nature microbiology

Article

https://doi.org/10.1038/s41564-025-02108-2

Phages with a broad host range are common across ecosystems

Received: 6 March 2025

Accepted: 4 August 2025

Published online: 19 September 2025

Check for updates

Amaury Bignaud[1,2,13], Devon E. Conti[1,2,3], Agnès Thierry[1], Jacques Serizay[1], Karine Labadie[4], Julie Poulain[5], Olivia Cheny[6], Maritrini Colón-González[7], Laurent Debarbieux[3], Marianna Guerrero-Osornio[7], Sophie Helaine[8], Peter Hill[8], Gwenaelle Le Tinier[9], Gael A. Millot[10], Lucia Morales[7], Andrés Parada[11,12], Nadia Riera[11], Gregorio Iraola[11], Romain Koszul[1] ✉ & Martial Marbouty[1,13] ✉

Phages are diverse and abundant within microbial communities, where they play major roles in their evolution and adaptation. Phage replication, and multiplication, is generally thought to be restricted within a single or narrow host range. Here we use published and newly generated proximity-ligation-based metagenomic Hi-C (metaHiC) data from various environments to explore virus–host interactions. We reconstructed 4,975 microbial and 6,572 phage genomes of medium quality or higher. MetaHiC yielded a contact network between genomes and enabled assignment of approximately half of phage genomes to their hosts, revealing that a substantial proportion of these phages interact with multiple species in environments as diverse as the oceanic water column or the human gut. This observation challenges the traditional view of a narrow host spectrum of phages by unveiling that multihost associations are common across ecosystems, with implications for how they might impact ecology and evolution and phage therapy approaches.

Viruses, the most abundant genomic entities across all habitats and a large reservoir of genetic diversity[1–3], are important drivers of bacterial community evolution both as predators and as agents of horizontal gene transfer[4,5]. Studying the role of viruses, and in particular those infecting bacteria (bacteriophages, or phages), in shaping natural microbial communities requires the ability to precisely characterize their relationships with their hosts and how specific these relationships are[6]. Although numerous in vitro works have shown that most phages infect a narrow range of hosts, a few recent studies have started to question the exclusivity of these relations, suggesting that some viruses exhibit a broader host spectrum in dense and complex microbial communities[7–9]. Today, the frequency at which a phage infects different bacterial species in a community, whether those are phylogenetically close, remains an open question. Indeed, current metagenomic approaches are limited when it comes to analysis of individual samples[10,11]. It is also difficult to characterize and validate complete viral sequences due to a lack

[1]Spatial Regulation of Genomes Group, Institut Pasteur, CNRS UMR 3525, Université Paris Cité, Paris, France. [2]Collège Doctoral, Sorbonne Université, Paris, France. [3]Bacteriophage Bacterium Host, Institut Pasteur, Université Paris Cité, CNRS UMR6047, Paris, France. [4]Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [5]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [6]Clinical Research Coordination Office, Institut Pasteur, Université Paris Cité, Paris, France. [7]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro, Mexico. [8]Department of Infectious Diseases, School of Immunology and Microbial Sciences, Guy's Hospital , King's College London, London, UK. [9]Médecine petite enfance, Orsay, France. [10]Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, Paris, France. [11]Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Montevideo, Uruguay. [12]Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. [13]These authors contributed equally: Amaury Bignaud, Martial Marbouty. ✉e-mail: romain.koszul@pasteur.fr; martial.marbouty@pasteur.fr

https://doi.org/10.1038/s41564-025-02108-2

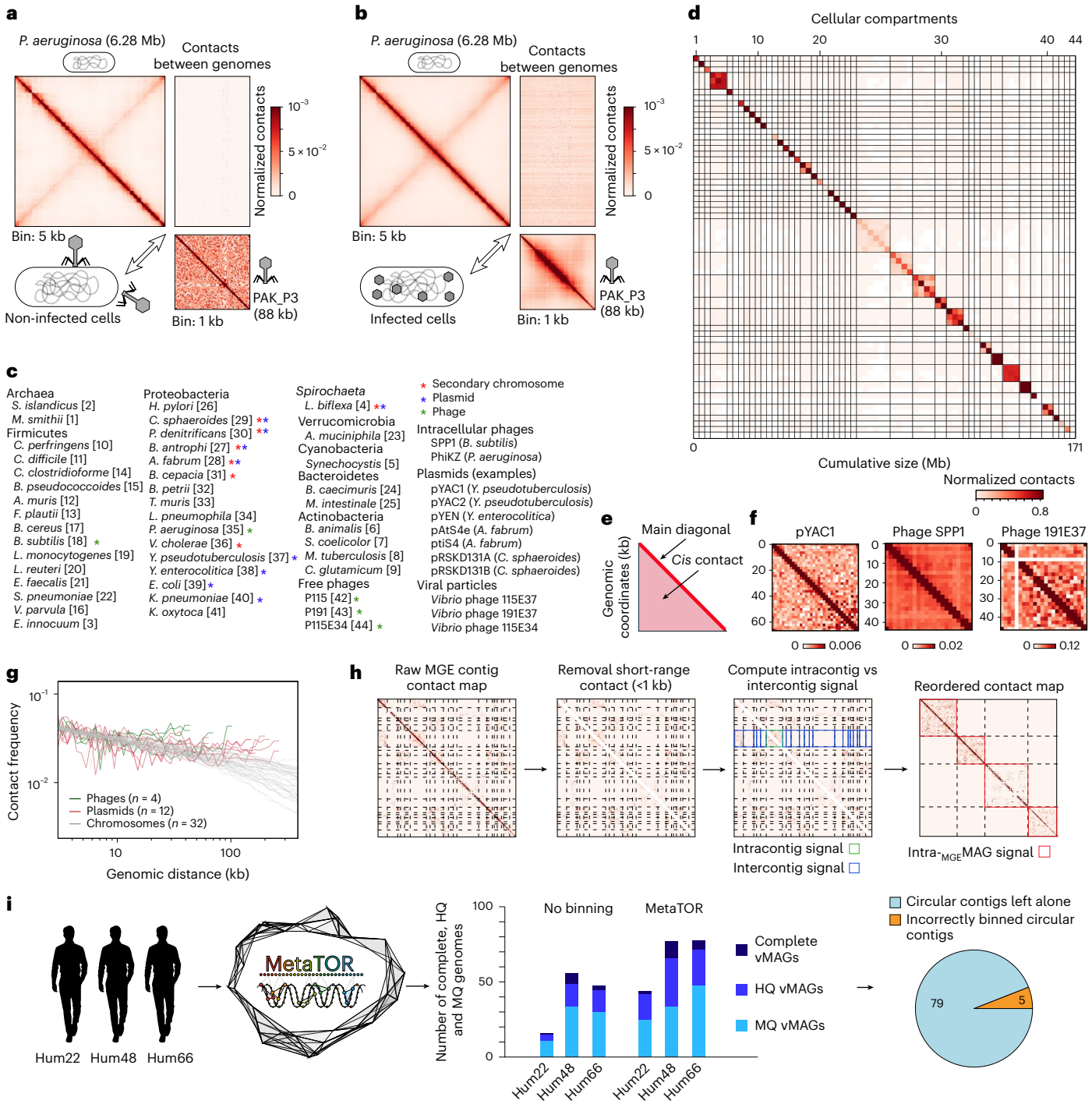


Fig. 1 | Principles of MGE binning and host association. a,**b**, Hi-C contact maps of *P. aeruginosa* (5 kb bins) and the phage PAK_P3 (1 kb bins) genomes before (**a**) and after (**b**) infection. **c**, Composition of the mock community. Stars indicate episomes (green, phage; blue, plasmid; red, secondary chromosome). Number in brackets indicates position of the corresponding cellular compartment in contact map in **d**. **d**, Normalized contact map (1 DNA molecule = 1 bin) of the mock community. Black lines delineate the 41 cellular and 3 viral compartments. **e**, Scheme of an unfiltered contact map exhibiting high signal in the narrow diagonal corresponding to very short-range contacts between pairs of adjacent DNA segments. **f**, Normalized contact map (2 kb bins) of MGEs in the mock community. **g**, Normalized contact signal as a function of the genomic distances for phages (green), plasmids (red) and bacterial chromosomes (grey) (log scale). **h**, Principles of the MGE binning module. From left to right: a raw contact map encompassing all the contigs assigned as MGE is built; short-range contacts (<1 kb) are removed; intra- and intercontig signals are computed; contigs exhibiting similar intra- and intercontig signals are binned together. **i**, Left: application of MetaTOR on three human gut datasets. Middle: bar plot of the number of complete, HQ and MQ viral genomes assessed by CheckV for the raw contigs and the vMAGs generated by MetaTOR. Right: pie chart of the proportion of circular viral contigs either left alone or binned with other contigs. Panels **a** and **i** adapted from SVG Silh under a CC 1.0 license.

of markers to assess their completeness[12–15]. Finally, inferring reliable phage–host relationships directly in microbial populations is challenging by conventional metagenomics approaches[6,16].

We and others demonstrated that physical collisions between DNA molecules sharing the same cellular compartment, quantified using the proximity-ligation-based method metaHiC, have the potential to solve these limitations[17–20]. For instance, it unveiled episome–host relationships between microbial genomes on the one hand, and plasmids[17,21] and viruses[22–24] on the other. However, the complexity of the large metaHiC networks impairs the application of the technique to large numbers of samples spanning multiple environments. Here we first applied metaHiC on a mock community to (1) show that free phage genomes are caught by the experiment and (2) implement a strategy to robustly and automatically deconvolve these data and reliably assign episomes to their hosts. We next performed an integrative analysis of 84 published and 27 unpublished metaHiC datasets generated over 5 natural ecosystems. The analysis resulted in 6,572 viral metagenomic assembled genomes (vMAGs) of medium quality or higher, including complete viral genomes larger than 400 kb from phyla as diverse as *Megaviricetes* or *Caudoviricetes*. The data also generated 4,975 microbial medium-quality (MQ) and high-quality (HQ) MAGs. Using three-dimensional (3D) contacts between MAGs and vMAGs, we inferred the hosts of 2,883 phages, unveiling how diverse families of phages interact with distantly related hosts. Our study reveals that a broad range of phages exhibit interactions with multiple hosts across bacterial phyla and throughout environments.

Results

Mobile genetic elements display specific 3D signatures

Contacts between plasmids and a host genome had experimentally been benchmarked in the past[17,18]. We initiated this study by quantifying DNA contacts from a known virus–host system, the virulent phage PAK_P3 infecting *Pseudomonas aeruginosa*[25]. After mixing at a phage:cell ratio of 25:1 to ensure homogeneous infection, two samples were taken at 0 and 5 min and processed with metaHiC. At $t = 0$ min, the PAK_P3 genome in the viral particle displays homogeneous *cis* contact and no *trans* contact with the *P. aeruginosa* genome (Fig. 1a). At $t = 5$ min after infection, the phage genome is now slightly decondensed and in close contact with the whole bacterial genome, as illustrated by the interchromosomal contact matrix (Fig. 1b). To dive further in complexity, we assembled and performed metaHiC on a mock community composed of 39 bacteria, 2 archaea and 3 free phage particles, including a total of 24 episomes made of 8 secondary chromosomes, 14 plasmids and 2 intracellular phages (Fig. 1c, Supplementary Table 1 and Methods). The resulting ~12 M paired-end (PE) metaHiC reads were processed to generate chromosomal contact maps of the community. The genome sequences represented along the *x* and *y* axes of the maps were either binned at the level of individual DNA molecules (Fig. 1d) or at a fixed resolution of 16 kb (Extended Data Fig. 1). The map is of high quality, as demonstrated by the low amount of *trans* contacts bridging genomes of different species ('noise' signal = 1.85%). Phage genomes embedded in particles remained isolated, in contrast with plasmid sequences and intracellular bacteriophage genomes that made important contacts, mostly with the genomes of their bacterial hosts (Fig. 1d).

These results were exploited to design a computational approach for characterizing mobile genetic elements (MGEs) in metaHiC data, taking advantage of a peculiar property of the MGE contact maps. Indeed, MGEs' *cis*-contact maps tend to display relatively uniform contacts besides a narrow diagonal of short-range contacts between pairs of adjacent DNA segments (that is, ≤2 kb apart) (black arrows, Fig. 1e,f). As a result, unordered contigs belonging to the same MGE genome display relatively similar *trans*-contact (that is, intercontigs) compared to *cis*-contact (that is, intracontigs) frequencies (excluding, of course, the short-range diagonal). This behaviour is not found in large bacterial genomes, where the contact decay as a function of genomic distance decreases gradually over genomic distance (Fig. 1g). We exploited these observations to develop a computational approach, integrated in an overhaul version of MetaTOR[22,26], that bins MGE-annotated contigs with comparable intra- and intercontig contact frequencies into MGE metagenome associated genomes ($_{MGE}$MAGs, either viral vMAGs or plasmid pMAGs) (Methods, Fig. 1h and Supplementary Fig. 1).

Virus binning from metagenome proximity-ligation data

We applied MetaTOR on complex natural datasets by seeking viruses in three human gut metaHiC libraries[11,22] (Hum22, Hum48 and Hum66; Supplementary Table 2). Following assembly, annotations and binning of each dataset independently, we detected an increased number of complete and HQ vMAGs compared with raw contigs (19 and 73 vs 11 and 34, respectively) (Fig. 1i), with 46 composed of 2 contigs or more. As another binning quality control, we sought for circular viral contigs that are probably complete and should not be pooled with extra contigs into a vMAG. Among the 84 viral contigs determined as circular by the assembly graph, only 5 were pooled with other contigs by MetaTOR (5.9%) (Fig. 1i). We further assessed the quality of the 46 HQ vMAGs made of multiple contigs by directly visualizing their individual contact maps[23]. Indeed, since control MGEs present a relatively uniform signal off their main diagonal (Fig. 1e,f), homogeneous contacts are expected within and between contigs belonging to the same genome and correctly pooled together. In contrast, an incorrect binning will result in contigs displaying different *cis* and *trans* contacts. We manually screened for aberrant signals in 2-kb-resolution contact maps for each of the 46 vMAGs: only one (2.1%) harboured an aberrant *trans*-contact signal (Supplementary Fig. 2, black arrow), while the others harboured a uniform (sometimes weak) pattern, suggesting a proper binning of viral contigs by the approach. These results are in sharp contrast with the highly contaminated vMAGs generated by SemBin[27] or ViralCC[28], alternate solutions that for instance respectively pooled 53 and 48 circular viral contigs (out of 84) with other contigs (Supplementary Fig. 3; see also Methods).

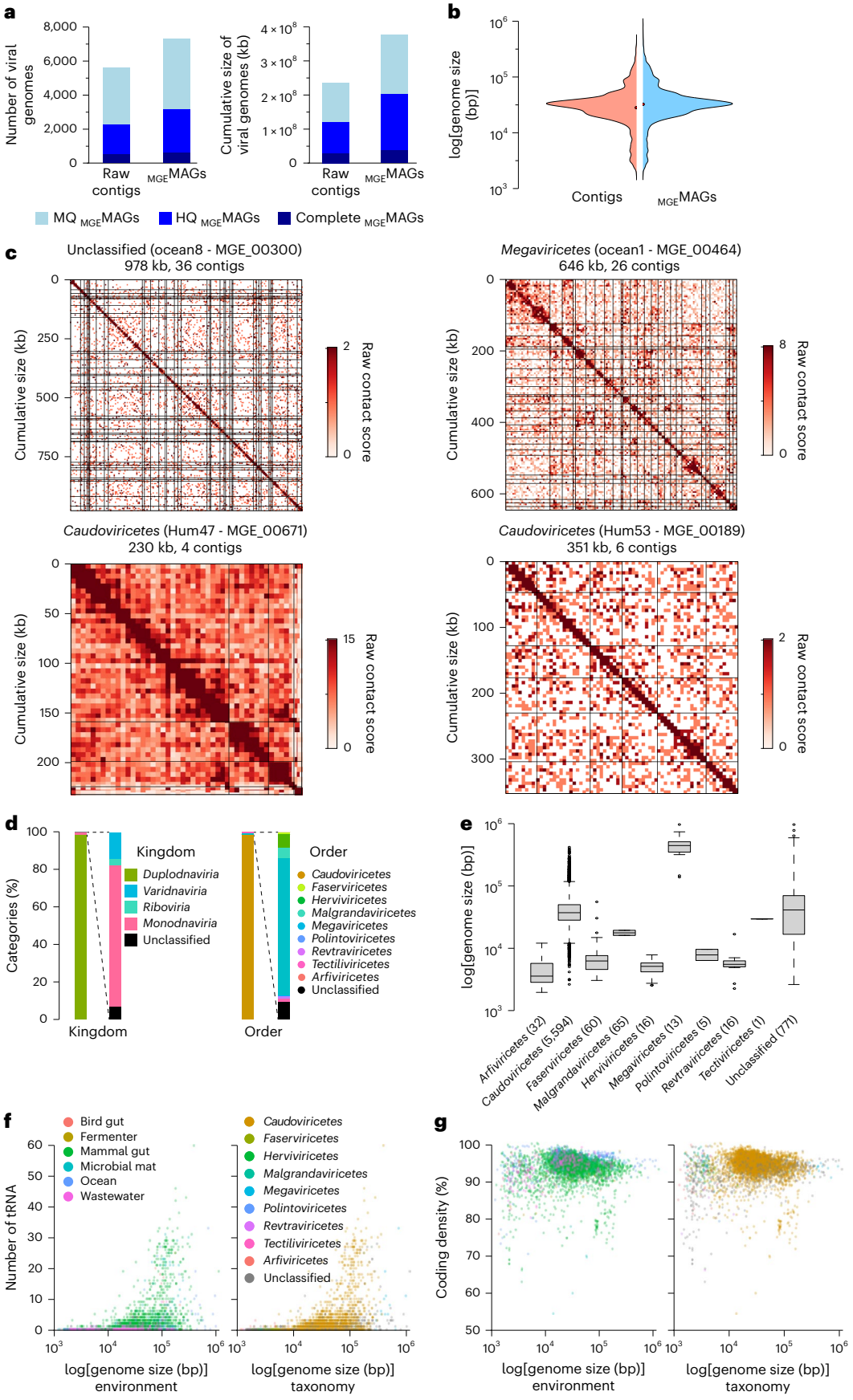
Altogether, these analyses demonstrate that MetaTOR generates dozens of vMAGs that have all the hallmarks of complete individual viral genomes from single metagenomics samples. It also stresses the value of performing visual inspection of contact maps to properly assess the quality of genome reconstruction from proximity-ligation datasets[23,29].

Reconstruction of virus genomes across 111 communities

We next investigated viruses and their potential hosts on metaHiC(-like) datasets from different ecosystems. A total of 84 datasets, corresponding to all published metagenomic samples ever processed by a proximity-ligation-like protocol, were collected[7,11,14,18,21–24,26–48]. We

Fig. 2 | MetaTOR MGE output from metagenomic datasets. a, Completeness of MGE contigs and MGE bins obtained using the MetaTOR MGE module according to CheckV in terms of number (left) and cumulative sequences (right). Only complete, HQ and MQ MGE genomes are shown (see also Supplementary Table 2). **b**, Violin plot of log(size) for the $_{MGE}$contigs (left) and $_{MGE}$MAGs (complete, HQ, MQ) (right). **c**, Raw contact map of vMAGs. Black lines delineate the different contigs. Annotation, sample, vMAG ID, size and number of contigs are indicated above the contact maps, while scale bars are indicated on the right. **d**, Taxonomic distribution at the kingdom and order level of the different characterized vMAGs according to geNomad software. **e**, Boxplot of the log(size) of the different vMAGs as a function of their taxonomic annotation at the order level (number of each representative is indicated in brackets; middle lines in the boxplots indicate mean, and boxplot limits indicate first and third quartiles). **f**, Plot of the number of tRNA genes detected as a function of viral genome size (log scale). Genomes are coloured as a function of their environment (left) or taxonomic annotation (right). **g**, Coding density computed as a function of genome size (log scale). The same as in **f**, viral genomes are coloured as a function of their environment (left) or their taxonomic annotation (right).

https://doi.org/10.1038/s41564-025-02108-2



broadened the diversity of the studied ecosystems by generating 27 new datasets, resulting in 111 samples from 5 different environments: animal gut (n = 89), oceanic filters (n = 8), hydrothermal mat (n = 10), wastewater (n = 3) and mezcal fermentation process (n = 1) (Extended Data Fig. 2 and Supplementary Table 2).

The 46.8 Gb dataset regrouping the different assemblies, including 597,064 MGE contigs, was independently segmented using MetaTOR into microbial (MAGs) and MGE ($_{MGE}$MAGs) genomes (Methods). First, 2,115 HQ and 2,860 MQ MAGs were recovered (1.35 Gb), consisting of 39 archaeal and 4,936 bacterial genomes spanning 18 phyla (Extended Data Fig. 3) (the overrepresentation of Firmicutes and Bacteroidetes reflects the prevalence of animal gut microbiomes in the datasets). Second, MetaTOR delivered 552,342 $_{MGE}$MAGs annotated and assessed for quality by geNomad[15] and CheckV[12]. Compared with the raw assembly and the assessment of single contigs, MetaTOR substantially increased the number of complete (555 vs 485; ~+15%), HQ (2,333 vs 1,540; ~+50%) and MQ (3,883 vs 2,824; ~+38%) viral genomes retrieved (Fig. 2a), resulting in 6,572 vMAGs of medium or higher quality kept for subsequent analysis. Compared with single contigs, we also noticed a net increase in the number of large vMAGs (from 321 contigs to 703 vMAGs above 100 kb), with genomes up to nearly 1 Mb in size (Fig. 2b). A careful examination of the contact map of different well-covered vMAGs showed that they encompass various numbers of contigs associated with coherent intercontigs signals, supporting the accuracy of the approach (Fig. 2c and Extended Data Fig. 4). Among these reconstructed vMAGs, 693 displayed plasmid annotation by geNomad, with 6 ranging in size between 500 and 978 kb. This result could stem from an incorrect annotation by geNomad or CheckV, but could also suggest the presence of plasmid prophages[40] and/or phage-plasmids[41] in the data. More analysis would however be required to further characterize the nature of these genomic entities that in the meantime remain included in the 6,572 vMAGs.

Reconstructed viruses exhibit wide diversity

Taxonomic annotation shows that more than 80% of the 6,572 vMAGs belong to the phylum *Duplodnaviria* and to the order *Caudoviricetes* (Fig. 2d). Among members of this clade of phages, we detected a small portion of *Crassvirales* (n = 138), while most others could not be assigned to a family. Among other phyla, we characterized different members (n = 208) belonging to the orders *Arfiviricetes*, *Malgrandaviricetes*, *Faserviricetes*, *Revtraviricetes* and *Megaviricetes*. Size distribution analysis of the reconstructed vMAGs according to their order showed that *Megaviricetes* have the largest genomes and *Arfiviricetes* have the smallest ones (Fig. 2e). We also binned several *Caudoviricetes* vMAGs with genomes larger than 100 kb (n = 80), up to 414 kb. Following functional annotation of the vMAGs using Pharokka[42], we identified a wide variety of genes involved in lysis, nucleic acid metabolism, structural proteins or host takeover, but also integration, indicating that we characterized potential temperate phages. The largest vMAGs encoded a high number of transfer (t)RNAs, with one *Caudoviricetes* vMAG displaying 60 tRNAs (Fig. 2f). Most of the vMAG genomes exhibited a high coding density (mean = 95.3%; Fig. 2g), with some notable exceptions (28 vMAGs with a coding density lower than 80% and down to 55%) suggesting that some of them could use a genetic code different from the standard one, a phenomenon rarely reported for viruses[43].

An important proportion of viruses exhibits multiple hosts

MetaHiC's main strength is its ability to assign vMAGs to their host within the same sample, thanks to the quantification of collisions between DNA molecules[23]. While existing pipelines use either raw contact[28] or binning output[22] for this task, we decided to exploit the proportion of normalized contact between vMAGs and MAGs, as it takes into account different biases such as coverage or bin length[44] (Methods). To do so, MetaTOR starts by computing for each sample a noise-to-signal ratio of contacts made by the contigs of each reconstructed MAG. It then extracts the subnetwork of contacts made by each vMAG with the different MAGs from the same sample and applies a series of filtering steps to retain the most likely and reliable connections. First, the vMAG must be bridged with one contact or more to at least 5 contigs from each considered MAG. Second, the contact score must be above the noise threshold computed above. The distribution of vMAG *trans* contacts is then characterized: (1) if most (>50%) normalized contacts involve a single MAG, the latter is considered the host; (2) if the vMAG makes between 10 and 50% of its normalized *trans* contacts with multiple MAGs, we consider several hosts; (3) finally, if *trans* contacts with any MAGs represent less than 10% of the total, we considered that the vMAG is not associated with a reliable host.

Among the 6,572 MQ/HQ vMAGs, 56% (n = 3,689) did not pass these thresholds and could not be accurately assigned to at least one host (Fig. 3a). This could be due to (1) a lack of coverage or poor contact data, (2) the absence of the host in the MAGs dataset (also eventually because of poor data), or (3) the fact that they are present as viral particles and therefore are not in contact with a microbial host, as metaHiC captures genomes in free viral particles (Fig. 1c–e).

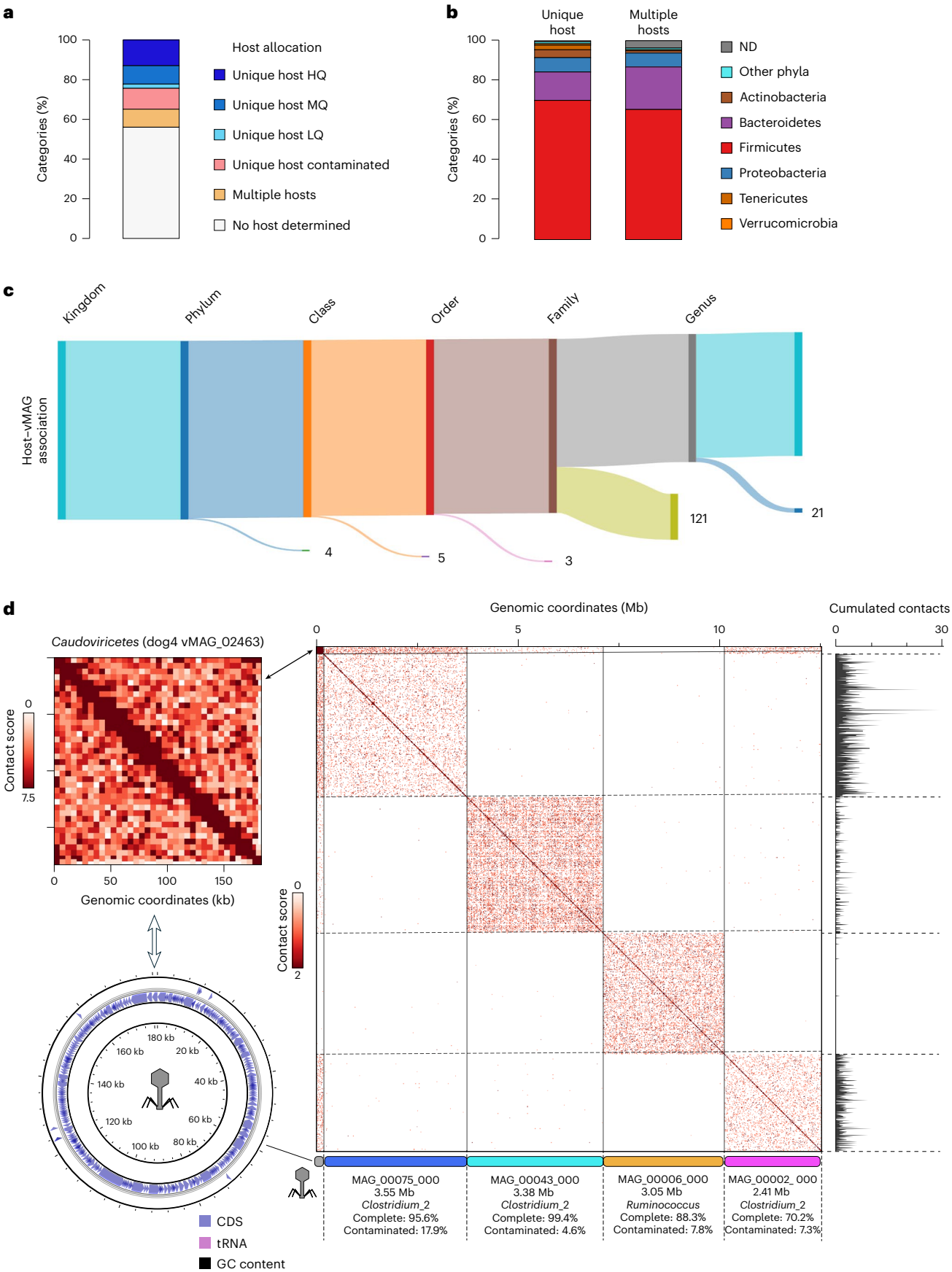
Of the remaining 2,883 host-associated vMAGs, 2,366 (83%) were unambiguously associated with a single HQ (n = 853), MQ (n = 609), LQ (n = 141) or contaminated (n = 689) MAG (note that coinfection events, when several vMAGs are bridged with the same MAG in a community, appear rare and remain under investigation). Firmicutes, Bacteroidetes and Proteobacteria were the most represented phyla, reflecting the datasets' compositions (Fig. 3b). Firmicutes and Proteobacteria hosts were overrepresented in the gut/fermenter and in oceanic/wastewater samples, respectively. A total of 754 vMAGs were assigned to MAG hosts characterized at the genus level, including 305 at the species level. The remaining 487 vMAGs (17% of all host-associated vMAGs) were associated with more than one bacterial MAG and up to 8 different hosts. Among those 487 vMAGs with multiple hosts in the same sample, 4, 5, 3, 121 and 21 interacted with bacteria from different phyla, classes, orders, families and genera, respectively (Fig. 3c). The increased proportion of vMAGs interacting with different families appears to be mainly due to the increased proportion of hosts for whom a genus was not assigned.

A pattern of multiple host interactions is a common feature

To validate these observations, we exploited the qualitative power of Hi-C contact maps. We generated the *cis*- and *trans*-contact maps of the

Fig. 3 | Host attribution to the vMAGs. a, Bar plot of the proportion of vMAGs as a function of their host attributions. The different categories are indicated by colours (white, no host attributed; orange, multiple hosts attributed; red, one contaminated host attributed; grey, one LQ host attributed; blue, one MQ characterized host attributed; dark blue, one HQ characterized host attributed). **b**, Taxonomy (at the phylum level) of the different hosts for the vMAGs with one host attributed (left) or multiple hosts attributed (right). **c**, Sankey plot representing the multiple host vMAGs and showing host divergence at the kingdom, phylum, class, order, family and genus level. For each level, the number of vMAGs with divergent hosts is given. For instance, the 121 indicates that 121 vMAGs have at least 2 hosts that do not belong to the same family level. **d**, Example of intra- and intercontact map obtained for a vMAG exhibiting multiple hosts. Left: contact map of the vMAG#02463 (sample, dog4; virus, *Caudoviricetes*) (2 kb bins) and composed of a unique contig of 181 kb and determined as complete by CheckV. A circular map of its genome is displayed below the contact map. Right: global contact map (50 kb bins) of the vMAG (top left) and four bacterial MAGs from the same sample. Dashed lines indicate borders of the different genomic entities. Schemes of the four MAGs as well as their identification number, size, taxonomic classification (order) and CheckM evaluation are presented under the contact map. The scale bar is indicated on the left. Histogram representing the raw contact scores of the vMAG with the different parts of each MAG is plotted on the right part of the contact map. CDS, coding sequence; ND, not determined. Panel **d** adapted from SVG Silh under a CC 1.0 license.

https://doi.org/10.1038/s41564-025-02108-2



144 complete and HQ vMAGs exhibiting at least 50 contacts with 2 or more MAGs of the same sample. While 55 maps exhibited low or ambiguous signals, the remaining 89 maps displayed discrete, non-ambiguous and homogeneous contact patterns with several microbial MAGs, fully consistent with the observed infection patterns experimentally characterized (Figs. 1a and 3d, and Extended Data Fig. 5). For instance, a large *Caudoviricetes* vMAG composed of a unique and circular 180 kb contig (Fig. 3d, left) exhibited multiple contacts with three different bacterial MAGs belonging to the *Clostridium* genus (Fig. 3d). These contacts probably reflect infection interactions in the same sample as supported by several observations: first, an absence of noisy and sporadic contacts between these 3 MAGs, suggesting that they have not contaminated each other; second, no important *trans* contacts between the MAGs; third, a uniform contacts distribution between the vMAG and MAG contigs discarding the possibility that the viral sequence is integrated within the bacterial genomes (Fig. 3d, cumulated contact). Multiple similar examples were found for other vMAGs (Extended Data Fig. 5), suggesting that multihost infection is frequent. These vMAGs tend to be slightly more abundant and are present in all environments except the hydrothermal mat, possibly reflecting the stringent filters applied and different levels of sequencing depth and complexity, respectively (Extended Data Fig. 6). To further explore this phenomenon, we assembled a proteomic tree of the 6,572 vMAGs using ViPTree[45] (Fig. 4a) (Methods) and analysed it with respect to their hosts. The resulting tree recapitulates viruses' taxonomic annotations as defined by the International Committee on Taxonomy of Viruses (ICTV), and confirms the general behaviour that related phages sharing protein features tend to (1) infect phylogenetically related hosts and (2) have relatively similar genome sizes, suggesting that genome sizes are somehow clade specific (Fig. 4a)[43]. On the other hand, the tree also reveals that some related phages can infect hosts across divergent bacterial genera (Fig. 4a, black rectangle; Fig. 4b). Interestingly, the 693 vMAGs annotated as plasmids by geNomad were distributed in one large cluster but also in different small clusters along the tree (Fig. 4a, geNomad annotation ring), potentially reflecting existing links between phages and plasmids[41]. Finally, in several instances, the same phage was able to interact with multiple hosts, and this is found all over the proteomic tree (Fig. 4a, red dots). Exceptions to this pattern were also distributed throughout the tree, notably the *Crassvirales* family. Indeed, the data contain 107 crass phages (Fig. 4a, red rectangle, and Extended Data Fig. 7), a phage family abundant in the human gut whose hosts have been identified as predominantly Bacteroidetes[46,47]. Several publications suggested that they exhibit pseudolysogenic[46] or phage-plasmid lifestyles[48], with potentially several hosts for the same virus. We only assigned 7 of the 107 crass phages (6%) to a Bacteroidetes host and none of them exhibits a multihost contact pattern (Extended Data Fig. 7). These results, therefore, do not confirm a multihost pattern for crass phages. They also suggest that crass phages are mostly found as viral particles in human faecal samples while the host remains in the intestine, or that they efficiently kill their hosts.

Overall, these analyses demonstrate that multihost phage interactions are common in the viral world and highlight the complexity of phage–bacterial interactions (Fig. 4a, red dots).

Discussion

This study sheds light on MGE diversity of complex microbial communities, offering a deeper insight into their genomic characteristics, host associations and potential ecological implications. We leverage computational analysis of proximity-ligation data to overcome existing challenges associated with virus genome binning and host assignment in single complex microbial samples. The low level of false positives of MetaTOR is readily apparent from the corresponding vMAG and MAG contact maps. The pipeline notably maintains the individuality of contigs that display characteristics of good-quality viral genomes, as illustrated by both the circularity of the assembly graph and the circularization signal in the contact map. MetaTOR was deployed without adaptation to diverse environmental datasets. The reconstruction of nearly 6,000 vMAGs of various sizes from different phyla, including *Megaviricetes*, highlights the importance of metaHiC in unveiling viral diversity in individual environmental samples. This catalogue of viral genomes represents a resource for future studies on viral ecology and evolution. Moreover, the identification of plasmid annotated sequences within vMAGs and the potential of the approach to scaffold plasmid MAGs pave the way for further analysis intending to unveil plasmids in environmental samples.

A relatively large proportion of phages (17% of all host-associated vMAGs) exhibit interactions with more than one host in the same community, sometimes infecting bacteria belonging to very different clades. Having multiple hosts is expected for plasmids[49,50] but less so for phages. A recent proximity-ligation-based work suggested that this could be the case in dense ecosystems such as biofilm[7], but stringent reprocessing of these data by MetaTOR did not support the claim, at least in these samples (Extended Data Fig. 6). In contrast, we observe multihost interactions in all other ecosystems, including gut, ocean and wastewaters. In addition, these patterns are not restricted to a subset of phage families but are widely distributed throughout the viral world. This result could be particularly relevant when considering horizontal gene transfer dynamics mediated by phages in natural communities. The analysis does not currently allow us to conclude that the interactions detected are associated with successful infection cycles and the production of viral particles. However, the study paves the way for in-depth investigation of the factors responsible for these broad spectra.

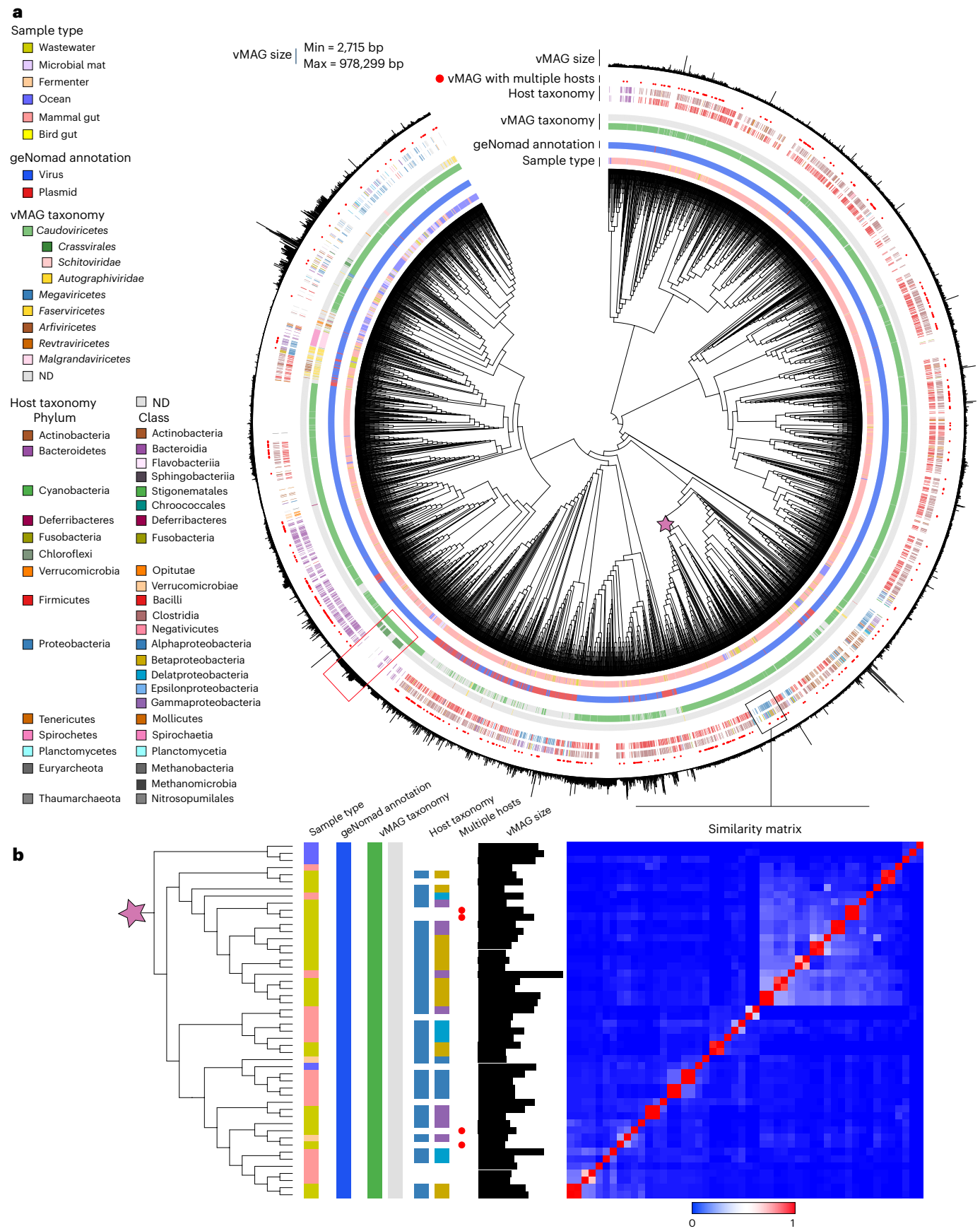
Our work also opens prospects for applications at several levels. First, the fact that covariance analysis is not needed brings new perspectives regarding work on rare ecosystems, as it allows drawing of strong MGE–host links from a single sample. This could for instance help identify phages targeting bacteria of interest in any natural environment of interest. Second, as phage infection spectra appear not so stringent, it could be interesting to apply metaHiC-type experiments in phage therapy approaches to identify potentially overlooked phage targets. More generally, contacts between DNA molecules would also make it possible not only to validate the correct delivery of a gene or plasmid vector to a target species or strain within a complex population, but also to identify potential off-targets in targeted delivery systems[51].

Future works will further deepen the exploitation of these datasets, as the characterization of vMAGs is just one aspect of the diversity of DNA episomes present in populations. In addition, our past work has shown that contact map analyses can be used to characterize prophage activation in bacterial genomes in microbial communities[17,23,52]. The future integration of all these analyses will enable us to take the exploration of complex population time series under antimicrobial or environmental stress to new levels of resolution.

We therefore anticipate that, by enabling us to better characterize the dynamics and balance of complex microbial populations, this

Fig. 4 | Proteomic tree of the characterized vMAGs. a, Proteomic tree of all the complete, HQ and MQ vMAGs as assessed by CheckV, computed using ViPTree and visualized using iTOL. The tree is decorated with different information (from the inside to the outside): (i) sample type, (ii) geNomad annotation (blue, virus; red, plasmid), (iii) vMAG taxonomy (order, genus), (iv) host classification (phylum, class, white if no host or multiple hosts attributed), (v) vMAG exhibiting multiple hosts (red circles) and (vi) vMAG genome size (min = 2.7 kb, max = 978 kb). The red rectangle indicates viruses from the *Crassvirales*. The black rectangle indicates a zoom inside the tree where related viruses infect divergent hosts (**b**).
b, Pruned proteomic tree corresponding to the genomes encompassed in the black rectangle. A similarity matrix of vMAGs computed by ViPTree is indicated on the right of the tree. A scale bar of the similarity matrix is also indicated.

https://doi.org/10.1038/s41564-025-02108-2



work will have broader implications for understanding the ecology and evolution of microbial communities.

Methods

Ethics approval and consent to participate

Faecal samples from a total of 1 mouse, 3 ducks and 4 dogs were recovered from collaborators in the Netherlands. The proposed activities were reviewed by the Medical Research Ethics Committee of UMC Utrecht (deposited under METC, protocol number 18-139/C). A total of 17 human stool samples from Uruguay were provided by G.I. in accordance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization. G.I. is registered in the National Registry of Researchers in Genetic Resources and Derivatives in the Environment Ministry. In the context of this registration, access to genetic material and derivatives is allowed for non-commercial purposes (Article 22, Law #17.283). The stool samples were collected with written informed consent from all participants. Before providing consent, all participants received comprehensive information about the project and had the opportunity to opt out of the study. The project was approved by the Ethics Committee and Institutional Review Board of CASMU (Centro Asistencial del Sindicato Médico del Uruguay, ref: FSGSK_1_2019_1_159735), and the study was conducted following national legislation (study performed from 2018 to 2020). The project was registered in the Ministry of Public Health (Ministerio de Salud Pública MSP, ref 39814). One infant faecal sample was recovered from the MetaKids cohort. The MetaKids study (No. ID-RCB: 2017-A00750-53, clinical trial: NCT03296631) received ethics approval from the ethics committee Comité de Protection des Personnes Sud Est 1 on 21 July 2017 as required by French regulation on clinical research.

The oceanic sample was recovered in the bay of Napoli during a sampling campaign in collaboration with Genoscope (Evry, France) and the Stazione Zoologica Anton Dohrm (Napoli, Italy) in agreement with national legislation. The Mezcal fermentation sample was recovered in Guadadupe hacienda in Mexico in collaboration with the University of Queretaro (UNAM). Samples were processed in Queretaro and the resulting data were transferred to Institut Pasteur for analysis.

The metadata associated with the different samples generated in the present study are provided in Supplementary Table 3.

Phage bacterial infection Hi-C libraries

The bacterial strain PAK (*P. aeruginosa*) was grown until mid-exponential phase (optical density (OD) = 0.3) in 80 ml of LB medium. Phage PAK-P3 was then added at a multiplicity of infection (MOI) of 25 (PhiKZ = 15). Aliquots (7 ml) of the phage bacterial culture were removed at two time points after the addition of the phage (0 and 5 min) and directly transferred to 50 ml falcon tubes containing 1 ml of formaldéhyde (35.5–37%). Tubes were then incubated for 30 min under shaking at r.t. followed by incubation at 4 °C during another 30 min. Formaldehyde was quenched by adding 2 ml 2.5 M glycine, followed by a 20-min incubation under shaking. Pellets were recovered by centrifugation (10 min, 10,000 × *g*, 4 °C), washed in PBS and recentrifuged using the same settings. Pellets were then frozen in dry ice and stored at −80 °C until use.

Mock community construction

The different bacterial strains used to construct the mock community (Supplementary Table 1) were grown in appropriate media and conditions until reaching mid-exponential phase. Bacteria (10^9) were recovered for each culture, formaldehyde was added (3% final concentration) and the mixture was incubated under agitation at r.t. for 1 h. Glycine (2.5 M stock solution) was then added to reach a final concentration of 125 mM, followed by incubation at r.t. for 20 min under gentle agitation. The solution was centrifuged (6,000 *g*, 10 min, 4 °C) and the pellet was washed in 1× PBS. After a similar centrifugation, the supernatant was removed, and the pellet was flash frozen and stored at −80 °C. Once the different pellets were obtained, they were thawed and mixed in random proportions to create the mock community. The large pellet obtained was then split in 10 aliquots that were stored at −80 °C until use.

Faecal sample collection

Faecal samples were recovered from human adults ($n = 17$), human child ($n = 1$), duck ($n = 2$), dog ($n = 4$) and mice ($n = 1$), flash frozen in liquid nitrogen, shipped with dry ice when necessary and stored at −80 °C at reception. For each sample, 50 mg was thawed directly in 50 ml of crosslinking solution (1×PBS supplemented with 3% formaldehyde) and incubated for 1 h at r.t. under strong agitation. Formaldehyde was quenched by adding glycine (125 mM) during 20 min at r.t. under gentle agitation. Pellets were recovered by centrifugation, washed with 10 ml 1× PBS, recentrifuged and split in two aliquots of 25 mg that were stored at −80 °C until use.

Oceanic sample collection

The oceanic sample was recovered in the bay of Napoli using 40 l of water, filtered through a 200 µm filter. The flowthrough was then sequentially filtered through filters of different sizes and using a peristaltic pump (20 µm, 3 µm and 0.2 µm). Filters (fraction 0.2–3 µm) were flash frozen in liquid nitrogen before being fixed independently in 50 ml of crosslinking solution (1×PBS supplemented with 3% formaldehyde) and incubated for 1 h at r.t. under strong agitation. The filter was removed from the tube and formaldehyde was quenched by adding glycine (125 mM) during 20 min at r.t. under gentle agitation. Samples were recovered by centrifugation, washed with 10 ml 1× PBS, recentrifuged and stored at −80 °C. Samples were shipped with dry ice and stored at −80 °C at reception until use.

Mezcal sample collection

Fermenter samples were recovered from Guadadupe hacienda in Mexico during the Mezcal process. A volume of 50 ml of the fermentation product was recovered at several time points after the addition of the pulque to the agave juice. Samples were immediately fixed independently in 200 ml of crosslinking solution (1×PBS supplemented with 3% formaldehyde) and incubated for 1 h at r.t. under strong agitation. Formaldehyde was quenched by adding glycine (125 mM) during 20 min at r.t. under gentle agitation. Samples were then stored at 4 °C until brought to the laboratory. Samples were then recovered by centrifugation, washed with 10 ml 1× PBS, recentrifuged and stored at −80 °C until processing.

MetaHiC library preparation

MetaHiC libraries were generated using the ARIMA Hi-C+ kit and following the manual provided by ARIMA with some changes. First, samples (between 20 and 50 mg depending on the samples) were resuspended in 1 ml of sterile water and transferred to 2 ml Precellys tubes containing glass beads of 0.1- and 0.5-mm diameter (Precellys, Bertin Technology). Cells and matrices were disrupted using the Precellys apparatus (Precellys Evolution) and the following programme (7,500 r.p.m., 6 cycles 30 s ON / 30 s OFF, 4 °C). Tubes were then centrifuged for 1 min at 1,000 *g*, and 700 µl of lysate was recovered and transferred to a new 1.5 ml Eppendorf tube. The tube was centrifuged for 20 min at 16,000 *g* at 4 °C. Supernatant was carefully removed and the pellet was resuspended in 45 µl of water. We then followed the ARIMA protocol. The Hi-C library was purified using AMPure beads and eluted in 130 µl of water before processing for sequencing as described previously[22]. Libraries were then sequenced on different Illumina apparatus (Nextseq, Novaseq).

Shotgun library preparation

DNA was extracted from the different samples using the ZymoBIOMICS DNA Microprep kit following manufacturer instructions First, all samples were homogenized and disrupted using a Precelllys apparatus

and the same process as described for the metaHiC libraries before purification. DNA quality was assessed by gel electrophoresis, quantified with the Qubit 3.0 fluorometer and processed for sequencing using the Colibri kit (ThermoFisher). Final shotgun library concentrations were quantified using the Qubit fluorometer and then sequenced using different Illumina apparatus (Nextseq, Novaseq).

Data acquisition

The different published metaHiC/3C libraries and their associated shotgun data were recovered from the SRA and ENA databases using the internet site, SRA explorer (https://sra-explorer.info/#) (Supplementary Table 2).

Metagenome assembly and annotation

Shotgun libraries coming from the same sample were pooled and cleaned using Trimmomatic 0.39 (ILLUMINACLIP: TruSeq3_PE_adapt.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36) and used as input for megahit (v.1.2.9)[53] with default parameters. The quality of the assembly was assessed using quast (v.5.2.0)[54]. geNomad software (v.1.4)[15] was then used to detect and annotate phage and plasmid contigs using the end-to-end pipeline and default parameters. All contigs annotated as MGEs were then evaluated using CheckV[12] and the ones annotated as prophages were removed for subsequent analysis.

MetaTOR v.1.2.8

The new version of MetaTOR is freely available at https://github.com/koszullab/metaTOR along with an extensive README file and a tutorial. MetaTOR is now easily installable through bioconda and pip (Supplementary Fig. 1). We have added an [mge] module to the new version of MetaTOR to bin MGE contigs and assign a host to the resulting bins. To bin MGE contigs, the pairs mapped onto the targeted contigs are extracted and pairs at a distance of less than 1 kb are removed. Inter-contig contacts are then normalized using the geometric mean of the intracontig contact, generating a score $S_{i,j} = x_{i,j}/\mathrm{sqrt}(x_{i,i}, x_{j,j})$ where $x_{i,j}$ are the contacts between the contigs i and j. The highest score is binned first, and the two binned contigs are processed as one to compute a new score based on the sum of their contacts with the others contigs. Contigs are binned together until there are no more scores above the 0.8 threshold. The resulting bins are the $_{MGE}$MAGs. We have also developed a quality check of the proximity-ligation library based on the 3D ratio, informative reads and an estimate of the noise signal between MAGs. Finally, to simplify downstream analysis, we connected our pipeline to hicstuff[55] to visualize the reconstructed MAGs contact matrices and to Anvi'o[66] to manually clean the MAGs and integrate multi-omics analysis. Therefore, the new version of our MetaTOR software offers an easy-to-use modulable pipeline integrated with other analysis platforms.

MAG binning and annotation

First, PE reads from metaHiC/3C libraries were cleaned using Trimmomatic 0.39 (ILLUMINACLIP: TruSeq3_PE_adapt.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36), followed by an in silico digestion step using the hicstuff cutsite module with the enzymes used in the corresponding experiment and the all-versus-all method (hicstuff (v.3.1.5)[55]) to further leverage the 3D signal (Extended Data Fig. 1). Indeed, since the sequencing fragments can be longer than the restriction fragments (for example, they are ~300 bp in length compared with a median size of 90 bp for the restriction fragments in our mock community), a sequencing read can contain multiple digestion–religation events[56]. Their digestion at the ligation sites breaks down these reads, generating new reads and ultimately increasing the number of relevant 3D contacts. Applying the digestion step to the data of our mock community, we observed an increase in the number of PE reads from 61 million to 143 million, resulting in a 20% increase in the number of informative reads without increasing noise signal (1.91% compared with 1.85% initially), demonstrating the usefulness of this digestion step. The digested reads and their corresponding assemblies were then used as input for MetaTOR v.1.2.8, with default parameters (100 Louvain iterations and overlapping threshold of 80% for the iterative procedure; 10 iterations and overlapping threshold of 90% for the recursive procedure). MAGs generated after the recursive step were then evaluated and taxonomically annotated using CheckM[57] and classified in terms of quality using the following criteria[58]: HQ MAGs, completion rate of at least 90% and contamination rate below 5%; MQ MAGs, completion rate of at least 50% and contamination rate below 10%; LQ MAGs, completion rate below 50% and contamination rate below 10%; contaminated MAGs, contamination rate above 10%.

vMAG binning, host association and annotation

Contigs annotated as viral or plasmid by geNomad were used as input for the mge module of MetaTOR using default parameters (binning score = 0.8, association score = 0.1). Resulting $_{MGE}$MAGs were then evaluated using CheckV. Each vMAG of medium quality or higher as assessed by CheckV was then annotated using geNomad and Pharokka[42].

Contact map generation and visualization

Contact maps were generated using the MetaTOR contact map module and hicstuff. The reads were aligned using bowtie2 (v.2.2.4.5)[59] with the very sensitive local mode, a mapping quality threshold of 30 was applied, and PCR duplicates were removed. Contact maps were then binned at the desired resolution and balanced using Cooler (v.0.8.11)[60]. Plots were displayed using different R packages[61] and a linear scale with a maximum value corresponding to 99% of the maximum value contained in the contact map.

MetaTOR benchmark

To evaluate MetaTOR, we compared it to existing tools ViralCC[28] and SemiBin[27]. We fed the different pipelines with viral contigs detected by geNomad for datasets Hum22, Hum48 and Hum66. We first evaluated the global output of the pipelines by assessing completion of the obtained vMAGs using CheckV (Supplementary Fig. 3a). To go beyond CheckV evaluation, we also assessed the quality of complete or/and HQ $_{MGE}$MAGs composed of more than one contig by directly visualizing individual vMAG contact maps. Indeed, contact maps of well-binned vMAGs should display a relatively uniform signal off the main diagonal (Fig. 1c). In contrast, an incorrect binning of DNA segments belonging to different genomes will pool together contigs displaying different *cis* and *trans* contacts. We generated 2-kb-resolution contact maps for each of the vMAGs composed of several contigs and obtained by ViralCC, SemiBin and MetaTOR. Also, in contrast to MetaTOR, ViralCC and SemiBin often clustered together contigs with one or more contigs displaying a circular signal and therefore are likely to already represent individual phage (Supplementary Fig. 3b,c). This strongly suggests contaminations between complete genomes that should rather be left as single entities. Indeed, among the 84 contigs detected as circular by Megahit (25 of them characterized as complete viral genome by CheckV), 48 and 53 were pooled with other contigs by ViralCC and SemiBin, respectively (compared with 5 by MetaTOR, including one complete genome). Altogether, these analyses demonstrate that MetaTOR is able to generate $_{MGE}$MAGs with little contamination.

vMAG proteomic tree

The proteomic tree was built using the latest version of ViPtree[45] with default parameters. The vMAGs proteomic tree was then annotated and visualized using iTOL (v.5)[62]. A tree for *Crassvirales* and associated viruses was computed the same way by solely adding reference Crass genomes in the input fasta file provided to ViPTree using data from ref. 47.

Statistics and reproducibility

No data were excluded from the analyses. Statistical tests were done using R.

Reporting summary

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

Data availability

Sequence data as well as raw assemblies generated in this study have been deposited in the NCBI under the BioProject number PRJNA1169672 (mock community) and PRJNA1169674 (metagenomic datasets). Publicly available datasets as well as newly produced data used in the present study are all listed in Supplementary Table 2 with their associated BioProject ID. All MAG and $_{MGE}$MAG data are provided as supplementary datasets and are available on Zenodo at https://doi.org/10.5281/zenodo.14851637 (ref. 63).

Code availability

Open-access versions of the programmes and pipelines used (Hicstuff, MetaTOR, HiContacts) are available online on GitHub: Hicstuff v.3.1.2 (https://github.com/koszullab/hicstuff), MetaTOR v.1.3.2 (https://github.com/koszullab/metaTOR) and HiContacts v.1.7.1 (https://github.com/js2264/HiContacts). Other mandatory programmes are also available online: Bowtie2 v.2.4.5 (http://bowtie-bio.sourceforge.net/bowtie2/), SAMtools v.1.9 (http://www.htslib.org/) and Cooler v.0.8.7–0.8.11 (https://cooler.readthedocs.io/en/latest/). The pipeline used in the present study to generate the data is available on GitHub at https://github.com/mmarbout/MetaHiC_pipeline (ref. 64).

References

1. Salmond, G. P. C. & Fineran, P. C. A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13**, 777–786 (2015).
2. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
3. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
4. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222 (2013).
5. Shousha, A. et al. Bacteriophages isolated from chicken meat and the horizontal transfer of antimicrobial resistance genes. *Appl. Environ. Microbiol.* **81**, 4600–4606 (2015).
6. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* https://doi.org/10.1093/femsre/fuv048 (2015).
7. 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.* **8**, 946–957 (2023).
8. Hedžet, S., Rupnik, M. & Accetto, T. Broad host range may be a key to long-term persistence of bacteriophages infecting intestinal Bacteroidaceae species. *Sci. Rep.* **12**, 21098 (2022).
9. Göller, P. C. et al. Multi-species host range of staphylococcal phages isolated from wastewater. *Nat. Commun.* **12**, 6965 (2021).
10. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
11. Gounot, J.-S. et al. Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nat. Commun.* **13**, 6044 (2022).
12. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
13. Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G. & Belcaid, M. CoCoNet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics* **37**, 2803–2810 (2021).
14. Johansen, J. et al. Genome binning of viral entities from bulk metagenomics data. *Nat. Commun.* **13**, 965 (2022).
15. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01953-y (2023).
16. Roux, S. et al. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
17. Marbouty, M. et al. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* **3**, e03318 (2014).
18. Beitel, C. W. et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
19. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* **4**, 1339–1346 (2014).
20. Marbouty, M. & Koszul, R. Metagenome analysis exploiting high-throughput chromosome conformation capture (3C) data. *Trends Genet.* **31**, 673–682 (2015).
21. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *ISME J.* https://doi.org/10.1038/s41396-019-0446-4 (2019).
22. Marbouty, M., Thierry, A., Millot, G. A. & Koszul, R. MetaHiC phage–bacteria infection network reveals active cycling phages of the healthy human gut. *eLife* **10**, e60608 (2021).
23. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host–virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* **3**, e1602105 (2017).
24. Chen, Y., Wang, Y., Paez-Espino, D., Polz, M. F. & Zhang, T. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. *Nat. Commun.* **12**, 5398 (2021).
25. Chevallereau, A. et al. Next-generation '-omics' approaches reveal a massive alteration of host RNA metabolism during bacteriophage infection of *Pseudomonas aeruginosa*. *PLoS Genet.* **12**, e1006134 (2016).
26. Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R. & Marbouty, M. MetaTOR: a computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (meta3C) libraries. *Front. Genet.* **10**, 753 (2019).
27. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep Siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).
28. Du, Y., Fuhrman, J. A. & Sun, F. ViralCC retrieves complete viral genomes and virus–host pairs from metagenomic Hi-C data. *Nat. Commun.* **14**, 502 (2023).
29. Marie-Nelly, H. et al. High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
30. Bickhart, D. M. et al. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol.* **20**, 153 (2019).
31. Yaffe, E. & Relman, D. A. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat. Microbiol.* https://doi.org/10.1038/s41564-019-0625-0 (2019).
32. Press, M. O. et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. Preprint at *bioRxiv* https://doi.org/10.1101/198713 (2017).

33. Kalmar, L. et al. HAM-ART: an optimised culture-free Hi-C metagenomics pipeline for tracking antimicrobial resistance genes in complex microbial communities. *PLoS Genet.* **18**, e1009776 (2022).
34. Varona, N. S. et al. Host-specific viral predation network on coral reefs. *ISME J.* **18**, wrae240 (2024).
35. DeMaere, M. Z. et al. Metagenomic Hi-C of a healthy human fecal microbiome transplant donor. *Microbiol. Resour. Announc.* **9**, e01523-19 (2020).
36. Rojas, C. A., Gardy, J., Eisen, J. A. & Ganz, H. H. Recovery of 52 bacterial genomes from the fecal microbiome of the domestic cat (*Felis catus*) using Hi-C proximity ligation and shotgun metagenomics. *Microbiol. Resour. Announc.* **12**, e0060123 (2023).
37. Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J. & Brito, I. L. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* **11**, 4379 (2020).
38. Ivanova, V. et al. Hi-C metagenomics in the ICU: exploring clinically relevant features of gut microbiome in chronically critically ill patients. *Front. Microbiol.* **12**, 770323 (2022).
39. Bickhart, D. M. et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* **40**, 711–719 (2022).
40. Piligrimova, E. G. et al. Putative plasmid prophages of *Bacillus cereus* sensu lato may hold the key to undiscovered phage diversity. *Sci. Rep.* **11**, 7611 (2021).
41. Pfeifer, E., Moura de Sousa, J. A., Touchon, M. & Rocha, E. P. C. Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).
42. Bouras, G. et al. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2022).
43. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
44. Du, Y. & Sun, F. HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome Biol.* **23**, 63 (2022).
45. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
46. Shkoporov, A. N. et al. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in *Bacteroides intestinalis*. *BMC Biol.* **19**, 163 (2021).
47. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
48. Schmidtke, D. T. et al. The prototypic crAssphage is a linear phage-plasmid. *Cell Host Microbe* **33**, 1347–1362.e5 (2025).
49. Beaulaurier, J. et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).
50. Ravi, A., Valdés-Varela, L., Gueimonde, M. & Rudi, K. Transmission and persistence of IncF conjugative plasmids in the gut microbiota of full-term infants. *FEMS Microbiol. Ecol.* **94**, fix158 (2018).
51. Brödel, A. K. et al. In situ targeted base editing of bacteria in the mouse gut. *Nature* **632**, 877–884 (2024).
52. Lamy-Besnier, Q. et al. Chromosome folding and prophage activation reveal specific genomic architecture for intestinal bacteria. *Microbiome* **11**, 111 (2023).
53. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
54. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
55. Matthey-Doret, C. et al. koszullab/hicstuff: use miniconda layer for docker and improved P(s) normalisation. *Zenodo* https://doi.org/10.5281/zenodo.4066363 (2020).
56. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
60. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2019).
61. Serizay, J., Matthey-Doret, C., Bignaud, A., Baudry, L. & Koszul, R. Orchestrating chromosome conformation capture analysis with Bioconductor. *Nat. Commun.* **15**, 1072 (2024).
62. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
63. Marbouty, M. Phages with a broad host range are common across ecosystems. *Zenodo* https://doi.org/10.5281/zenodo.14851637 (2025).
64. Bignaud, A., Serizay, J., Baudry, L., Matthey-Doret, C. & Marbouty, M. Metagenomic Tridimensional Organisation-based Reassembly. *GitHub* https://github.com/mmarbout/MetaHiC_pipeline (2024).

Acknowledgements

We thank P. Moreau for assistance during experimental work, the different teams of the Microbiology department from Institut Pasteur, and especially J. Czarnecki for providing bacterial pellets; D. d'Alelio from the Stazione Zoologica Anton Dohrm for help with the oceanic sample; and the hacienda Vergel de Gaudalupe for allowing us to sample their fermenter. This research was funded, in whole or in part, by Agence nationale pour la recherche ANR-20-CE92-0048 to M.M. and L.D. and ANR-16-JPEC-0003–05 to R.K. and by a grant from the French government, managed by the Agence Nationale de la Recherche under the France 2030 programme (ANR-23-CHBS-0002) to R.K. The Biomics Platform, C2RT, Institut Pasteur, Paris, France, is supported by France Génomique (ANR-10-INBS-09) and IBISA. A.B. was supported by an ENS fellowship from the French Ministry of Higher Education, Research and Innovation. D.E.C. is supported by a PhD grant from the PhastGut project. A.B. and D.E.C. belong to Ecole Doctorale Complexité du vivant ED515 of Sorbonne Université. L.M., M.G.-O. and M.C.-G. were supported by funding from Programa de Apoyo a Proyectos de Investigacion e Innovacion Technologica (DGAPA-UNAM – IN212524). Sequencing and library preparation was supported by Agencia Nacional de Investigación e Innovación (ANII-Uruguay) grant number FSGSK_1_2019_1_159735 and Fondo para la Convergencia Estructural del MERCOSUR (FOCEM). Illustrations used in the present publication (Figs. 1 and 3 and Extended Data Fig. 2) were obtained from the Internet under a Creative Commons CC0 license (https://svgsilh.com/fr/). A CC-BY public copyright license has been applied by the authors to the present document and on all subsequent versions up to the author-accepted manuscript version arising from this submission, in accordance with the grants' open-access conditions.

Author contributions

A.B., R.K. and M.M. conceptualized the study. A.B., R.K. and M.M. designed the methodology. A.B. and M.M. designed software.

M.M. performed validation. M.M. conducted investigations, with contributions from J.S., G.L.T., O.C., N.R., D.E.C., A.T., M.G.-O., M.C.-G., J.P. and K.L. A.B. and M.M. conducted formal analysis, with contributions from A.P., G.A.M. and J.S. M.M. and A.B. curated data. D.E.C., G.I., N.R., M.M., K.L., L.M., A.T., A.B., J.P., P.H., D.E.C., G.L.T., M.G.-O. and M.C.-G. procured resources. M.M. and R.K. performed visualization. M.M., R.K. and A.B. wrote the original draft. All authors edited the paper. M.M., R.K., S.H., L.M., G.I., L.D., G.L.T. and O.C. supervised the project. M.M., G.I., L.D. and R.K. acquired funding. M.M. and R.K. administered the project.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-025-02108-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-025-02108-2.

Correspondence and requests for materials should be addressed to Romain Koszul or Martial Marbouty.

Peer review information *Nature Microbiology* thanks Louis-Patrick Haraoui and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

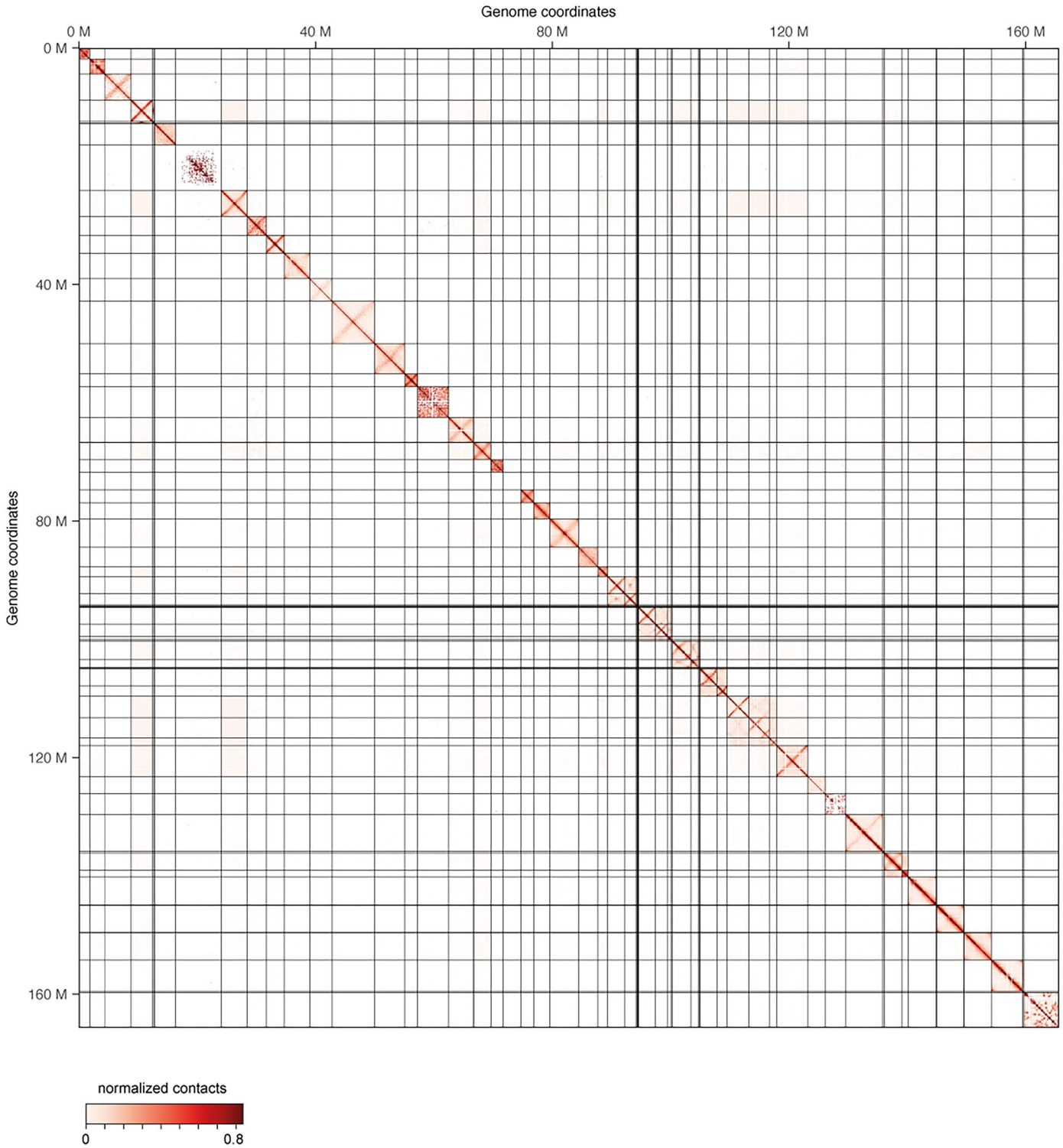
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

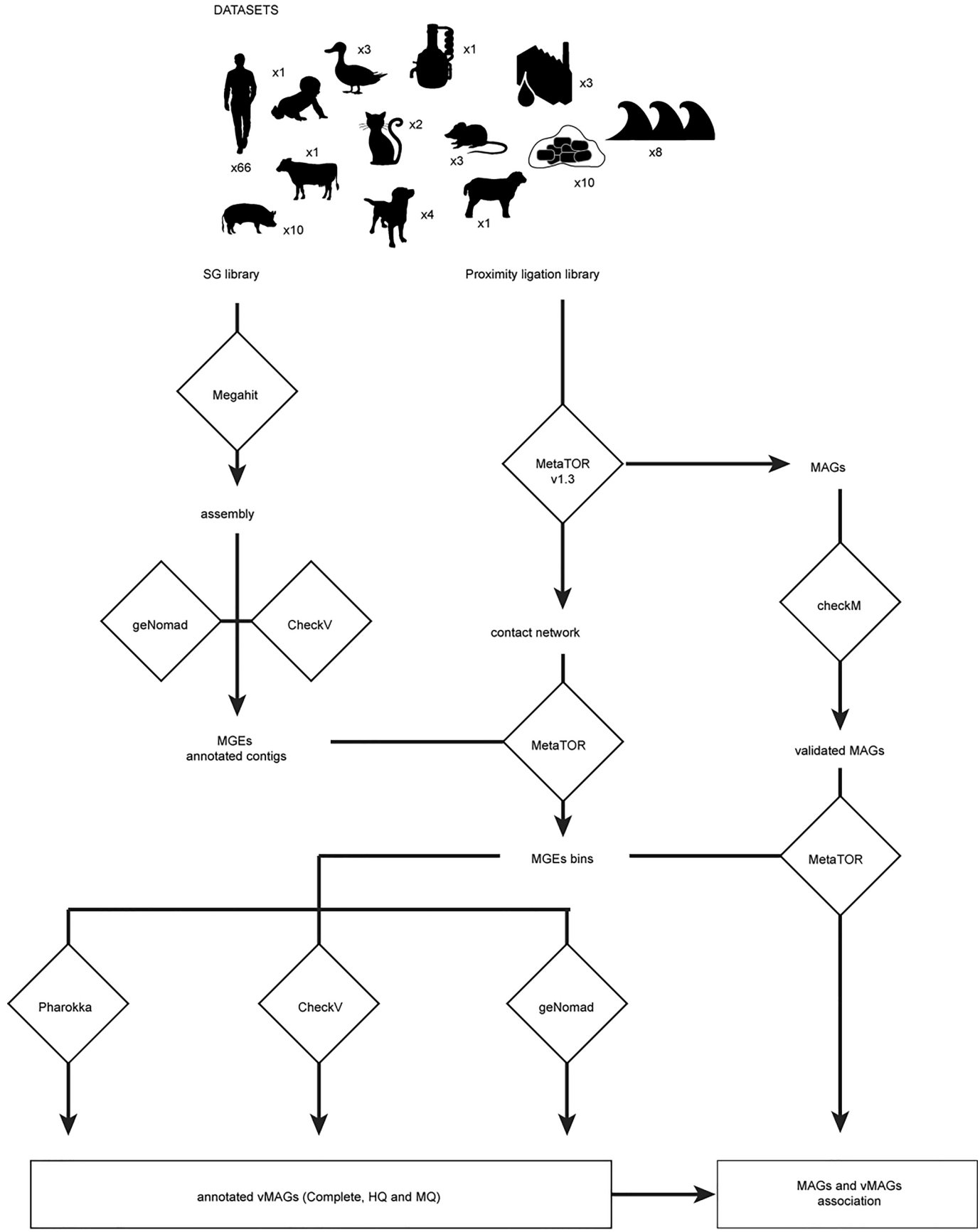
© The Author(s), under exclusive licence to Springer Nature Limited 2025

https://doi.org/10.1038/s41564-025-02108-2



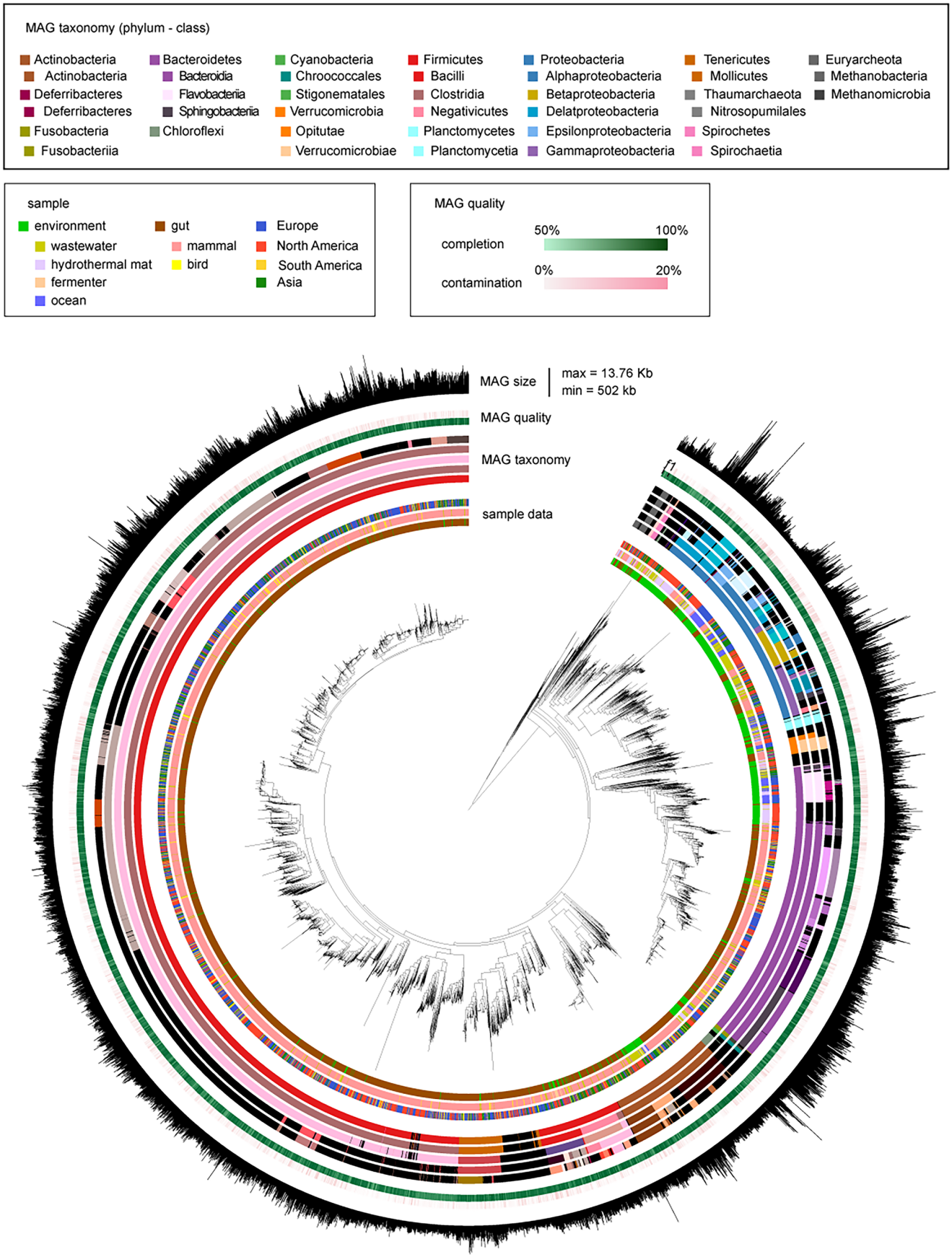
Extended Data Fig. 1 | contact map of the mock community. Normalized contact map (bin = 16 kb) of the mock community. Black lines delineate the different DNA molecules. Genomic coordinates are indicated on the sides of the contact map and the scale bar is indicated under. Some bacteria exhibit an abundance below our detection threshold (~0.1%) and therefore appear as white squares in the contact map.

https://doi.org/10.1038/s41564-025-02108-2



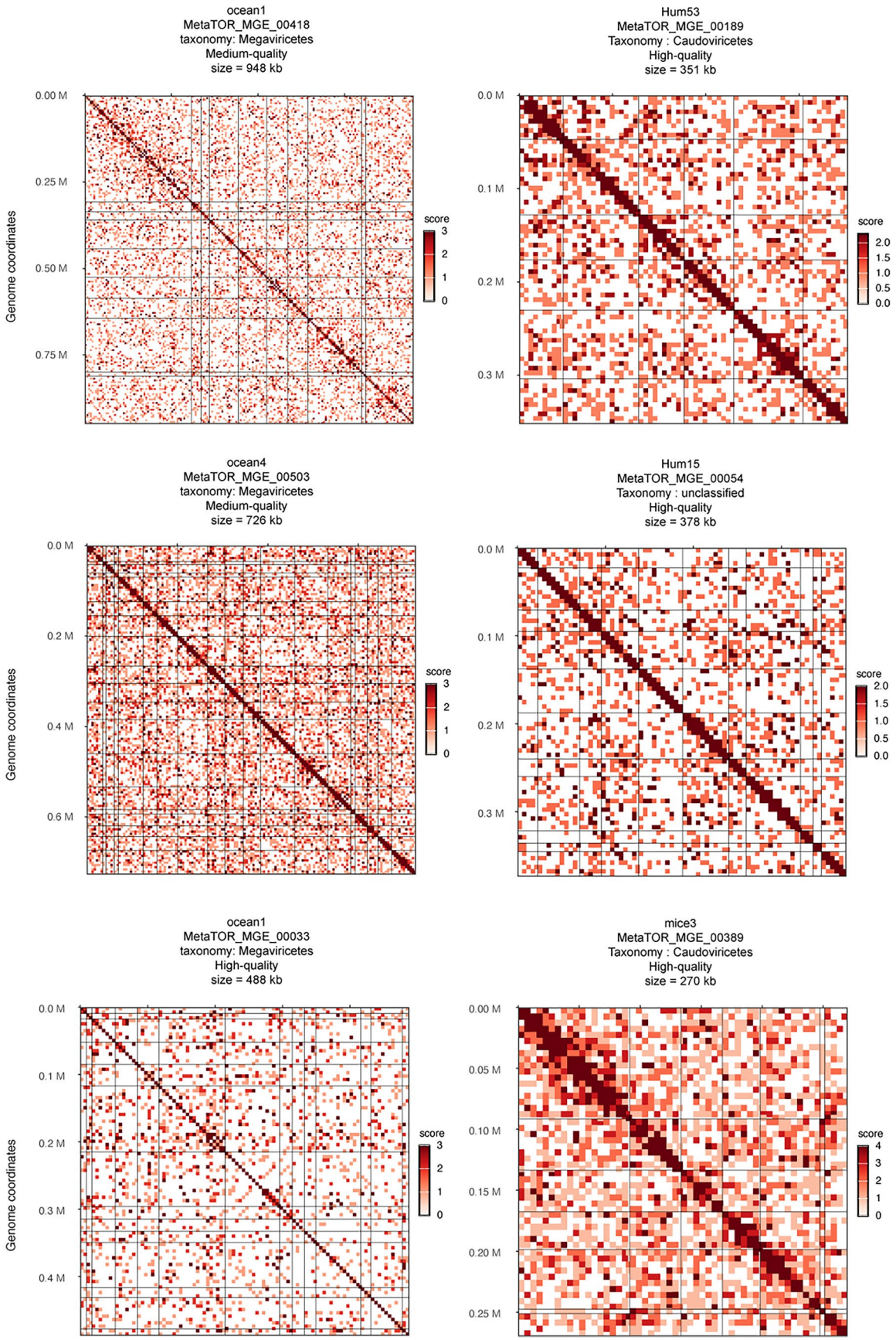
Extended Data Fig. 2 | datasets and pipeline used in the present study. Datasets used in the present study. The number of each type of sample is indicated next to each drawing. The different softwares used to generate and analyze the data are indicated in diamond. (SG : Shotgun). Figure adapted from SVG Silh under a CC 1.0 license.

https://doi.org/10.1038/s41564-025-02108-2



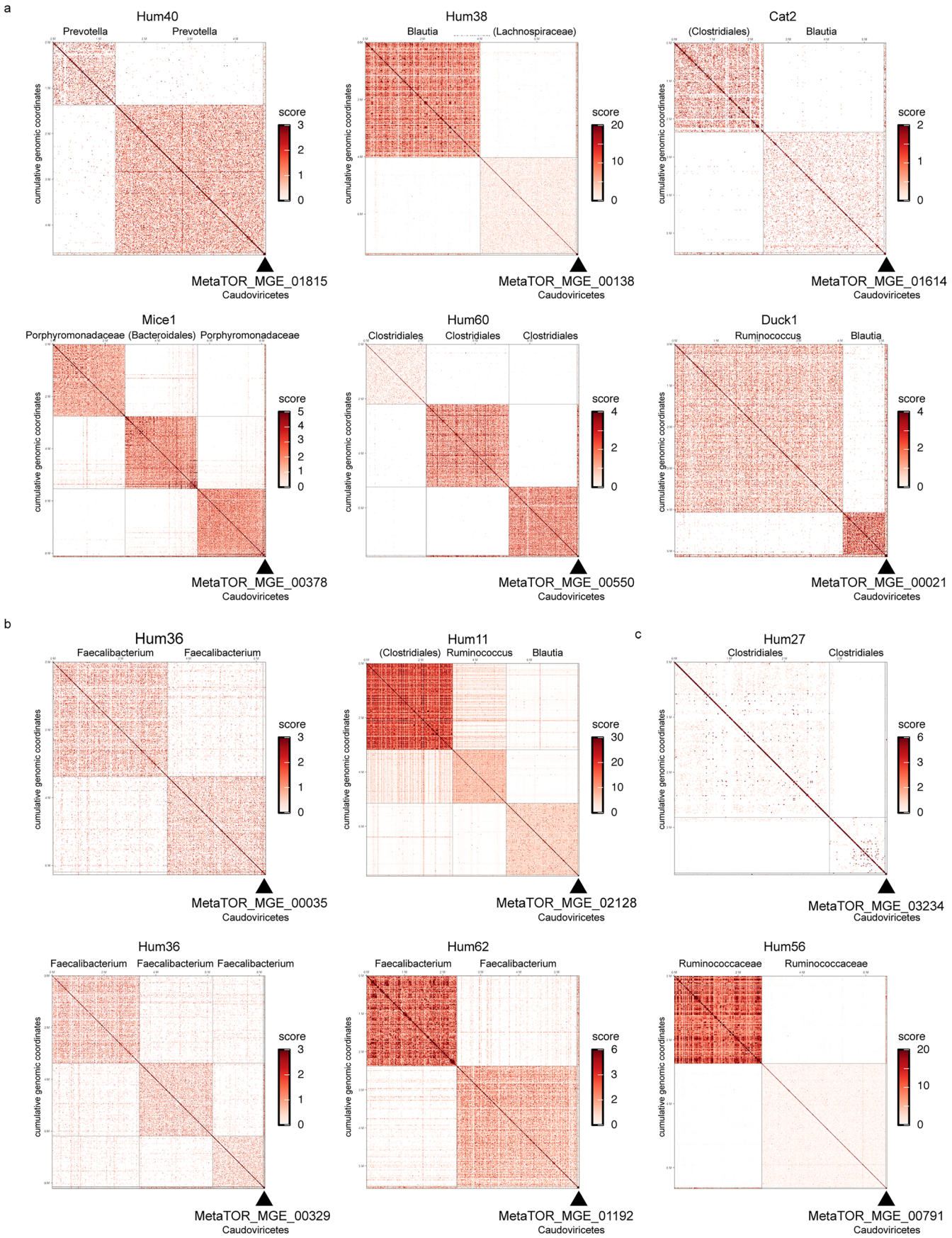
Extended Data Fig. 3 | Phylogenetic tree of the MAGs. Phylogenetic tree of medium and high-quality MAGs obtained across all datasets using MetaTOR. The tree is decorated with different annotations. From inner to outer: sample data (1- sample type: environment or gut; 2- ecosystem: wastewater, hydrothermal mat, fermenter, ocean, mammal, bird; 3- geographic location: Europe, North America, South America, Asia), MAG taxonomy (1- Phylum, 2- Class, 3- Order, 4- Family, 5- Genus), MAG quality (completion - green, contamination - red), MAG size in bp. The different legends for the annotations are indicated above the tree.

https://doi.org/10.1038/s41564-025-02108-2



Extended Data Fig. 4 | Contigs contact map of different large vMAGs. Raw contact map of six vMAGs. Black lines delineate the boundaries of the different contigs. Sample (environment), vMAG ID, taxonomic annotation, quality and size are indicated above contact maps while scale bars are present on the left.

https://doi.org/10.1038/s41564-025-02108-2

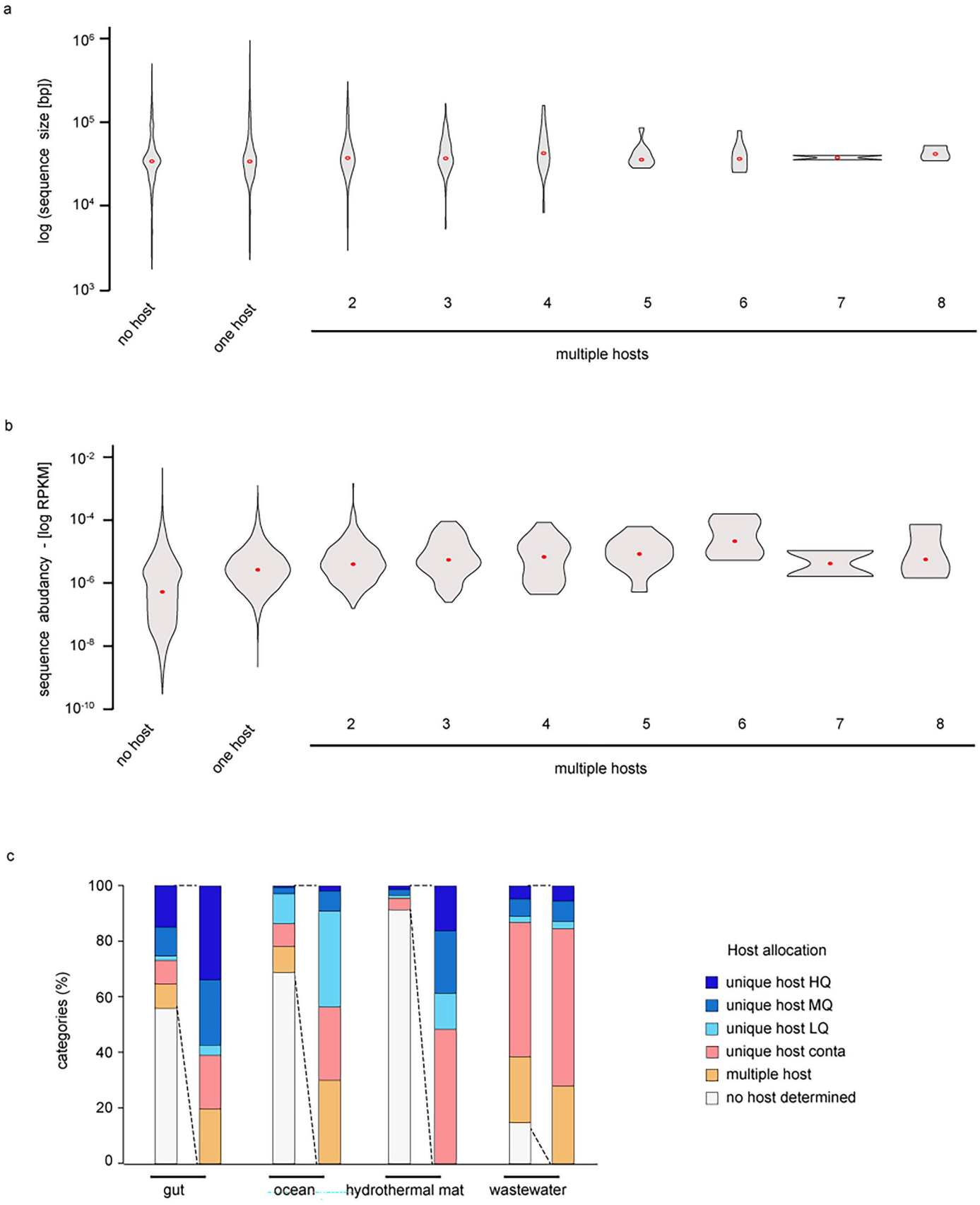


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41564-025-02108-2

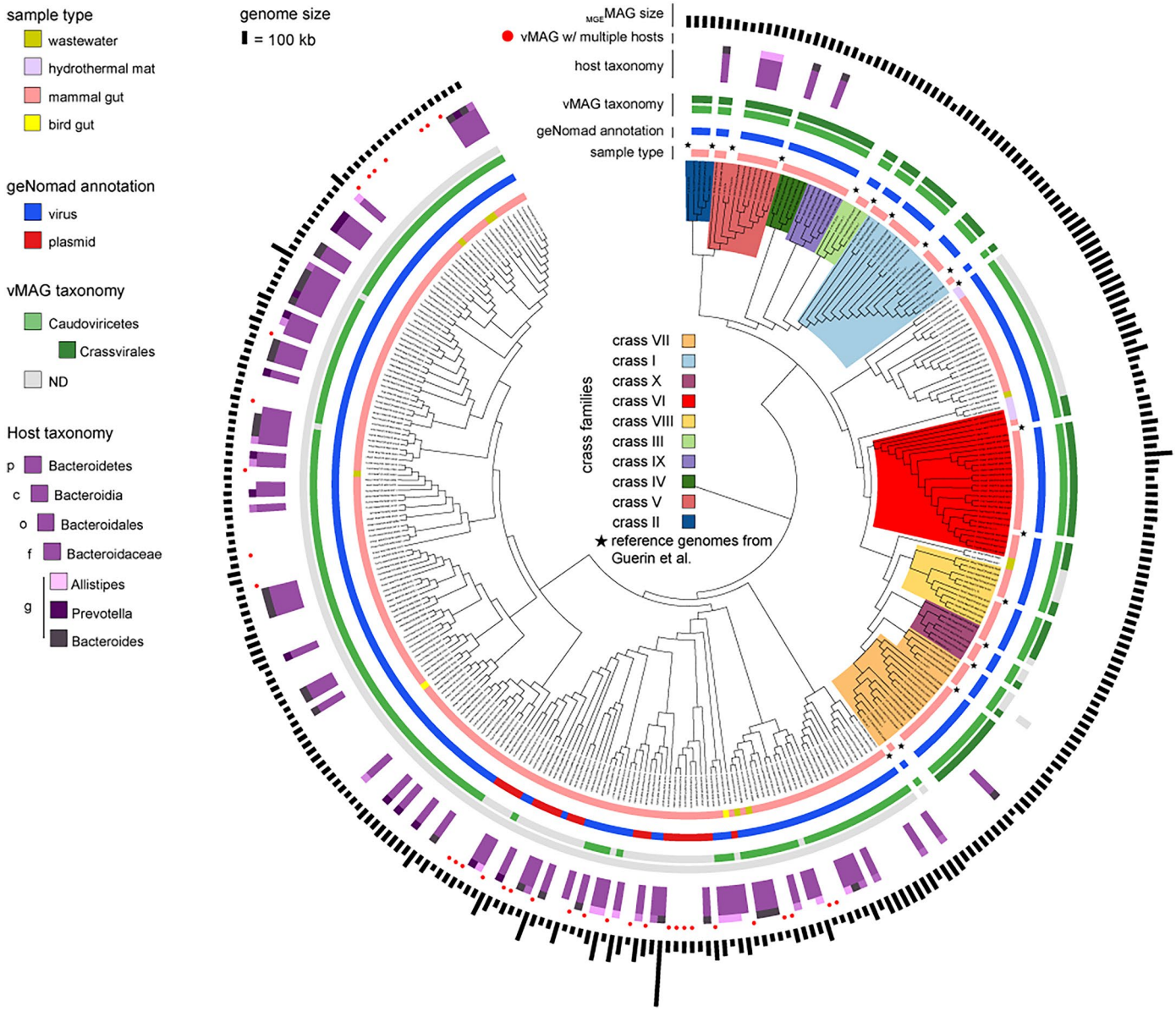
Extended Data Fig. 5 | Contact maps of $_{MGE}$MAGs and their different hosts. Intra- and inter- contact maps obtained of vMAG exhibiting multiple hosts (50kb bins). Sample (environmental) origin and vMAG references are indicated above and below each contact map. Dark lines indicate boundaries of the different genomic entities. Black triangles point at vMAGS. Scale bars (raw scores) are indicated aside each contact map. **a**. vMAGs with clear contact with multiple microbial MAGs that do not display noise signals between them. **b**. vMAGs with clear contact with multiple microbial MAGs exhibiting noise signals between them. **c**. vMAGs with low contact signal with multiple microbial MAGs.

https://doi.org/10.1038/s41564-025-02108-2



Extended Data Fig. 6 | Multihost vMAGs features. a. Violin plot of the log(size) of vMAGs as a function of their host number. **b**. Violin plot of the log(RPKM) of vMAGs as a function of their host number (RPKM: Reads Per Kilobase Million). **c**. Bar plot of vMAGs proportion as a function of their host assignment for the different processed environments encompassing a sufficient number of vMAGs (gut, ocean, hydrothermal mat and wastewater). The different categories are indicated by colors (white = no host assigned, orange = multiple host assigned, red = one contaminated [conta] host assigned, grey = one LQ host assigned, blue = one MQ characterized host assigned, darkblue = one HQ host assigned).

https://doi.org/10.1038/s41564-025-02108-2



Extended Data Fig. 7 | Proteomic tree of the Crassphages family and related phages infecting *Bacteroidetes*. Proteomic tree of the different *Crassvirales* and related phages characterized in the present study. The Tree also encompasses different representative reference Crass genomes from Guerin et al.[57], indicated by a black star. The different crass families are indicated by colored areas over the branches. The tree is decorated with different informations (from the inside to the outside): *i*) sample type, *ii*) reference genomes (black stars), *iii*) geNomad annotation (blue = virus; red = plasmid), *iv*) vMAG taxonomy (order, genus), *v*) host taxonomy (phylum, class, order, family, genus) white if no host or multiple host attributed), *vi*) vMAG exhibiting multiple hosts (red circles), *vii*) vMAG genome size (scale bar = 100 kb).

nature portfolio

Corresponding author(s): Marbouty Martial

Last updated by author(s): Jul 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://sra-explorer.info/#
Data analysis	Open-access versions of the programs and pipelines used (Hicstuff, MetaTOR, HiContacts) are available online on the github account of the Koszul lab Hicstuff (https://github.com/koszullab/hicstuff) version 3.1.2, MetaTOR (version 1.3.2 available online at https://github.com/koszullab/metaTOR), HiContacts (version 1.7.1 available online at https://github.com/js2264/HiContacts). Other mandatory programs are also available online: Bowtie2 (version 2.4.5 available online at http://bowtie-bio.sourceforge.net/bowtie2/), SAMtools (version 1.9 available online at http://www.htslib.org/), and Cooler (versions 0.8.7–0.8.11 available online at https://cooler.readthedocs.io/en/latest/). The pipeline used in the present study to generate the data is available at the following address: https://github.com/mmarbout/MetaHiC_pipeline.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequence data as well as raw assemblies generated in this study have been deposited in the NCBI under the BioProject number PRJNA1169672 (mock community) and PRJNA1169674 (metagenomic datasets).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not applicable
Reporting on race, ethnicity, or other socially relevant groupings	not applicable
Population characteristics	not applicable
Recruitment	Participants were recruited on a voluntary basis and informed about the study. We only enroled healthy particpant that had not take antibiotics in the last two month. For the Metakids cohort, infant were recruited in the day-care center of Orsay (France) at their entrance in the day-care ccenter. Parents were informed about the study.
Ethics oversight	A total of one mouse, three ducks and four dogs fecal samples were recovered from collaborators in Netherlands. The proposed activities have been reviewed by the Medical Research Ethics Committee of UMC Utrecht (deposited under METC-protocol number 18-139/C). A total of 17 human stool samples from Uruguay were provided by Gregorio Iraola in accordance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization. Gregorio Iraola is registered in the National Registry of Researchers in Genetic Resources and Derivatives in the Environment Ministry. In the context of this registration, access to genetic material and derivatives is allowed for non-commercial purposes (Article 22, Law #17.283). The stool samples were collected with written informed consent from all participants. Before providing consent, all participants received comprehensive information about the project and had the opportunity to opt out of the study. The project was approved by the Ethics Committee and Institutional Review Board of CASMU (Centro Asistencial del Sindicato Médico del Uruguay – ref : FSGSK_1_2019_1_159735), and the study was conducted following national legislation (study performed from 2018 to 2020). The project was registered in the Ministry of Public Health (Ministerio de Salud Pública MSP - ref 39814). One infant fecal sample was recovered from the MetaKids cohort. The MetaKids study (N° ID-RCB : 2017-A00750-53 – clinical trial : NCT03296631) received ethics approval from the ethics committee Comité de Protection des Personnes Sud Est 1 on July 21, 2017 as required by French regulation on clinical research. Oceanic sample was recovered in the bay of Napoli during a sampling campaign in collaboration with the Genoscope (Evry, France) and the Stazione Zoologica Anton Dohrm (Napoli, Italy) in agreement with the national legislation. Mezcal fermentation sample was recovered in Guadadupe hacienda in Mexico in collaboration with university of Queretaro (UNAM). Samples were processed in Queretaro and the resulting data was transferred to Institut Pasteur for analysis.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The present study explores MGE - host interactions at the sample level by processing 84 existing and 27 newly generated proximity-ligation-based MetaHiC data using a dedicated computational pipeline. This analysis reconstructed 4,975 microbial, 5,879 viral and 693 plasmid genomes of medium quality or higher. The contact network between genomes further allowed assigning approximately

	half of viral genomes to their bacterial hosts, revealing that a substantial proportion of bacteriophages interact with multiple species.
Research sample	a set of Metagenomic HiC data from various environmental samples
Sampling strategy	no strategy needed for this publication
Data collection	data were collected from SRA database using the sra explorer
Timing and spatial scale	no time series in the sample
Data exclusions	no data were excluded from the analaysis
Reproducibility	not applicable
Randomization	not applicable
Blinding	not applicable

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	not applicable
Location	oceanic sample: Napoli , Italy - 2meter depth fermenter samples, Queretaro Mexico
Access & import/export	all samples or data were obtained in accordance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization.
Disturbance	nothing to declare

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*