nature communications

Article

https://doi.org/10.1038/s41467-024-52450-y

Prophage-encoded antibiotic resistance genes are enriched in human-impacted environments

Received: 24 October 2023

Accepted: 7 September 2024

Published online: 27 September 2024

Check for updates

Hanpeng Liao[1,12], Chen Liu[1,12], Shungui Zhou[1] ✉, Chunqin Liu[1], David J. Eldridge[2], Chaofan Ai[1], Steven W. Wilhelm[3], Brajesh K. Singh[4], Xiaolong Liang[5], Mark Radosevich[6], Qiu-e Yang[1], Xiang Tang[1], Zhong Wei[7], Ville-Petri Friman[8], Michael Gillings[9], Manuel Delgado-Baquerizo[10] ✉ & Yong-guan Zhu[11] ✉

The spread of antibiotic resistance genes (ARGs) poses a substantial threat to human health. Phage-mediated transduction could exacerbate ARG transmission. While several case studies exist, it is yet unclear to what extent phages encode and mobilize ARGs at the global scale and whether human impacts play a role in this across different habitats. Here, we combine 38,605 bacterial genomes, 1432 metagenomes, and 1186 metatranscriptomes across 12 contrasting habitats to explore the distribution of prophages and their cargo ARGs in natural and human-impacted environments. Worldwide, we observe a significant increase in the abundance, diversity, and activity of prophage-encoded ARGs in human-impacted habitats linked with relatively higher risk of past antibiotic exposure. This effect was driven by phage-encoded cargo ARGs that could be mobilized to provide increased resistance in heterologous *E. coli* host for a subset of analyzed strains. Our findings suggest that human activities have altered bacteria-phage interactions, enriching ARGs in prophages and making ARGs more mobile across habitats globally.

Viruses are ubiquitous in the biosphere, playing critical ecological roles owing to their high abundance and diversity[1,2]. Phages (viruses infecting prokaryotes) have two main life cycles: lytic and lysogenic, which play distinctive roles in shaping the bacterial communities. Upon infection, lytic phages enter a productive replication cycle, promptly killing the host cell and exerting significant control on host population densities[3]. In addition to lytic cycle, lysogenic phages can integrate their DNA into the bacterial genome and enter prophage stage without causing host cell lysis[1]. While in prophage stage, integrated phages can expand the repertoire of functional genes available

[1]Fujian Provincial Key Laboratory of Soil Environmental Health and Regulation, College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou, China. [2]Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia. [3]Department of Microbiology, The University of Tennessee, Knoxville, TN, USA. [4]Global Centre for Land-Based Innovation, Western Sydney University, Penrith, NSW, Australia. [5]Key Laboratory of Pollution Ecology and Environmental Engineering, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, Liaoning Province, China. [6]Department of Biosystems Engineering and Soil Science, The University of Tennessee, Knoxville, TN, USA. [7]Jiangsu Provincial Key Lab of Solid Organic Waste Utilization, Nanjing Agricultural University, Nanjing, Jiangsu, China. [8]Department of Microbiology, University of Helsinki, Helsinki, Finland. [9]ARC Centre of Excellence in Synthetic Biology, Macquarie University, Sydney, NSW, Australia. [10]Laboratorio de Biodiversidad y Funcionamiento Ecosistémico, Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS), Consejo Superior de Investigaciones Científicas, Seville, Spain. [11]State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China. [12]These authors contributed equally: Hanpeng Liao, Chen Liu. ✉e-mail: sgzhou@fafu.edu.cn; m.delgado.baquerizo@csic.es; ygzhu@iue.ac.cn

to the host bacterial cells via lysogenic conversion, potentially enhancing the host fitness[4]. While several case studies have demonstrated such beneficial effects especially under different environmental stresses[5,6], we still poorly understand the distribution of prophage-encoded cargo genes at the pangenome level. Moreover, it is unclear if these patterns are shaped by human activities, such as antibiotic usage, that often create strong selection for bacterial survival and carriage of prophage-encoded antibiotic resistance genes (ARGs).

The fact that at least nearly half of sequenced bacterial genomes contain one or more prophages[7] suggest that prophages are likely to encode important ecological functions for bacteria in microbial communities[4,5]. In contrast to the exploitative relationship between lytic phage and their hosts, prophages can have a mutually beneficial symbiotic relationship with bacteria, where they enhance their host survival[8]. For example, prophages actively participate in host stress adaptation and elemental cycling by enhancing or altering host metabolic functions through the expression of cargo genes[4,9]. In addition, prophages can regulate the expression of host genes in marine bacteria, helping their hosts to adapt to the deep-sea environment[4]. Therefore, phage-host dynamics can serve as an indicator of ecological functions in response to environmental conditions they reside in refs. 1,10. From the clinical perspective, the presence of ARGs in prophages allows bacteria to persist and adapt to antibiotic exposure, contributing to the development of antibiotic-resistant strains[11,12] that cannot be controlled using traditional antibiotic treatments. While phage-host interactions in relation to ARG-carriage have been studied at the level of bacteria-phage pairs[13,14] we still lack a global understanding of the role of phage-encoded ARGs in relation to the global antibiotic resistance crisis[3,14–16].

Understanding the mobilization and proliferation of ARGs, particularly under the selection pressures imposed by antibiotics, is critical for the global public health management[17]. Most studies on the horizontal gene transfer of ARGs, have focused on bacteria and plasmids[18,19]. However, the extent to which phages, particularly prophages, mediate this ARG movement in complex community remains less well understood, despite previous evidence of phage-mediated transduction of bacterial DNA with various bacterial species[20–22]. Recently, prophages of pathogenic bacteria have been found to harbor abundant ARGs associated with enhanced survival under antibiotic exposure[6,23,24]. While this suggests that prophages may serve as an overlooked reservoir of ARGs, the distribution and activity of prophages across different microbial habitats with different degrees of human impact remains largely unknown[25,26].

Here, we investigate this by conducting a global genomic analysis of ARGs carried by prophages across different environments with varying degree of human impact. Our analysis consists of 38,605 bacterial genomes (covering 50 phyla) and 1432 metagenome samples collected across 12 habitat types, which were classified into low and high antibiotic exposure environments based on global antibiotic consumption data and Random Forest modelling using the metagenome samples (see Methods and Results). We first examine the bacterial genomes and metagenomes across contrasting habitats to determine the effect of human impact on the distribution and transmission of ARGs in prophages. Additionally, we create a global database of 1186 metatranscriptomes to investigate the transcriptional activity of prophage-encoded ARGs (pARGs) under low and high antibiotics exposure environments, and experimentally validate the functioning of pARGs with a subset of bacterial isolates covering four major phyla and 32 genera. Our findings suggest that human activities are enriching ARGs in prophage genomes that show higher transmission risk and a wider distribution across environmental habitats. This work improves our understanding of the role of prophages in the evolution of bacterial pangenomes and horizontal gene transfer of ARGs due to human use of antibiotics.

Results

Classification of bacterial genomes into low- and high-level antibiotic exposure habitats

A global database was compiled, which included 38,605 completed bacterial genomes from 12 contrasting habitats, representing varying degrees of human-impacted environments and potential prior exposure to antibiotics globally (see Methods). We then explored the effect of environmental type on the occurrence, composition, and distribution of prophages across different bacterial taxa (Supplementary Data 1). The isolation habitats of bacterial genomes represented 12 different environments: human gut (32.7%), domestic animals (9.7%), processed food (5.4%), wildlife (3.6%), aquatic organisms (3.6%), insects (2.1%), soils (9.7%), plants (7.0%), surface water (4.8%), seawater (1.6%), sediments (2.0%), and unclassified habitats (19.4%, Fig. 1a). Using the data from global antibiotic consumption and Random Forest modelling using the metagenome samples (see Methods)[27], these habitats were classified into two broad categories based on potential prior exposure to antibiotics due to human activities. Low antibiotic exposure habitats (LH) included natural environments, which have likely experienced relatively lower levels of antibiotic exposure due to less frequent human activities: wildlife, aquatic organisms, insects, soils, plants, surface water, seawater, and sediments. High antibiotic exposure habitats (HH) included genomes derived from human gut, farmed animals, and processed foods, which receive over 95% of the world's antibiotics[28]. While this classification is not perfect, as antibiotics are often leaked to natural aquatic and terrestrial environments, HH habitats are relatively more often exposed to antibiotics creating potentially stronger selection for the maintenance of ARGs in these environments. Bacterial taxa were distributed across 50 phyla, of which the most dominated six phyla were *Pseudomonadota*, *Bacillota*, *Actinomycetota*, *Bacteroidota*, *Spirochaetota*, and *Mycoplasmatota* (Fig. 1b). The bacterial genomes represented 1341 genera, with ten genera belonging to common human and animal commensals and pathogens (Supplementary Data 1). The number of bacterial genomes (57% in HH vs. 43% in LH) and taxonomic composition (at phyla level) of sampled genomes from HH to LH habitats was similar (Supplementary Data 1), indicating that our genome collection represents well the taxonomic diversity found in microbiomes across chosen ecosystems.

Antibiotic exposure selectively enriches prophages on a global scale

We identified predicted lysogens (bacterial cells which were predicted to encode one or more prophages) in all genomes using DEPhT—a stand-alone prophage finder[29]. Across all bacterial genomes and habitats, we identified a total of 11,736 lysogens that spanned 18 phyla and 635 genera (Fig. S1 and Supplementary Data 2). Interestingly, 98.7% of lysogens were found in just four bacterial phyla, *Bacillota*, *Pseudomonadota*, *Actinomycetota*, and *Bacteroidota*, and the proportion of lysogens clearly differed among different phyla (Fig. 1c). Habitat type was an especially important factor in determining prophage presence (Fig. 1d). For example, bacteria isolated from processed food had the highest proportion of lysogens (42%), followed by human gut (38%) and domestic animals (38%). In contrast, lysogens were identified the least often in the genomes of bacteria isolated from seawater (13%, Fig. 1d). Crucially, bacteria isolated from HH habitats carried more often lysogens (38%) compared to bacteria isolated from LH habitats across the whole data (22%, Fig. 1e).

Overall, we identified 26,858 potential prophage elements among all the lysogens, with genome lengths ranging from 20 kb to 623 kb (Supplementary Data 3). Only 30% of predicted prophages could be assigned taxonomically with known viruses using PhaGCN2 (Fig. S2 and Supplementary Data 3). The prophage hosts were distributed across 18 bacterial phyla and 635 genera (Supplementary Data 3), and the mean number of prophages per host was 2.3 (ranging between 1

https://doi.org/10.1038/s41467-024-52450-y

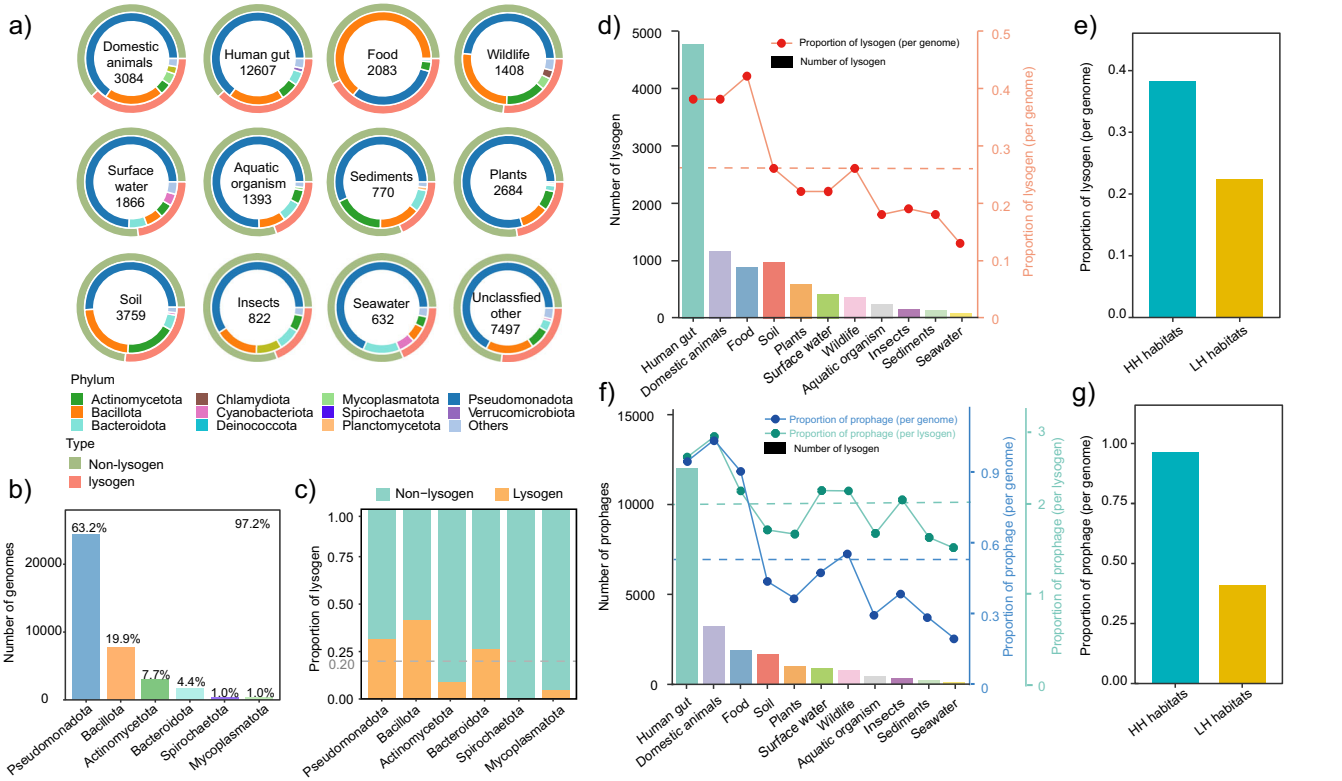

Fig. 1 | Distribution of lysogens and prophages in different taxa across different habitats. a The composition of bacterial isolates at phylum level (inner circle) and the proportion of lysogens (outer circle) in different habitats. The number in the center of circle, represent the number of bacterial genomes collected from different habitats. **b** The number of dominant bacterial genomes grouped by phylum across all habitats. **c** The proportion of dominant lysogens at phylum level across all habitats. **d** The number (bar plot) and proportion (line plot) of lysogenic bacteria isolated from different habitats. **e** The mean proportion of lysogenic bacteria between highly antibiotic exposure habitats (HH) and low antibiotic exposure habitats (LH). **f** The number (bar plot) and proportion (line plot) of prophages in different habitats. **g** The mean proportion of genomes containing prophages between HH habitats and LH habitats. In (**d**) and (**f**), the dotted line represents the mean value across all habitats.

and 14). Prophage distributions were clearly influenced by host taxonomy (Fig. S3) and habitat type (Fig. 1f). Overall, the proportion of genomes containing prophages was about 96% in HH habitats, while in contrast, only 41% of bacterial genomes from LH environments carried prophages (Fig. 1g). Overall, 63.6% of prophages originated from HH habitats and only 20.2% of prophages originated from LH habitats (Supplementary Data 3). This analysis suggests that bacteria isolated from HH habitats were enriched with prophages compared to bacteria isolated from LH habitats.

Prophage-encoded ARGs are enriched in HH habitats

To assess to what extent prophages carry ARGs, prophage genomes were examined using the RGI tool against The Comprehensive Antibiotic Resistance Database (CARD)[30] under strict parameter control (see Methods). Using all prophage elements ($n = 26,858$), we identified a total of 11,543 ARGs that confer resistance to 42 classes of antibiotic drugs (Fig. S4a and Supplementary Data 4). After removing duplicate genes, 397 non-redundant ARG subtypes were detected in prophages across all habitats. We then analyzed the distribution of phage-encoded ARGs (pARGs) among different environments. Interestingly, the majority of ARGs (77.8%, 8983 of 11,543) were found in prophages isolated from HH habitats, including human gut ($n = 6071$), domestic animal ($n = 2335$) and processed food ($n = 577$, Fig. 2a) samples.

The variation in pARG contents between different environments was compared after normalization of pARGs per bacterial genome and per prophage. Overall, the pARG contents per genome were consistent with pARGs across habitats, except for surface water, wildlife, aquatic organisms, and sediment habitats, where the pARG content was higher than ARG content per genome (Fig. 2a). This result suggests that these environments experienced an enrichment of ARGs in phage regions compared to the bacterial genome. Moreover, we found that mean content of pARGs per lysogen was over five-fold higher in HH compared to LH habitats (Fig. 2b, c). More specifically, 248 pARGs were exclusively detected in HH habitats, while only 63 pARGs were exclusively detected in LH habitats, while 110 pARGs were shared between the HH and LH habitats (Fig. 2d). A significant correlation ($R = 0.92$, $p < 0.0001$) between mean number of pARGs and prophages in bacteria (normalized to per prophage and per genome) was observed across HH and LH habitats (Fig. S5), suggesting that most prophages carried ARGs. To compare the relative contribution of prophages to lytic viruses for ARG carriage, all lytic virus genomes available in the IMG/VR database ($n = 627,970$, v4.1) were subjected to ARG detection using the same tools and parameters as with pARG analysis (Fig. S4b). We found that the proportion of ARGs in lysogenic viruses (42.98%, 11,543 of 26,858) was enriched by over three orders of magnitude compared to ARGs carried by lytic viruses (<0.01%, 67 of 627,970, Supplementary Data 5). Together, these results suggest that ARGs are enriched in prophages, which are more commonly found in antibiotic-impacted habitats globally.

Antibiotic exposure facilitates the ARGs movement across habitats and between bacterial taxa

To track the potential movement of phages and their ARGs between different habitats and bacterial hosts, we analyzed the presence of CRISPR-spacer regions in bacterial and prophage genomes. Overall, 460 prophages showed evidence of cross-genera transmission in terms of matching spacer sequences (Supplementary Data 6), while 32 prophages showed evidence of between bacterial phyla transmission

https://doi.org/10.1038/s41467-024-52450-y

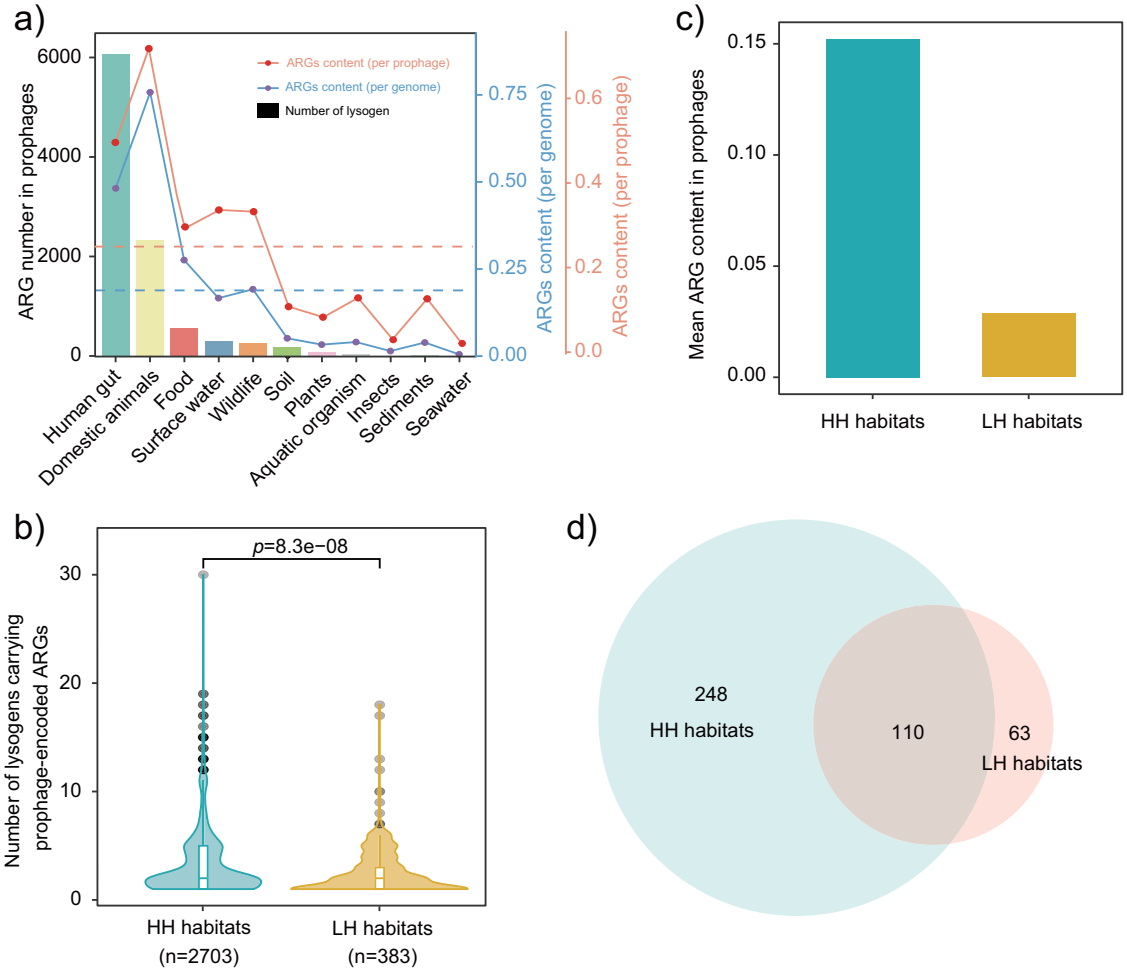

Fig. 2 | Prophage-encoded ARGs are more dominant in HH habitats impacted by humans. a The number (bar plot) and proportion (line plot) of prophage-encoded ARGs in bacterial genomes isolated from different habitats. The dotted line represents the mean content of prophage-encoded ARGs across all habitats. **b** Changes in number of lysogens carrying prophage-encoded ARGs in highly antibiotic exposure habitats (HH, $n = 2703$) and low antibiotic-exposure habitats (LH, $n = 383$). **c** Changes in the mean content of ARGs in prophages per lysogen in both HH and LH habitats. **d** The distribution of individual prophage-encoded ARG subtypes among HH and LH habitats. In (**b**), the significant differences between two groups were analyzed based on nonparametric Wilcoxon test ($p < 0.05$, two-sided). Box plots encompass 25–75th percentiles, whiskers show the minimum and maximum values, and the midline shows the median.

(Fig. 3a). Moreover, 497 ARGs (distributed in 229 prophages) showed between-species transmission potential, including 58 cases of between-genera and 6 cases of between-phyla transmission (Supplementary Data 6). Overall, 11.9% of prophages (3200 of 26,858) could be matched with 10,161 bacterial genomes using CRISPR-spacer matching (excluding the original prophage hosts), suggesting that these prophages could potentially move between different bacterial species (Supplementary Data 6). Among the predicted prophage hosts, 62.8% (6378/10161) were from HH habitats, while only 16.3% of hosts (1656/10161) were from LH habitats (Fig. 3b). Moreover, 62.3% of prophages with between-species transmission potential (1992 of 3200; for more information sees Methods) were detected in HH habitats, while only 23.1% of prophages (740 of 3200) isolated from LH habitats showed between-species transmission potential (Fig. 3c). We also examined prophage-host pairs derived from different habitats to obtain potential dissemination ranges. All prophages from human gut and domestic animals from HH habitats showed between-habitat type transmission potential, while prophages from sediment and seawater LH habitats showed 70% between-habitat type transmission potential (Fig. S6a). In other words, prophages originating from HH habitats were more often detected in all other types of habitats (Supplementary Data 7), indicative of their relatively higher transmission potential.

We further tracked the transmission potential of pARGs based on movement of their prophages. Overall, 377 ARG-carrying prophages showed evidence of past horizontal gene transfer events, based on CRISPR spacer matching (Supplementary Data 7). Among these mobile prophages, 72.4% (273 of 377) were from HH habitats, while only 11.7% (44 of 377) originated from LH habitats (Fig. 3d). For example, pARG-containing prophages from human gut and domestic animals could be considered as critical hotspots for ARG dissemination, with more than 90% of prophages showing transmission potential between habitats. In contrast, pARG-containing prophages from LH habitats, such as sediment and seawater, did not exhibit any between-habitat transmission potential (Fig. S6b). Taken together, prophages and their ARGs showed more frequent movement when originating from HH environments.

The pARGs are enriched and have a wider geographical distribution when originating from HH than LH metagenomes

To overcome potential sampling biases in bacterial genome collection, we also analyzed 1432 metagenomes available in different databases from the same 11 environmental habitat types (at least 100 metagenomes per habitat: human gut, domestic animals, processed food, wildlife, insects, plant, freshwater, seawater, soil, and sediments;

https://doi.org/10.1038/s41467-024-52450-y

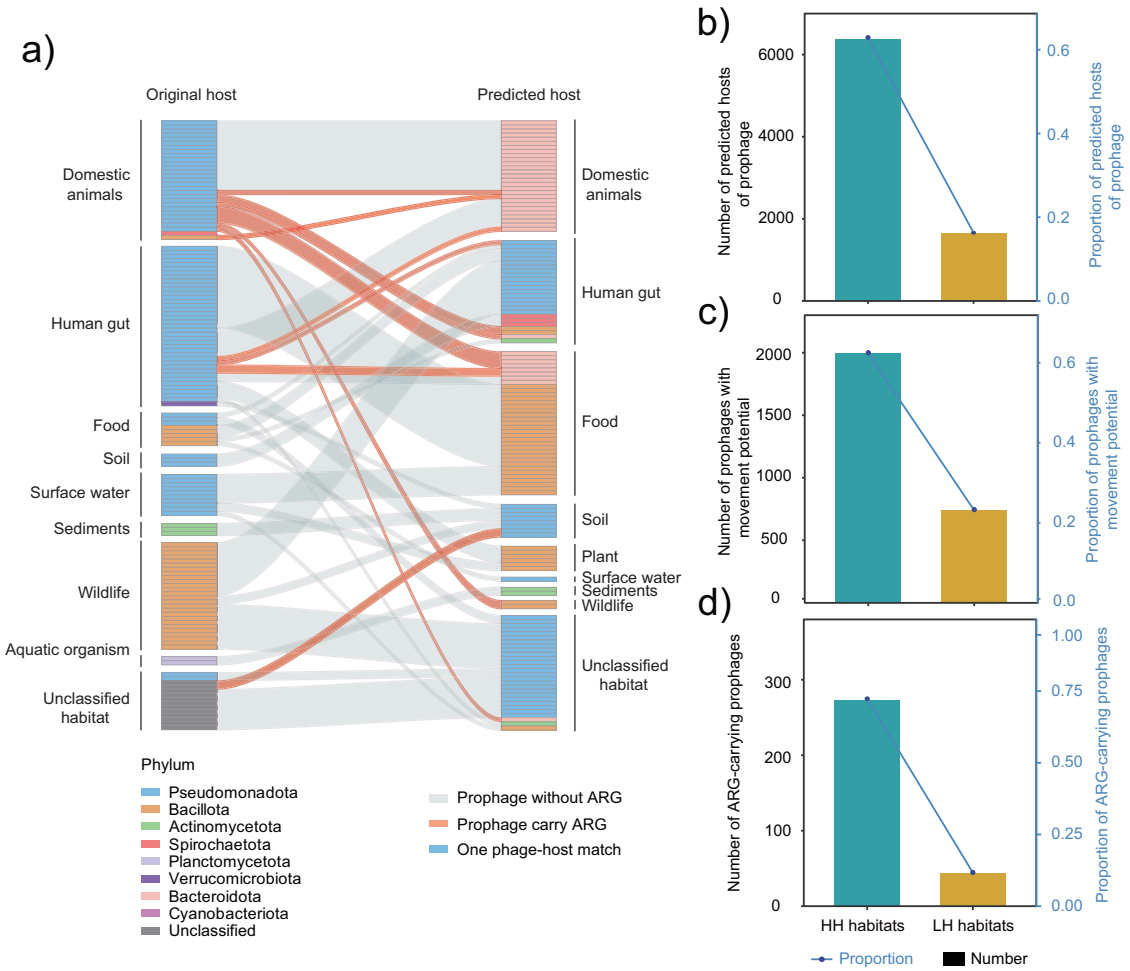

Fig. 3 | Human activity facilitates the ARGs movement across habitats and taxa based on CRISPR spacers matching. a Sankey plot depicting association of original hosts of prophages with the predicted hosts identified by CRISPR spacer matching across different habitats and host taxa at phylum level. One small grid represents one virus-bacterium pair, while different colors show the phylum of lysogen. **b** The distribution and proportion of prophage hosts between highly antibiotic-exposure habitats (HH) and low antibiotic-exposure habitats (LH). **c** The distribution and proportion of prophages with transmission potential between HH habitats and LH habitats. **d** The distribution and proportion of ARGs-carrying prophages between HH and LH habitats.

Fig. 4a and Supplementary Data 8). Overall, 95.1% of pARGs found in bacterial genome collection (10,982 of 11,543) could also be detected in the global metagenome collection, even though the detection frequency (Df) of these pARGs varied more between different habitat types (Fig. S7). For example, pARGs in HH metagenomes, such as human gut and domestic animals, showed more than an average of 70% Df, while LH metagenomes, including seawater, soil, sediment, and plants had only an average of 40% Df (Fig. S8). This result supports our bacterial genome analysis, demonstrating that pARGs are enriched also in HH compared to LH metagenomes.

To study the geographic distribution of pARGs in more detail, we divided all pARGs into two groups per isolation location: HH and LH metagenomes using the same habitat classification criteria as previously. The pARGs showed clearly different patterns in their abundance (Student's t test, $p < 0.001$) and composition (PERMANOVA test, $R^2 = 0.042$, $p = 0.001$) between HH and LH metagenomes (Fig. 4b, c). Overall, pARGs-originating from HH metagenomes had significantly higher abundances compared to LH metagenomes (Student's test, $p < 0.001$; Fig. S9). This indicates that pARGs-originating from HH habitats have spread across the globe, while pARGs-originating from the LH habitats have not diffused into other environments. Moreover, pARGs-originating from HH habitats were especially abundant in densely inhabited regions such as Southeast Asia, Eastern Australia, North and Southeast Africa, Western Europe, and Midwest North America (Fig. 4c), indicative of their association with humans. To further explore the effect of environmental habitat on the transmission potential of pARGs (see details in Methods), we compared the transmission risk of pARGs in HH and LH habitats. We found that the habitat types significantly impacted the global transmission risk of pARGs (Fig. S10): on average, prophages originating from HH habitats had a relatively higher transmission risk (0.49) compared to LH habitats (0.14) (Fig. S11) on average, and hence, a higher prevalence and transmission risk compared to that in LH environments.

We next investigated the effect of human activity on the potential linkages between phages and their host based on 25,858 prophage-host pairs using the CRISPR spacer matching. Each predicted phage-host linkage was further investigated by analyzing the prophage and host abundances in different habitats based on metagenome data. The abundances of prophages between HH and LH metagenomes clearly separated into two distinctive groups (Fig. 4d). Overall, there was a significant (Student's test, $p < 0.0001$) difference in lineage-specific virus-host ratios (estimated using host and prophage data) between HH and LH environments (Fig. 4e). In other words, antibiotic exposure might alter the phage-host dynamics at the community level, by selecting for ARG localization into prophages, which could potentially enhance host survival. The rate of pN scaled by the rate of pS can provide evidence on the selective forces driving the evolution of pARGs. While the pN/pS ratio of most pARGs was less than one, the pN/

https://doi.org/10.1038/s41467-024-52450-y

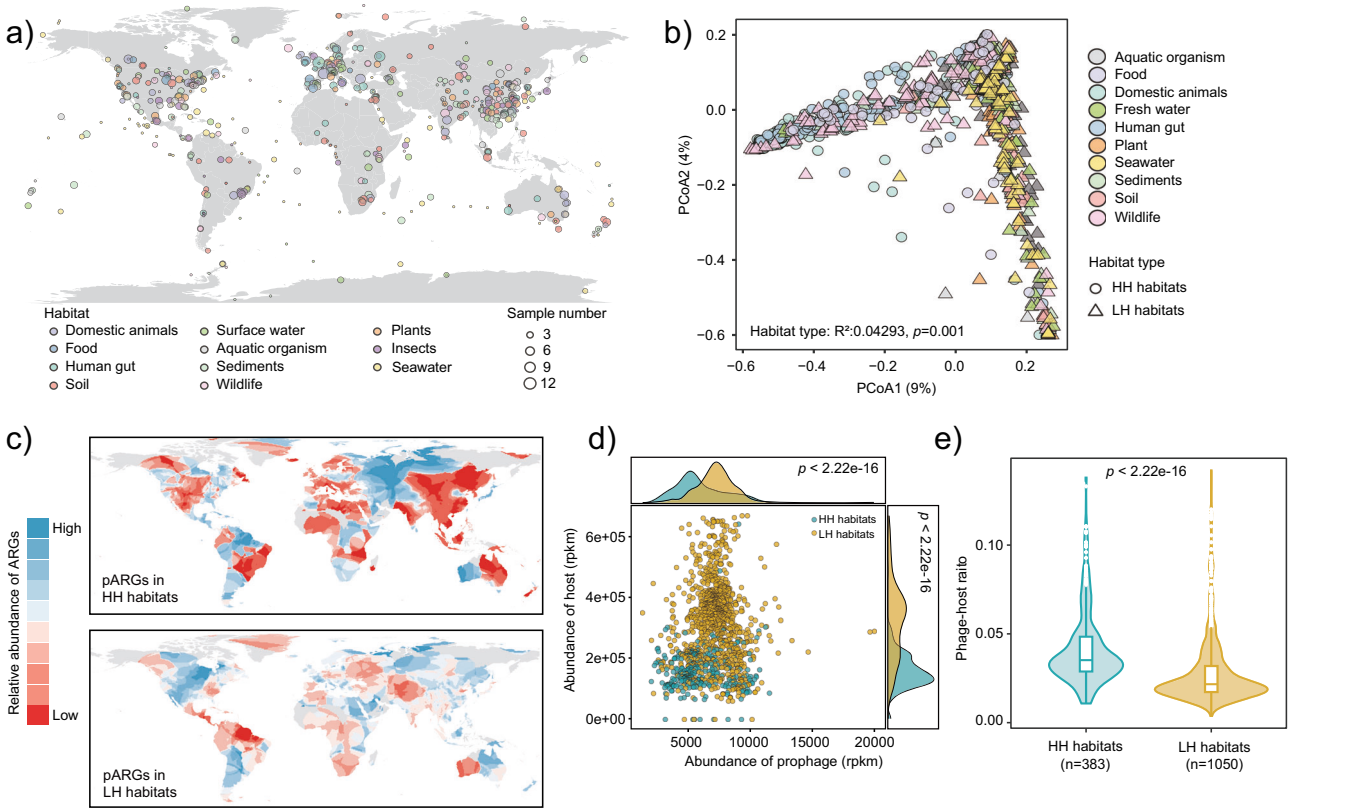

Fig. 4 | The global distribution and abundance of prophage-encoded ARGs (pARG) based on metagenomics across different environments. a Global map shows the 1432 metagenomics sample sites across different habitats. **b** PCoA analysis showing the effects of habitats on the global distribution of pARGs based on distance dissimilarity. Non-parametric PERMANOVA (Adonis function, 999 permutations) was used to determine the significance of habitats on the pARGs composition. **c** The global abundance of pARGs from highly antibiotic exposure habitats (HH) and low antibiotic exposure habitats (LH) based on mapping of pARGs to metagenomic samples collected worldwide (except of ocean samples). The maps in the (**c**) were generated using ArcGIS Pro v3.0.2 software. **d** The global distribution patterns, based on prophage and corresponding host abundances, encompass all metagenomic samples worldwide. **e** The change in phage-host ratio (estimated using host and prophage abundances) between HH habitats ($n = 2703$) and LH habitats ($n = 383$) based on all metagenomic samples. In (**d**) and (**e**), asterisks indicate significant differences between different groups based on nonparametric Wilcoxon test ($p < 0.05$, two-sided). Box plots encompass 25–75th percentiles, whiskers show the minimum and maximum values, and the midline shows the median.

pS ratio of pARGs was significantly higher in HH compared to LH habitats[31]. This suggests that pARGs accumulate more non-synonymous mutations under human impact (Fig. S12), indicative of diversifying selection on pARGs in HH habitats.

To explore if pARGs were potentially active, we mapped all pARG sequences against 1186 publicly available metatranscriptomes to estimate their transcriptional activity (covering 11 similar habitat types included in previous datasets: 26.8% HH vs. 73.1% LH; Fig. S13 and Supplementary Data 9). ~76% of pARGs could be mapped back to metatranscriptomic reads, suggesting that most pARGs are likely to be transcriptionally active. Specifically, pARGs from HH habitats had a higher Df and transcriptional activity compared to those from LH environments even though their sample number was overall lower (data normalized with sample group sizes; Fig. 5). This pattern was particularly clear in densely populated areas with potential high human activities such as East Asia, Central Europe, East-Central North America (Fig. 5).

To experimentally validate the prophage induction and the resistance of pARGs conferred to host bacteria, we selected 41 genome-sequenced strains (spanning four phyla and 32 genera) for prophage induction assays using mitomycin C[32]. We found that 35% of prophages (20 of 58) in 17 strains could be induced to produce phages, suggesting that these prophages have the potential to produce virions and transfer ARGs when these virions reinfect other hosts (Supplementary Data 11 and Fig. S14). The intact phage particles produced by prophages from a few representative strains ($n = 4$) were confirmed by scanning electron microscopy (Fig. S15). Furthermore, we randomly selected six different types of ARGs in six prophages to directly test if they provide *Escherichia coli* DH5α strain resistance when cloned and expressed in plasmids. We found that three pARGs significantly (all $p < 0.001$, Student' test) increased the antibiotic tolerance to streptomycin (*aadA2*), chloramphenicol (*catII*), and trimethoprim (*dfrC*) in *E. coli* bacterium compared to control treatment with empty vectors (Figs. S16, 17). These results suggest that bioinformatically identified pARGs can be functionally active, conferring resistance to heterologous host bacterium.

Discussion

In this study, we investigated the global distribution, abundance, and activity of prophages and their encoded ARGs in different habitats through extensive genomic and metagenomic analyses. Our results reveal, for the first time, a significant enrichment of lysogens and ARG-carrying prophages in habitats with likely higher antibiotic exposure risk due to human activity. Furthermore, both the abundance, transmission risk, and transcriptional activity of pARGs were enriched in HH compared to LH environments. These results suggest that human antibiotic use may affect phage-host interactions by selecting for localization of ARGs in prophage genomes, playing a critical role in the global spread of ARGs.

Our investigation revealed that prophages serve as globally important, hidden reservoir for ARGs. While previous reports have identified that several prophages of pathogenic bacteria can carry

https://doi.org/10.1038/s41467-024-52450-y

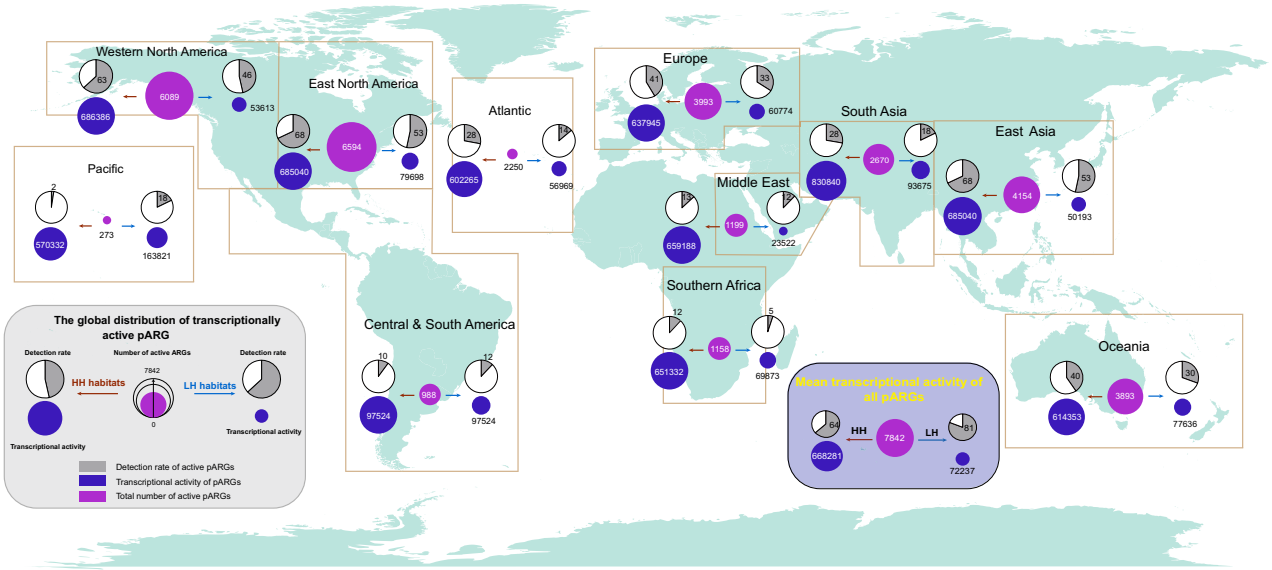

Fig. 5 | The global distribution of transcriptionally active prophage-encoded ARG (pARG) based on metatranscriptomes across different regions globally. Metatranscriptomic sample location and individual regions are shown on the basemap. For each region, the circles in the center refer to the total number of pARGs with transcriptional activity across HH and LH habitats. To the left and right, the circles show the relative changes in the detection rate of active pARGs (upper circle) and their relatively transcriptional activity (lower circle) from highly antibiotic exposure habitats (HH) habitats and low antibiotic exposure habitats (LH), respectively.

ARGs[23,33], we show that this pattern is widespread across different environments and holds at the global level across bacterial taxa. In contrast to previous studies, we found that 30% of the sequenced bacterial genomes carried prophages, which is a lower number than previously detected[34]. One likely reason for our conservative estimate is that we used highly stringent identification parameters to detect prophages. Regardless, almost half of the identified prophages contained ARGs, suggesting a significant enrichment of ARGs in prophages in contrast to lytic viruses, which are much less likely to carry ARGs. Previous studies have yielded inconsistent results on whether viruses carry ARGs[35–37], possibly because they have not taken into account the lifestyle of the viruses. Here we show that prophages have a much higher gene load than virulent phages due to the significantly longer prophage genomes (Fig. S18). Moreover, we found that ARGs were enriched in prophage regions relative to bacterial genomes, supporting the idea that ARGs are mainly located in bacterial accessory genome and likely often mobilized by phages.

Crucially, we found that bacteria in HH environments contained a higher proportion of prophages that can deliver a larger gene cargo to the host, in agreement with previous studies[38,39]. Human-associated antibiotic use could hence positively select for pARG carriage as this is likely to help their hosts to resist antibiotic stress, resulting in potentially mutually beneficial phage-host relationship[39,40]. Tracking the transmission of prophages based on CRISPR-spacer matching between phages and hosts, revealed that prophages had often be able to move between different bacteria and environments; 12% of prophages could be linked with more than two host taxa, suggesting that these prophages could have been moving genetic materials between different bacterial taxa. In particular, 32 prophages showed evidence of being able to infect different bacterial phyla, which is consistent with previous studies[41]. It should be noted that the CRISPR spacer-based method for determining prophage movement requires further experimental verification and might not tell if identified taxa can still interact.

Based on the global metagenomic analysis, prophages and their pARGs had a higher transmission risk between different environments if they originated from HH compared to LH habitats. For example, pARGs from HH habitats could be detected across all different environments, while pARGs from marine, soil, and sediment had much slower global Df of less than 40%. The likely reason for this is that bacteria from HH habitats might have a greater capacity for movement along with human activity, or human-associated bacterial taxa, whereas bacterial and phage mobility in natural environments could be more limited by physical distance and lack of suitable vectors[42]. In addition, we found that pARGs were more likely to move between the same type of environments. This could potentially be explained by the microbiome similarity between these environments, which is critical for prophage and pARGs movement and locating suitable host taxa. While more research is needed to validate phage and bacterial movement between environments, this hypothesis is consistent with previous findings where horizontal gene transfer via plasmids and transposons has been found to occur more likely between microbial lineages with small phylogenetic distances[43].

To explore the potential gene expression activity of pARGs across different environments, comparative metatranscriptomics analysis was conducted, which showed higher Df and transcriptional activity of prophages in HH compared to LH environments. This analysis adds more support to previous comparative genomics results, suggesting that human activities can significantly affect the abundance and activity of pARGs. In the future, it would be important to verify if the observed transcriptional activity results in significant increase in antibiotic resistance using experimental approaches and proteomics for example.

Moreover, we found that the pN/pS ratio of pARGs in HH environments was significantly higher than in LH habitats, suggesting that pARGs in HH environments might be under diversifying selection, which could be indicative of evolutionary response to antibiotic selection in these environments. Antibiotic exposure is known to play a critical selective role in the evolution of chromosomally encoded bacterial resistance[44,45], and this finding suggests that similar selective pressures may also apply to prophages. Further work linking this variation with antibiotic resistance is however required to test if this could be an adaptive signal. While we did not test this specifically, we conducted additional experimental work to test whether prophages can

be induced from bacterial genomes and if a subset of identified pARGs could provide resistance to antibiotics. We found that around 35% of tested prophages could be induced and three out of six tested pARGs could increase resistance to antibiotics when cloned in *E. coli* model host. While this work was done only for a subset of strains due to practical constraints, it suggests that pARGs identified in comparative genomic analyses can indeed be active and potentially selected for under antibiotic stress.

In conclusion, our study provides evidence that human activities could be altering phage-bacterial relationships, affecting the global spread of ARGs via enrichment of pARGs. In particular, pARGs originating from HH environments show a higher prevalence and wider spread across different environments globally. While our investigation focused mainly on prophages in sequenced bacterial genomes and metagenomes, a large number of unculturable bacteria and other types of MGEs that could move ARGs between bacteria and bacterial populations remain to be explored. Further experimental verification is also required, as most of the results are based on the analysis of sequencing data, and more information on the biological relevance of pARGs is needed. Moreover, our classification of environments to LH and HH habitats was relatively crude due to missing metadata in databases, which prevented more detailed analysis and assessment of the level of human impacts on the pARG prevalence. As a result, more detailed and finer-scale studies and experiments are needed to causally link the potential health risks of pARGs transmission for the development of antibiotic resistance in the clinic, veterinary, and natural environments. This will help to better understand the responses of bacteria and phages across antibiotic gradients, providing new insights into virus-host dynamics and transmission of antibiotic resistance via phage cargo genes.

Methods

Datasets #1: bacterial and viral genome data

We retrieved fully-sequenced and complete bacterial genomes (assembled into one chromosome) larger than 2 Mbp from the National Center for Biotechnology Information database (NCBI, Jan. 2023). After deleting duplicate genomes, a total of 38,605 fully completed bacterial genomes from 50 bacterial phyla and contrasting habitats were collected. Our final database included genomes from 12 common habitat types, including human gut, domestic animals, processed food, wildlife, insects, plant, freshwater, seawater, soil, sediments, and unclassified habitats (habitat not known or samples without metadata of source of isolation). The habitat of domestic animals included different body tissues, such as skin, respiratory tract, and gut. The habitat information of bacterial genomes was obtained based on the metadata provided in NCBI. In order to explore the impact of human activities (except for unclassified habitat) types were categorized into two main groups based on global antibiotic use and random forest modelling analysis[27,28] of metagenomics data from 11 different habitats, where we analyzed the relationship between 13 optimal representative anthropogenic correlates of global human activities and bacterial ARGs as an indicative signal of human-associated antibiotic exposure (Fig. S19; for more detail, please see the "Statistical analysis" section in the Methods). Our analysis shows that the human factors such as antibiotic usage were more clearly associated with human impacted (including human gut, farmed animals, and processed food) compared to natural habitats (including wildlife, insects, plants, freshwater, seawater, soil, sediments; Fig. S19). These results provide further evidence that habitats associated with humans, farmed animals and processed foods are more likely to be influenced by human-associated antibiotic input and exposure. As the use of antibiotics and disinfectants in humans, farmed animals and processed foods covers over 95% of all antibiotic use globally (https://resistancemap.onehealthtrust.org/About.php, Antibiotic Consumption Data)[28], these habitats were considered as "high antibiotic exposure habitats" (HH), while others natural environments (except of unclassified habitat) were considered as "low antibiotic exposure habitats" (LH). Of course, natural environments may also be contaminated with antibiotics, but this was not considered in this study because no information on antibiotics use in natural environments was not available[46,47]. Moreover, we included a collection of 627,970 high-confidence lytic virus genomes from the IMG/VR viral database for a subset of analyses (v4.1, downloaded Jan. 2023). The lytic viruses were confirmed using both VIBRANT v1.2.1[48] and CheckV v1.0.1[49] by detection of non-lysogeny-associated genes (i.e., integrase, recombinase, transposase, and excisionase, CI/Cro repressor, and *parAB*)[50]. Detailed information for selected bacterial and viral genomes is included in Supplementary Data 1 and 10, respectively.

Detection of putative prophage in bacterial genomes

A multimodal tool for potential prophage sequence discovery and extraction called DEPhT[29] was used to identify prophages in bacterial genomes using the normal mode (-s 10 -p 2). A bacterium was considered a potential lysogen if at least one prophage could be detected in the bacterial genome based on the above tool[7]. We identified 27,253 putative prophages from 38,605 bacterial genomes. To minimize the risk of false positives, we only retained viral predictions presenting at least one viral hallmark gene or viral-like genes in the prophage based on the CheckV v1.0.1 output[51]. After removing non-viral contigs, 26,858 potential prophages were finally obtained. The genome quality of prophage consisted of 45.0% of compete-quality, 25.7% of high-quality, 24.3% of medium-quality, and 4.8% of low quality as assessed by CheckV. Under certain conditions such as mitomycin C treatment, prophages can be induced to resume a lytic lifestyle, resulting in the production of viral particles[32]. To experimentally confirm the accuracy of prophage identification based on the DEPhT tool, 41 genome-sequenced isolates (spanning 32 genera and four phyla) that had at least one predicted prophage element, were subjected to prophage induction using mitomycin C (Supplementary Data 11). Among the 41 bacterial strains, we computationally predicted 57 potential prophages across all genomes. Based on PCR (see later in the Methods), we could induce and recover 20/57 prophages (35%) from the filtrates after mitomycin C treatment[32]. In other words, 17 lysogenic bacteria (41%) could produce an active and lytic phage and around 35% of predicted prophages based on DEPhT tool could be induced indicative of phage activity (Supplementary Data 11). It should be noted that the absence of detectable prophage activity does not indicate complete prophage inactivity as not all phages can be inducted under mitomycin C treatment[32].

Taxonomic classification of prophages and lytic viruses and detection of ARGs

Taxonomic assignment of prophages was performed using PhaGCN2 based on the latest ICTV classification tables[52]. Open reading frames of prophages and lytic viruses were predicted using Prodigal with default parameter[53] and then run through The CARD; http://arpcard.mcmaster.ca, v3.2.5 to detect ARGs using the resistance gene identifier (RGI v5.2.1) software with strict parameters to reduce false positives[30].

pN/pS ratio and nucleotide diversity analyses

The rate of accumulation of non-synonymous polymorphism (pN) relative to the rate of synonymous polymorphism (pS) provides an opportunity to assess if selection is driving diversification of a protein-coding sequence. Thus, genes with a high pN/pS (i.e. >1) ratio are likely to be evolving under the influence of positive selection[31,54]. For pN/pS ratio calculation, the representative metagenomic samples (average of 24 metagenomes per environment) were mapped to an indexed database of the pARGs sequences using Bowtie2 to produce the BAM files (v2.2.5; default parameters) due to server computing resource

constraints. Mapping files were then taken as input by inStrain (v1.3.1; default parameters, 'profile') to calculate the nucleotide diversity and pN/pS ratio at the gene level.

Using shared CRISPR spacers to track the movement of prophages and their ARGs among the bacteria

A good match between prophage sequences and bacterial CRISPR spacers indicates that a bacterial strain or taxon has previously encountered that phage, and consequently could be a potential host[55,56]. Consequently, shared CRISPR spacers between bacteria and prophages can be used to track virus transmission events[57]. Local alignments of extracted spacers from bacterial genomes with lengths greater than 25 bp were searched against prophage genomes using "blastn-short". Only BLAST matches with 100% alignment coverage and at most one mismatch were considered as high-confidence protospacer-to-spacer matches[55]. CRISPR spacers were recovered from all bacterial genomes with CRT (v1.2) with default parameters[58]. The transmission rate was defined as the shared number of spacers and matching protospacers between prophage and bacteria (log2). If a phage CRISPR spacer could be matched with more than two host bacterial species, this phage was referred as to having between-species transmission potential.

Datasets #2: a global database of prophages-encoded ARGs using metagenomes

To assess the global distribution and abundance of pARGs in different environments, we collected 1432 metagenome datasets from 11 similar habitats as with the full genome data (at least 100 metagenomes per habitat, including human gut, domestic animals, processed food, wildlife, insects, plant, freshwater, seawater, soil, and sediment) from the NCBI (Supplementary Data 8, Jan. 2023). The habitat types of metagenomes were obtained based on the metadata information provided by the submitter on NCBI. All metagenomic samples were grouped into "low antibiotic exposure habitats" (LH) and "high antibiotic exposure habitats" (HH) in a similar way as with bacterial genomes. We excluded the samples that were clearly affected by antibiotics or chemicals to keep the analysis as conservative as possible. The relative abundance of pARGs in the 1432 metagenome datasets was quantified using the CoverM pipeline[55] (v0.61, https://github.com/wwood/CoverM). Briefly, to calculate the relative abundance of each pARG, quality-controlled reads from each metagenome were mapped to the set of all ARGs sequences with CoverM pipeline using the "rpkm" calculation method (reads per kilobase of exon per million reads mapped). RPKM[59] is recommended for relative abundance comparisons with metagenomic datasets, because RPKM normalizes the data based on both sequence depth (per million reads) and sequence length (in kilobases). For details, reads after quality control were first mapped to viral contigs using "make" command in CoverM (v0.6.1), to make BAM files, after "filter" command was used to remove low-quality alignments with read identity ≤95% and aligned percent ≤75% (parameters: --percentage_id 0.95 --percentage_aln 0.75). Filtered bam files were used as input in CoverM to generate coverage profiles across samples (parameters: --trim-min 0.10 --trim-max 0.90 --min-read-percent-identity 0.95 --min-read-aligned-percent 0.75-m rpkm).

To investigate the phage-host ratio in metagenomes, the relative abundance of prophage and corresponding hosts were analyzed based on 25,858 prophage-host pairs using the CoverM pipeline with the same parameters as described above. For prophages, clean reads were first mapped to prophage sequences using "make" command in CoverM to make BAM files, after "filter" command was used to remove low-quality alignments. Filtered bam files were used as input in CoverM to generate coverage profiles across samples. For the prophage hosts, the abundance of host bacteria in metagenomes was estimated based on 16S rRNA gene of each bacterium. Briefly, the 16S rRNA gene of each phage host was extracted using Barrnap (v0.90, https://github.com/tseemann/barrnap/tree/master) under default parameters after the same calculation protocol was used as with prophage.

We used the frequency of detection and relative abundance in one habitat compared to other habitats to assess the risk of transmission of pARGs. We calculated a Df of pARGs derived from a given habitat in different metagenomic samples. Df represents the proportion of the number of pARGs detected in metagenomic sample to the total metagenomic samples and was calculated as:

$$\text{Detection frequency (Df)} = \text{Number}_{ARG}/\text{Number}_{sum}*100$$

where Number $_{ARG}$ and Number $_{sum}$ represent the number of detected pARGs in a metagenomic sample and the total number of metagenomic samples for any one habitat.

We first calculated the total abundance of pARGs derived from each habitat from different metagenome sources. The transmission ratio was calculated based on the relative proportion of detected pARGs in any one habitat relative to pARGs detected in other habitats as follows:

Relative abundance (Ra) = total abundance of pARG from one environmental habitat/total abundance of pARG measured in other habitats * 100.

Finally, the transmission risk was calculated based on the following:

Transmission risk = Df * Ra

Geographic distribution analysis of prophage-encoded ARG abundances based on metagenomics

We used the sampled metagenome dataset to generate a global map of pARG distributions geographically. Specifically, spatial distributions and abundance of pARGs were calculated using Empirical Bayesian Kriging (EBK) following previously developed procedures[60]. The EBK technique, which is a geostatistical technique available on the ArcGIS Desktop (ArcGIS Pro v3.0.2) was used to map pARGs distribution. The EBK method is a more practical geostatistical technique compared to other forms of kriging methods[60]. The principles governing the technique include the interpolation of a mapped property to any specific point (pixel). The variogram model was estimated from the data, and at each of the input data locations, a new value is simulated which then generates a new semivariogram model estimated from the simulated data using the Bayesian rule.

Datasets #3: analysis of transcriptional activity of prophage-encoded ARGs using metatranscriptome data

To assess the transcriptional activity of pARGs in different environments, we used 1186 metatranscriptome datasets collected from around world available in NCBI databases (Supplementary Data 9). The metatranscriptomic samples from 11 habitat types are similar to metagenomes include human gut, domestic animals, processed food, wildlife, insects, plants, freshwater, seawater, soil, and sediments. The majority of samples (83%) came from human gut, domestic animals, freshwater, and soil samples due to biases in metatranscriptomic dataset availability in current databases. The relative transcriptional abundance of pARGs in the 1,186 metatranscriptome datasets was quantified using the CoverM pipeline in the same way as the previous metagenome[55] (v0.61, https://github.com/wwood/CoverM). Metatranscriptomic reads were quality filtered via Trimmomatic (v0.39) using the following parameter (score > 30 and length > 36 bases)[55]. Moreover, SortMeRNA (v4.3.4)[61] was used to remove non-coding RNA sequences (tRNA, tmRNA, 5S, 16S, 18S, 23S, and 28S rRNA sequences) from the metatranscriptomic reads. The remaining total mRNA reads were mapped back to pARGs sequences to identify gene expression activity based on the average coverage of transcripts per using minimap2[62] of the CoverM pipeline. During the mapping, we set the threshold very high (read identity >95% and alignment percentage

>95%) to reduce the likelihood of false mapping errors (parameters: --percentage_id 0.95 --percentage_aln 0.95). To calculate the relative activity of each pARG, quality-controlled reads from each metatranscriptome were mapped to the set of all ARG sequences with CoverM pipeline using the "tpm" calculation method (Transcripts Per Kilobase Per Million Mapped Reads). The relative activity of pARGs in each environment type was standardized by the number of samples. All activity was quantified at the level of ARGs in prophage were deemed as active when the "tpm" values were larger than 0 according to previously studies[51].

Datasets #4: experimental analysis of prophage-encoded ARGs functioning and transmission potential

Prophage induction from a subset if isolated strains. To experimentally validate the accuracy of prophage identification by DEPhT[29], we randomly choose 41 isolates with sequenced genomes to prophage induction experiments under mitomycin C treatment. Overnight cultures of prophage host strains (including four phyla and 32 genera, stored in our lab, Supplementary Data 11) were prepared from glycerol stocks in 5 mL Luria Broth (LB). After overnight incubation at 37 °C at 180 rpm/min, the cultures were diluted 1:100 and grown again in LB. At the exponential phase of growth (OD600 = 0.8), all strain cultures were split into two sub-cultures, and 10 μL of the mitomycin C was added to another subset culture (1.5 μM at final concentration) with a final volume of 2 mL (other culture was used as negative control). The cultures were further incubated at 37 °C at 180 rpm. After 12 h, 1 mL was centrifuged at 1000 *g* at 4 °C for 10 min (5 biological replicates per strain). The supernatant was collected, sterile filtered through 0.2 μm membrane filters, and stored at 4 °C.

PCR sample preparation for prophage detection from induced filtrates. We took advantage of the fact that DNA packed in capsids is well protected from nucleases and can thus be differentiated from free genomic DNA of disrupted bacterial cells[32]. The cell-free supernatants of induced cultures were DNase treated to digest genomic DNA of lysed cells, whereas DNA packed inside phage particles would remain intact. DNase was then inactivated, and the capsids were disrupted by a heat denaturation step. Subsequently, diagnostic fingerprint regions were amplified by PCR using sequence-specific primers (Supplementary Data 11). For this experiment, the steps of phage induction and propagation were conducted as described above[32]. The software DEPhT (v1.1.3)[29] was used to map the prophage-like regions in the genome of isolates. For the complete phages, the major capsid protein genes were selected for amplification. The primers designed for each of the prophage were synthesized by Sangon Biotech Co., Ltd. (Shanghai, China). Genomic DNA was used as positive control and noninduced samples served correspondingly as negative controls in the assay. For each treatment, five biological replicates were performed. Initially, the phage supernatant was transferred into a new tube, DNase (10 mg/mL, Solarbio, Beijing) was added, and cultures were incubated for two more hours at room temperature until there was no bacterial DNA contamination based on16S rRNA gene PCR (27F: AGAGTTTGATCCTGGCTCAG; 1492R: TACCTTGTTACGACTT). To inactivate the DNase, samples were incubated at 75 °C for 5 min. The PCR solution contained: 7 μL milli-Q H_2O, 1 μL sample or genomic DNA, 1 μL forward primer, 1 μL reverse primer, and 10 μL PCR master mix (Sangon Biotech, Shanghai). The information on primers is listed in Supplementary Data 11.

Enumeration of phage particles by fluorescence and transmission electron microscopy

The harvested phage particles were treated with glutaraldehyde (0.5% final concentration) as a fixative at 4 °C for 20 min prior to staining, then this viral suspension was vacuum filtered through a 0.02-μm-pore-size Anodisc Al_2O_3 filter. The inverted fluorescence microscope (Olympus BX53, Japan) was used to observe phage particles stained with SYBR Gold fluorescent dyes (phenylenediamine as antifade) as previously described[63]. Viral particles were verified by transmission electron microscopy (Hitachi, HT7700, Japan) with the phosphotungstic acid counterstaining method described previously[5]. It should be mentioned that the enrichment method of viral nucleic acid in this study excluded RNA viruses.

Cloning, expression, and antimicrobial susceptibility tests of prophage-encoded ARGs

To validate the functioning of viral ARGs in prophages, we randomly choose six different types of pARGs located in six different prophages to conduct pARG expression in *Escherichia coli* DH5α (*aadA2* confers resistance to aminoglycosides, *catII* confers resistance to phenicols, *CRP* confers resistance to fluoroquinolones, *CTX-M-15* confers resistance to cephalosporin, *dfrC* confers resistance to diaminopyrimidines, and *emrK* confers resistance to tetracyclines, Supplementary Data 12). For the ARG to be considered a 'high-confidence' pARG, we only chose pARGs that were surrounded by viral structural genes, terminases or integrases either upstream or downstream of the pARGs. The genes encoding for a putative pARG sequence were chemically synthesized (Beijing Tsingke, Beijing, China) and inserted into the plasmids (PACYCDuet-1 for *aadA2* and pET-28a for other genes, plasmids have own promoter without induction). The recombinant plasmids were used to transform chemically competent *Escherichia coli* DH5α (Beijing Tsingke, Beijing, China) from which 1 mL 15% glycerol stocks (LB media, OD600 = 0.8) were prepared from a single colony and frozen (−80 °C) for future use.

The minimum inhibitory concentrations (MICs) of viral ARGs were assessed with PCR to derive their sequences using the primer pairs listed in Supplementary Data 12. The PCR products were double-digested with *BamHI* and *SalI*, and the digested DNAs were cloned into corresponding restriction enzyme-digested pET-28a (+) and PACYCDuet-1 (+) vectors (Novagen, Madison, WI, USA). Recombinant plasmids were transformed into *E. coli* BL21 to test the antibiotic tolerance (Beijing Tsingke, Beijing, China). Recombinant *E. coli* and quality control strains (*E. coli* ATCC 25922) were incubated overnight in LB medium to reach ~OD600 = 0.6, and after the strains were titrated onto a series of different antibiotic plates (including streptomycin, chloramphenicol, ceftazidime, ciprofloxacin, trimethoprim and tetracycline) along with antibiotic concentration gradients (1–43 μg/ml). MICs were determined after 24 h of incubation at 37 °C using a microbroth dilution method[64]. Control strains containing empty vector without cloned pARGs were used as controls for MIC determination. The MIC of strains with pARGs increased significantly compared to the negative control, suggesting that these pARGs can increase antibiotic resistance in this bacterial host.

Statistical analyses

Data was statistically analyzed using the *R* platform (v4.30, https://www.r-project.org/)[65]. ANOVA and PERMANOVA (Adonis function, 999 permutations) combined with principal components analysis (PCA) that differentiated the composition of ARG gene among varied habitats and continents were conducted by vegan and ggplot2 package. In most cases, the overall mean differences between two groups were analyzed using Student' *t* test using the p value < 0.05 as significance threshold. If the data do not meet a normal distribution, nonparametric Wilcoxon test was used. According to previous studies[47], the global distribution of bacterial ARGs has been significantly influenced by human activities (over 95% of antibiotic use in the available database comes from humans, farmed animals, and food production systems). The human gut, farmed animals and processed food can be considered to represent the high antibiotic impact habitats (HH), while other included environments can be considered to represent low antibiotic impact habitats based on global antibiotic consumption data

(wildlife, insects, plant, freshwater, seawater, soil, sediments). To test the validity of this assumption, we performed a random forest modelling analysis using 1432 metagenomes from 11 different habitats (HH and LH habitats) and analyzed the relationships between anthropogenic correlates of human activities (obtained from various databases; see below) and bacterial ARGs as an indicative signal of human-associated antibiotic exposure in HH and LH habitats[47,51]. As anthropogenic activity cannot be narrowed down to one variable, we collected data on 38 anthropogenic factors from public databases and satellite observations (Supplementary Data 13). These factors cover the agricultural, industrial, and economic aspects of human activity such as antibiotics usage, pesticide usage, air pollution, level of economic development, energy production, mining industry, sewage treatment, agricultural crops, and land use and cover change. All datasets comprising of 38 anthropogenic factors were first normalized (log-transformed as needed) and standardized using Z-score transformation using the scale package in R. The rotated PCA was performed on the standardized 38 factors to minimize multicollinearity among predictor variables using IBM SPSS Statistics 25[47]. This resulted in 13 principal components associated with human activity after dimensionality reduction based on the magnitude of the eigenvalues using variance maximizing rotation method[47]. The retention of these optimal principal components was determined by the Kaiser-Guttman rule, which requires that the eigenvalues of the principal components exceed one[47]. We assessed the relative importance of identified principal components on bacterial ARGs as an indicative signal of anthropogenic impact through the variable importance tool using Random Forest model. Briefly, two Random Forest models were constructed with same parameters based on the values of the above 13 principal components and total abundance of bacterial ARGs (Euclidean distance dissimilarity matrices) to quantify the impact of human activities on the geographical distribution of ARGs in HH and LH habitats. The Random Forest models were performed in R using randomForest and rfPermute packages, with the random seed set to 123 with otherwise default parameters[51]. To optimize the parameters, the random forest model was initially trained on 70% of the data using the randomForest package. The remaining 30% of the data served as a validation set to assess the model's accuracy. After optimizing the parameters, the final model was constructed using all data based on following parameters: importance = TRUE, ntree = 500, and nrep = 1000. The significance of the models and cross-validated R^2 values were assessed based on 1000 permutations using all datasets with the "rfPermute" package in *R*. In the Random Forest model, a higher percentage of mean squared error (MSE) indicates a higher importance of a given factor[66]. In the Random Forest model, a higher percentage of MSE indicates a higher importance of a given factor[67]. The MSE for every decision tree with out-of-bag estimates based on Random Forest model was produced using rfPermute package, which assesses the relative importance of each predictor variable. All the scripts for Random Forest model analysis are available in GitHub (see the "Code availability" section). As shown in Fig. S19, we observed that anthropogenic factors (including antibiotic usage) had higher and more often statistically significant MSE values with ARG abundances in HH compared to LH habitats (12 factors vs. 2 factors). This analysis provides more support to our initial LH and HH habitat classification based on global antibiotics consumption data. Livestock production (including buffalo, goat, cattle, horse, chicken, pig, ducks, and sheep) was attained from http://fao.org/livestock-systems/global-distributions/en/. Crop yields (wheat, rice, maize, barley, cotton, sorghum, pearl, soybean, alfalfa, and tea yields) were collected from CGIAR-CSI (https://cgiarcsi.community). Human influence index, development threat index, human modification of terrestrial systems, and pesticide use (chlorothalonil, paraquat, glufosinate, glyphosate, chlorpyrifos, dicamba) were available from EarthData (https://beta.sedac.ciesin.columbia.edu/search/data). Human development index was acquired from Dryad (https://datadryad.org/stash/dataset/doi:10.5061/dryad.dk1j0). Antibiotic use in clinical settings and food animals was available in ResistanceMap (https://resistancemap.onehealthtrust.org/About.php). Energy production (unconventional oil, conventional oil, natural gas extraction, and global coal mining industry) and mining production (metal mining and non-metal mining) were available from SEDAC. Other 10 anthropogenic factors (including sewage treatment capacity, anthropogenic biomes of the world, GDP, particulate matter 2.5, global freshwater availability, nitrogen fertilizer application, population density, human footprint, nitrogen in manure production, and phosphorus fertilizer application) were extracted from EarthData (https://sedac.ciesin.columbia.edu), Food and Agriculture Organization of the United Nations (https://data.apps.fao.org), and OneHealth Trust (https://resistancemap.onehealthtrust.org). The metadata of all human activities is based on the latitude and longitude of each metagenome sample. The abundance of bacterial ARGs in metagenomes was analyzed using local ARG-OAP (v 3.0) against the SARG database at the cutoff of 10^{-7} *E*-value, 80% identity and 80% coverage[68].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data for this study are provided in the Supplementary Information files. Source data are provided with this paper.

Code availability

All the scripts and codes for machine learning, statistical analysis, and visualization used in this study are available online at https://zenodo.org/records/13301199.

References

1. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
2. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
3. Jansson, J. K. & Wu, R. Soil viral diversity, ecology, and climate change. *Nat. Rev. Microbiol.* **21**, 296–311 (2022).
4. Yi, Y. et al. A systematic analysis of marine lysogens and proviruses. *Nat. Commun.* **14**, 6013 (2023).
5. Tang, X. et al. Lysogenic bacteriophages encoding arsenic resistance determinants promote bacterial community adaptation to arsenic toxicity. *ISME J.* **17**, 1104–1115 (2023).
6. Wendling, C. C., Refardt, D. & Hall, A. R. Fitness benefits to bacteria of carrying prophages and prophage-encoded antibiotic-resistance genes peak in different environments. *Evolution* **75**, 515–528 (2021).
7. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact, and ecology of temperate phages. *ISME J.* **11**, 1511–1520 (2017).
8. Huang, D. et al. Adaptive strategies and ecological roles of phages in habitats under physicochemical stress. *Trends Microbiol.* **32**, 902–916 (2024).
9. Tang, X. et al. Bacteriophages from arsenic-resistant bacteria transduced resistance genes, which changed arsenic speciation and increased soil toxicity. *Environ. Sci. Technol. Lett.* **6**, 675–680 (2019).
10. Haak, B. W. & Wiersinga, W. J. Uncovering hidden antimicrobial resistance patterns within the hospital microbiome. *Nat. Med.* **26**, 826–828 (2020).
11. Wang, M. et al. Role of enterotoxigenic Escherichia coli prophage in spreading antibiotic resistance in a porcine-derived environment. *Environ. Microbiol.* **22**, 4974–4984 (2020).

12. Lucidi, M. et al. Phage-mediated colistin resistance in Acinetobacter baumannii. *Drug Resist. Update* **73**, 101061 (2024).
13. Kauffman, K. M. et al. Resolving the structure of phage–bacteria interactions in the context of natural diversity. *Nat. Commun.* **13**, 372 (2022).
14. Piel, D. et al. Phage–host coevolution in natural populations. *Nat. Microbiol.* **7**, 1075–1086 (2022).
15. Wright, R. C. T., Friman, V.-P., Smith, M. C. M. & Brockhurst, M. A. Cross-resistance is modular in bacteria–phage interactions. *PLOS Biol.* **16**, e2006057 (2018).
16. Moniruzzaman, M. et al. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* **8**, 16054 (2017).
17. Liu, J., Gefen, O., Ronin, I., Bar-Meir, M. & Balaban, N. Q. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science* **367**, 200–204 (2020).
18. Yang, Q. E. et al. Interphylum dissemination of NDM-5-positive plasmids in hospital wastewater from Fuzhou, China: a single-center, culture-independent, plasmid transmission study. *Lancet Microbe* **5**, e13–e23 (2024).
19. Castañeda-Barba, S., Top, E. M. & Stalder, T. Plasmids, a molecular cornerstone of antimicrobial resistance in the One Health era. *Nat. Rev. Microbiol.* **22**, 18–32 (2024).
20. Gabashvili, E. et al. Phage transduction is involved in the intergeneric spread of antibiotic resistance-associated blaCTX-M, Mel, and tetM loci in natural populations of some human and animal bacterial pathogens. *Curr. Microbiol.* **77**, 185–193 (2020).
21. Sun, R., Yu, P., Zuo, P. & Alvarez, P. J. J. Bacterial concentrations and water turbulence influence the importance of conjugation versus phage-mediated antibiotic resistance gene transfer in suspended growth systems. *ACS Environ. Au* **2**, 156–165 (2022).
22. Chen, J. et al. Genome hypermobility by lateral transduction. *Science* **362**, 207–212 (2018).
23. Kondo, K., Kawano, M. & Sugai, M. Distribution of antimicrobial resistance and virulence genes within the prophage-associated regions in nosocomial pathogens. *mSphere* **6**, e00452–00421 (2021).
24. Huang, J. et al. Conjugative transfer of streptococcal prophages harboring antibiotic resistance and virulence genes. *ISME J.* **17**, 1467–1481 (2023).
25. Coban, O., De Deyn, G. B. & van der Ploeg, M. Soil microbiota as game-changers in restoration of degraded lands. *Science* **375**, abe0725 (2022).
26. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
27. Van Boeckel, T. P. et al. Reducing antimicrobial use in food animals. *Science* **357**, 1350–1352 (2017).
28. Tang, K. L. et al. Restricting the use of antibiotics in food-producing animals and its associations with antibiotic resistance in food-producing animals and human beings: a systematic review and meta-analysis. *Lancet Planet. Health* **1**, e316–e327 (2017).
29. Gauthier, C. H. et al. DEPhT: a novel approach for efficient prophage discovery and precise extraction. *Nucleic Acids Res.* **50**, e75–e75 (2022).
30. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, 517–525 (2020).
31. Dong, X. et al. Phylogenetically and catabolically diverse diazotrophs reside in deep-sea cold seep sediments. *Nat. Commun.* **13**, 4885 (2022).
32. Jancheva, M. & Böttcher, T. A metabolite of Pseudomonas triggers prophage-selective lysogenic to lytic conversion in Staphylococcus aureus. *J. Am. Chem. Soc.* **143**, 8344–8351 (2021).
33. Castillo, D. et al. Widespread distribution of prophage-encoded virulence factors in marine Vibrio communities. *Sci. Rep.* **8**, 9973 (2018).
34. Touchon, M., Bernheim, A. & Rocha, E. P. C. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* **10**, 2744–2754 (2016).
35. Enault, F. et al. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).
36. Debroas, D. & Siguret, C. Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J.* **13**, 2856–2867 (2019).
37. Billaud, M. et al. Analysis of viromes and microbiomes from pig fecal samples reveals that phages and prophages rarely carry antibiotic resistance genes. *ISME Commun.* **1**, 55 (2021).
38. Dragoš, A. et al. Phages carry interbacterial weapons encoded by biosynthetic gene clusters. *Curr. Biol.* **31**, 3479–3489 (2021).
39. Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpena, N. & Novick, R. P. Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* **23**, 171–178 (2015).
40. Shkoporov, A. N., Turkington, C. J. & Hill, C. Mutualistic interplay between bacteriophages and bacteria in the human gut. *Nat. Rev. Microbiol.* **20**, 737–749 (2022).
41. Hwang, Y., Roux, S., Coclet, C., Krause, S. J. E. & Girguis, P. R. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat. Microbiol.* **06**, 946–957 (2023).
42. Zhu, Y.-G. et al. Microbial mass movements. *Science* **357**, 1099–1100 (2017).
43. Redondo-Salvo, S. et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
44. Xiong, W. et al. Antibiotic-mediated changes in the fecal microbiome of broiler chickens define the incidence of antibiotic-resistance genes. *Microbiome* **6**, 34 (2018).
45. Lopatkin, A. J. et al. Antibiotics as a selective driver for conjugation dynamics. *Nat. Microbiol* **1**, 1–8 (2016).
46. Buelow, E., Ploy, M.-C. & Dagot, C. Role of pollution on the selection of antibiotic resistance and bacterial pathogens in the environment. *Curr. Opin. Microbiol.* **64**, 117–124 (2021).
47. Zheng, D. et al. Global biogeography and projection of soil antibiotic resistance genes. *Sci. Adv.* **8**, eabq8015 (2022).
48. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
49. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
50. Luo, X.-Q. et al. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).
51. Liao, H. et al. Mesophilic and thermophilic viruses are associated with nutrient cycling during hyperthermophilic composting. *ISME J.* **17**, 916–930 (2023).
52. Jiang, J. Z. et al. Virus classification for viral genomic fragments using PhaGCN2. *Brief. Bioinform.* **24**, bbac505 (2023).
53. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
54. Axelsson, E. et al. Natural selection in avian protein-coding genes expressed in brain. *Mol. Ecol.* **17**, 3008–3017 (2008).
55. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
56. Zhang, L. et al. CRISPR arrays as high-resolution markers to track microbial transmission during influenza infection. *Microbiome* **11**, 136 (2023).
57. Kim, M.-S. & Bae, J.-W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* **12**, 1127–1141 (2018).
58. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 209 (2007).

59. Li, Z. et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J.* **15**, 2366–2378 (2021).
60. Krivoruchko, K & Gribov, A. Pragmatic Bayesian kriging for non-stationary and moderately non-Gaussian data. *Mathematics of Planet Earth*. In: *Proc. 15th Annual Conference of the International Association for Mathematical Geosciences*) 61–65 (Springer, 2014).
61. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
63. Huang, D. et al. Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress. *Microbiome* **9**, 150 (2021).
64. Wiegand, I., Hilpert, K. & Hancock, R. E. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3**, 163–175 (2008).
65. Team, R. C. R: a language and environment for statistical computing. *R. Found. Stat. Comput.* **201**, 12 (2019).
66. Jiao, S. et al. Soil microbiomes with distinct assemblies through vertical soil profiles drive the cycling of multiple nutrients in reforested ecosystems. *Microbiome* **6**, 1–13 (2018).
67. Liaw, A. & Wiener, MJRn. Classification and regression by randomForest. *R. N.* **2**, 18–22 (2002).
68. Yin, X. 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* **34**, 2263–2270 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (42277357 to H.P.L.), Outstanding Youth Science Foundation of Fujian Province (2022J06016 to H.P.L.), Simons Foundation (735077 to S.W.W.), TED2021-130908B-C41/AEI/10.13039/501100011033/Unión Europea NextGenerationEU/PRTR and the Spanish Ministry of Science and Innovation for the I+D+i project PID2020-115813RA-I00 funded by MCIN/AEI/10.13039/501100011033 (M.D.B.). V.P.F. is funded by Research Council of Finland, Finnish Research Impact Foundation, and Novo Nordisk Foundation.

Author contributions

Conceptualization: H.P.L., M.G., S.G.Z.; Methodology: H.P.L., C.L., C.Q.L., D.J.E., X.L.L.; Investigation: H.P.L., C.A., X.T., V.P.F., M.D.B., Y.G.Z., M.D.; Visualization: C.L., H.P.L., Z.W., Q.Y.; Funding acquisition: L.H.P., M.D.B., V.P.F.; Project administration: S.G.Z., H.P.L.; Supervision: S.G.Z., H.P.L., M.D.B.; Writing—original draft: H.P.L., S.G.Z., M.D.B; Writing—review and editing: H.P.L., Y.G.Z., M.G., M.R., S.W.W., B.K.S., M.D.B., D.J.E., V.P.F.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52450-y.

Correspondence and requests for materials should be addressed to Shungui Zhou, Manuel Delgado-Baquerizo or Yong-guan Zhu.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.