A curated data resource of 214K metagenomes for characterization of the global resistome

Short title: Abundance data of 214K metagenomes of the global resistome

Authors:

Hannah-Marie Martiny[*1],

Patrick Munk[1],

Christian Brinch[1],

Frank M. Aarestrup[1],

Thomas N. Petersen[1]

Affiliations:

[1]Research Group for Genomic Epidemiology, Technical University of Denmark, Kongens Lyngby, Denmark

[*] Corresponding author

E-mail: hanmar@food.dtu.dk (HMM)

Abstract

The growing threat of antimicrobial resistance (AMR) calls for new epidemiological surveillance methods, as well as a deeper understanding of how antimicrobial resistance genes (ARGs) have transmitted around the world. The large pool of sequencing data available in public repositories provides an excellent resource for monitoring the temporal and spatial dissemination of AMR in different ecological settings. However, only a limited number of research groups globally have the computational resources allowing analyses of such data. We retrieved 442 Tbp of sequencing reads from 214,095 metagenomic samples from the European Nucleotide Archive (ENA) and aligned them using a uniform approach against ARGs and 16S/18S rRNA genes. Here, we present the results of this extensive computational analysis and share the counts of reads aligned. Over $6.76 \cdot 10^8$ read fragments were assigned to ARGs and $3.21 \cdot 10^9$ to rRNA genes, where we observed distinct differences in both the abundance of ARGs and the link between microbiome and resistome compositions across various sampling types. This collection is another step towards establishing a global surveillance of AMR and can serve as a resource for further research into the environmental spread and dynamic changes of ARGs.

Introduction

The vast amount of genomic data available in public data repositories is a unique and potentially important resource for doing research and genomic surveillance of antimicrobial resistance (AMR). Using these datasets collected from locations all over the world across different years and from various sampling sources might further aid our understanding of the emergence and distribution of antimicrobial resistance genes (ARGs).

The sharing of genomic sequence data to one of the available repositories is today a major and often mandatory step in peer-reviewed journals, for which several repositories were created by the members of the International Nucleotide Sequence Database Collaboration (INSDC)[1], including the European Nucleotide Archive (ENA)[2]. The number of sequencing data available at ENA continues to increase with an estimated doubling time of 18 months (https://www.ebi.ac.uk/ena/browser/about/statistics, accessed 2022-03-08).

Several approaches for how to analyze genomic data depending on the sample types are already well established.

However, the exploration of these resources are often restricted to a few research groups only, since both sufficient skills in bioinformatics and access to high-performing computer resources are needed to handle the large amount of available data.

Existing collections of analyzed datasets tend to focus on either specific sample sources, such as humans[3,4], marine[5], or urban sewage[6,7], or focus on specific genera[8]. Especially the COVID19 pandemic has highlighted the value of data sharing to trace the spread and evolution of the virus[9]. Despite the attempts to standardize the analysis workflows of these databases, they are limited in their ability to generalize across environments and locations.

A recent study[10] has shared a searchable collection of 661K bacterial genomes for exploring the global bacterial diversity across different origins, providing an easy-to-access resource for genomic research. While this is an impressive data-sharing effort, the authors did not include metagenomic samples in their pipeline. Metagenomic techniques aim to sequence all DNA in a sample and can be used to characterize the microbiome in different environments[11,12], discover novel organisms[13], monitor disease[14,15] and specific genes, such as ARGs[5,6,16].

Here, we present a large-scale metagenomic analysis of 214,095 metagenomic samples retrieved from ENA. We have carried out an assembly-free approach by aligning sequencing reads against ARGs and 16S/18S ribosomal RNA genes. We have previously published an in-depth analysis of the distribution of mobilized colistin resistance[17] based on those data. Now we both share the entire collection of mapping results and showcase how to characterize the global resistome and microbiome with this dataset. The curated metadata and mapping results are available at https://doi.org/10.5281/zenodo.6519844 and documentation at https://hmmartiny.github.io/mARG/Tables.html.

Materials and Methods

Retrieval of metagenomes

We retrieved metagenomic datasets from the European Nucleotide Archive (ENA)[2] uploaded between 2010-01-01 and 2020-01-01 that had library source as 'METAGEOMIC'

and library strategy of 'WGS'. We collected 214,095 sequencing runs from 146,732 samples from 6,307 projects corresponding to 442 Tbp of raw reads taking up 300 TB of storage. The associated metadata for each sample was also retrieved.

Preprocessing and mapping of sequencing reads

The retrieved raw FASTQ reads were trimmed and aligned against reference sequences, as outlined in Martiny (2022)[17]. In brief, we used FASTQC v.0.11.15 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for read quality checking and BBduk2 v.36.49[18] for trimming the raw sequencing reads. With the k-mer based alignment tool KMA 1.2.21[19], the trimmed reads were mapped against reference sequences from two different databases: The AMR gene database ResFinder20 (downloaded 25-01-2020), which contained 3,085 sequences, and the ribosomal rRNA Silva21 gene database (version 138, downloaded 16-01-2020), which had 2,225,272 template sequences with more than 88% of them being 16/18S rRNA genes. Data retrieval, quality checking, trimming, and read alignments were done using the Danish National Supercomputer for Life Sciences (https://www.computerome.dk/).

Standardization of metadata

The following attributes for each metagenome were standardized: sampling location, sampling host or environment (referred to as a host below), and sampling date.

To standardize the label for sampling locations, we looked at the values entered in the two fields ´country´ and ´location´. First, the latitude and longitude coordinates were mapped to a country using the Python library Shapely 1.7.1[22] to find the matching area defined in one of the three public domain map datasets (countries, marine, and lakes) available in the Natural Earth Data collection. If the lookup failed or the coordinates were not given, the second step was to match the text attribute in the country label to ISO 3166 country codes with a fuzzy search with the Python library PyCountry 20.7.3 (https://github.com/flyingcircusio/pycountry). Finally, if the two lookup searches did not yield a match, we did a manual lookup of the country labels to standardize the text.

For the standardization of host labels, we mapped the taxonomic id given by the attribute ´host_tax_id´ to the NCBI Taxonomy database[23], or if the feature was missing, the ´tax_id` was used instead.

Since the only way to curate entered collection dates is to look up suspicious dates in published studies manually, and that was deemed too time-intensive, we decided to replace dates entered as later than 2020-01-01 in the sample attribute field ´collection_date´ with the missing value NULL.

Measuring the abundance of ARGs

Since we report the fragment count aligned to each reference gene, the mapping results are compositional and should be treated as such[24]. In the simplest form, the ARG abundance for a sample or sample group can be calculated as the log-ratio of the count of reads, n_i, aligned to each ARG i over the total sum of rRNA read fragments n_B:

$$x = [n_1, n_2, \ldots, n_D, n_B] \ , \qquad i = 1..D$$

$$\text{Abundance}(x) = \left[\log\frac{n_1}{n_B}, \log\frac{n_2}{n_B}, \ldots, \log\frac{n_D}{n_B}\right]$$

where D is the number of ARGs and $n_B = \frac{\sum_j^{D_B} n_j}{1{\cdot}10^6}$ with D_B being the number of rRNA genes. Each ARG count n_i has been adjusted with the length of the gene in kilobases.

The relative abundance resistance classes were calculated as the proportion of ARG resistance assigned to different classes and scaled with $\kappa = 100$:

$$\text{Relative abundance}(x) = \frac{\kappa}{\sum n_i} n_i$$

Diversity measurements

Besides the read abundance values, we report the species richness, Shannon diversity index[25], and the Gini-Simpson[26] diversity index of read counts of ARGs, genera, and phyla per sample. Species richness is the number of different genes or taxonomic groups present in the sample with at least one read fragment aligned.

The Shannon index (H') were calculated using the proportions of reads $p_i = \frac{n_i}{\sum n}$:

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

whereas the Gini-Simpson index (GS) was calculated using the read counts $n = [n_1, \ldots, n_D]$ and $N = \sum n$ is the total count of reads for the group:

$$GS = 1 - \frac{\sum n_i \cdot (n_i - 1)}{N \cdot (N - 1)}$$

Together with these two indices, we also report the sample-wise unique number of templates or taxonomic groups matched.

Code and data availability

The code to produce the figures is available at https://github.com/hmmartiny/mARG. The data has been deposited at https://doi.org/10.5281/zenodo.6519844, and documentation of the various tables be accessed at https://hmmartiny.github.io/mARG.

Results

Here, we present a large-scale mapping of 442 Tbp of raw reads of 214,095 metagenomic samples suitable for analyzing the distribution of acquired antimicrobial resistance genes and 16S/18S rRNA genes. Furthermore, we have spent considerable effort standardizing three main sample attributes: sampling date, location, and source. To facilitate easy access and usage, we have shared the mapping results and corrected metadata in three different data formats (TSV, HDF, and MySQL dumps). We also provide tutorials with code examples in R and Python on using the data in different scenarios. Data files are all available at https://doi.org/10.5281/zenodo.6519844.

By collecting the sequencing reads from ENA, we could also verify the inherited bias of specific sample types or sources being overrepresented simply due to the availability in the public repository. While the 214,095 metagenomic datasets were collected from 797 different hosts, most were either of human or marine origin (Figure 1a). A similar skewed geographical distribution towards European and North American countries was observed in the sampling locations (Figure 1b). The distribution of samples according to the sampling year reveals that a considerable number were collected between 2010 and 2020 (Figure 1c).

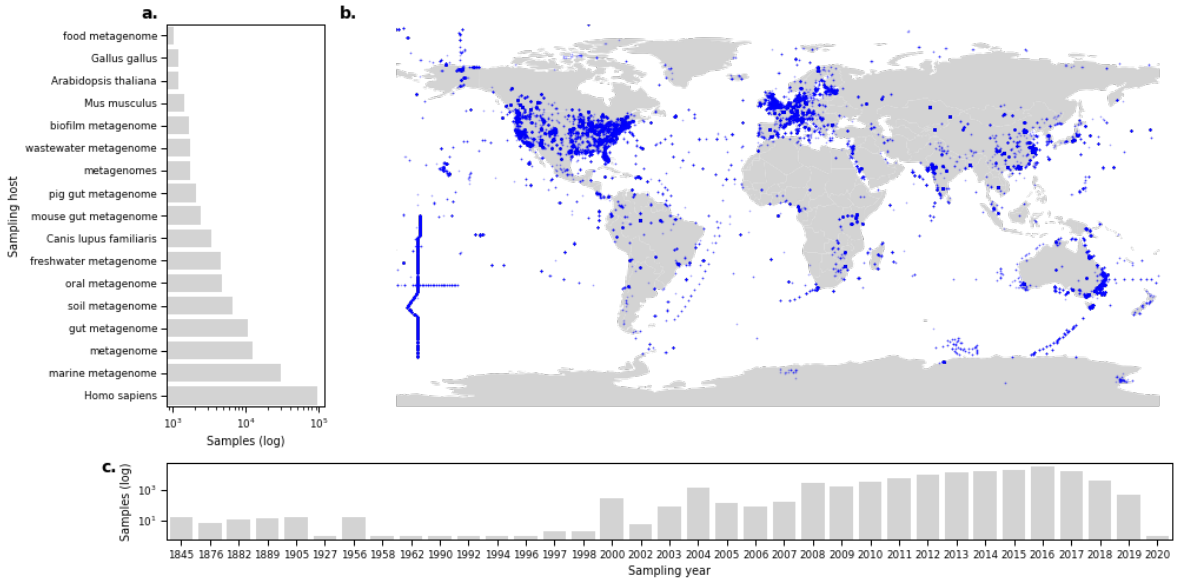

Figure 1: ***Distribution of metagenomes reveals the overrepresentation of samples from specific sources****.* ***a****. Number of samples grouped per sampling host, where only hosts with more than 1000 samples are plotted.* ***b****. Sample locations for metagenomes with available GPS coordinates; each marker is a sample.* ***c.*** *Year of which a sample was collected.*

Of the more than $1.8 \cdot 10^{12}$ raw sequencing reads, corresponding to 442.1 Tbp, 93% of the reads were generated using Illumina sequencing technologies (Figure S1). We mapped over $1.69 \cdot 10^{12}$ trimmed read fragments, with a median of 784,748 fragments per sample (range 1 – 916,901,400) (Figure 2a). 0.04% of all read fragments could be aligned to ARGs, and 0.19% to rRNA genes. The number of ARG fragments aligned increased with the number of aligned rRNA fragments, although for 34% of the samples, we did not find any ARGs despite having read fragments aligning to 16S rRNA genes (Figure 2b). The microbial differences in the different sampling origins were highlighted in the number of aligned fragments (Figure S3).

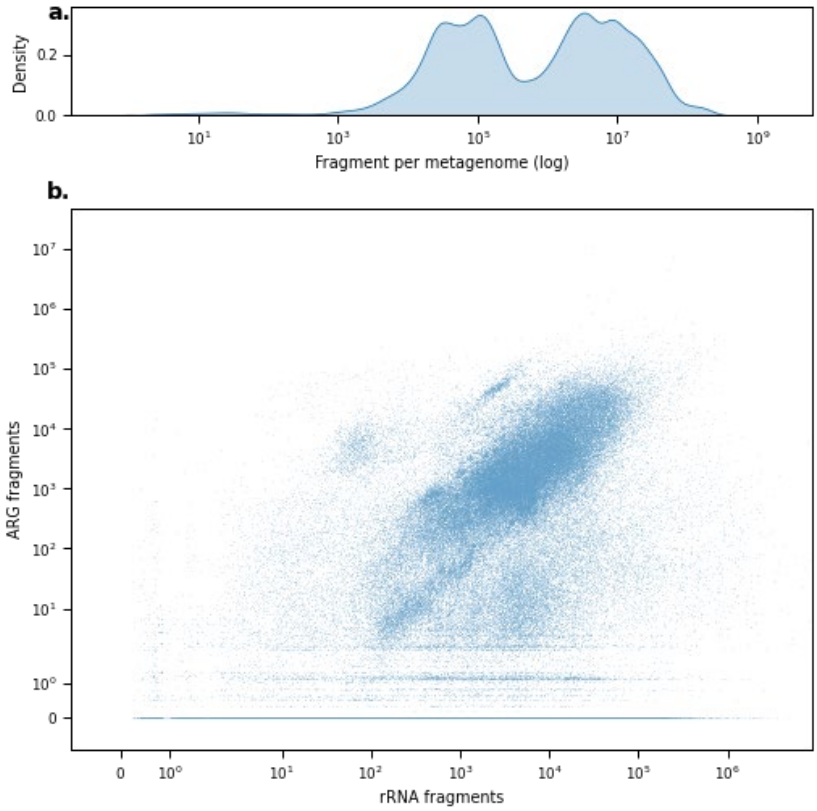

Figure 2: ***Distribution of available and aligned fragments. a.*** *Density distribution of available fragments per sample.* ***b.*** *The distribution compares the number of fragments mapped to rRNA genes and ARGs.*

The global abundance of antimicrobial resistance

To measure the global distribution of ARGs and the composition of the resistome, we calculated the abundance of ARGs as the log-ratio of ARG fragments over summed rRNA sequence fragments. Almost all of the template sequences from the ResFinder database had at least one fragment aligned, and only 94 ARGs had no hits (Figure S2). The median observed resistance load per metagenomic sample was 11.74 (log range: -1.45 to 23.52) (Figure 3a), which appeared to be mainly dependent on the geographic origin and environment (Figure 3b-d) and not on which year the sample was taken. For example, samples originating from locations within Europe showed similar abundance levels for most of the samples but with several outliers, whereas multiple samples from locations in the Oceania region had a much broader load distribution with few outliers (Figure 3c).

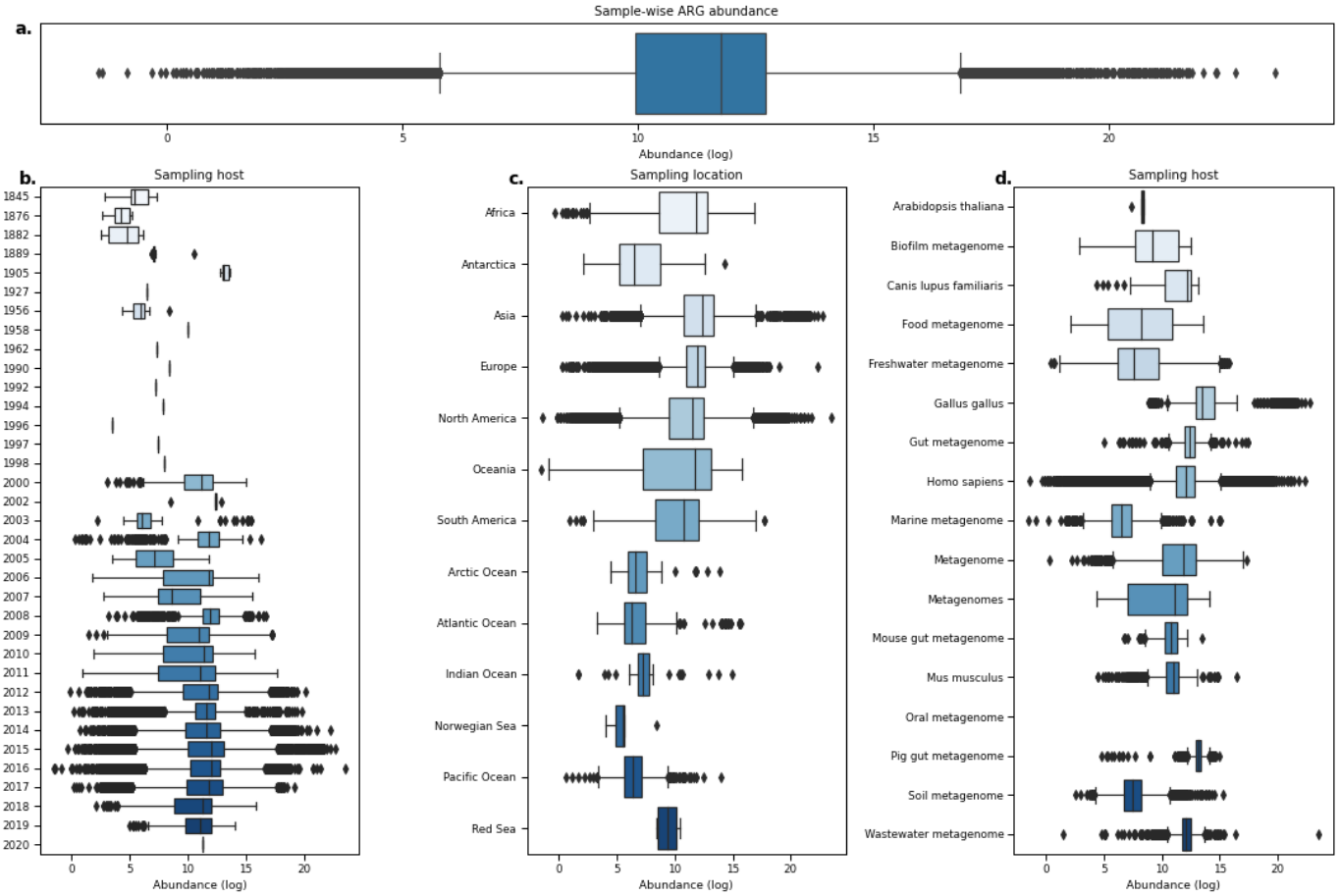

Figure 3: ***Boxplots of ARG abundances in metagenomic samples show that levels vary across different origins.*** *a. Distribution of ARG abundance per sample. b. Distribution of sample-wise ARG abundance grouped by sampling year. c. Sample-wise ARG abundance per sampling location. d. Sample-wise ARG abundance grouped by hosts. Only hosts with more than 1000 metagenomes analyzed are shown.*

While the distribution of sample-wise resistance loads illustrates the high variability in this data collection (Figure 3), we saw that once we stratified the relative ARG read proportions per resistance class and sample type, there were clear separations between different groups (Figure 4). For the sampling years with a considerable number of samples available (2004-2019), the relative proportion of classes was relatively consistent, with Tetracycline reads being the most common, except for a spike of Beta-lactam reads in 2017 (Figure 4a). When looking at the geographic regions, we observed that reads stemming from samples collected from large water bodies had more reads aligned to Aminoglycosides and Beta-lactam classes than land regions, which had more diverse class distributions (Figure 4b), possibly due to that samples from land regions had higher resistance loads overall (Figure 3a). Once we stratified by sampling host or source, the distribution of resistance classes was very dependent on the group, as seen by the high proportion of read fragments aligned to, for

example, Phenicol for marine and soil samples and Tetracycline reads being highly prevalent in mice (*Mus musculus*) samples (Figure 4c).

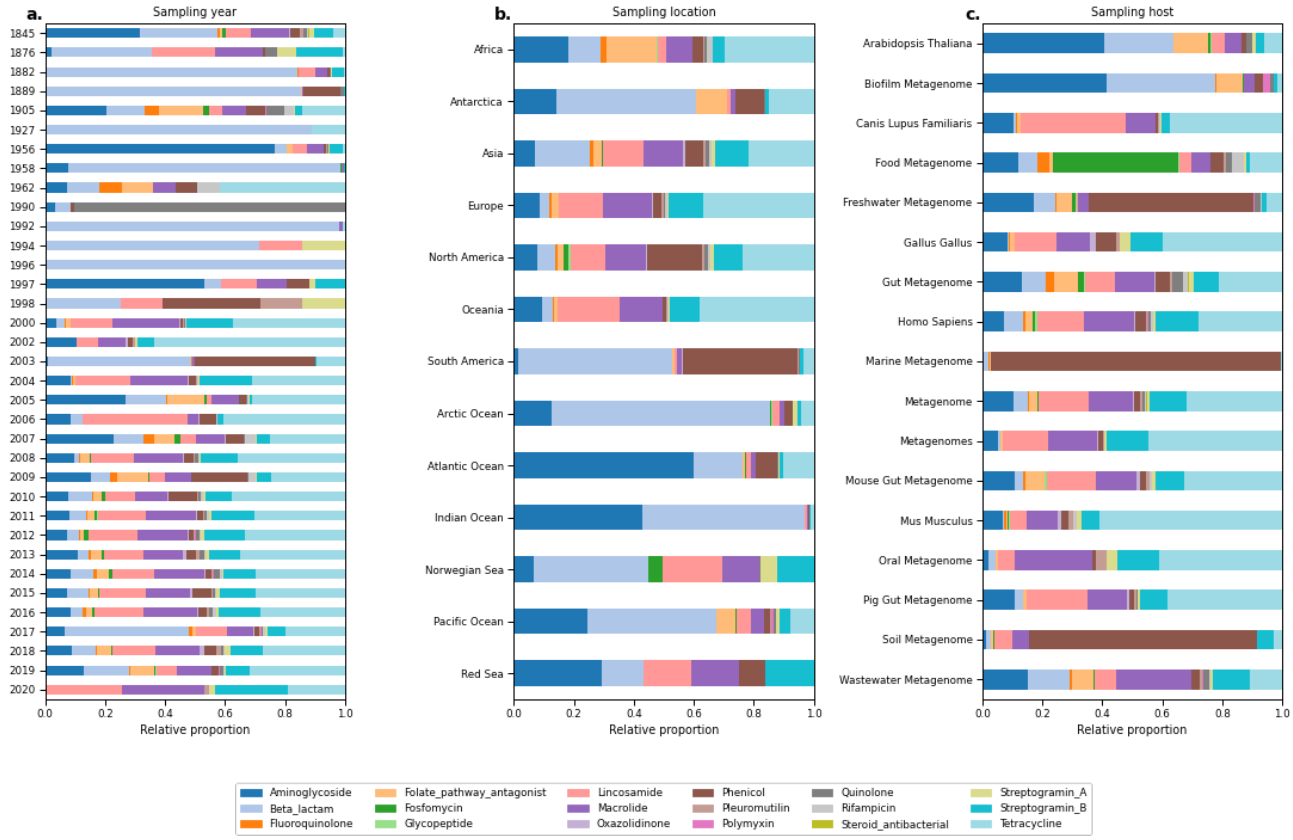

Figure 4: ***Composition of reads assigned to ARGs from different resistance classes grouped by sampling origin. a.*** *Grouped by sampling year.* ***b.*** *Grouped per sampling location.* ***c.*** *Grouped per sampling host. Only hosts with more than 1000 metagenomes analyzed are shown.*

Linking the microbiome diversity with resistance diversity

The relationship between the diversity of the microbiome and the resistance genes was quantified by calculating the species richness and two alpha diversity measurements (Shannon and Gini-Simpson) on ARG levels and phyla and genera taxonomic levels. We saw a general trend of increased diversity of the microbiome also meant an increased ARG diversity (Figure 5, Figure S4). Although, the relationship between genera and ARG diversity indexes further characterized by sampling sources revealed a higher differentiation, suggesting that increased diversity of microbes in, for example, soil samples does not necessarily lead to a higher diversity of resistance genes. Contrarily, the chicken (*Gallus gallus*) samples showed that they still had elevated ARG diversity despite having lower microbial diversity (Figure 5).

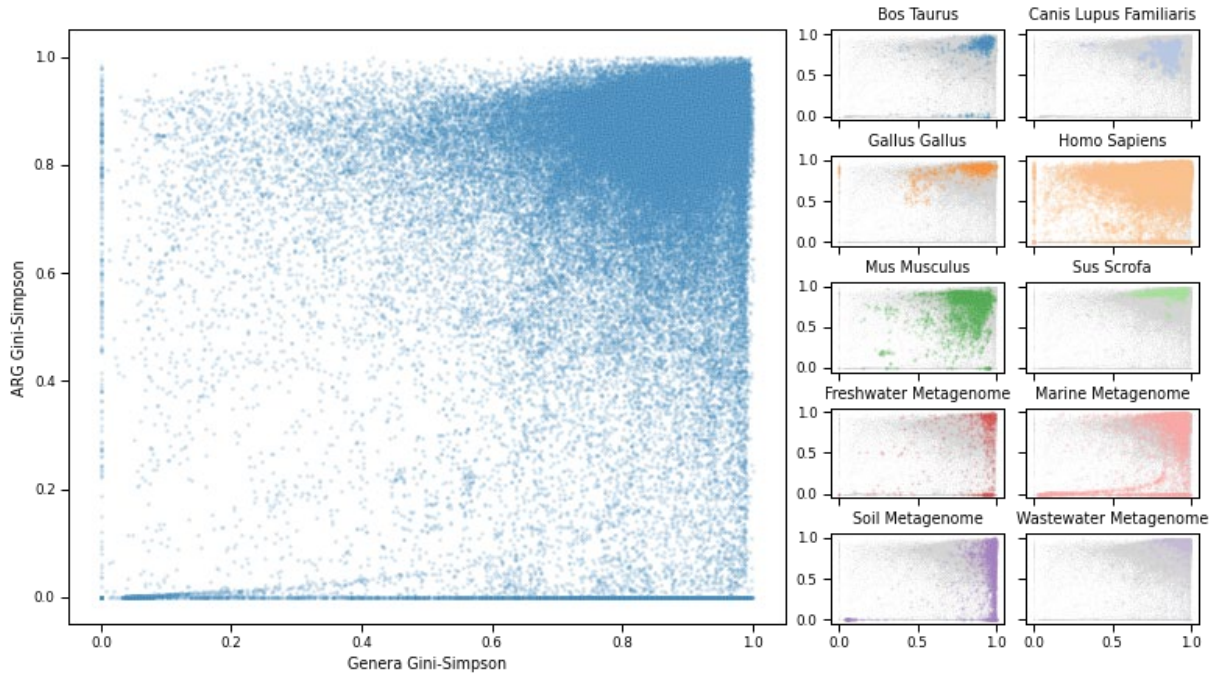

*Figure 5: **The genus-ARG diversity relationship for all metagenomic samples**. The Gini-Simpson diversity indexes were calculated on genus categories (x-axis) compared to ARG levels (y-axis). Left: Scatterplot of all samples. Right: Samples colored by selected host or environmental origins.*

Discussion

Global surveillance of AMR based on genomics continues to become more accessible due to the advancement in NGS technologies and the practice of sharing raw sequencing data in public repositories. Standardized pipelines and databases are needed to utilize these large data volumes for tracking the dissemination of AMR. We have uniformly processed the sequencing reads of 214,095 metagenomes for the abundance analysis of ARGs.

Our data sharing efforts enable users to perform abundance analyses of individual ARGs, the resistome, and the microbiome across different environments, geographic locations, and sampling years.

We have given a brief characterization of the distribution of ARGs according to the collection of metagenomes. However, in-depth analyses remain to be performed to investigate the influence of temporal, geographical, and environmental origins on the dissemination and evolution of antimicrobial resistance. For example, analyzing the spread of specific ARGs across locations and different environments could reveal new transmission routes of resistance and guide the design of intervention strategies to stop the spread. Another use of the data collection could be to explore how the changes in microbial abundances affect and are affected by the resistome. Furthermore, our coverage statistics

of reads aligned to ARGs could be used to investigate the rate of new variants occurring in different reservoirs. Even though we have focused on the threat of antimicrobial resistance, potential applications of this resource can be to look at the effects of, e.g., climate changes on microbial compositions.

We recommend that potential users consider all the confounders present in this data collection in their statistical tests and modeling workflows, emphasizing that the experimental methods and sequencing platforms dictate the obtained sequencing reads. Furthermore, it is important to consider the compositional nature of microbiomes[27]. The reads do not depend on the distribution of genetic material in the sample but the capacity of the sequencing platform[24,28]. Various statistical methods already exist that consider the compositionality[24,29,30].

With this data resource, we have taken a step towards enabling the scientific community to utilize the wealth of information in these metagenomic samples to broaden our understanding of the dissemination of antimicrobial resistance and changes in microbiomes at both local and global scales through time and environments.

References

1. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121 (2021).
2. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* **39**, 44–47 (2011).
3. Shao, L., Liao, J., Qian, J., Chen, W. & Fan, X. MetaGeneBank: a standardized database to study deep sequenced metagenomic data from human fecal specimen. *BMC Microbiol.* **21**, 1–12 (2021).
4. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
5. Cuadrat, R. R. C., Sorokina, M., Andrade, B. G., Goris, T. & Dávila, A. M. R. Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples. *Gigascience* **9**, 1–12 (2020).
6. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on

metagenomics analyses of urban sewage. *Nat. Commun.* (2019). doi:10.1038/s41467-019-08853-3

7. Fresia, P. *et al.* Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. *Microbiome* **7**, 1–9 (2019).
8. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
9. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly.* **3**, 1049–1051 (2021).
10. Blackwell, G. A. *et al.* Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biol.* **19**, e3001421 (2021).
11. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21390–21395 (2012).
12. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science (80-.).* **312**, 1355–1359 (2006).
13. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
14. Nieuwenhuijse, D. F. *et al.* Setting a baseline for global urban virome surveillance in sewage. *Sci. Rep.* **10**, 1–13 (2020).
15. Liu, P., Chen, W. & Chen, J. P. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica). *Viruses 2019, Vol. 11, Page 979* **11**, 979 (2019).
16. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509**, 612–616 (2014).
17. Martiny, H.-M. *et al.* Global distribution of mcr gene variants in 214,095 metagenomic samples. *mSystems* (2022). doi:10.1128/msystems.00105-22
18. Bushnell, B. BBMap. (2014).
19. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 1–8 (2018).
20. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
21. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data

processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).

22. Gillies, S. & Others, A. Shapely: manipulation and analysis of geometric objects. (2007).

23. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).

24. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).

25. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

26. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).

27. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **44**, 139–160 (1982).

28. Quinn, T. P. *et al.* A field guide for the compositional analysis of any-omics data. *Gigascience* **8**, 1–14 (2019).

29. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS One* **8**, (2013).

30. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8**, 1–11 (2012).

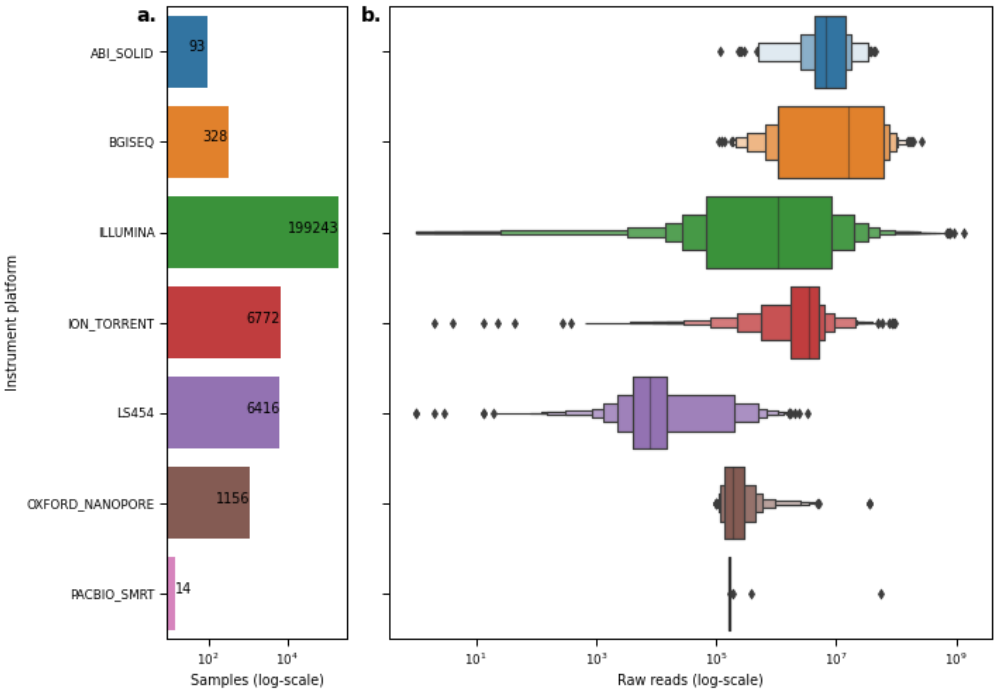

Figure S1: ***Distribution of samples per sequencing instrument platform****.* ***a.*** *Sample count per platform.* ***b.*** *Distribution of raw sequencing read counts per platform.*

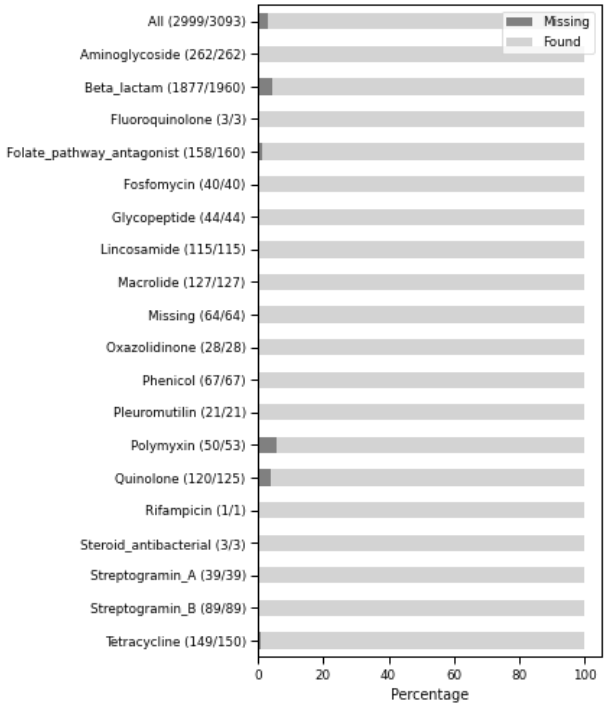

Figure S2: ***More than 96% of ARG templates had at least one aligned fragment.*** *The bars illustrate the percentage of ARGs per resistance class without and with at least one aligned fragment. The parenthesis after each class label contains the number of genes found out of the total available templates.*

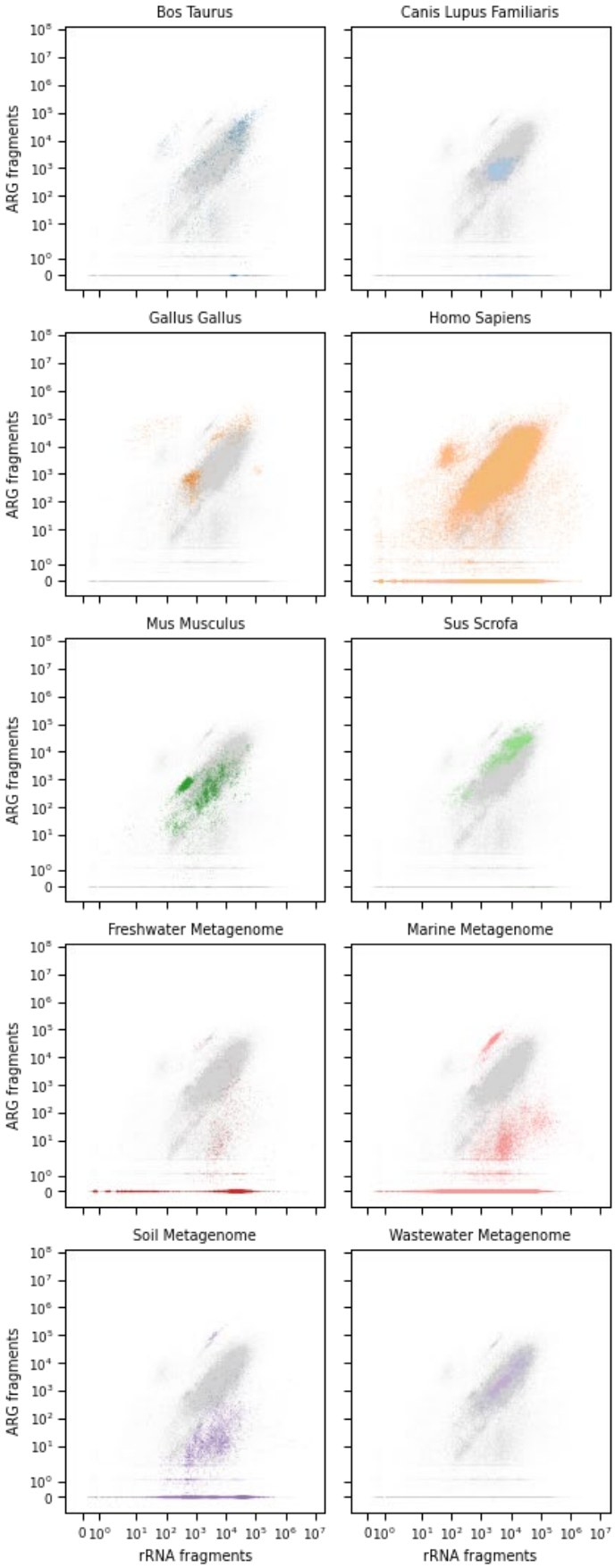

Figure S3: ***The sample-wise distribution of aligned rRNA fragments and ARG*** *fragments, colored by selected host and environmental sources.*

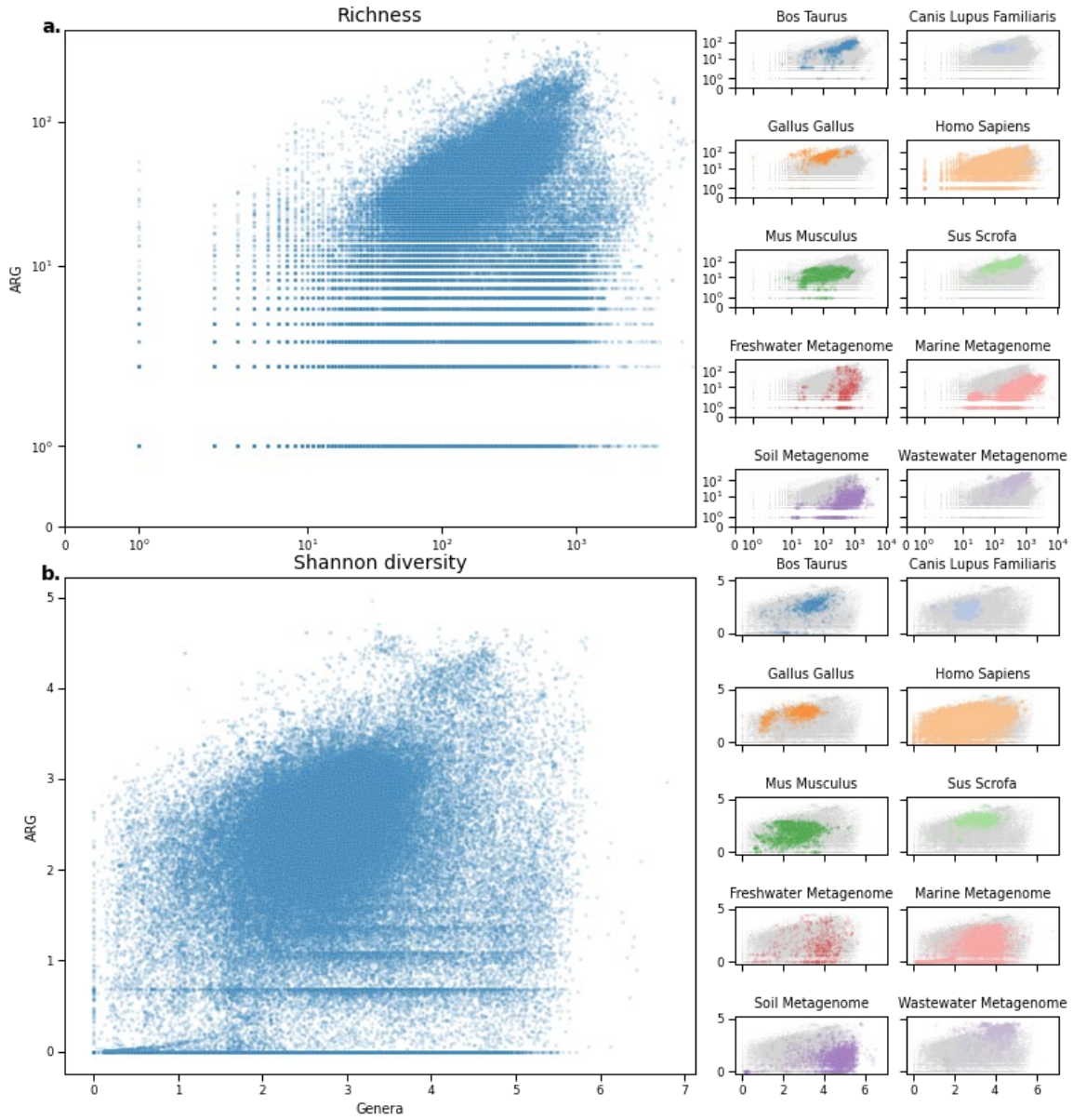

Figure S4: ***Additional distributions showing the relationship between ARGs and genera for all metagenomic samples. a.*** *The richness of genus groups (x-axis) vs. ARG richness (y-axis).* ***b.*** *The relationship between Shannon diversity index calculated on genus level (x-axis) and ARGs(y-axis). Right: Samples colored by selected host or environmental origins.*