden Markov Model (HMM) [39] against the similarity-based searching approaches of BLAST [40], USEARCH [41] and DIAMOND [42]. Another major obstacle for the parallel comparison was the use of different reference databases. Even for the top two highly-cited ARG databases, the Comprehensive Antibiotic Resistance Database (CARD; http://arpcard.mcmaster.ca) [43] and Antibiotic Resistant Database (ARDB) [44], bias could be raised during comparison because they only share a small portion of reference sequences [45]. Recently, a widely-applied and well-curated ARG database named the Structured ARG reference database (SARG; http://smile.hku.hk/SARGs) [38, 45], was constructed by integrating both the CARD and ARDB, which was selected as a standard database in this study. The metagenomic datasets from diverse habitats should

be re-analyzed using a standardized pipeline for ARG identification and annotation, from which the ecological distribution of ARGs could be comprehensively drawn.

Although metagenomic analysis could uncover the distribution and abundance of ARGs in a habitat, the information it can provide is limited. Further information describing the host, the mobility and the gene arrangement of ARGs is critical and necessary to investigate the frontier scientific questions about the origin, the evolution, the spreading and the co-selection of ARGs in different environments. This problem can be partially alleviated through the use of assembled metagenomes and nanopore sequencing techniques , but such studies are still limited to only a few environmental samples [35]. On the contrary, such details can be provided with precision and certainty through analyzing the collection of bacterial whole genomes. It was demonstrated by some previous researches [46, 47] via mining the co-occurrence patterns of ARGs and metal resistant genes (MGEs) in the collection of bacterial complete genomes. To construct a global profile of ARGs, the integration of whole genomes and metagenomes is a promising attempt.

Many public databases publish the information in a well-organized way for the convenient search and download by the user ends, such as the IMG/VR

(https://img.jgi.doe.gov/vr/) [48]. However, to the best of our knowledge, such application was not introduced to the field of ARG research, especially to provide information on the phylogenetic and ecological distribution of antimicrobial resistance. Moreover, the processing of the big datasets of whole genomes and metagenomes is time- and resource- consuming, and is unnecessary repeated by individual researchers all over the world. In this study, a global profile of ARGs is constructed by a standard pipeline and is presented in the form of an online searching platform for ARGs (ARGs-OSP, http://args-osp.herokuapp.com/), serving as a valuable resource for future studies.

Data and Webpage Description

To construct a global ARG profile covering the information of their phylogenetic and ecological distribution, the ARGs were identified and quantified by searching two collections of bacterial whole genomes and metagenomes using a standard pipeline, the ARGs online analysis platform (ARGs-OAP v1.0) [45]. The occurrence and abundance of ARGs were summarized and organized with the metadata information into mothertables, which were published on the ARGs-OSP (http://args-osp.herokuapp.com/). On ARGs-OSP, search and download functionality

was designed for users to retrieve the occurrence of ARGs in different taxonomy and the abundance of ARGs in different habitats. The availability and convenience of this platform could meet the requirements of versatile research interests, such as: the current host range of some specific resistant genes on the bacterial phylogenetic tree, the dissemination of some specific resistant genes in both the natural and anthropogenic habitats, the antibiotic resistome currently detected in some specific taxa or specific environments, and the comparison of the ARG profiles of a local sample to the global profile. Through data sharing, ARGs-OSP is expected to motivate and facilitate future studies into mining new information and knowledge from the combined data, without making repeated efforts in dataset processing.

Methods

Two collection of datasets

The Whole Genome Database (WGD) containing 54,718 bacterial genomes (7,770 complete genomes and 46,948 draft genomes with medium and high quality of more than 50% completeness [49]) was downloaded from the NCBI genome database [50] (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria,

ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt) (on July 8, 2017), as summarized in Table S1. The potential pathogenicity of bacterial genomes was obtained by matching their taxonomic information with a published database covering the taxonomy of currently recognized human bacterial pathogens [51] (Table S1). If either one of the taxonomic annotations of genus, species or strain level was matched to the human pathogen list, the genome was labeled as a potential human pathogen [10, 52].

The Metagenome Database (MGD) totaling 854 metagenomic datasets were downloaded from NCBI SRA database [50] and MG-RAST[53], which were all generated through Illumina shotgun sequencing. The habitat information of the metadata of all samples was manually organized to categorize them into totally 25 eco-subtypes of 7 eco-types by integrating the guidance of previous studies [54, 55], covering both the natural environments (water, sediment, soil, and permafrost) and the anthropogenic environments (WWTPs, animal feces, and human feces). Thus, this collection of MGD is expected to represent a wide and comprehensive ecological diversity. The quality control of raw reads was conducted with Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) with the minimum quality score of Q20 within at least 90% of bases, resulting in in the total number of clean reads varied

from 0.1 million to 91 million. All raw reads were trimmed equally to the length of 100bp to allow more accurate inter-sample comparison.

Identification and annotation of the ARGs and intI1 gene

All the coding sequences (CDS) of WGD were extracted from the genbank files to be searched against SARG [45] and the class 1 integrases (*intI1*) database (manuscript under review). Those CDS meeting the criteria of e-value 1e-5, 90% amino acid (aa) identity over 80% aa hit-ratio against the SARG, and e-value 1e-3, 80% aa identity over 50% aa hit-ratio against the *intI1* database, were annotated as the ARGs and *intI1* genes, respectively. The ARGs and *intI1* genes in MGD were also investigated by these two databases. An effective and time-saving searching process was conducted by two-step sequence-based methods, first through usearch v8.0.1623_i86linux64 [41] and followed by BLASTX 2.2.28+ [40]. The abundance of the ARGs and *intI1* genes was calculated and transformed to different units of ppm, copy per 16S and copy per cell. Furthermore, the abundance of each unit was specifically calculated under different combination of searching criteria (e-value, identity and hit-length). ARGs-OSP provides 60 combinations of searching criteria for each abundance unit, that is, totally 180 mothertables for more flexible usage by future studies. The cutoff

used in previous studies [7, 45] (manuscript under review) for metagenomes was e-value 1e-7, 80% aa identity over 75% aa hit-length of the SARG and the *intI1* database, which was adopted in this study as a standard cutoff for further analysis. Also, the standard abundance unit for the ARGs and *intI1* genes in MGD was set as copy per cell, which was comparable to copy per genome in WGD.

Mothertable analysis and visualization

The mothertables were organized by combining the ARG profile with the phylogenetic and ecological information, via self-written Python 2.7.6 and R scripts using R 3.3.2[56] (packages'dplyr','ggalt','ggthemes', 'ggplot2', and 'plyr') and Python 2.7 (https://www.python.org/). The rarefaction curves[57] of WGD (based on each genome) and MGD (based on each raw read) were conducted by randomly subsampling the genomes or raw reads without replacement [58, 59], that is, each genome or raw read was sampled only once. The step of the raw read number for the entire MGD and for each eco-type was set at the total number of raw reads (for each dataset) divided by 10,000, which ensured that each rarefaction curve was plotted with 10,000 points. All the networks were visualized with Cytoscape 3.3.0 [60] in *Tree* or *Hierarchic* layout or with R 3.3.2[56] (package 'ggplot2').

Results and Discussions

The representativeness and coverage of WGD and MGD

In WGD, 54,718 bacterial genomes were downloaded from the NCBI genome database in total [50], covering 32 bacterial phyla, 162 classes, 299 orders, 643 families, 1,986 genera, and 3,654 species, without counting the unclassified taxonomy (Fig 1a and Table S1). Approximately 88.9% of the bacterial genomes were derived from phyla of *Proteobacteria*, *Firmicutes*, and *Actinobacteria*, indicating that the WGD might be biased by the over sequencing of some specific taxa, especially those taxa of medical importance. To avoid such bias, individual genomes of the same species were merged together, resulting in a curated percentage of 23.6% bacterial species obtained from these three dominant phyla. Still, some genera displayed high prevalence of pathogenic species, such as *Klebsiella* (57.1%), *Enterobacter* (55.6%), and *Escherichia* (25.0%), compared to the average 8.6% of pathogenic species within *Bacteria*.

The MGD collected in this study covered a wild range and diversity of both the anthropogenic (WWTPs, animal feces and human feces) and natural (water, sediment, soil, and permafrost) habitats (Fig 1b and Table S2). The classification of the ecosystems was manually conducted based on a two-tier hierarchical classification system[61]. For a more comprehensive analysis and specific detailed comparison, each habitat (eco-type) was further classified into 3-5 eco-subtypes, except for the natural permafrost habitat category. The datasets in MGD are both geographically and ecologically distinct, collecting across different countries and continents, together to draw a global map.

Rarefaction curves [57] are adopted to evaluate the representativeness and coverage of these two datasets about the ARGs in the bacterial life tree and in the environments. The rarefaction curve for WGD was plotted by the consecutive addition of one random genome extracted from the collection of 54,718 bacterial genomes. The number of unique ARGs (not detected in those genomes previously added into the pool) provided by the new genome was counted. After the inclusion of the last genome, the rarefaction curve for WGD gradually reached a plateau of 2,625 unique ARGs (Fig S1a). Additional inclusion of new bacterial genomes in the future was expected to contribute a mild increase of 1 novel ARG per 170 genomes (minimum

number of new genomes), indicating the representativeness of the current collection of bacterial whole genomes.

Similarly, for MGD, all the raw reads from 854 samples were pooled together to construct a rarefaction curve. The raw reads were randomly selected from the entire MGD pool one by one, to be evaluated against the SARG (as described in the section of methods). The total number of unique ARGs was added by one if the reference ARG assigned to this raw read was not identified previously. It seemed that after sampling 1.6E+10 raw reads from the entire MGD pool, the rarefaction curve illustrated a trend of a flat slope, and finally reached a plateau of 3,821 unique ARGs with the inclusion of all 2.4E+10 raw reads. It was predicted that one novel ARG could be expected with an extra sample of at least 1.4E+8 raw reads, also demonstrating the representativeness of the current collection of metagenomic datasets. Besides, the coverage of the MGD to the ARGs harbored by individual habitats was specifically evaluated by drawing the rarefaction curves for each eco-type. Generally speaking, the rarefaction curves for the anthropogenic habitats of animal feces, human feces and WWTPs tend to have reached their plateaus of 2,951, 2,815, and 3,131 unique ARGs (Fig S2), respectively. In contrast, the rarefaction curves were still growing gently for the natural environments of water (2,514 ARGs),

permafrost (1,897 ARGs), soil (1,618 ARGs), and sediment (1,181 ARGs) after fully sampling all the raw reads. However, when calculating the new raw reads required to obtain one novel ARG, it was suggested that the current MGD had a higher representativeness to the habitats of WWTP (8.7E+7), human feces (2.4E+7), natural sediment (2.1E+7) than the habitats of animal feces (8.2E+6), natural water (4.2E+6), natural permafrost (3.8E+6), and natural soil (1.2E+6). The disagreement between two observations could be caused by the different patterns of the richness and density of the antibiotic resistome within 7 habitats. For example, the first inclusion of 1.0E+8 raw reads contributed to an average increase of 1,182 unique ARGs in the anthropogenic environments, which almost double the results (601 unique ARGs) of the natural environments. Another example was that at the level of 7.9E+8 raw reads (the lowest sampling depth), the anthropogenic environments recovered 2,250 unique ARGs while only 1,421 unique ARGs were detected in the natural environments. These two observations indicated the higher richness and higher density of antibiotic resistance in those anthropogenic habitats.

Overall, both current versions of WGD and MGD were evaluated to have high representativeness of the antibiotic resistance in the bacterial life tree and in the

environments. However, additional sampling is expected to enrich the ARG profiles in individual natural habitats, especially the categories of water, permafrost, and soil.

ARGs-OSP (antibiotic resistant genes-online searching platform)

The ARGs-OSP provides the search and download functionality of the occurrence and abundance of the ARGs and *intI1* genes retrieved by a global investigation in this study. The mothertables were separated into two Modules for the WGD and MGD, where the information of the hosts and the habitats can be specifically offered. In order to facilitate the diverse requirements of future studies, the identification and quantification of the target genes can be constrained by the customized cutoff selected by the users, including the identity, hit-length or hit-ratio, and e-value. Without any input, the ARGs-OSP returns the results of all target genes detected in all whole genomes or metagenomes, applying the default search cutoff described in the section of the methods. In each Module, three classes of inquiry factors are listed at the top middle panel, including the ARGs (*"Sequence"*, *"Subtype"* and *"Type"*), the host information (*"Genome"*, *"Accession"*, *"Organism"*, *"Assembly_level"*, *"Phylum"*, *"Class"*, *"Order"*, *"Family"*, *"Genus"*, *"Species"*, *"Strain"*, and *"Pathogen"*) for Module 1,

the habitat information (*"Accession"*, *"Eco-subtype"* and *"Eco-type"*) for Module 2, and the searching criteria (*"Identity"*, *"Hit-length"* or *"Hit-ratio"* and *"E-value"*).

In Module 1 (Fig 2), the ARGs can be viewed and searched in all the whole genomes under a given cutoff, and all the details about their host taxa were summarized based on the annotation in GenBank and NCBI genome database. This functionality would be extremely beneficial for those research interests into the occurrence of ARGs in some specific bacterial lineages or the phylogenetic distribution of some specific ARGs. To meet versatile demands at the user end, the host information supports the details of the taxonomy information (from phylum level to strain level), the accession number (*"Genome"* for genbank assembly accession and *"Accession"* genbank sequence accession number), the status of the genomes (*"Assembly_level"* referring to the genome completeness), and the potential pathogenicity (*"Pathogen"*), which can all be set to filter the search results simultaneously. The searching criteria in Module 1 were transformed into aa-based *"Identity"* and *"Hit-ratio"* against the reference SARG, and the *"Hit-ratio"* referred to the percentage of hit-length to the reference length. Without imputing any searching criteria, ARGs-OSP will output all the inquiries that meet the default cutoff (≥ 90% aa identity, ≥ 80% aa hit-ratio and ≤ e-value 1e-5). The allowance of searching criteria is defined in the range of 50%-100% aa identity,

50%-100% aa hit-ratio and e-value smaller than 1.0E-1. For more comprehensive downstream analysis by the user end, the complete output table can be easily downloaded as a local file.

Module 2 (Fig 3) was constructed by investigating the ARGs in the collection of metagenomes under different combinations of cutoff, combined with the habitat information of each metagenomic dataset. This functionality would be extremely helpful for users inquiring the abundance of some specific ARGs in all the environments or the habitats of their interests, or for users to compare the ARG profile of local samples to a global collection. For reliable parallel comparison, users are recommended to process the ARG identification and annotation using the ARGs-OAP v1.0 [38, 45], which is compatible to the ARGs-OSP. In details, the habitat information covered the *"Accession"* (run accession number of NCBI SRA or MG-RAST databases), *"Eco-subtype"* and *"Eco-type"*, which was classified and curated manually. The searching criteria in Module 2 were also based on the aa *"Identity"* and *"Hit-length"* against the reference databases. The default criteria was set as ≥ 80% aa identity, ≥ 75% aa hit-ratio and ≤ e-value 1e-7, for empty cutoff input. The cutoff range allowed for searching is defined for the aa identity (60%, 70%, 80%, 90%, and 100%), aa hit-length (50%, 75%, and 100%) and e-value (1.0E-6, 1.0E-7, 1.0E-8,

1.0E-9). However, compared to the identity, the hit-length and e-value were evaluated to have little influence on the ARG profile [45]. Since all the metagenomic datasets were trimmed into a standard read length of 100bp, the aa hit-length of 50%, 75% and 100% was expressed in the form of 17aa, 25aa and 33aa. As mentioned before, three abundance units (copy per cell, copy per 16S and ppm)[45] are provided for the flexible comparison at the user end, which could be easily switched by the buttons at the top left panel of output mothertable. A bottom panel is designed to customize the table layout and turn the pages.

A global profile of antibiotic resistome

In this study, a global profile of antibiotic resistome was constructed and presented by integrating the phylogenetic and ecological distribution of ARGs from both WGD and WGD. In WGD, 2,625 ARGs (764 genotypes and 22 phenotypes) were detected in 809 bacterial species from 13 phyla. Within all the ARG-carrying species, 26.8% were identified as pathogenic species, which were almost 3 times the prevalence of pathogenic species in the WGD. This observation indicated that ARGs could be selected by the ecological fittings provided with strong selection force of anthropogenic pollution [62-66], where human pathogens are universally present.

Among all the phenotypes, the multidrug (the occurrence of 29.4%), beta-lactam (15.5%), aminoglycoside (10.5%) and tetracycline (8.7%) were universally detected in bacterial species, which was consistent with a previous study investigating 2,500 complete genomes [47]. The ARG genotypes of *tetA*, *tetM*, *acrB*, *aph(3')-I*, *aadA*, *mdtK*, *TolC*, and class A beta-lactamase resistant to tetracycline, aminoglycoside, multidrug and beta-lactam were frequently identified in 1.5% of bacterial species, covering a wide spectrum of taxonomy lineage and possessing more than 50% prevalence in human pathogens. These ARGs have successfully invaded across the phylogenetic barrier of the bacterial phylum level, especially into human microbiome, which should raise substantial alarm in both the medical and environmental fields.

3,821 ARGs from 993 subtypes/genotypes and 24 types/phenotypes were identified in diverse environments of the MGD in total. Even though the anthropogenic environments were illustrated to have ARGs of higher density and richness than the natural environments (Fig S2), after normalizing against the bacterial cell number in each sample, the abundance of total ARGs showed no significant difference (less than 10 folds) among the 7 eco-types (Fig S3). This observation suggested that those divergences of the density and richness of antibiotic resistant profiles could be caused

by the density and richness of the microbial community among different eco-types, while the resistant level within the bacterial cells seemed to be quite stable.

Generally speaking, ARGs are widely distributed in almost all natural and anthropogenic habitats with the average abundance varying from 5.1E-2 copy per cell to 6.7E-1 copy per cell. It was not surprising since antibiotic resistance is originally harbored by natural microorganisms and most antibiotics used nowadays are produced by natural antibiotic producers [19]. The animal feces harbored the highest abundance of ARGs (6.7E-1 copy per cell) under the strong selective pressure of over-dosing antibiotics [67-69], which may promote the dissemination of ARGs [70]. The other two anthropogenic environments, the human feces and WWTPs, displayed similar level of ARGs of 3.7E-1 and 1.7E-1 copy per cell. It was notable that high abundance of ARGs was hosted by the microbial communities in the natural environments of permafrost (6.1E-1 copy per cell) and soil (2.6E-1 copy per cell), even denser than the WWTPs as the hotspot of ARGs [13, 22]. Within the soil eco-type, the eco-subtype of rural soil could be contaminated by the agricultural usage of animal manure, the other three soil eco-subtypes of city, amazon catchment and prairie displayed equivalent level of antibiotic resistant. A relative low level of ARGs was detected in the natural water (1.0E-1 copy per cell) and sediment (6.7E-2 copy per cell), while two obvious

trends of heterogeneous and homogeneous distribution were respectively observed in the water and soil eco-types.

To make the results comparable to previous studies, the abundance of ARGs was transformed into the units of copy per 16S and ppm, and provided on the ARGs-OSP. Most habitats exhibited similar resistance level compared to previous studies [7, 54, 71] except for a much higher detection of ARGs in soil (2.6E-1 copy per 16S) and surface water (1.0E-1 copy per 16S) compared with the abundances of 2.0E-2 to 9.0E-3 copy per 16S and 2.0E-2 to 8.0E-3 copy per 16S, respectively. This difference could be contributed by the heterogeneity within the soil environments and divergence within the surface water environments, and a large sample size with multiple eco-subtypes in this study would help construct a more comprehensive and representative global profile.

Besides the abundance of total ARGs, the antibiotic resistome of different habitats also displayed their divergence regarding to their composition and diversity. Instead of directly counting the unique reference sequences detected in one eco-type, ARGs conducting the same resistant mechanisms were classified into one genotype. The sediment environment harbored a deficient collection of 281 ARG genotypes,

accounting for less than 30% of the global profile. Approximately 50% of the global resistant profile was recovered in the soil (404 genotypes), permafrost (513 genotypes) and human feces (604 genotypes). The animal feces, water and WWTP environments provided a rich pool of antibiotic resistance, covering 714, 722 and 761 genotypes, respectively. Those genotypes that have long been the focus of ARG-related researches were found to be widespread and abundant in all 7 eco-types, including aminoglycoside resistance gene *aph(3)-I*, the beta-lactam resistant gene *TEM*, sulfonamide resistance genes of *sul1* and *sul2*, MLS resistant genes of *macA*, *macB*, tetracycline resistant genes of *tetA* and *tetM*, multidrug resistant genes of *acrA* and *acrB*, and vancomycin resistant gene *vanA* [14, 54, 72]. Overall, the anthropogenic pollution appeared to have a weak influence on the total antibiotic resistant level within the bacterial cell, but may cast a relatively strong impact on the diversity and density of the antibiotic resistome within the habitat.

Co-occurrence of the ARGs and intI1 genes: a demonstration of ARGs-OSP

The *intI1* genes were considered as a potential indicator for anthropogenic pollution [73] because of its high abundance and universal occurrence in human-related environments, such as wastewater treatment plants and animal feces [74, 75].

Previous studies on the correlation between *intI1* genes and human pollution mainly focused on three directions: 1) *intI1* genes tended to have high abundance in human-related environments, and cannot be effectively removed from WWTPs [76-78]; 2) the co-selection of *intI1* genes with ARGs, metal resistant genes (MRGs) and disinfectant resistant genes [79-81]; 3) the abundance of *intI1* genes increases with anthropogenic pollution, such as heavy metal, disinfectants, antibiotics, and pesticides [82-84]. Besides, some ARGs (*sul1*, *sul2* and *tetM*) were evaluated to have strong and positive correlation with *intI1* genes in polluted sediment samples [85]. However, the co-occurrence of the *intI1* genes and ARGs was not comprehensively evaluated before. Moreover, this co-occurrence could be casually caused by their co-selection by the same selective pressure or by physically linked on the same transposons and plasmids, and thus resulted in quite inconsistent relationship.

To evaluate whether the *intI1* genes could be an indicator for the anthropogenic pollutant of total ARGs and in which habitats there could be a strong correlation, the mothertables of ARGs and *intI1* genes were downloaded from ARGs-OSP. The abundances of the total ARGs and total *intI1* genes were summed up for each sample, and the samples were pooled into 7 eco-types. The second question was raised in this study that if the *intI1* genes were a weak proxy for total ARGs, which subgroup of

ARGs could be indicated by the *intI1* genes? One subgroup of 107 ARGs found on class 1 integrons in WGD (manuscript under review) was proposed as a potential target.

Overall, the correlation of *intI1* genes to two groups of ARGs (total ARGs and ARGs on class 1 integrons) in 7 habitats showed that higher abundance of ARGs corresponding to higher abundance of *intI1* genes (Table 1). However, it was interesting to find that high abundance of total ARGs was detected in many samples, while no *intI1* gene was identified (light blue nodes in Fig 4). This indicated that *intI1* genes were not a universal marker for the presence of ARGs, not even for the anthropogenic environments where *intI1* gene was absent in a large portion of animal fecal and human fecal samples. This observation was also supported by the co-occurrence of *intI1* genes and ARGs in the WGD, where *intI1* genes were highly conserved in the class of *Gammaproteobacteria* (red nodes in Fig 5). Nevertheless, the ARGs were discovered to be widely distributed across different classes (light blue nodes in Fig 5), and the antibiotic resistance carried by those bacterial species outside the spectrum of *Gammaproteobacteria* were not likely to be indicated by the *intI1* genes.

Besides an overall picturing of their correlation, the linear regressions were fitted to the *intI1* genes and total ARGs in 7 eco-types, which displayed week linear relation with the R^2 varied from 0.03-0.47 (Table S1). Here, to avoid the biased caused by sequencing depth, samples with no presence of *intI1* gene were not considered during assessment. The low R^2 of WWTP (0.03), soil (0.06) and permafrost (0.08), indicated that the *intI1* genes had a poor linear relationship to the total ARGs in these habitats and their role as an indicator should be treated with caution. Moreover, 7 habitats were found to comply with different linear relations, regarding to their slopes of 0.53 to 2.10. The anthropogenic environments were expected to have higher slope of linear relations because of higher level of anthropogenic pollution, while this trend was not clear in this study. Thus, the correlation between *intI1* and the total ARGs was both weak (in terms of R^2) and inconsistent (in terms of slopes) for different environments.

The *intI1* genes were fit with better linear relationships (in terms of R^2) to the subgroup of ARGs on class 1 integrons in all habitats except for the sediment (Table 1 and Fig S4). The *intI1* genes could be proposed as a good indicator for ARGs on class 1 integrons, especially for animal feces (R^2 of 0.75). Even though, this relationship may only be applied to samples in soil, water, human feces and animal feces ($R^2 \geq$ 0.35), and the linear relationship should also be specifically tuned for each habitat.

Conclusions

In this study, a global profile of the antibiotic resistome was constructed by integrating two big datasets of the WGD (54,718 bacterial genomes) and MGD (854 metagenomes of 7 habitats). Both the WGD and MGD were evaluated to have good representativeness and comprehensive coverage of ARGs in bacterial genomes and metagenomes, serving as the fundamental bases to investigate the phylogenetic and ecological distribution of antibiotic resistance. Moreover, all ARGs were identified and quantified using a standardized pipeline for reliable parallel comparison. Most importantly, a user-friendly and well-organized online searching platform, the ARGs-OSP, was designed to publish all the mothertables obtained in this study, making the data easily accessible for other researchers. The ARGs-OSP can serves as valuable sources and references for future studies with versatile research interests, while avoiding unnecessary re-computations. Finally, the potential of the ARGs-OSP was demonstrated by evaluating whether the *intI1* genes could be a good proxy for the anthropogenic pollution of ARGs by investigating their co-occurrence.

The major limitation of this study lies in the two datasets of WGD and MGD. The WGD could be biased by over-sequencing those bacterial species of research interest and medical importance, which may not represent all the environmental microbes. For the MGD, the current sampling scheme mainly focused on the representative eco-types, and the extreme environments were not included in this version. Besides, the availability of qualified samples was limited for some important environments, such as air. However, with the rapid development and decreasing cost of sequencing techniques, datasets of high quality and massive quantity is promising for the expansion of WGD and MGD. Further updates of ARGs-OSP is expected to be enhanced responding to the continuously growing number of bacterial genomes and environmental metagenomes in public database. Also, more flexible search and visualization functionality will be continuously complemented in future versions of ARGs-OSP, for more convenient and versatile usage.

Declarations

Availability of supporting data and materials

The datasets supporting the results of this article are available on the ARGs-OSP (XXXX)

List of abbreviations

ARGs: antibiotic resistant genes; ARGs-OAP: ARGs online analysis pipeline; ARGs-OSP: ARGs online searching platform; *intI1*: class 1 integrases; WGD: whole genome database; MGD: metagenome database.

Competing Interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank Hong Kong General Research Fund (XXX) for financial support. Miss A. N. Zhang acknowledges University of Hong Kong for postgraduate studentship. Mr. Bryan Hou would like to thank University of Hong Kong for research assistant fellowship.

Author's Contributions

A. N. Zhang downloaded the datasets, analyzed the data, designed the platform and wrote the manuscript. Chen-Ju Hou constructed the platform. L. L. Guan provided suggestions in the data analysis and manuscript preparation. T. Zhang guided webpage development and revised this manuscript.

References

1. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nature communications. 2013;4:2151.
2. Sommer MO, Church GM and Dantas G. The human microbiome harbors a diverse reservoir of antibiotic resistance genes. Virulence. 2010;1 4:299-303.
3. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proceedings of the National Academy of Sciences. 2011;108 Supplement 1:4578-85.
4. Diaz-Torres ML, Villedieu A, Hunt N, McNab R, Spratt DA, Allan E, et al. Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. FEMS microbiology letters. 2006;258 2:257-62.
5. Seville LA, Patterson AJ, Scott KP, Mullany P, Quail MA, Parkhill J, et al. Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA. Microbial Drug Resistance. 2009;15 3:159-66.
6. Devirgiliis C, Barile S and Perozzi G. Antibiotic resistance determinants in the interplay between food and gut microbiota. Genes & nutrition. 2011;6 3:275.
7. Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, et al. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. The ISME journal. 2015;9 11:2490-502.
8. Chambers L, Yang Y, Littier H, Ray P, Zhang T, Pruden A, et al. Metagenomic analysis of antibiotic resistance genes in dairy cow feces following therapeutic administration of third generation cephalosporin. PLoS One. 2015;10 8:e0133764.
9. Wichmann F, Udikovic-Kolic N, Andrew S and Handelsman J. Diverse antibiotic resistance genes in dairy cow manure. MBio. 2014;5 2:e01017-13.
10. Ma L, Xia Y, Li B, Yang Y, Li L-G, Tiedje JM, et al. Metagenomic assembly reveals hosts of antibiotic resistance genes and the shared resistome in pig, chicken, and human feces. Environmental science & technology. 2015;50 1:420-7.
11. Czekalski N, Berthold T, Caucci S, Egli A and Bürgmann H. Increased levels of multiresistant bacteria and resistance genes after wastewater treatment and their dissemination into Lake Geneva, Switzerland. Frontiers in Microbiology. 2012;3:106.
12. Bouki C, Venieri D and Diamadopoulos E. Detection and fate of antibiotic

resistant bacteria in wastewater treatment plants: a review. Ecotoxicology and environmental safety. 2013;91:1-9.

13. Rizzo L, Manaia C, Merlin C, Schwartz T, Dagot C, Ploy M, et al. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review. Science of the total environment. 2013;447:345-60.
14. Martinez JL, Coque TM and Baquero F. What is a resistance gene? Ranking risk in resistomes. Nat Rev Microbiol. 2015;13 2:116-23. doi:10.1038/nrmicro3399.
15. Levy SB. Microbial resistance to antibiotics. An evolving and persistent problem. Lancet. 1982;2:83-8.
16. Turnbull F, Wallace A, Stewart S and Crofton J. Streptomycin resistance after treatment with PAS alone. British Medical Journal. 1953;1 4822:1244.
17. Tadesse DA, Zhao S, Tong E, Ayers S, Singh A, Bartholomew MJ, et al. Antimicrobial drug resistance in Escherichia coli from humans and food animals, United States, 1950–2002. Emerging infectious diseases. 2012;18 5:741.
18. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, et al. Characterization of a new metallo-β-lactamase gene, blaNDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrobial agents and chemotherapy. 2009;53 12:5046-54.
19. Alonso A, Sanchez P and Martinez JL. Environmental selection of antibiotic resistance genes. Environmental microbiology. 2001;3 1:1-9.
20. Gothwal R and Shashidhar T. Antibiotic Pollution in the Environment: A Review. CLEAN - Soil, Air, Water. 2015;43 4:479-89. doi:10.1002/clen.201300989.
21. Li WC. Occurrence, sources, and fate of pharmaceuticals in aquatic environment and soil. Environ Pollut. 2014;187:193-201. doi:10.1016/j.envpol.2014.01.015.
22. Michael I, Rizzo L, McArdell C, Manaia C, Merlin C, Schwartz T, et al. Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: a review. Water research. 2013;47 3:957-95.
23. Czekalski N, Díez EG and Bürgmann H. Wastewater as a point source of antibiotic-resistance genes in the sediment of a freshwater lake. The ISME journal. 2014;8 7:1381.
24. Gonzalez-Plaza JJ, Šimatović A, Milaković M, Bielen A, Wichmann F and Udikovic-Kolic N. Functional repertoire of antibiotic resistance genes in

antibiotic manufacturing effluents and receiving freshwater sediments. Frontiers in microbiology. 2017;8:2675.

25. Chu BT, Petrovich ML, Chaudhary A, Wright D, Murphy B, Wells G, et al. Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. Applied and environmental microbiology. 2018;84 5:e02168-17.
26. Su JQ, Wei B, Xu CY, Qiao M and Zhu YG. Functional metagenomic characterization of antibiotic resistance genes in agricultural soils from China. Environment international. 2014;65:9-15.
27. Nesme J, Cécillon S, Delmont TO, Monier J-M, Vogel TM and Simonet P. Large-scale metagenomic-based study of antibiotic resistance in the environment. Current biology. 2014;24 10:1096-100.
28. Lau CH-F, van Engelen K, Gordon S, Renaud J and Topp E. Novel antibiotic resistance determinants from agricultural soil exposed to antibiotics widely used in human medicine and animal farming. Applied and Environmental Microbiology. 2017:AEM. 00989-17.
29. Rodriguez-Mozaz S, Chamorro S, Marti E, Huerta B, Gros M, Sànchez-Melsió A, et al. Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. Water research. 2015;69:234-42.
30. Amos G, Zhang L, Hawkey P, Gaze W and Wellington E. Functional metagenomic analysis reveals rivers are a reservoir for diverse antibiotic resistance genes. Veterinary microbiology. 2014;171 3-4:441-7.
31. Tang J, Bu Y, Zhang X-X, Huang K, He X, Ye L, et al. Metagenomic analysis of bacterial community composition and antibiotic resistance genes in a wastewater treatment plant and its receiving surface water. Ecotoxicology and environmental safety. 2016;132:260-9.
32. Perron GG, Whyte L, Turnbaugh PJ, Goordial J, Hanage WP, Dantas G, et al. Functional characterization of bacteria isolated from ancient arctic soil exposes diverse resistance mechanisms to modern antibiotics. PLoS One. 2015;10 3:e0069533.
33. Rascovan N, Telke A, Raoult D, Rolain JM and Desnues C. Exploring divergent antibiotic resistance genes in ancient metagenomes and discovery of a novel beta-lactamase family. Environmental microbiology reports. 2016;8 5:886-95.
34. D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, et al. Antibiotic resistance is ancient. Nature. 2011;477 7365:457.
35. Ma L, Li B, Jiang X-T, Wang Y-L, Xia Y, Li A-D, et al. Catalogue of antibiotic resistome and host-tracking in drinking water deciphered by a large scale

survey. Microbiome. 2017;5 1:154.

36. Devirgiliis C, Zinno P, Stirpe M, Barile S and Perozzi G. Functional screening of antibiotic resistance genes from a representative metagenomic library of food fermenting microbiota. BioMed research international. 2014;2014.
37. Bengtsson-Palme J. Antibiotic resistance in the food supply chain: where can sequencing and metagenomics aid risk assessment? Current Opinion in Food Science. 2017;14:66-71.
38. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2. 0 with an Expanded SARG Database and Hidden Markov Models for Enhancement Characterization and Quantification of Antibiotic Resistance Genes in Environmental Metagenomes. Bioinformatics. 2018;1:8.
39. Eddy SR. Profile hidden Markov models. Bioinformatics (Oxford, England). 1998;14 9:755-63.
40. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. doi:10.1186/1471-2105-10-421.
41. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26 19:2460-1.
42. Buchfink B, Xie C and Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12 1:59.
43. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic acids research. 2016:gkw1004.
44. Liu B and Pop M. ARDB—antibiotic resistance genes database. Nucleic acids research. 2008;37 suppl_1:D443-D7.
45. Yang Y, Jiang X, Chai B, Ma L, Li B, Zhang A, et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics. 2016; doi:10.1093/bioinformatics/btw136.
46. Li L-G, Xia Y and Zhang T. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. The ISME journal. 2017;11 3:651-62.
47. Pal C, Bengtsson-Palme J, Kristiansson E and Larsson DJ. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. BMC genomics. 2015;16 1:964.
48. Paez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic acids research. 2016:gkw1030.

49. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017; doi:10.1038/s41564-017-0012-7.
50. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the national center for biotechnology information. Nucleic acids research. 2012;40 D1:D13-D25.
51. Woolhouse ME and Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. Emerging infectious diseases. 2005;11 12:1842.
52. Li L-G, Xia Y and Zhang T. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. The ISME journal. 2016.
53. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic acids research. 2015;44 D1:D590-D4.
54. Pal C, Bengtsson-Palme J, Kristiansson E and Larsson DG. The structure and diversity of human, animal and environmental resistomes. Microbiome. 2016;4 1:54. doi:10.1186/s40168-016-0199-5.
55. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551 7681.
56. Team RC. R: A language and environment for statistical computing. 2013.
57. Gotelli NJ and Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology letters. 2001;4 4:379-91.
58. Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology. 2012;5 1:3-21. doi:10.1093/jpe/rtr044.
59. Deng C, Daley T and Smith AD. Applications of species accumulation curves in large-scale biological data analysis. Quant Biol. 2015;3 3:135-44. doi:10.1007/s40484-015-0049-7.
60. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13 11:2498-504.
61. Li L-G, Yin X and Zhang T. Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. Microbiome. 2018;6 1:93.
62. Baquero F, Martínez J-L and Cantón R. Antibiotics and antibiotic resistance in

water environments. Current opinion in biotechnology. 2008;19 3:260-5.

63. Martinez JL. Environmental pollution by antibiotics and by antibiotic resistance determinants. Environmental pollution. 2009;157 11:2893-902.
64. Wright GD. Antibiotic resistance in the environment: a link to the clinic? Current opinion in microbiology. 2010;13 5:589-94.
65. Vaz-Moreira I, Nunes OC and Manaia CM. Bacterial diversity and antibiotic resistance in water habitats: searching the links with the human microbiome. FEMS microbiology reviews. 2014;38 4:761-78.
66. Ji X, Shen Q, Liu F, Ma J, Xu G, Wang Y, et al. Antibiotic resistance gene abundances associated with antibiotics and heavy metals in animal manures and agricultural soils adjacent to feedlots in Shanghai; China. Journal of hazardous materials. 2012;235:178-85.
67. Li C, Chen J, Wang J, Ma Z, Han P, Luan Y, et al. Occurrence of antibiotics in soils and manures from greenhouse vegetable production bases of Beijing, China and an associated risk assessment. Science of the total environment. 2015;521:101-7.
68. Aust M-O, Godlinski F, Travis GR, Hao X, McAllister TA, Leinweber P, et al. Distribution of sulfamethazine, chlortetracycline and tylosin in manure and soil of Canadian feedlots after subtherapeutic use in cattle. Environmental Pollution. 2008;156 3:1243-51.
69. Hu X, Zhou Q and Luo Y. Occurrence and source analysis of typical veterinary antibiotics in manure, soil, vegetables and groundwater from organic vegetable bases, northern China. Environmental Pollution. 2010;158 9:2992-8.
70. Beaber JW, Hochhut B and Waldor MK. SOS response promotes horizontal dissemination of antibiotic resistance genes. Nature. 2004;427 6969:72.
71. Bengtsson-Palme J, Angelin M, Huss M, Kjellqvist S, Kristiansson E, Palmgren H, et al. The human gut microbiome as a transporter of antibiotic resistance genes between continents. Antimicrobial agents and chemotherapy. 2015;59 10:6551-60.
72. Berendonk TU, Manaia CM, Merlin C, Fatta-Kassinos D, Cytryn E, Walsh F, et al. Tackling antibiotic resistance: the environmental framework. Nat Rev Microbiol. 2015;13 5:310-7. doi:10.1038/nrmicro3439.
73. Gillings MR, Gaze WH, Pruden A, Smalla K, Tiedje JM and Zhu YG. Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. The ISME journal. 2015;9 6:1269-79. doi:10.1038/ismej.2014.226.
74. Ma L, Li A-D, Yin X-L and Zhang T. The Prevalence of Integrons as the Carrier of Antibiotic Resistance Genes in Natural and Man-Made Environments.

Environmental Science & Technology. 2017;51 10:5721-8.

75. Pruden A, Arabi M and Storteboom HN. Correlation between upstream human activities and riverine antibiotic resistance genes. Environmental science & technology. 2012;46 21:11541-9.
76. Stalder T, Barraud O, Jové T, Casellas M, Gaschet M, Dagot C, et al. Quantitative and qualitative impact of hospital effluent on dissemination of the integron pool. The ISME journal. 2014;8 4:768.
77. Ma L, Zhang X-X, Zhao F, Wu B, Cheng S and Yang L. Sewage treatment plant serves as a hot-spot reservoir of integrons and gene cassettes. Journal of environmental biology. 2013;34 2 suppl:391.
78. Du J, Ren H, Geng J, Zhang Y, Xu K and Ding L. Occurrence and abundance of tetracycline, sulfonamide resistance genes, and class 1 integron in five wastewater treatment plants. Environmental Science and Pollution Research. 2014;21 12:7276-84.
79. Khan GA, Berglund B, Khan KM, Lindgren P-E and Fick J. Occurrence and abundance of antibiotics and resistance genes in rivers, canal and near drug formulation facilities–a study in Pakistan. PLoS One. 2013;8 6:e62712.
80. Seiler C and Berendonk TU. Heavy metal driven co-selection of antibiotic resistance in soil and water bodies impacted by agriculture and aquaculture. Frontiers in microbiology. 2012;3:399.
81. Graham DW, Olivares-Rieumont S, Knapp CW, Lima L, Werner D and Bowen E. Antibiotic resistance gene abundances associated with waste discharges to the Almendares River near Havana, Cuba. Environmental science & technology. 2010;45 2:418-24.
82. Chen B, Yang Y, Liang X, Yu K, Zhang T and Li X. Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. Environmental science & technology. 2013;47 22:12753-60.
83. Lehmann K, Bell T, Bowes MJ, Amos GC, Gaze WH, Wellington EM, et al. Trace levels of sewage effluent are sufficient to increase class 1 integron prevalence in freshwater biofilms without changing the core community. Water research. 2016;106:163-70.
84. Yang Y, Xu C, Cao X, Lin H and Wang J. Antibiotic resistance genes in surface water of eutrophic urban lakes are related to heavy metals, antibiotics, lake morphology and anthropic impact. Ecotoxicology. 2017;26 6:831-40.
85. Lv B, Cui Y, Tian W, Li J, Xie B and Yin F. Abundances and profiles of antibiotic resistance genes as well as co-occurrences with human bacterial pathogens in ship ballast tank sediments from a shipyard in Jiangsu Province, China.

Ecotoxicology and environmental safety. 2018;157:169-75.

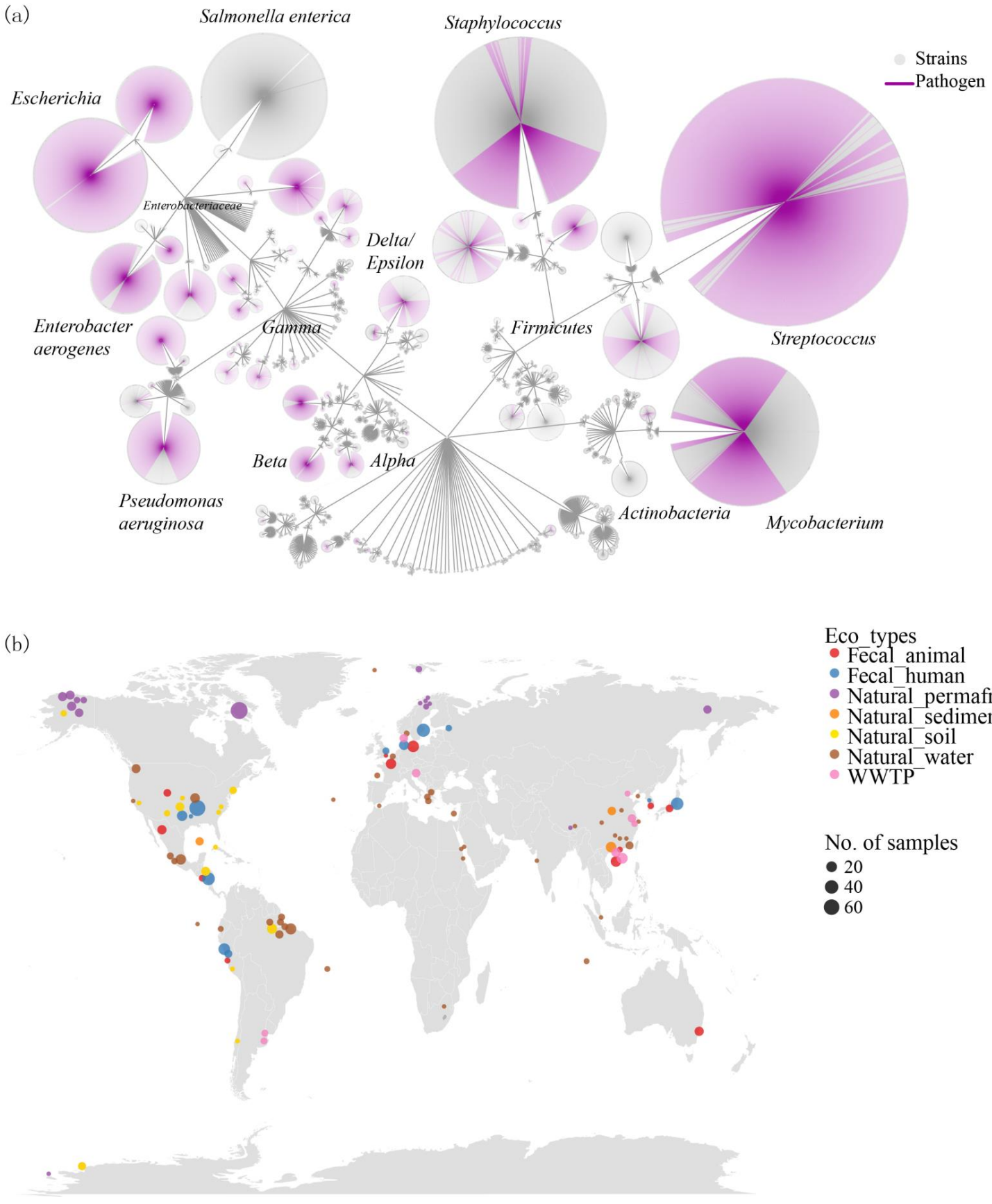


Fig.1. Overview of the Whole Genome Database (WGD) and Metagenome Database (MGD). (a) The phylogenetic relationship of all 54,718 bacterial genomes of 45 phyla in WGD and the occurrence of ARGs (blue nodes) and Rank I ARGs (red nodes). The pathogenic strains were indicated by purple edges. (b) The global map of metagenomic datasets in MGD. The size of the datasets (nodes) was proportional to the number of samples and the color was differentiated by the eco-type.

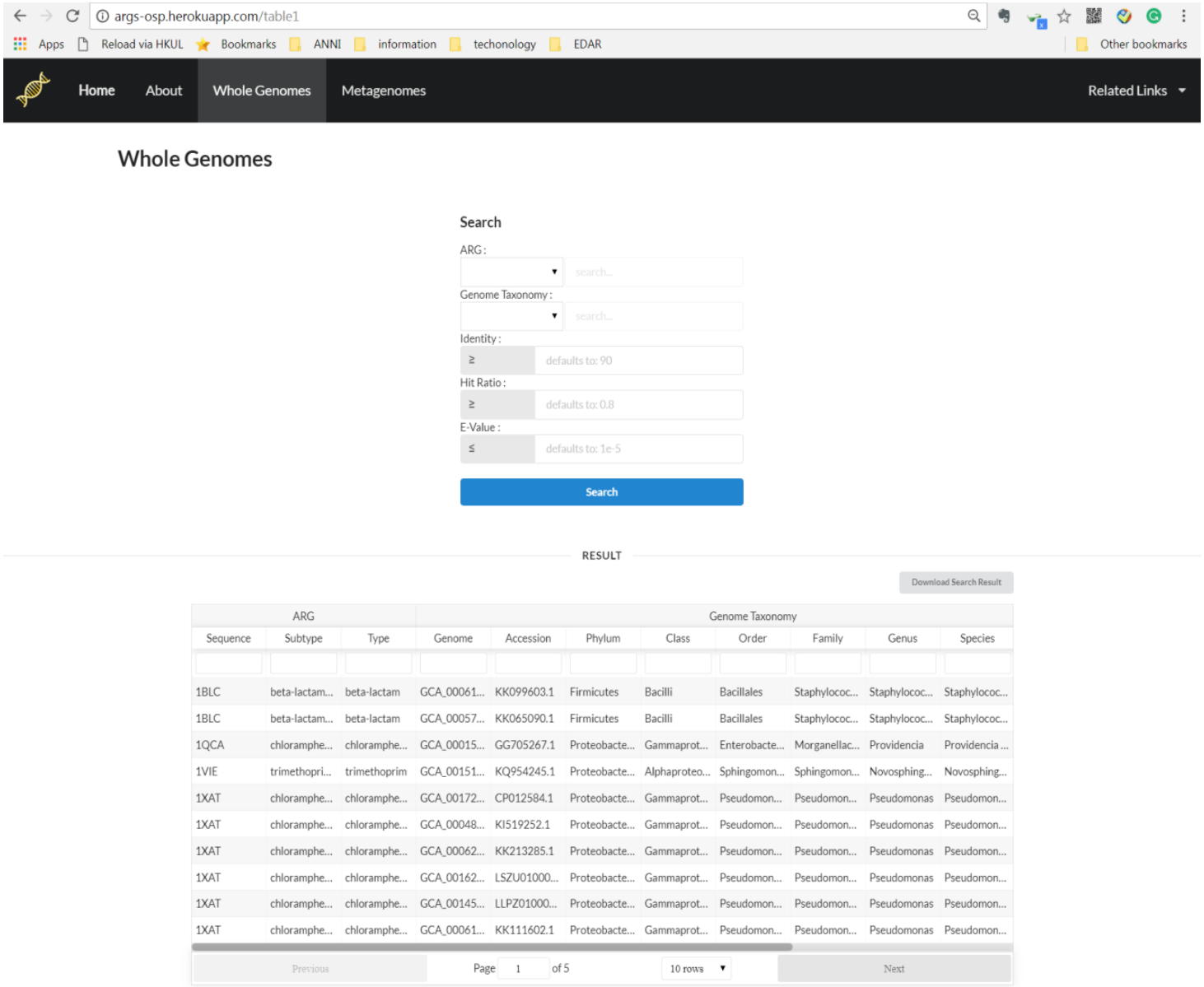


Fig 2. The layout of the ARGs-OSP Whole Genomes.

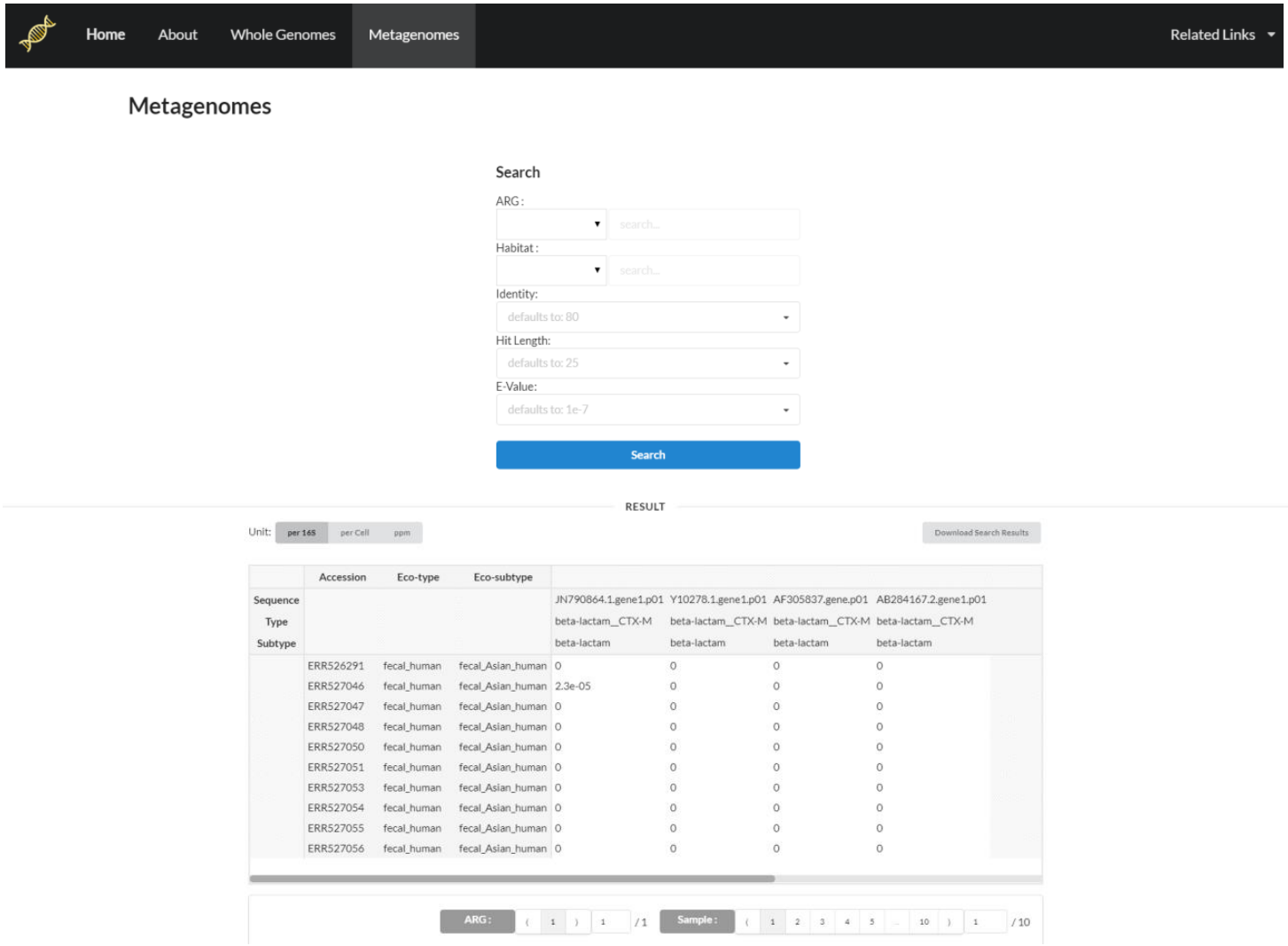


Fig 3. The layout of the ARGs-OSP Metagenomes.

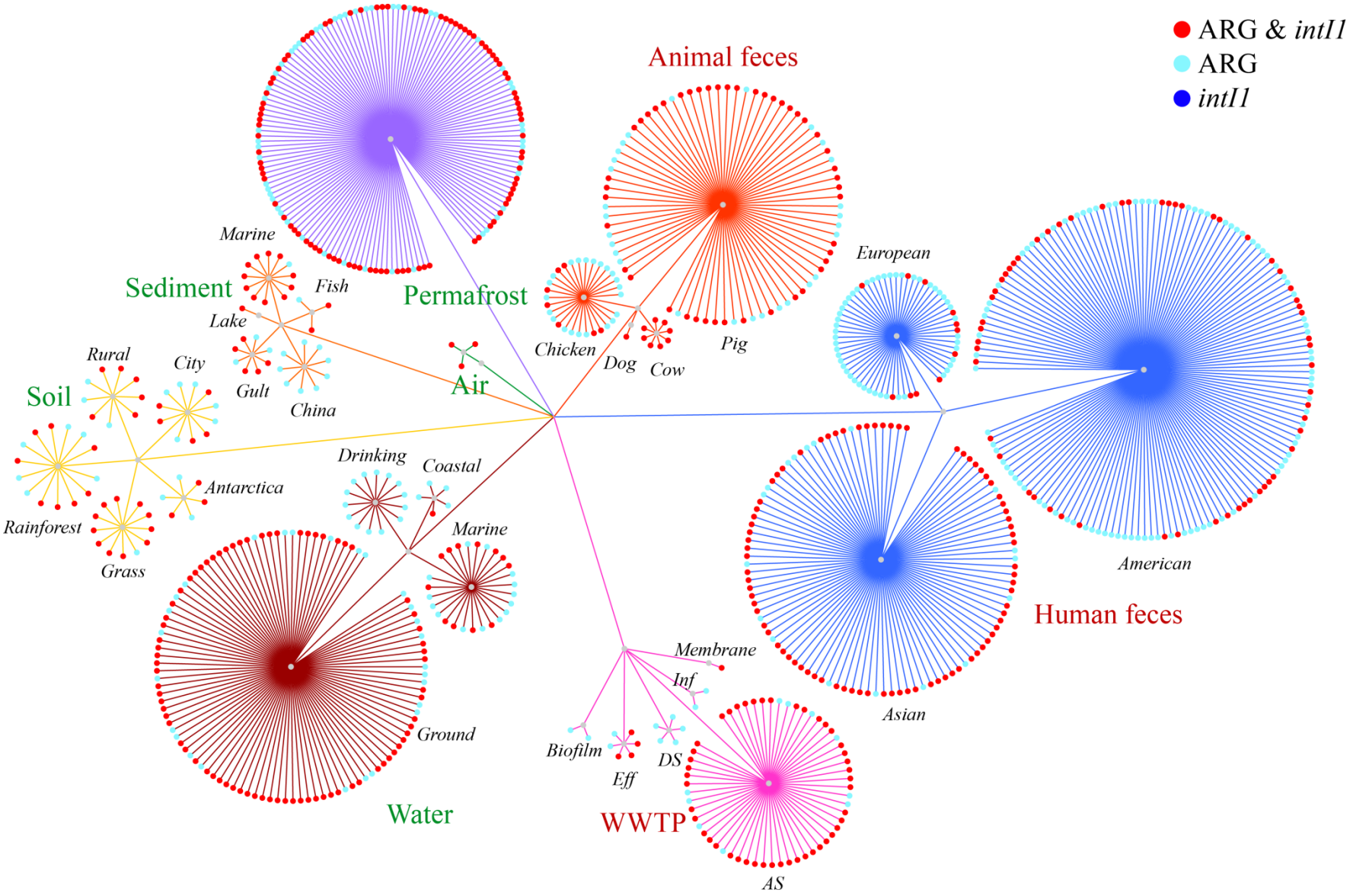


Fig 4. The ecological co-occurrence of class 1 integrases and total ARGs in 7 eco-types.

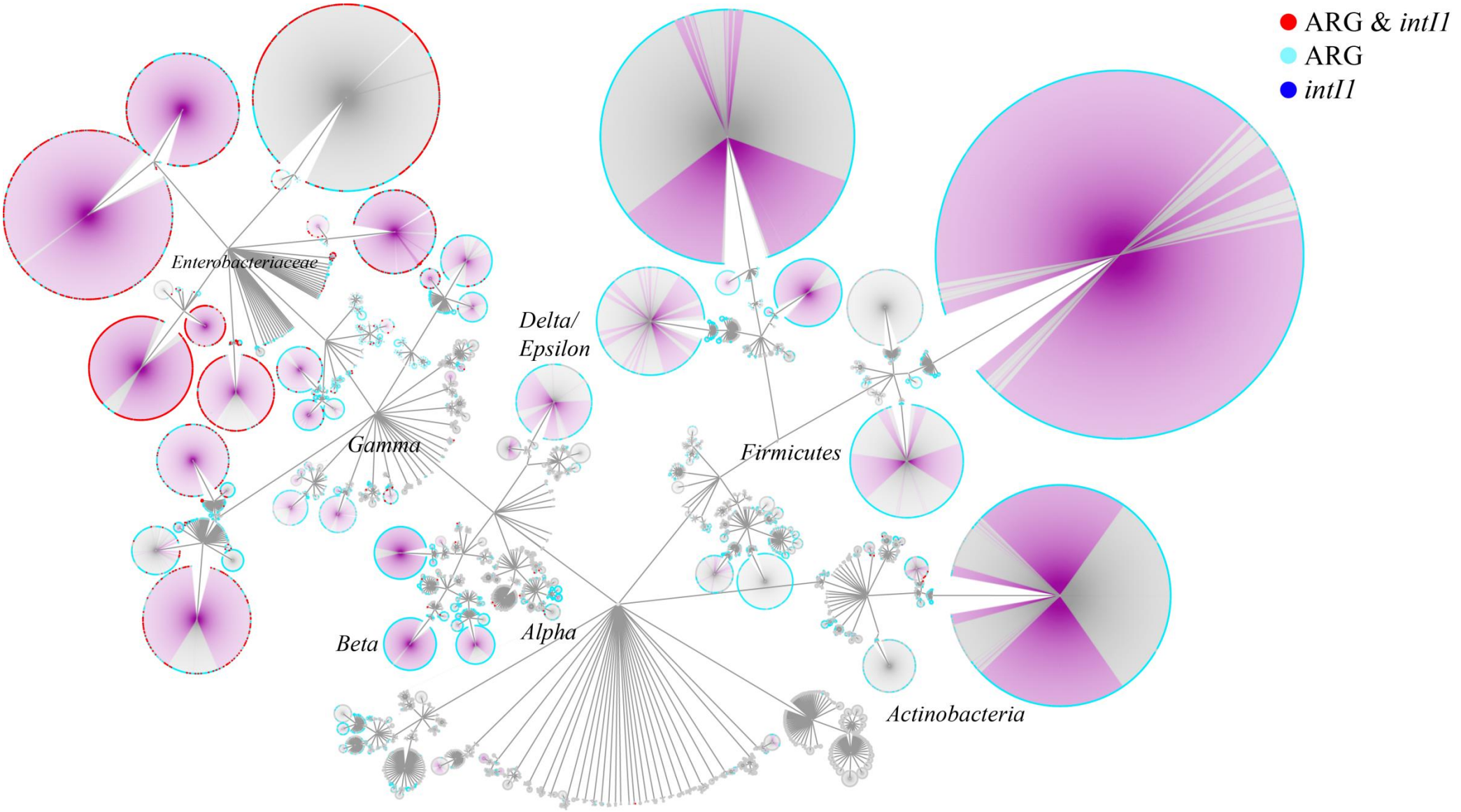
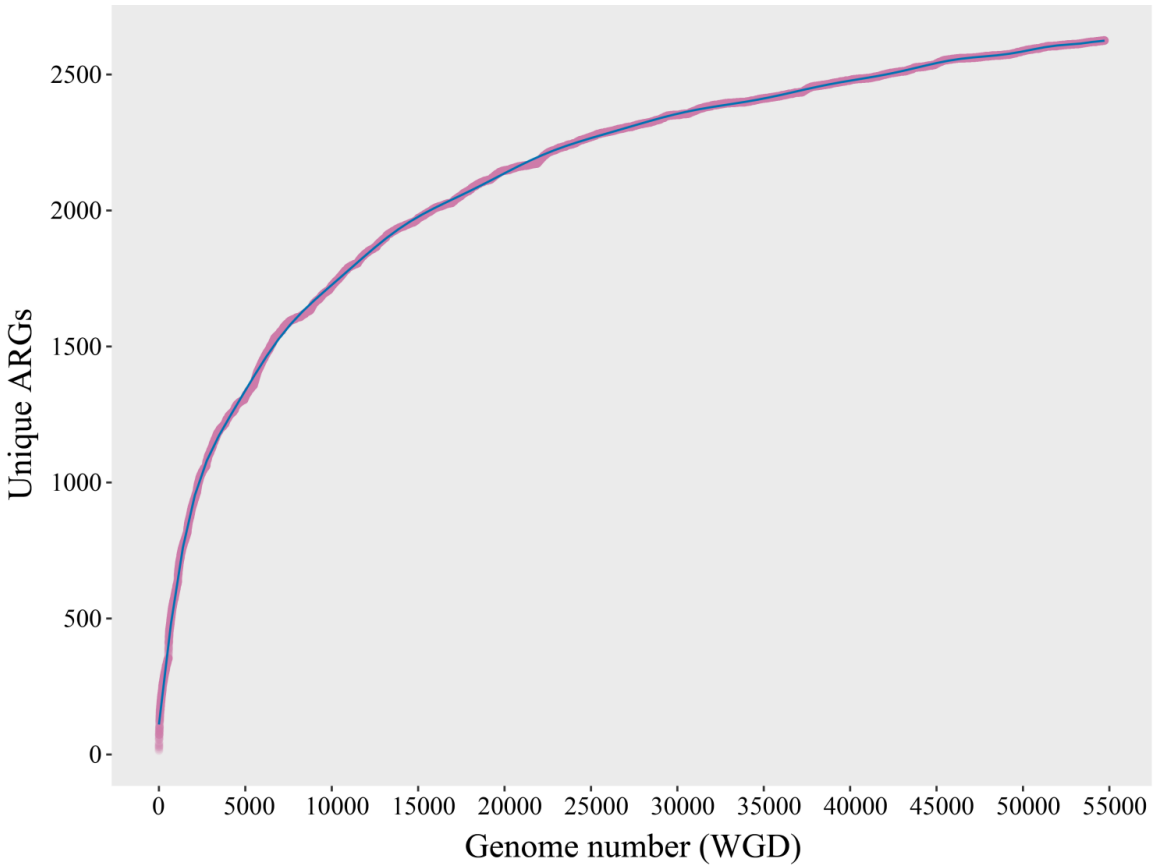


Fig 5. The phylogenetic co-occurrence of class 1 integrases and total ARGs in bacterial life tree.



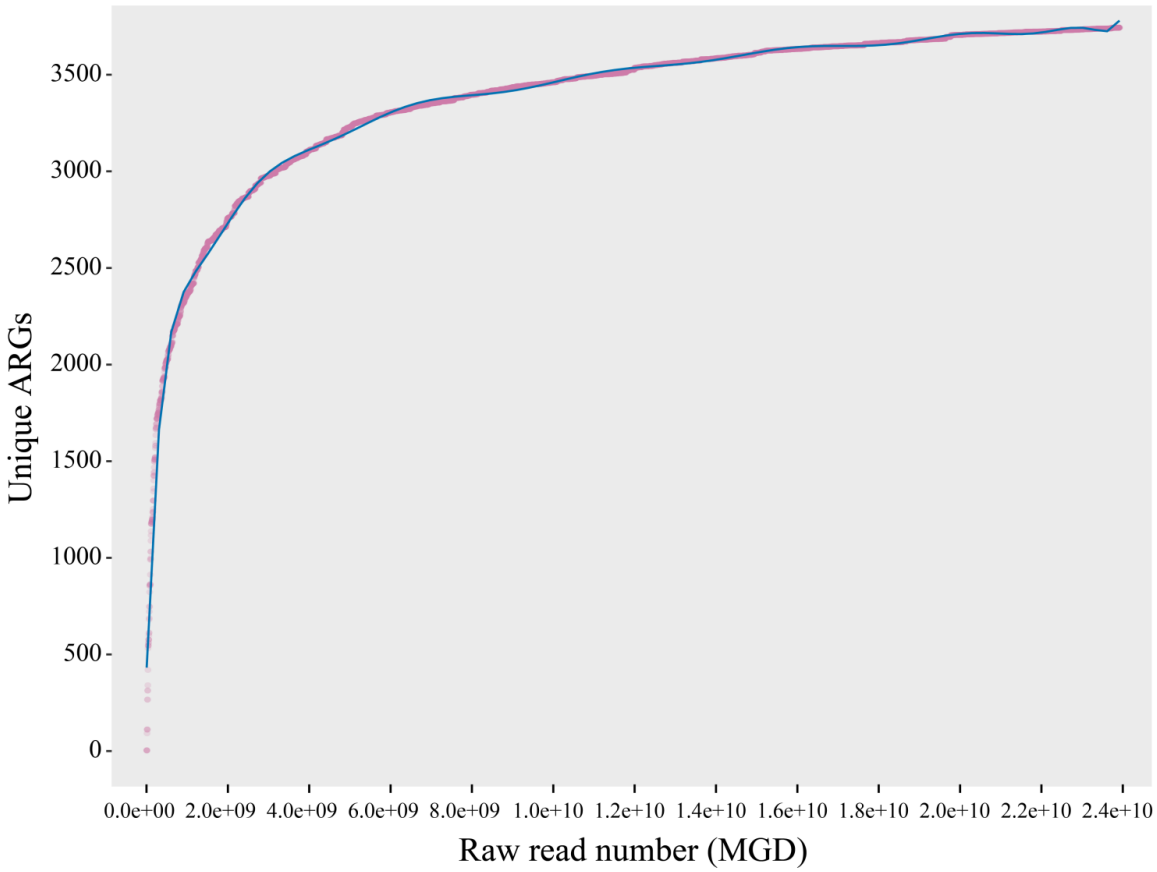


Fig S1. The rarefaction curves of the unique ARGs in the Whole Genome Database (WGD, each genome) and Metagenome Database (MGD, each raw read), constructed by randomly subsampling without replacement.

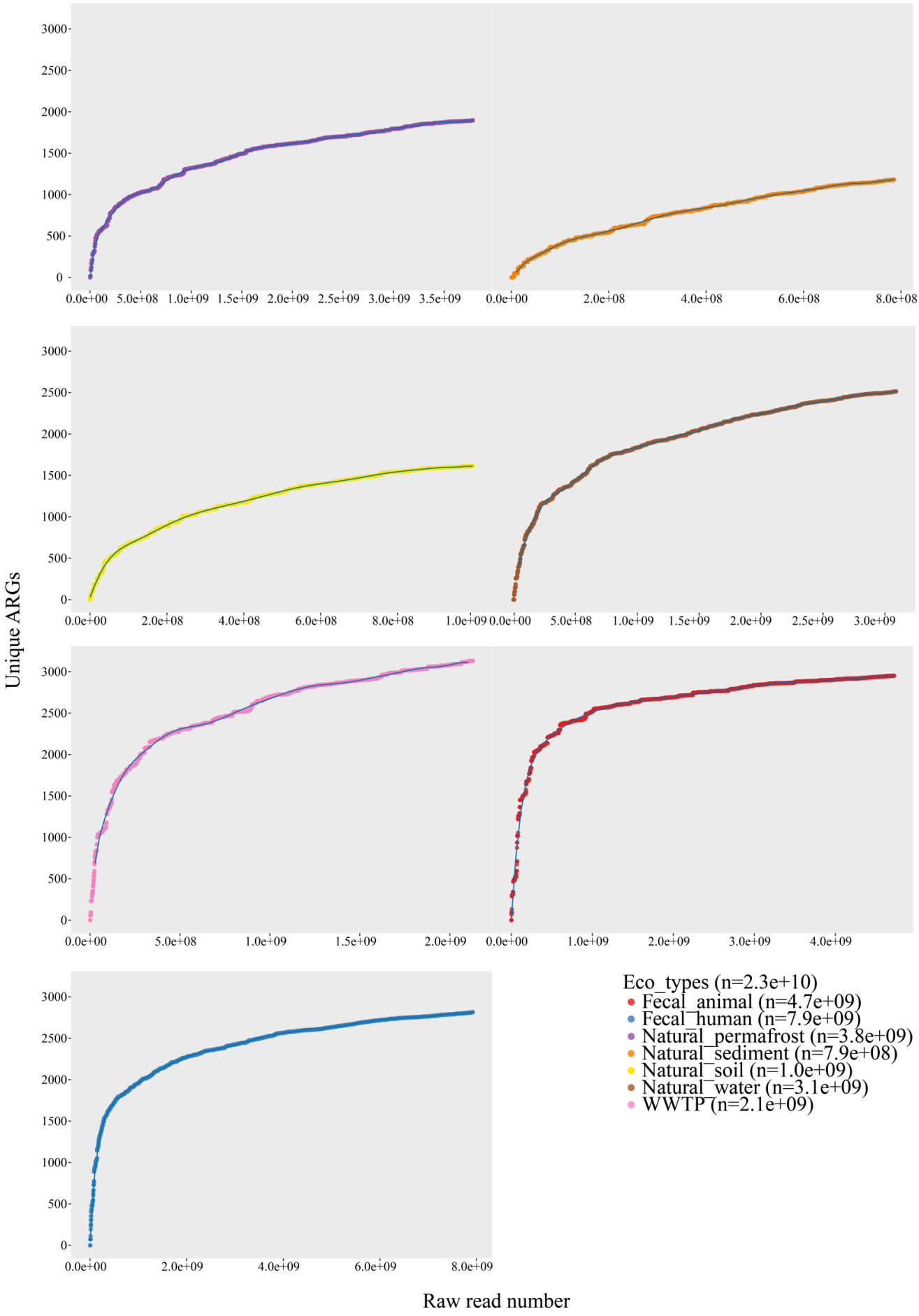


Fig S2. The rarefaction curves of the unique ARGs constructed specifically for 7 eco-types of Metagenome Database (MGD, each raw read) by randomly subsampling without replacement.

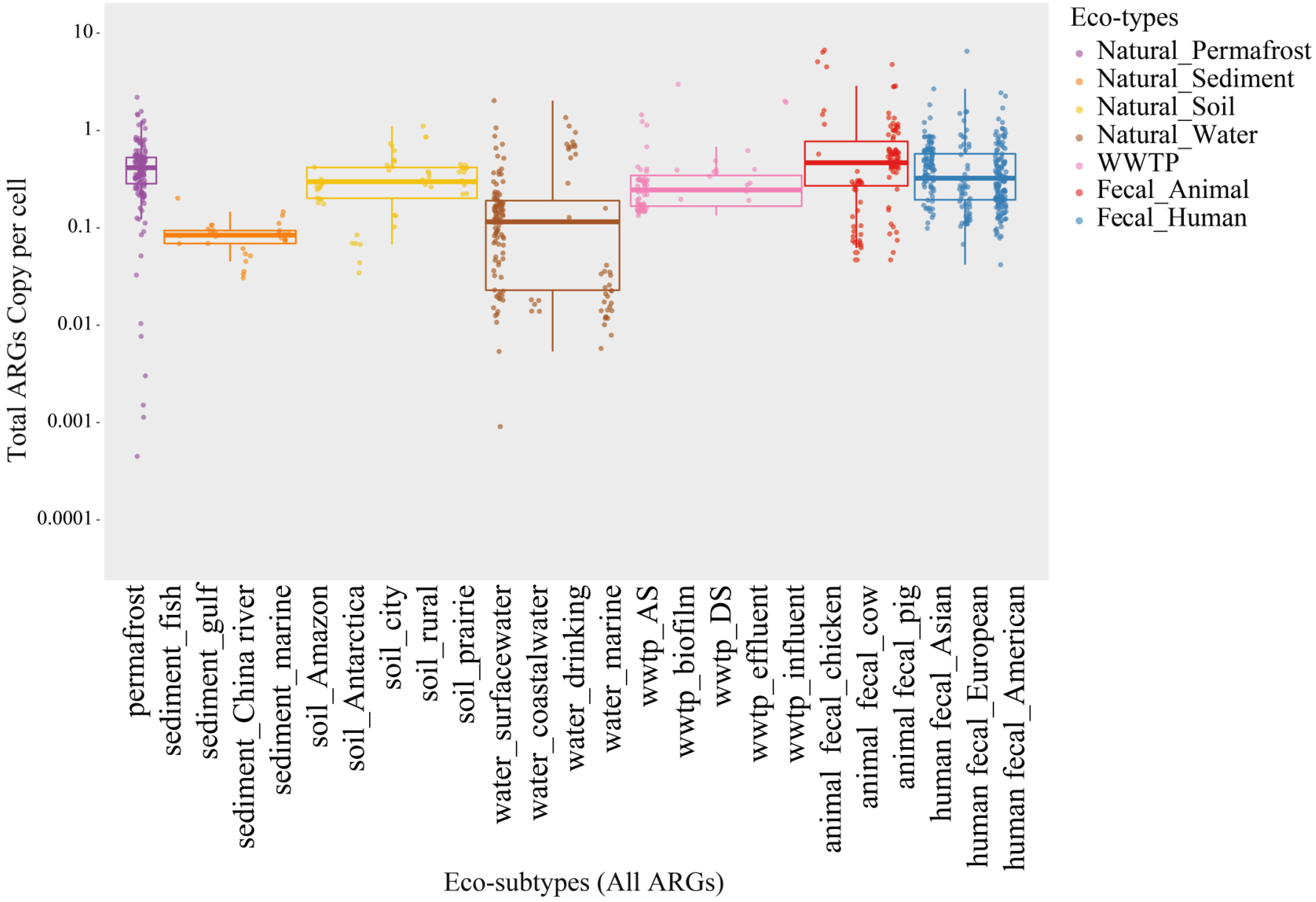


Fig S3. The ecological distribution and abundance (copy per cell) of total ARGs in 25 eco-subtypes.

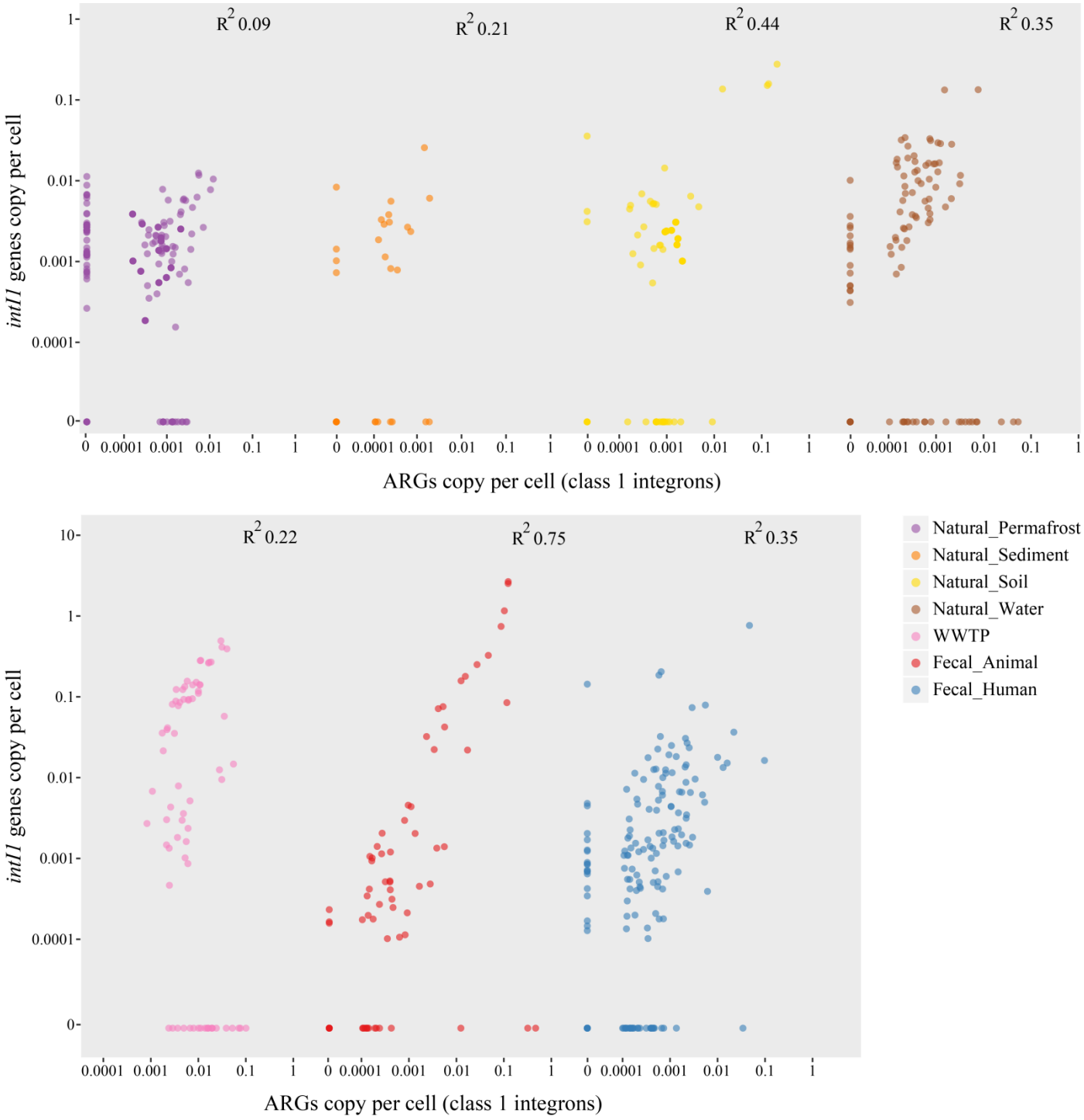


Fig S4. The co-occurrence of class 1 integrases and a subgroup of ARGs that were discovered on class 1 integrons in 7 eco-types, fitted by linear regression (R^2).

Table 1. The linear regression relations of the abundance of *intI1* genes to the abundance of total ARGs (Total) and 107 ARGs detected on the class 1 integrons (Class I).

Eco-type	R^2		Slope	
	intI1-Total	*intI1*-Class I	*intI1*-Total	*intI1*-Class I
Natural_Permafrost (144)	0.08	0.09	0.69	0.28
Natural_Sediment (33)	0.37	0.21	2.00	0.39
Natural_Soil (52)	0.06	0.44	0.53	0.61
Natural_Water (146)	0.34	0.35	0.83	0.64
WWTP (69)	0.03	0.22	0.62	0.95
Fecal_Animal (110)	0.47	0.75	2.10	1.10
Fecal_Human (300)	0.30	0.35	1.40	0.63
All (854)	0.04	0.35	0.40	0.65