FEMS Microbiology Ecology, 96, 2020, fiaa107

doi: 10.1093/femsec/fiaa107
Advance Access Publication Date: 30 May 2020
Research Article

RESEARCH ARTICLE

Online searching platform for the antibiotic resistome in bacterial tree of life and global habitats

An Ni Zhang[1,2,4], Chen-Ju Hou[2], Mishty Negi[2], Li-Guan Li[2] and Tong Zhang[1,2,3,*]

[1]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, China, [2]Environmental Microbiome Engineering and Biotechnology Laboratory, The University of Hong Kong, Hong Kong, China, [3]School of Public Health, The University of Hong Kong, Hong Kong, China and [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, USA

*Corresponding author: Environmental Microbiome Engineering and Biotechnology Laboratory, The University of Hong Kong, Hong Kong. Tel: +852-2857 8551; Fax: +852-2859 8987; E-mail: zhangt@hku.hk

One sentence summary: Online searching platform for antibiotic resistome in bacterial tree of life and global habitats by big data mining into 54 718 bacterial genomes, 15 738 bacterial plasmids, 3000 bacterial integrons and 854 environmental metagenomes.

Editor: Kornelia Smalla

ABSTRACT

Metagenomic analysis reveals that antibiotic-resistance genes (ARGs) are widely distributed in both human-associated and non-human-associated habitats. However, it is difficult to equally compare ARGs between samples without a standard method. Here, we constructed a comprehensive profile of the distribution of potential ARGs in bacterial tree of life and global habitats by investigating ARGs in 55 000 bacterial genomes, 16 000 bacterial plasmid sequences, 3000 bacterial integron sequences and 850 metagenomes using a standard pipeline. We found that >80% of all known ARGs are not carried by any plasmid or integron sequences. Among potential mobile ARGs, tetracycline and beta-lactam resistance genes (such as *tetA*, *tetM* and class A beta-lactamase gene) distribute in multiple pathogens across bacterial phyla, indicating their clinical relevance and importance. We showed that class 1 integrases (*intI1*) display a poor linear relationship with total ARGs in both non-human-associated and human-associated environments. Furthermore, both total ARGs and *intI1* genes show little correlation with the degree of anthropogenicity. These observations highlight the need to differentiate ARGs of high clinical relevance. This profile is published on an online platform (ARGs-OSP, http://args-osp.herokuapp.com/) as a valuable resource for the most challenging topics in this field, i.e. the risk, evolution and emergence of ARGs.

Keywords: genomic analysis; metagenomic analysis; environmental selection; global antibiotic resistome; anthropogenicity

INTRODUCTION

Due to the intensive use of antibiotics, antibiotic resistance genes (ARGs) are emerging in almost all environments, becoming one of the top global concerns. Through high-throughput sequencing and metagenomic analysis, ARG-related studies have been conducted in various habitats. They are found widely distributed across both human-associated habitats (Tadesse *et al.* 2012; Martinez, Coque and Baquero 2015) and natural and polluted non-human-associated environments (Alonso, Sanchez and Martinez 2001), such as sediment (Gonzalez-Plaza *et al.* 2017; Chu *et al.* 2018: 201), soil (Nesme *et al.* 2014; Lau *et al.* 2017), surface water (Rodriguez-Mozaz *et al.* 2015; Tang *et al.* 2016), marine water (Nesme *et al.* 2014: 20) and permafrost (D'Costa *et al.* 2011; Rascovan *et al.* 2016). The identification of ARGs in drinking water (Ma *et al.* 2017b) and human food (Bengtsson-Palme 2017) further reveals the potential for their direct exposure to the human microbiome. All these

Received: 8 May 2019; **Accepted:** 27 May 2020

metagenomes are valuable sources for the construction of a global profile of the antibiotic resistome, including the phylogenetic and ecological distribution, diversity, abundances, host pathogenicity and potential gene mobility.

However, lacking a standard method to identify and annotate ARGs makes it difficult to conduct an equal comparison across samples (Gupta, Tiwari and Cytryn 2020). Currently, different studies use various searching methods, searching criteria, ARG databases and quantitative units. For example, the abundances of ARGs were found to have a 5–20-fold difference (Yin *et al.* 2018) when comparing the domain-based searching method of the Hidden Markov Model (HMM) (Eddy 1998) and the similarity-based searching methods of BLAST (Camacho *et al.* 2009), USEARCH (Edgar 2010) or DIAMOND (Buchfink, Xie and Huson 2015). In addition, advanced bioinformatics tools have been developed to annotate ARGs in metagenomes, such as using gene-capture platforms (Lanza *et al.* 2018) and a structure-based model (Ruppé *et al.* 2019). Another obstacle for direct comparison is the different ARG reference databases used because different databases suit different analytical scenarios (Gupta, Tiwari and Cytryn 2020). Even the two most-cited ARG databases, the Comprehensive Antibiotic Resistance Database (CARD; http://arpcard.mcmaster.ca) (Jia *et al.* 2016) and the Antibiotic Resistance Genes Database (ARDB) (Liu and Pop 2008), share only 10–25% of reference sequences (Yang *et al.* 2016). This highlighted the need to re-analyse the metagenomes from diverse habitats using a standard pipeline to depict the ecological distribution of ARGs.

Still, metagenomic analysis alone could hardly answer the frontier scientific questions in this field, such as the evolution, emergence and risk of ARGs. These knowledge gaps could be partially filled by bacterial genome sequences (Pal *et al.* 2015; Li, Xia and Zhang 2017), including the phylogenetic distribution, gene mobility and gene arrangement of ARGs. The integration of whole genome analysis and metagenome analysis is a promising attempt to construct a comprehensive profile of the antibiotic resistome.

However, processing big datasets is time- and resource-consuming and is unnecessarily repeated by individual researchers all over the world. Many researchers publish their processed results on online platforms for convenient usage by other researchers, such as IMG/VR (https://img.jgi.doe.gov/vr/) (Paez-Espino *et al.* 2016). Thus, in this study, we constructed an updated and comprehensive profile of the distribution of potential ARGs on an ARGs online searching platform (ARGs-OSP, http://args-osp.herokuapp.com/), to serve as a valuable resource to guide diverse research topics and inspire new research interests in this area.

MATERIALS AND METHODS

Collecting various bacterial datasets

We collected 54 718 quality-filtered bacterial genome sequences with >50% completeness and <10% contamination (Parks *et al.* 2017) (Table S1, see online supplementary material, and Fig. 1A) from NCBI Genome (Sayers *et al.* 2012), 15 738 bacterial plasmids from NCBI Genome (Sayers *et al.* 2012) and 2977 bacterial integrons (Zhang *et al.* 2018) as Whole Genome Dataset (WGD). All sequences in WGD were curated by Plasfow (Krawczyk, Lipinski and Dziembowski 2018) and a well-curated plasmid database (Brooks, Kaze and Sistrom 2019) into plasmids and non-plasmids. The taxonomy of bacterial genomes was downloaded from NCBI Genome to predict potential pathogenicity by mapping the taxonomy to a list of potential human bacterial pathogens (Woolhouse and Gowtage-Sequeria 2005). In summary, the genomes in WGD covered in total 3654 bacterial species from 32 phyla. We collected 854 metagenomic datasets from NCBI SRA (Sayers *et al.* 2012) and MG-RAST (Wilke *et al.* 2015) (Fig. 1B) as the Metagenome Dataset (MGD). The metadata were obtained from the original publications of all metagenomes and were manually categorized by the guidelines from previous studies, such as the abundance of clinical class 1 integrases (*intI1*) (Gillings 2018) and habitat information (Pal *et al.* 2016; Li, Xia and Zhang 2017; Thompson *et al.* 2017). Briefly, all metagenomes were classified into 25 habitat-subtypes of seven habitat-types, including animal feces (chicken, cow and pig); human feces (American, Asian and European); activated sludge, reactor biofilm, digested sludge, effluent and influent from wastewater treatment plants (WWTPs); permafrost soil; sediment (river sediment, fishpond sediment, gulf sediment and marine sediment); soil (Amazon soil, Antarctica soil, city soil including grasslands and parks, prairie soil and rural soil); water (coastal water, drinking water, marine and surface water) (for metadata and accession data see Table S2 in online supplementary material). Here, we defined animal feces, human feces and WWTPs as human-associated habitats, and permafrost soil, sediment, soil and water as non-human-associated habitats. The non-human-associated samples were further curated by <0.05% fecal contamination, which was represented by the total percentage of reads covered by three taxa as fecal markers, including uncultured crAssphage (Karkman, Pärnänen and Larsson 2019), *Bacteroides* (Savichtcheva, Okayama and Okabe 2007) and *Escherichia coli* (Carlos *et al.* 2010) (Fig. S1, see online supplementary material). The percentage of reads covered by the three taxa was calculated by the number of reads mapping to the three taxa divided by the total number of reads mapping to all taxa by kraken (Wood and Salzberg 2014) using the default library and settings. Non-human-associated samples with >0.05% fecal contamination were excluded.

Identifying and annotating ARGs and *intI1* genes

We used the Structured ARG reference database (SARG) (Yang *et al.* 2016), a well-curated database integrating CARD and ARDB, as a standard database for identifying and annotating ARGs (details in Supplementary methods, see online supplementary material). All the coding sequences (CDS) were extracted from GenBank files of WGD and were searched against SARG and the *intI1* database (Zhang *et al.* 2018). In MGD, ARGs and *intI1* genes were identified by a two-step similarity-search method, through USEARCH v8.0.1623_i86linux64 (Edgar 2010) and BLASTX 2.2.28+ (Camacho *et al.* 2009). The abundances of genes were calculated into units of ppm, gene copies per bacterial 16S and gene copies per bacterial cell (Yang *et al.* 2016). Briefly, the copy number of a gene was firstly normalized against its reference sequence length and then normalized to the total number of raw reads, total number of bacterial 16S rRNA genes and total number of bacterial cells, respectively (Yin *et al.* 2018). The total number of bacterial cells in metagenomes was inferred from the average coverage of 30 universal bacterial essential single copy genes (Nayfach and Pollard 2015).

Analysis and visualization

We combined the abundances of ARGs with the phylogenetic and ecological information of samples via self-written scripts using Python 3.6 (https://www.python.org/) and R 3.3.2 (Team

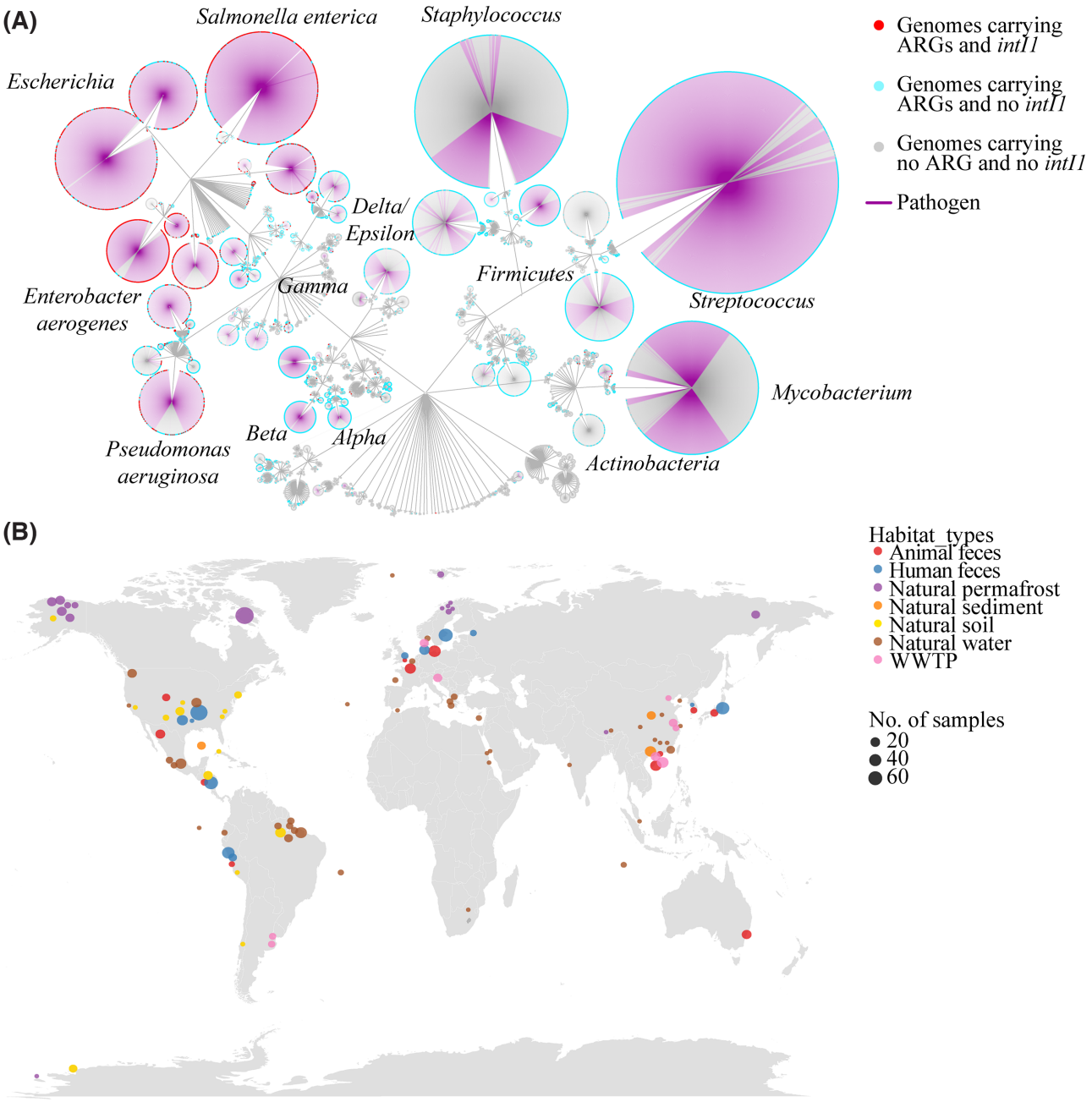

Figure 1. Overview of the Whole Genome Dataset (WGD) and Metagenome Dataset (MGD). (**A**) The phylogenetic relationship of 54 718 bacterial genomes of 45 phyla in WGD. Each genome is represented by a node and the pathogenic genomes are indicated by purple edges. The color of the node represents the co-occurrence of class 1 integrases (*intI1*) and ARGs in that genome. To be specific, a red node indicates that the genome carries both ARGs and *intI1* genes; a blue node indicates that the genome carries ARGs, but not *intI1* genes; a grey node indicates that the genome carries neither ARGs nor *intI1* genes. We found no genomes carrying *intI1* genes but no ARGs. (**B**) The global map of metagenomic datasets in MGD. The size of the node is proportional to the number of samples and the color of the node represents the habitat-type.

2013). The rarefaction curves (Gotelli and Colwell 2001) of WGD and MGD were conducted by randomly subsampling genomes or metagenomic reads without replacement (Colwell *et al.* 2012; Deng, Daley and Smith 2015). All the networks were visualized with Cytoscape 3.3.0 (Shannon *et al.* 2003) in Tree or Hierarchic layout.

Availability and implementation

All data, results and scripts are available on ARGs-OSP (http://args-osp.herokuapp.com/) and Github (https://github.com/caozhichongchong/ARGs-OSP and https://github.com/bryanhou1/argosp).

RESULTS AND DISCUSSION

A global profile of the antibiotic resistome

We developed ARGs-OSP (http://args-osp.herokuapp.com/) to publish a global profile of the antibiotic resistome covering the phylogenetic and ecological distribution, diversity, abundances, host range and host pathogenicity of ARGs in WGD and MGD (Fig. 1). By constructing the rarefaction curves for WGD and MGD, we evaluated that both datasets provide good representation of the ARGs currently found in bacterial tree of life and global habitats (Fig. S2, see online supplementary material). Users can search and download the prevalence and abundances of ARGs across taxonomies and habitats from ARGs-OSP. The accessibility and convenience of this platform could meet the requirements of versatile research interests, such as host ranges of ARGs, dissemination of ARGs between undisrupted non-human-affected and human-affected habitats, and comparing ARGs in local samples to all global samples.

Most known ARGs are non-mobile and the mobile ARGs conferring tetracycline and beta-lactam resistance are widely distributed across bacterial species and highly prevalent in human pathogens. In whole bacterial genomes, 65% of ARGs in SARG were detected in 809 bacterial species across 13 phyla of a total of 3654 bacterial species across 32 phyla, and 27% of all ARG-carrying species are genomes of potential human pathogens. Even though human pathogens could be overrepresented in WGD, we found that genomes carrying an ARG are three times as likely to come from a pathogen as from an average genome in WGD. This observation indicates that ARGs could be strongly positively selected by antibiotics in human pathogens than non-pathogens (Ji *et al.* 2012; Vaz-Moreira, Nunes and Manaia 2014). Of all known ARGs identified in WGD, <20% are potentially mobile, i.e. were found on mobile genetic elements (MGEs) such as plasmids and integrons. ARGs conferring multidrug resistance (with an occurrence of 29.4%), beta-lactam resistance (15.5%), aminoglycoside resistance (10.5%) and tetracycline resistance (8.7%) were universally detected in bacterial species, which is consistent with a previous study investigating 2500 complete bacterial genomes (Pal *et al.* 2015). Some ARGs are frequently carried by a wide lineage of >50 bacterial species, i.e. *tetA*, *tetM*, *acrB*, *aph(3′-I*, *aadA*, *mdtK*, *tolC*, and class A beta-lactamase gene, with >50% prevalence in human pathogens. Some of these ARGs encoding efflux pumps, i.e. *acrAB* and *mdtK*, could be non-mobile ARGs (Nishino and Yamaguchi 2001; Piddock 2006) as they were not found on any MGEs. Instead, *tetA*, *tetM*, *aadA*, *tolC* and class A beta-lactamase gene were found on multiple MGEs, indicating that they are more likely to function mainly in antibiotic resistance and to be strongly selected by antibiotics. These mobile ARGs are widely distributed across bacterial phyla and the human microbiome, suggesting their importance in both medical and environmental fields.

The total abundance of ARGs is not significantly higher in human-associated habitats than in non-human-associated environments. We found that >90% of known ARGs are present in MGD. We observed a higher density and richness of ARG reads in human-associated environments of human feces, animal feces and WWTPs than in non-human-associated environments of water, sediment, permafrost and soil (Fig. 2). However, the abundances of total ARGs show little difference across habitats (<10-fold difference and *P*-value of 0.2 by Kruskal–Wallis test) after normalizing the total number of ARGs against the total number of bacterial cells per sample (Fig. S3 and supplementary results, see online supplementary material). This means that even though different habitats harbor different compositions of bacterial communities, the average level of antibiotic resistance within bacterial cells is quite stable (Ruppé *et al.* 2019). Thus, the high richness and density of ARGs in human-associated habitats could be mainly contributed by a high density of bacterial communities in these habitats (Pal *et al.* 2016). Since most known ARGs are non-mobile, this indicates that the diversity and richness of ARGs that are detected in MGD are tightly linked with the native bacterial community composition.

To compare to other studies, we normalized the abundances into ARG copies per bacterial 16S and summed up the total relative abundance of all ARGs in a metagenome sample. In detail, the animal feces samples harbor the highest abundance of ARGs (0.4 copies per bacterial 16S), where the sub-therapeutic usage of antibiotics for growth promotion (Aust *et al.* 2008; Hu, Zhou and Luo 2010) may promote the selection and dissemination of ARGs (Beaber, Hochhut and Waldor 2004). Non-human-associated permafrost and soil also display high abundances of ARGs (0.3 and 0.2 copies per 16S, respectively), which is at a similar level to human feces and WWTPs (0.2 and 0.1 copies per 16S, respectively). However, >90% of ARG reads in permafrost share a low similarity to ARG reads in other habitats (<90% amino-acid similarity), suggesting that most permafrost ARGs could be sequence variants or ancestors of the ARGs of clinical relevance in human-associated environments. Besides, the majority (>90%) of ARGs in human-associated metagenomes and on plasmids share a high similarity (>90% amino-acid similarity) to ARG reference sequences. These observations indicate the importance for current ARG databases to differentiate the sequence variants of clinically relevant ARGs and to use a strict similarity cutoff, i.e. 90% amino-acid similarity, to identify clinically relevant ARGs. Within the soil, the habitat-subtype of rural soil has the highest level of ARGs (0.3 copies per 16S), indicating that these samples could be contaminated by the agricultural use of animal manure. The non-human-associated water and sediment display the lowest ARG abundances of 0.05 and 0.02 copies per 16S, respectively. Most habitats exhibit a similar resistance level to previous studies (Bengtsson-Palme *et al.* 2015; Li *et al.* 2015; Pal *et al.* 2016) except for a 10–100-fold higher detection of ARGs in soil and surface water. These differences could be partially explained by the pollution history of the soil environment (3-fold ARG abundance in rural soil compared with Antarctica soil) and divergent types of surface water environment. Thus, we collected a large sample size of metagenomes from various habitat-subtypes to provide a more representative global profile.

Anthropogenicity appears to have a weak correlation with the total abundances of ARGs, but could increase the diversity of ARGs in human-associated environments. The function of ARGs, including the ARG families and their target antibiotics, was annotated by SARG. The sediment harbors the smallest diversity, <30%, of all ARG families. Even though the sediment habitat has the smallest sample size (32 metagenomes), its rarefaction curve appears to be leveling off (Fig. 2, and online supplementary results). Bacteria in soil, permafrost and human feces cover ~50% of all ARG families. Bacteria from animal feces, water and WWTP environments display a rich diversity covering almost 75% of all ARG families. The ARG families that have long been the focus of ARG-related research are widely distributed in all seven habitats, such as aminoglycoside resistance gene *aph(3)-I*, beta-lactam resistance gene *TEM*, sulfonamide resistance genes *sul1* and *sul2*, macrolides, lincosamides, streptogramines (MLS) resistance genes *macAB*, tetracycline resistance genes *tetA* and *tetM*, multidrug resistance genes *acrAB*, and vancomycin resistance gene *vanA* (Berendonk *et al.* 2015; Pal *et al.* 2016). *sul1*, *sul2*,

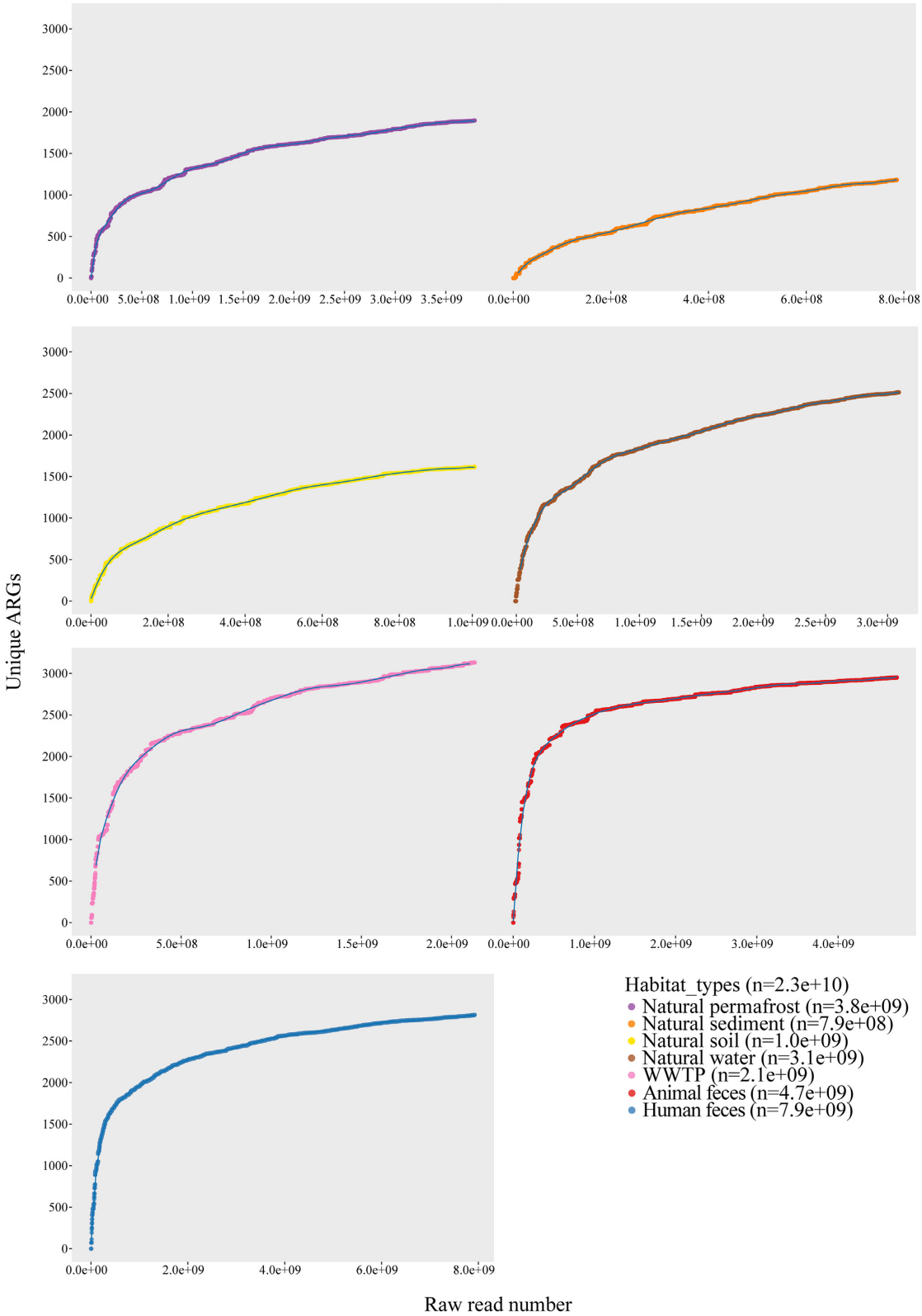

Figure 2. The rarefaction curves of the number of unique ARGs for seven habitat-types of the Metagenome Dataset (MGD) obtained by randomly subsampling metagenomic reads without replacement.

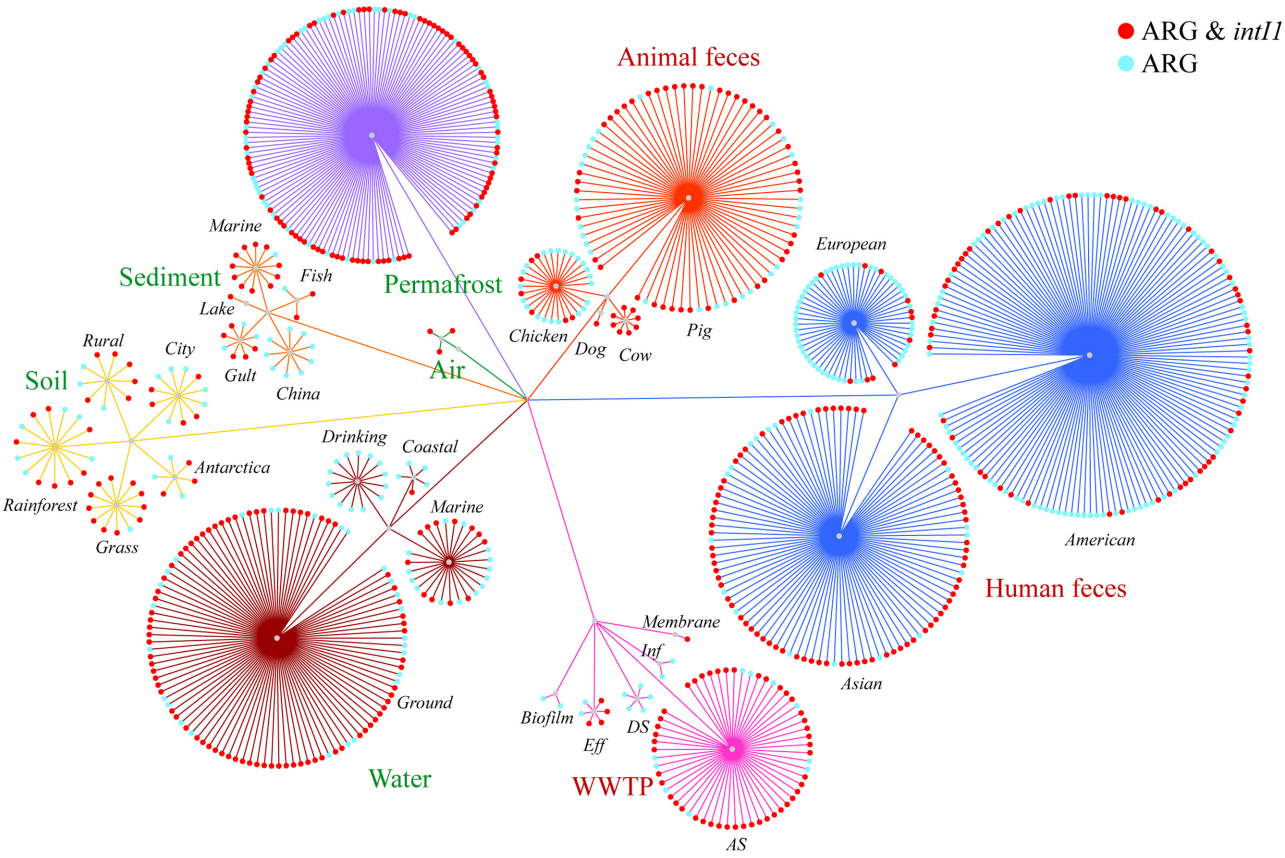

Figure 3. The ecological co-occurrence of class 1 integrases (*intI1*) and ARGs in seven habitat-types. Each node represented a metagenomic sample and the color represents the co-occurrence of *intI1* and ARGs. To be specific, a red node indicates that the sample harbors both ARGs and *intI1*; a blue node indicates the sample harbors ARGs but not *intI1*. We found no samples harboring *intI1* but no ARGs. The colors of the edges indicate the habitat-types.

aph(3)-I and *TEM* were found on many MGEs, indicating that they could be strongly selected under the wide usage of antibiotics. *macAB* and *acrAB*, on the other hand, were not found on any MGEs and may convey ecological functions other than antibiotic resistance.

Co-occurrence of ARGs and *intI1* genes

Although high abundances of *intI1* genes and ARGs have been reported in many human-associated environments, the co-occurrence of *intI1* genes and total ARGs has not been well-evaluated using metagenomes. The clinical *intI1* genes are considered as a potential indicator of anthropogenic pollution (Gillings *et al.* 2015) because of their high abundance and universal occurrence in human-associated environments (Pruden, Arabi and Storteboom 2012; Ma *et al.* 2017a). Previous studies on the correlation between *intI1* genes and human pollution mainly focus on: (i) bacteria carrying *intI1* genes are highly abundant in human-associated environments and cannot be effectively removed from WWTPs (Du *et al.* 2014: 2); (ii) *intI1* genes and ARGs are co-selected (Seiler and Berendonk 2012; Khan *et al.* 2013) and significantly correlated (i.e. *qacE/qacE*Δ 1, but not *sul1*) in habitats with or without human impact (Jechalke et al. 2014, 2015); (iii) *intI1* genes increase with anthropogenic pollution, such as heavy metals, disinfectants, antibiotics and pesticides (Lehmann *et al.* 2016; Yang *et al.* 2017). Additionally, some ARGs (*sul1*, *sul2* and *tetM*) are evaluated to have a strong correlation with *intI1* genes in polluted environments (Lv *et al.* 2018). Thus, it is critical to evaluate the co-occurrence of *intI1* genes and ARGs by our datasets.

The *intI1* genes display a weak linear relationship with all ARGs, but a stronger one with ARGs carried on integrons. Overall, the total abundances of *intI1* genes (including clinical *intI1* genes and their environmental variants) have a positive correlation with the total abundances of ARGs in all environments (Table S3, see online supplementary material, and Fig. 4). However, *intI1* genes were not detected in a large portion of animal feces and human feces samples (light blue nodes in Fig. 3), including samples with high abundances of ARGs. Moreover, *intI1* genes are highly conserved in *Gammaproteobacteria* (red nodes in Fig. 1A) (Zhang *et al.* 2018), whereas ARGs are widely distributed across bacterial phyla (red and blue nodes in Fig. 1A). This indicates that *intI1* genes are more likely to be correlated to a subset of ARGs, i.e. ARGs carried on integrons, than total ARGs. This also explains the relatively low abundance of *intI1* genes in some animal feces (such as chicken feces), where *Gammaproteobacteria* is usually not dominant. Furthermore, the total abundances of *intI1* genes and ARGs display a weak linear relationship in most habitats (R^2 = 0.003–0.47), especially in WWTPs, soil and permafrost (R^2 = 0.003–0.008), excluding samples without detection of *intI1* genes (Table S3). Different habitats were found to comply with different linear relationships, with slopes ranging from 0.53 to 2.10. The *intI1* genes fit better with linear relationships (in terms of R^2) to a group of ARGs carried on integrons (integron-carried ARGs), which amounted to a total of 105 ARGs (accession data are listed in Table S4, see online supplementary material) that were found among all available clinical

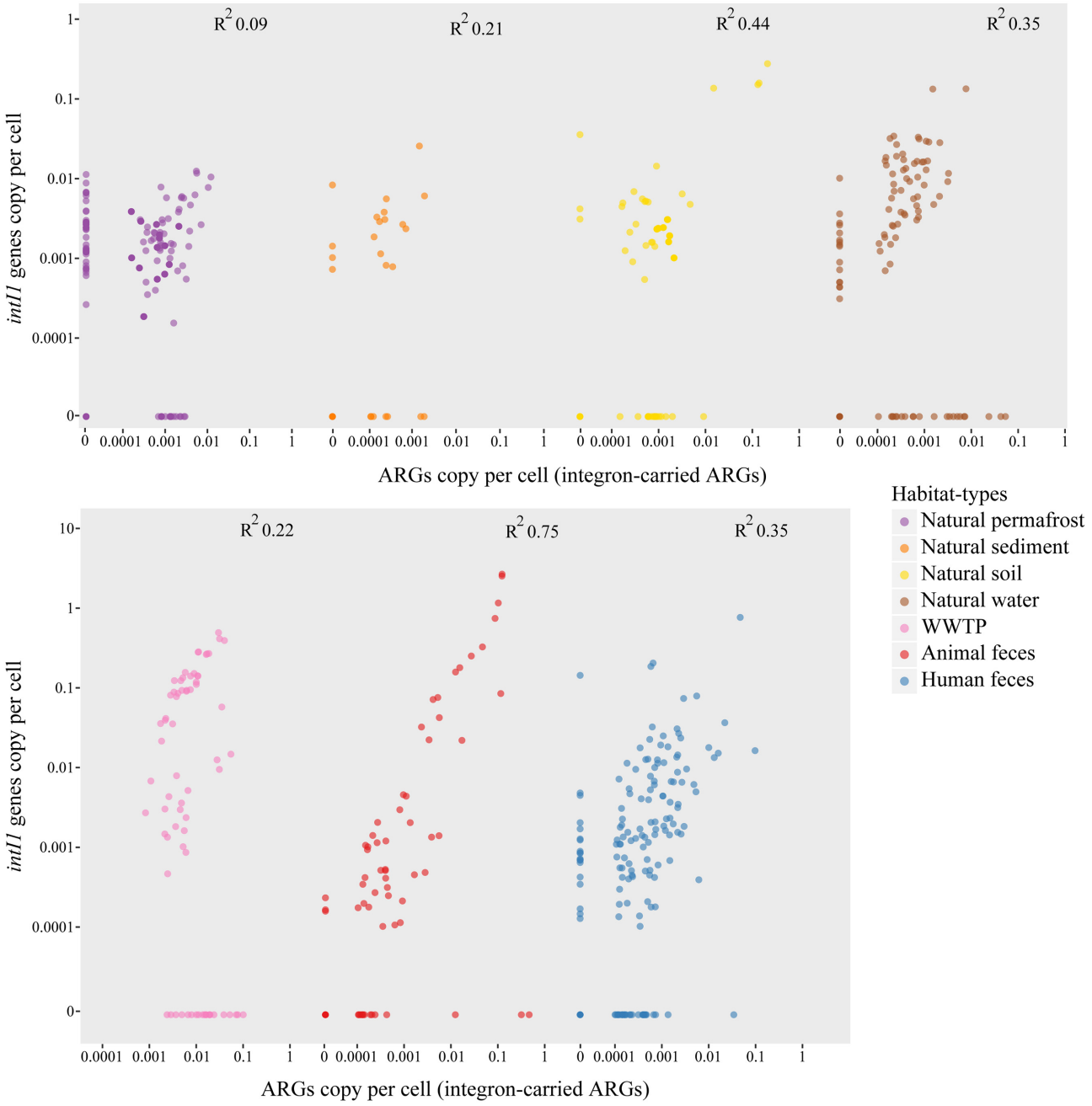

Figure 4. The co-occurrence of class 1 integrases (*intI1*) and a group of 105 ARGs found among all available class 1 integrons (integron-carried ARGs) in seven habitat-types, fitted by linear regression (R^2). Samples without detection of *intI1* genes were excluded when calculating linear regression.

and non-clinical class 1 integrons (Zhang *et al.* 2018), especially in animal feces ($R^2 = 0.75$).

CONCLUSIONS

In this study, a comprehensive profile of the antibiotic resistome was constructed by investigating the ARGs in WGD (54 718 bacterial genomes, 15 738 bacterial plasmids and 2977 bacterial integrons) and MGD (854 metagenomes). Both WGD (3654 species covering 32 bacterial phyla) and MGD (25 habitat-subtypes covering seven habitats) were evaluated to have good representativeness and coverage of known ARGs that reached a plateau of 2625 and 3821 unique ARGs, respectively, in rarefaction curves. Thus, they could serve as good resources to investigate the phylogenetic and ecological distribution of ARGs. All ARGs were identified and annotated using a standardized pipeline for equal comparison among different samples. An online searching platform, ARGs-OSP, was designed to publish all results obtained in this study, making the data easily accessible to other researchers without unnecessary re-computations. ARGs-OSP could provide a valuable resources for versatile research interests.

By analysing the profile of the antibiotic resistome, we found that >80% of ARGs in the SARG database are less likely to be mobile, as they were not found in any plasmid or integron collected in WGD. Mobile ARGs conferring tetracycline and beta-lactam resistance were detected in multiple human pathogens

across bacterial phyla, indicating their importance for future research and intervention. Moreover, both total ARGs and *intI1* genes show little correlation to the degree of anthropogenicity. These observations highlighted the need to differentiate ARGs based on properties indicating their potential risk to human health, such as gene mobility, correlation to anthropogenicity, host range and host pathogenicity, to help prioritize ARGs of clinical relevance and importance. However, although in this study, clinical *intI1* genes were not distinguished from environmental variants of *intI1* genes, sequence variance between clinical and environmental *intI1* genes should be considered in future studies.

The major limitation of this study lay in the database of ARGs and the datasets of WGD and MGD. No ARGs were found among a wide lineage of the environmental microbiome (80% of bacterial species). Although WGD is populated by bacterial species of research interest and medical importance, almost 30% of ARGs in SARG were only detected in MGD, but not WGD, indicating that the current collection of bacterial genomes could not fully represent the environmental microbiome. Plasmid assembly from short read sequences is often problematic, therefore investigating the association of ARGs on plasmids is problematic in both MGD and WGD (Blau *et al.* 2018; Reid *et al.* 2020). MGD mainly covers the representative habitat-types, but not extreme environments. Moreover, ARGs detected in MGD largely represent the natural resistome of native bacterial communities. In contrast, the analysis of WGD has numerous advantages and would allow many questions to be addressed, such as the co-occurrence of ARGs with metal resistance genes (Li, Xia and Zhang 2017), *intI1* genes (Zhang *et al.* 2018) and MGEs (Reid *et al.* 2020). However, with the decreasing cost of sequencing, new datasets of high quality and massive quantity will facilitate the continual expansion of WGD and MGD. In addition, the development of new technologies such as long-read sequencing and single-cell sequencing will also supply genomes, plasmids and metagenomes of high quality. Periodic updating and flexible visualization are expected in ARGs-OSP for more convenient and versatile usage in future studies.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

ACKNOWLEDGMENTS

A.N.Z. acknowledges the University of Hong Kong for a postgraduate studentship. C.J.H. would like to thank University of Hong Kong for a research fellowship. The authors appreciate the help of Miss Vicky Fung and Miss Lilian Y.L. CHAN for technical assistance with the High Performance Computing & Grid Computing System.

FUNDING

The authors would like to thank National Key R&D Program of China (grant No. 2018YFC0310600) and Hong Kong Theme-based Research Scheme (T21-711/16-R) for financial support.

AUTHOR CONTRIBUTIONS

A.N.Z. and T.Z. developed this research project. A.N.Z. downloaded the datasets, analysed the data, designed the platform and wrote the manuscript. C.J.H. and M.N. constructed the platform. L.L.G. provided suggestions in the data analysis and manuscript preparation. T.Z. guided webpage development and revised the manuscript.

REFERENCES

Alonso A, Sanchez P, Martinez JL. Environmental selection of antibiotic resistance genes. *Environ Microbiol* 2001;**3**:1–9.

Aust M-O, Godlinski F, Travis GR *et al.* Distribution of sulfamethazine, chlortetracycline and tylosin in manure and soil of Canadian feedlots after subtherapeutic use in cattle. *Environ Pollut* 2008;**156**:1243–51.

Beaber JW, Hochhut B, Waldor MK. SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* 2004;**427**:72.

Bengtsson-Palme J, Angelin M, Huss M *et al.* The human gut microbiome as a transporter of antibiotic resistance genes between continents. *Antimicrob Agents Chemother* 2015;**59**:6551–60.

Bengtsson-Palme J. Antibiotic resistance in the food supply chain: where can sequencing and metagenomics aid risk assessment? *Curr Opin Food Sci* 2017;**14**:66–71.

Berendonk TU, Manaia CM, Merlin C *et al.* Tackling antibiotic resistance: the environmental framework. *Nat Rev Microbiol* 2015;**13**:310–7.

Blau K, Bettermann A, Jechalke S *et al.* The transferable resistome of produce. *MBio* 2018;**9**:e01300–18.

Brooks L, Kaze M, Sistrom M. A Curated, Comprehensive Database of Plasmid Sequences. Dunning Hotopp JC (ed.). *Microbiol Resour Announc* 2019;**8**:e01325–18.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59.

Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

Carlos C, Pires MM, Stoppe NC *et al.* Escherichia coli phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiol* 2010;**10**:161.

Chu BT, Petrovich ML, Chaudhary A *et al.* Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. *Appl Environ Microbiol* 2018;**84**:e02168–17.

Colwell RK, Chao A, Gotelli NJ *et al.* Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* 2012;**5**:3–21.

D'Costa VM, King CE, Kalan L *et al.* Antibiotic resistance is ancient. *Nature* 2011;**477**:457.

Deng C, Daley T, Smith AD. Applications of species accumulation curves in large-scale biological data analysis. *Quant Biol* 2015;**3**:135–44.

Du J, Ren H, Geng J *et al.* Occurrence and abundance of tetracycline, sulfonamide resistance genes, and class 1 integron in five wastewater treatment plants. *Environ Sci Pollut Res* 2014;**21**:7276–84.

Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.

Gillings MR, Gaze WH, Pruden A *et al.* Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *ISME J* 2015;**9**:1269–79.

Gillings MR. DNA as a pollutant: the clinical class 1 integron. *Current Pollution Reports* 2018;**4**:49–55.

Gonzalez-Plaza JJ, Šimatović A, Milaković M *et al*. Functional repertoire of antibiotic resistance genes in antibiotic manufacturing effluents and receiving freshwater sediments. *Front Microbiol* 2017;**8**:2675.

Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* 2001;**4**:379–91.

Gupta CL, Tiwari RK, Cytryn E. Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes. *Environ Int* 2020;**138**:105667.

Hu X, Zhou Q, Luo Y. Occurrence and source analysis of typical veterinary antibiotics in manure, soil, vegetables and groundwater from organic vegetable bases, northern China. *Environ Pollut* 2010;**158**:2992–8.

Jechalke S, Broszat M, Lang F *et al*. Effects of 100 years wastewater irrigation on resistance genes, class 1 integrons and IncP-1 plasmids in Mexican soil. *Frontiers in Microbiology* 2015;**6**:163.

Jechalke S, Schreiter S, Wolters B *et al*. Widespread dissemination of class 1 integron components in soils and related ecosystems as revealed by cultivation-independent analysis. *Frontiers in Microbiology* 2014;**4**:420.

Jia B, Raphenya AR, Alcock B *et al*. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2016;**45(D1)**:D566-73.

Ji X, Shen Q, Liu F *et al*. Antibiotic resistance gene abundances associated with antibiotics and heavy metals in animal manures and agricultural soils adjacent to feedlots in Shanghai; China. *J Hazard Mater* 2012;**235**:178–85.

Karkman A, Pärnänen K, Larsson DJ. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat Commun* 2019;**10**:1–8.

Khan GA, Berglund B, Khan KM *et al*. Occurrence and abundance of antibiotics and resistance genes in rivers, canal and near drug formulation facilities–a study in Pakistan. *PLoS One* 2013;**8**:e62712.

Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;**46**:e35.

Lanza VF, Baquero F, Martínez JL *et al*. In-depth resistome analysis by targeted metagenomics. *Microbiome* 2018;**6**:11.

Lau CH-F, van Engelen K, Gordon S *et al*. Novel antibiotic resistance determinants from agricultural soil exposed to antibiotics widely used in human medicine and animal farming. *Appl Environ Microbial* 2017;**83**:e00989–17.

Lehmann K, Bell T, Bowes MJ *et al*. Trace levels of sewage effluent are sufficient to increase class 1 integron prevalence in freshwater biofilms without changing the core community. *Water Res* 2016;**106**:163–70.

Li B, Yang Y, Ma L *et al*. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* 2015;**9**:2490–502.

Li L-G, Xia Y, Zhang T. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. *ISME J* 2017;**11**:651–62.

Liu B, Pop M. ARDB—antibiotic resistance genes database. *Nucleic Acids Res* 2008;**37**:D443–7.

Lv B, Cui Y, Tian W *et al*. 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. *Ecotoxicol Environ Saf* 2018;**157**:169–75.

Ma L, Li A-D, Yin X-L *et al*. The Prevalence of Integrons as the Carrier of Antibiotic Resistance Genes in Natural and Man-Made Environments. *Environ Sci Technol* 2017a;**51**:5721–8.

Ma L, Li B, Jiang X-T *et al*. Catalogue of antibiotic resistome and host-tracking in drinking water deciphered by a large scale survey. *Microbiome* 2017b;**5**:154.

Martinez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. *Nat Rev Microbiol* 2015;**13**:116–23.

Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015;**16**:51.

Nesme J, Cécillon S, Delmont TO *et al*. Large-scale metagenomic-based study of antibiotic resistance in the environment. *Curr Biol* 2014;**24**:1096–100.

Nishino K, Yamaguchi A. Analysis of a complete library of putative drug transporter genes in Escherichia coli. *J Bacteriol* 2001;**183**:5803–12.

Paez-Espino D, Chen I-MA, Palaniappan K *et al*. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res* 2016; **45(D1)**:D457-65.

Pal C, Bengtsson-Palme J, Kristiansson E *et al*. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics* 2015;**16**:964.

Pal C, Bengtsson-Palme J, Kristiansson E *et al*. The structure and diversity of human, animal and environmental resistomes. *Microbiome* 2016;**4**:54.

Parks DH, Rinke C, Chuvochina M *et al*. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;**2**:1533–42.

Piddock LJ. Multidrug-resistance efflux pumps? not just for resistance. *Nat Rev Microbiol* 2006;**4**:629.

Pruden A, Arabi M, Storteboom HN. Correlation between upstream human activities and riverine antibiotic resistance genes. *Environ Sci Technol* 2012;**46**:11541–9.

Rascovan N, Telke A, Raoult D *et al*. Exploring divergent antibiotic resistance genes in ancient metagenomes and discovery of a novel beta-lactamase family. *Environ Microbiol Rep* 2016;**8**: 886–95.

Reid CJ, Blau K, Jechalke S *et al*. Whole Genome Sequencing of Escherichia coli From Store-Bought Produce. *Frontiers in Microbiology* 2020;**10**:3050.

Rodriguez-Mozaz S, Chamorro S, Marti E *et al*. Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. *Water Res* 2015;**69**:234–42.

Ruppé E, Ghozlane A, Tap J *et al*. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature microbiology* 2019;**4**:112–23.

Savichtcheva O, Okayama N, Okabe S. Relationships between Bacteroides 16S rRNA genetic markers and presence of bacterial enteric pathogens and conventional fecal indicators. *Water Res* 2007;**41**:3615–28.

Sayers EW, Barrett T, Benson DA *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2012;**40**:D13–25.

Seiler C, Berendonk TU. Heavy metal driven co-selection of antibiotic resistance in soil and water bodies impacted by agriculture and aquaculture. *Front Microbiol* 2012;**3**: 399.

Shannon P, Markiel A, Ozier O *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.

Tadesse DA, Zhao S, Tong E *et al*. Antimicrobial drug resistance in Escherichia coli from humans and food animals, United States, 1950–2002. *Emerg Infect Dis* 2012;**18**:741.

Tang J, Bu Y, Zhang X-X *et al*. Metagenomic analysis of bacterial community composition and antibiotic resistance genes in a wastewater treatment plant and its receiving surface water. *Ecotoxicol Environ Saf* 2016;**132**:260–9.

Team RC. *R: A Language and Environment for Statistical Computing*. 2013.

Thompson LR, Sanders JG, McDonald D *et al*. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63.

Vaz-Moreira I, Nunes OC, Manaia CM. Bacterial diversity and antibiotic resistance in water habitats: searching the links with the human microbiome. *FEMS Microbiol Rev* 2014;**38**:761–78.

Wilke A, Bischof J, Gerlach W *et al*. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res* 2015;**44**:D590–4.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:R46.

Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. *Emerg Infect Dis* 2005;**11**:1842.

Yang Y, Jiang X, Chai B *et al*. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics* 2016;**32**:2346–51.

Yang Y, Xu C, Cao X *et al*. Antibiotic resistance genes in surface water of eutrophic urban lakes are related to heavy metals, antibiotics, lake morphology and anthropic impact. *Ecotoxicology* 2017;**26**:831–40.

Yin X, Jiang X-T, Chai B *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.

Zhang A-N, Li L-G, Ma L *et al*. Conserved phylogenetic distribution and limited antibiotic resistance of class 1 integrons revealed by assessing the bacterial genome and plasmid collection. *Microbiome* 2018;**6**:130.