Phages with a broad host range are common across ecosystems

Amaury Bignaud, Devon Conti, Agnès Thierry, Jacques Serizay, Karine Labadie, Julie Poulain, Olivia Cheny, Maritrini Colón-González, Laurent Debarbieux, Marianna Guerrero-Osornio, et al.

Phages with a broad host range are common across ecosystems

Amaury Bignaud*[1,2], 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[9], Gwenaelle Le Tinier[10], Gael A. Millot[11], Lucia Morales[7], Andrés Parada[12,13], Nadia Riera[12], Gregorio Iraola[12], Romain Koszul♀[1], Martial Marbouty♀*[1]

Abstract

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 and this, in environments as diverse as oceanic water column or the human gut. This observation challenges the traditional view of a narrow host spectrum of phages by unveiling that multi-host associations are common across ecosystems, with implications how they might impact ecology and evolution and phage therapy approaches.

Affiliations

[1] Institut Pasteur, CNRS UMR 3525, Université Paris Cité, Spatial Regulation of Genomes Group, Paris F-75015 France. [2] Sorbonne Université, Collège Doctoral, Paris, France. [3] Institut Pasteur, Université Paris Cité, CNRS UMR6047, Bacteriophage Bacterium Host, F-75015 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, 2 Rue Gaston Crémieux, 91057, Evry, France.. [6] Clinical Research Coordination Office, Institut Pasteur, Université Paris Cité, 75015 Paris, France. [7] Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro, Mexico. [8] Department of Microbiology, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [9] Department of Infectious Diseases, School of Immunology and Microbial Sciences, Guy's Hospital, King's College London, London, SE1 9RT, UK. [10] Médecine petite enfance, Orsay, France. [11] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France. [12] Institut Pasteur de Montevideo, Microbial Genomics Laboratory, Montevideo, Uruguay. [13] Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Uruguay.

* These authors contributed equally to this work

♀ Corresponding authors:

Martial Marbouty
Spatial Regulation of Genomes Unit, Department of Genomes and Genetics
25-28 Rue du Dr Roux, Institut Pasteur
Paris F-75015 France.
martial.marbouty@pasteur.fr

Romain Koszul
Spatial Regulation of Genomes Unit, Department of Genomes and Genetics
25-28 Rue du Dr Roux, Institut Pasteur
Paris F-75015 France.
romain.koszul@pasteur.fr

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 analyze individual samples [10,11]. It is also difficult to characterize and validate complete viral sequences, due to a lack of markers to assess their completeness [12–15]. Finaly, inferring reliable phages-hosts 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 episomes-hosts 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) metagenomic assembled genomes (MAGs). Using 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 bacteria 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

homogenous 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 inter-chromosomal 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 eight secondary chromosomes, 14 plasmids and two intra-cellular phages (**Fig. 1c**) (**Supplementary Table 1**; **Methods**). The resulting ~12M PE metaHiC reads were processed to generate chromosomal contact maps of the community. The genome sequences represented along the x and y axis 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%). Phages genomes embedded in particles remained isolated, in contrast with plasmid sequences and intracellular bacteriophage genomes that made significant contacts with, and mostly with, the genomes of their bacterial hosts (**Fig. 1d**).

These results were exploited to design a computational approach for characterizing 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 (*i.e.* $\leq$ 2 kb apart) (black arrowheads, **Fig. 1e, f**). As a result, unordered contigs belonging to the same MGE genome display relatively similar trans-contact (i.e. inter-contigs) compared to cis-contact (i.e. intra-contigs) frequencies (excluding, of course, the short-range diagonal). This behavior is not found for 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 bin MGE-annotated contigs with comparable intra- and inter-contig contact frequencies into MGE metagenomes associated genomes ($_{MGE}$MAGs, either viral vMAGs or plasmid pMAGs) (**Methods, Fig. 1h and Supplementary Fig. 1**).

Viruses binning from metagenomes 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 to raw contigs (19 and 73 vs. 11 and 34, respectively) (**Fig. 1i**), with 46 composed of two contigs or more. As another binning quality control, we sought for circular viral contigs that are likely complete and should not be pooled with extra contigs into a vMAGs. 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**).

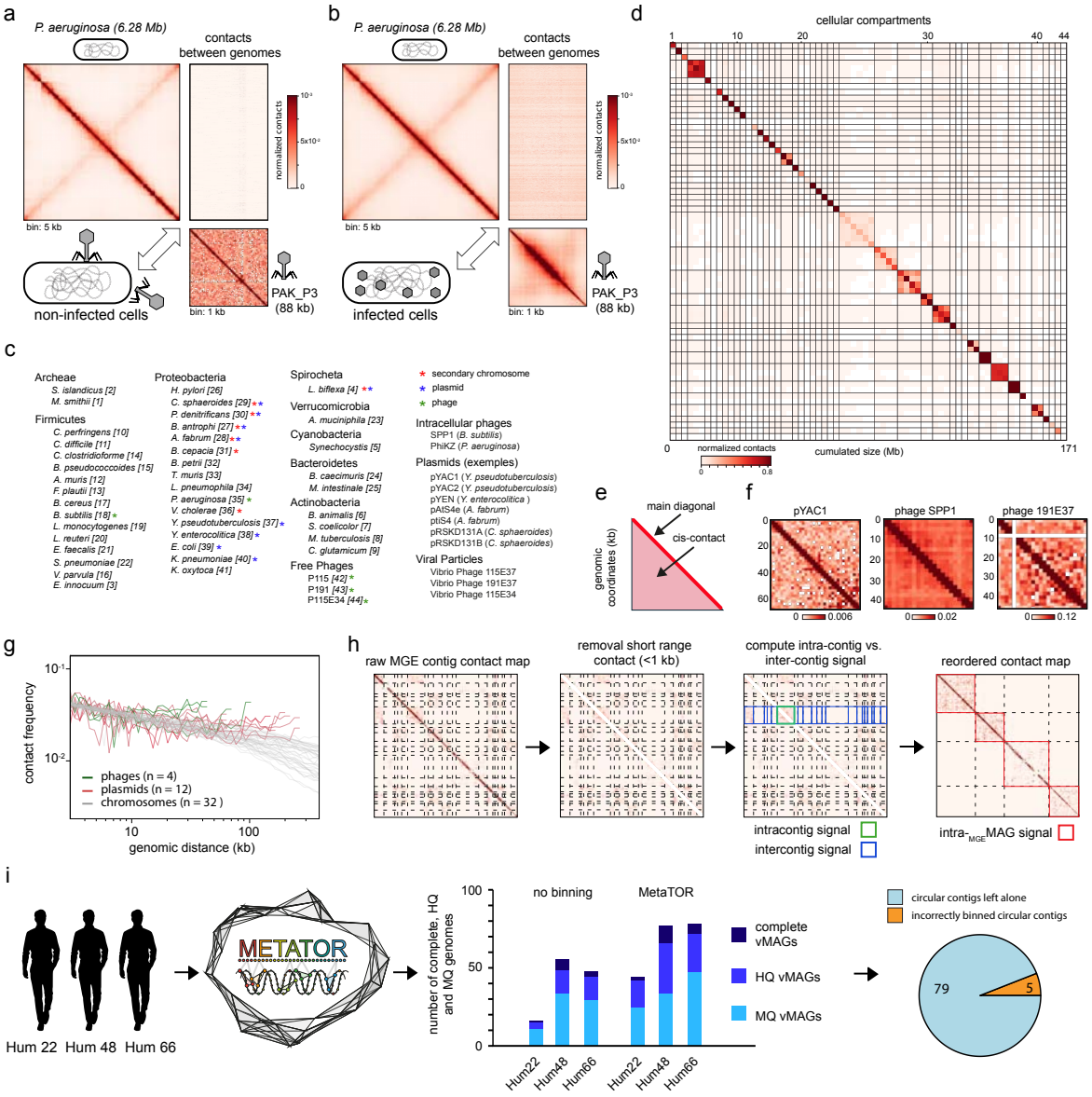


Figure 1: principles of MGE binning and host association.

a-b. Hi-C contact maps of *Pseudomonas aeruginosa* (5kb bins) and the phage PAK_P3 (1kb bins) genomes before (a) and after infection (b). **c**. Composition of the mock community. Stars: episomes (green: phage, blue: plasmid, red: secondary chromosome). Number in brackets: 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 (2kb bins) of MGEs in the mock community. **g**. Normalized contact signal as function of the genomic distances for phages (green), plasmids (red), bacteria 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 (<1kb) are removed; intra and inter-contigs signals are computed; contigs exhibiting a similar intra and inter signal are binned together. **i**. Left: Application of MetaTOR on three human gut datasets. Middle: bar plot of the number of complete, high quality (HQ) and medium quality (MQ) viral genomes assessed by CheckV for the raw contigs and the vMAGs generated by MetaTOR is presented. Right: pie chart of the proportion of circular viral contigs either left alone or binned with other contigs.

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%) harbored an aberrant trans-contact signal (**Supplementary Fig. 2** - black arrow), while the others harbored 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 pooled 53 and 48 circular viral contigs (out of 84) with other contigs, respectively (**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 viruses genomes across 111 communities.

We next investigated of viruses and their potential hosts on metaHiC(-like) datasets from different ecosystems. 84 datasets, corresponding to all published metagenomic samples ever processed by a proximity-ligation like protocol, were collected [7,11,21–24,26,30–39] [14,18,27–30,32,39–48]. We broadened the diversity of the studied ecosystems by generating 27 new datasets, resulting in 111 samples from five 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; Supplementary Table 2**). The 46.8 Gb dataset regrouping the different assemblies, including 597,064 MGE contigs, were 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 to 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 to 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**).

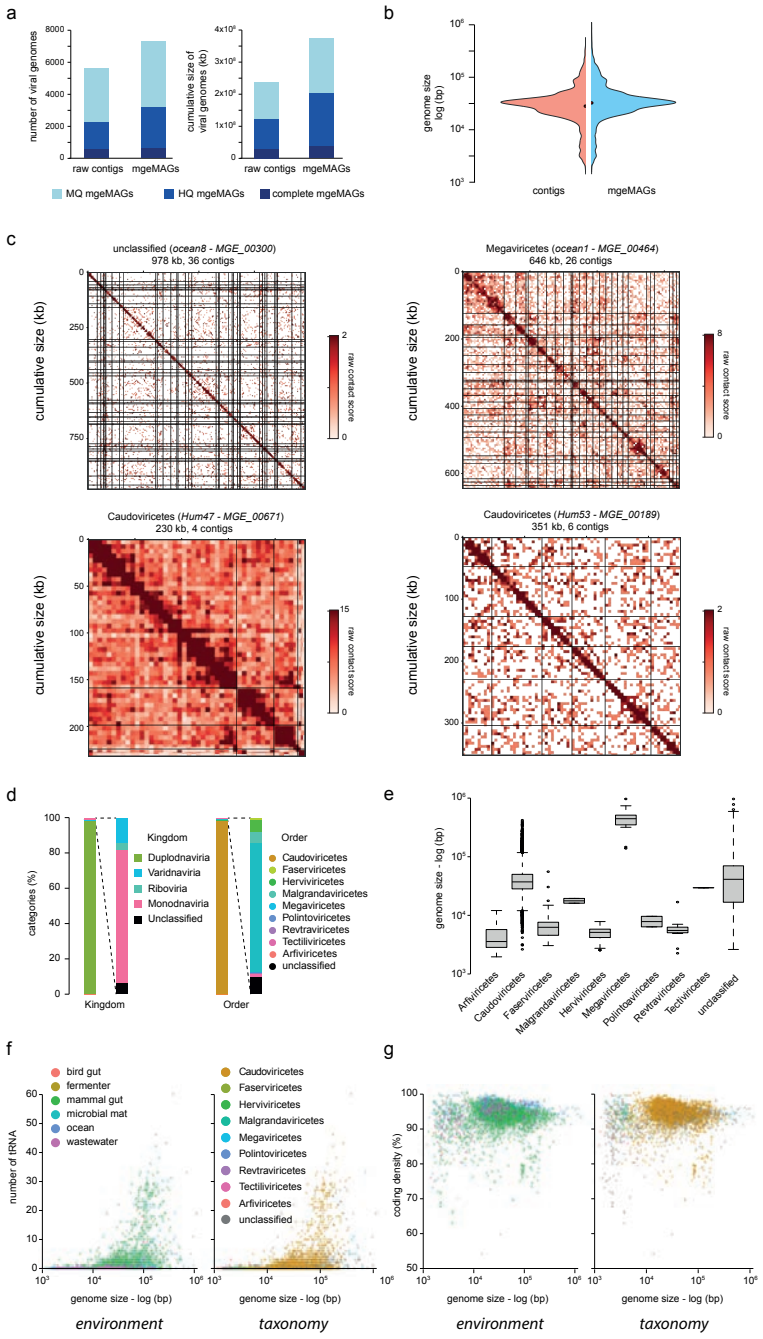


Figure 2: MetaTOR MGE output from metagenomic datasets.

a. Completeness of MGE contigs and MGE bins obtained using MetaTOR MGE module according to the CheckV in terms of number (left) and cumulative sequences (right). Only complete, high quality (HQ) and medium quality (MQ) MGE genomes are shown (see also Supplementary Table 2). **b.** Violin plot of the log(size) for the MGEcontigs (left) and MGEMAGs (complete, HQ, MQ) (right). **c.** Raw contact map of vMAGs. Black lines delineate the different contigs. Annotation, Sample, vMAG ID, size, number of contigs are indicated above contact maps while scale bars are present below. **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 in function of their taxonomic annotation at the order level (number of each representative are indicated in brackets; middle lines in the boxplots indicate mean and boxplot limits indicate first and third quartile). **f.** Plot of the number of tRNA genes detected in function of the viral genome size (log scale). Genomes are colored in function of their environment (left) or taxonomic annotation (right). **g.** Coding density computed in function of the genome size (log scale). As for (a) viral genomes are colored in function of their environment (left) or their taxonomic annotation (right).

A careful examination of the contact map of different well covered vMAGs showed that they encompass a various number of contigs associated with coherent inter-contigs signals, supporting the accuracy of the approach (**Fig. 2c**; **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 result 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 a wide diversity.

Taxonomic annotation shows that more than 80% of the 6,572 vMAGs belong to the phylum of *Duplodnaviria* and to the order of the *Caudoviricetes* (**Fig. 2d**). Among 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 of *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 the *Arfiviricetes* 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 metabolism, structural proteins or host takeover, but also integration, indicating that we characterized potential temperate phages. The largest vMAGs encoded a high number of tRNAs, with one *Caudoviricetes* vMAG displaying 60 tRNAs (**Fig. 2f**). Most of the vMAGs 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 different genetic code than the standard one, a phenomenon rarely reported for viruses [43].

A significant 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 like 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 MAGs. 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: *i*) if most (>50%) normalized contacts involve a single MAG, the latter is considered the host; *ii*) if the vMAG makes between 10 and 50% of its normalized trans-contacts with multiple MAGs, we consider several hosts; *iii*) 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 medium/high quality 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 *i*) a lack of coverage or poor contact data, *ii*) the absence of the host in the MAGs dataset (also eventually because of poor data), or *iii*) 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 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. 754 vMAGs were assigned to MAGs hosts characterized at the genus level, including for 305 at the species level. The remaining 487 vMAG (17% of all host-associated vMAGs) were associated with more than one bacterial MAGs, 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, class, order, family, and genus respectively (**Fig. 3c**). The increased proportion of vMAGs interacting with different families appears mainly due to the increased proportion of hosts for whom a genus was not assigned.

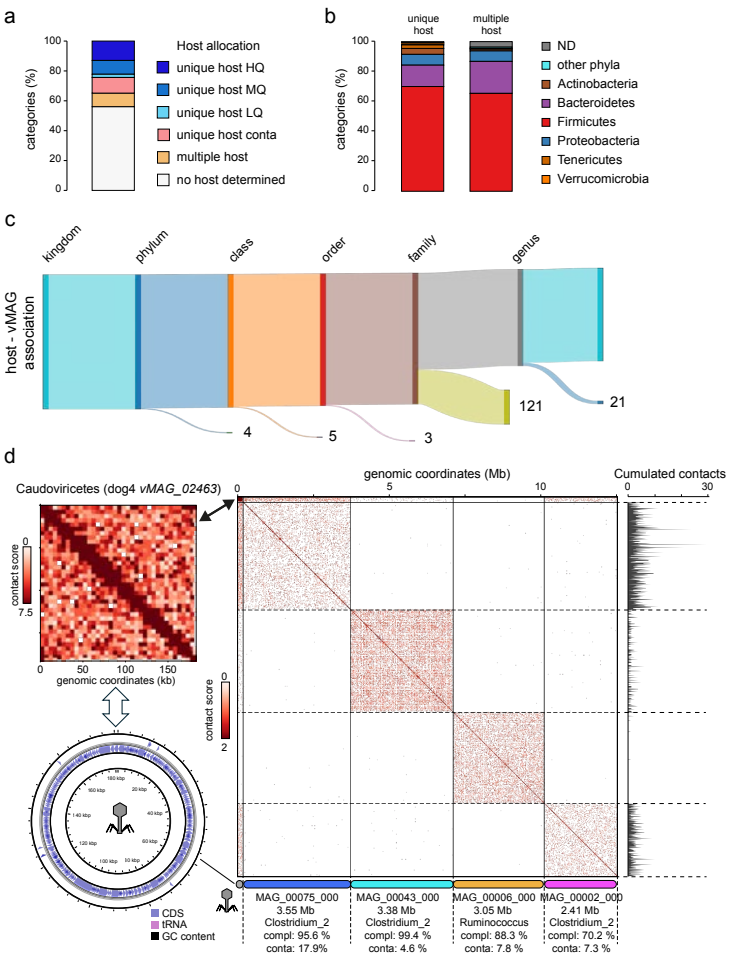


Figure 3: host attribution to the vMAGs

a. bar plot of the vMAGs proportion in function of their host attributions. The different categories are indicated by colors (white = no host attributed, orange = multiple host attributed, red = one contaminated host attributed, grey = one LQ host attributed, blue = one MQ characterized host attributed, darkblue = 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 two hosts that do not belong to the same family level. **d.** Example of intra- and inter-contact map obtained for a vMAG exhibiting multiple hosts: *left* - contact map of the vMAG#02463 (sample = dog4, virus, *Caudoviricetes*) (bins = 2kb) 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 (50kb bins) of the vMAG (upper left) and 4 bacterial MAGs from the same sample. Dashed lines indicate borders of the different genomic entities. Schemes of the 4 MAGs as well as their identification number, size, taxonomic classification (order) and CheckM evaluation are present under the contact map. The scale bar is indicated on the left. Histogram representing the raw contact scores of the vMAG with the different part of each MAG is plotted on the right part of the contact map.

Multiple host interactions pattern 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 144 complete and HQ vMAGs exhibiting at least 50 contacts with two or more MAGs of the same sample. While 55 maps exhibited low or ambiguous signals, the remaining 89 maps displayed discrete, non-ambiguous and homogenous contact patterns with several microbial MAGs, fully consistent with the observed infection patterns experimentally characterized (**Fig. 1a**) (**Fig. 3d; 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* genera (**Fig. 3d**). These contacts likely 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 they have not contaminated each other's; second, no important trans-contacts between the MAGs; third, a uniform contacts distribution between the vMAG and MAGs 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 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 analyzed it with respect to their hosts. The resulting tree recapitulates viruses' taxonomic annotations as defined by ICTV, and confirms the general behavior 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 all over the proteomic tree (**Fig. 4a**, red dots**)**.

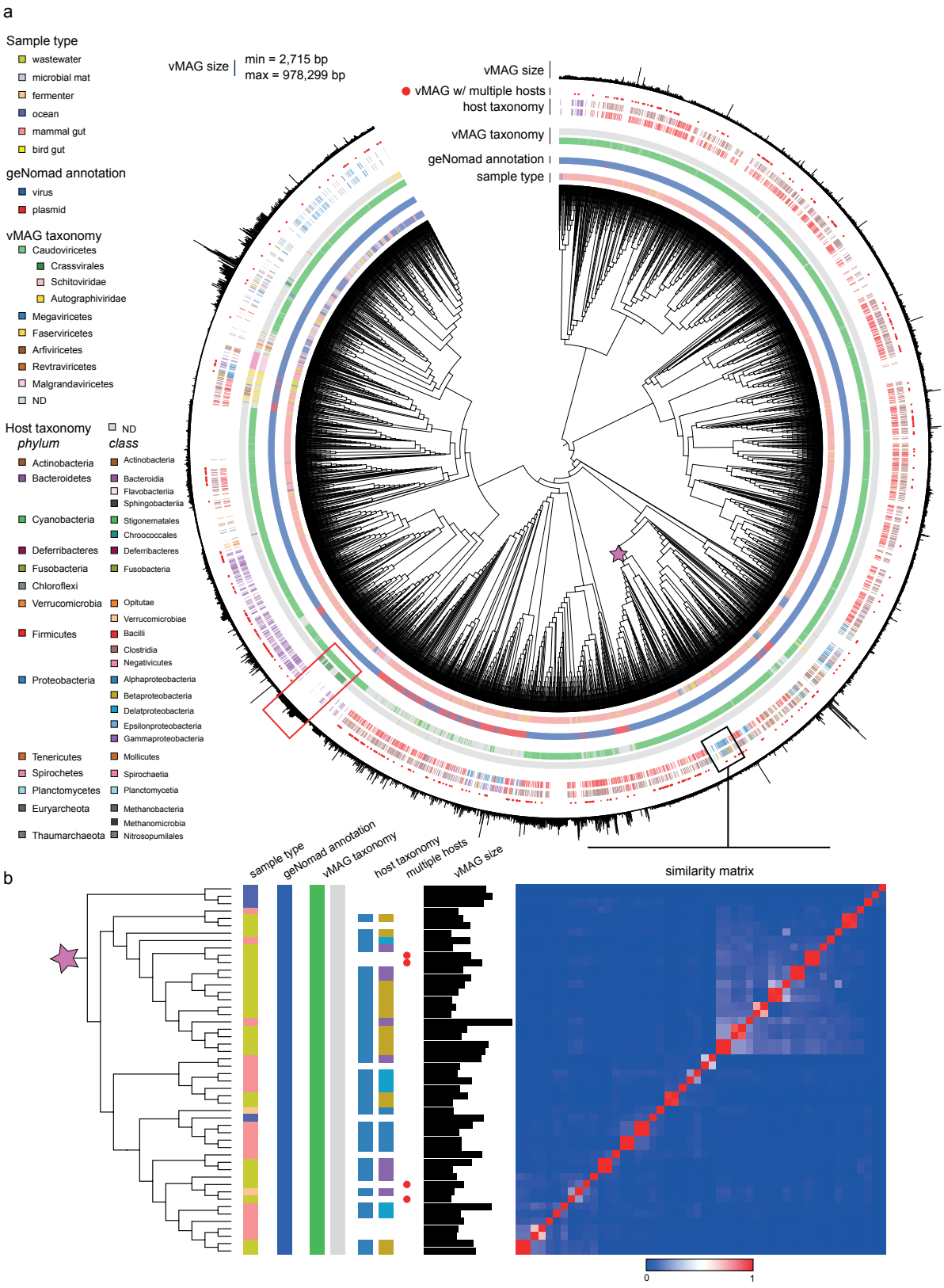


Figure 4: Proteomic tree of the characterized vMAGs.

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 informations (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), *vi*) vMAG genome size (min = 2.7 kb - max = 978 kb). The red rectangle points to 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 side of the Tree. Scale bar of the similarity matrix is also indicated

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 pseudo-lysogenic [46] or phage-plasmid lifestyles [48], with potentially several hosts for the same virus. We only assigned seven crass phages (7 / 107 = 6%) to a *Bacteroidetes* host and none of them exhibit a multihost contact pattern (**Extended Data Fig. 7**). These results, therefore, do not confirm a multihost pattern of Crassphages. They also suggest that Crassphages are mostly found as viral particles in human fecal 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 complex microbial communities MGEs diversity, 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 vMAGs and MAGs 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*, highlight the interest of metaHiC to unveil viral diversity in individual environmental samples. This catalog 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 paves 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. Multihost 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 to consider considering horizontal gene transfer dynamics mediated by phages in the 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 to the 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 to draw 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 to validate not only 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 the 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 work has broader implications for the understanding of the ecology and evolution of microbial communities.

Methods

Ethics approval and consent to participate

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

Phage - bacteria infection HiC libraries

The bacterial strain PAK (*P. aeruginosa*) was grown until mid-exponential phase (OD = 0.3) in 80 mL of LB. Phage PAK-P3 was then added at a MOI of 25 (PhiKZ =15). 7 mL aliquots of the phage bacterial culture were removed at two time points after the addition of the phage (0 and 5 minutes) 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 RT followed by incubation at 4°C during another 30 min. Formaldehyde was quenched by adding 2 mL of Glycine 2.5 M followed by an incubation of 20 min under shaking. Pellets were recovered by a centrifugation of 10 min at 10,000 x g and at 4°C, washed in PBS and re-centrifuge 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 the middle of the exponential phase. 10^9 bacteria were recovered for each culture, formaldehyde was added, (3% final) and the mixture was incubated under agitation at room temperature for 1h. Glycine (2.5 M stock solution) was then added to reach a final concentration of 125 mM and followed by an incubation at room temperature for 20 min under gentle agitation. The solution was centrifuged (6,000 g, 10 min, 4°C) and the pellet was washed in 1X PBS. After a similar centrifugation, the supernatant was removed, the pellet was flash frozen and stored at -80°C. Once the different pellets were obtained, they were thawed and mixed in random proportion to create the mock community. The large pellet obtained was then split in 10 aliquots that were stored at -80°C until use.

Feces sample collection

Feces 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 (1X PBS supplemented with 3% formaldehyde) and incubated for 1 hour at room temperature under strong agitation. Formaldehyde was quenched by adding glycine (125 mM) during 20 min at room temperature under gentle agitation. Pellets were recovered by centrifugation, washed with 10 mL 1X PBS, re-centrifuged, and split in two aliquots of 25 mg that were stored at -80°C until use.

Oceanic sample collection
Oceanic sample was recovered in the bay of Napoli using 40 L of water and filtered through a 200 µm filter. Flowthrough was then sequentially filtered through filters of different size 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 (1X PBS supplemented with 3% formaldehyde) and incubated for 1 hour at room temperature under strong agitation. Filter was removed from the tube and formaldehyde was quenched by adding glycine (125 mM) during 20 min at room temperature under gentle agitation. Samples were recovered by centrifugation, washed with 10 mL 1X PBS, re-centrifuged, 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. 50 mL of fermentation were recovered at several time points after the adding of the pulque to the agave juice. Samples were immediately fixed independently in 200 mL of crosslinking solution (1X PBS supplemented with 3% formaldehyde) and incubated for 1 hour at room temperature under strong agitation. Formaldehyde was quenched by adding glycine (125 mM) during 20 min at room temperature under gentle agitation. Samples were then stored at 4°C until back to the laboratory. Samples were then recovered by centrifugation, washed with 10 mL 1X PBS, re-centrifuged, and stored at -80°C until process.

MetaHiC library preparation
MetaHiC libraries were generated using the ARIMA HiC+ kit and following the manual provided by ARIMA company 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 program (7500 rpm, 6 cycles 30 sec ON / 30 sec OFF, 4°C). Tubes were then centrifuged for 1 min at 1000 g and 700 µL of lysate was recovered and transferred to a new 1.5 mL Eppendorf tube. Tube was centrifuged for 20 min at 16,000 g, 4°C. Supernatant was carefully removed and pellet was resuspended in 45 µL of water. We then followed the ARIMA protocol. HiC library was then purified using AMPure beads and eluted in 130 µL of water before processing for sequencing as described previously [28,62]. 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 the manufacteres' 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 libraries concentrations were quantified using the Qubit 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 SRA and ENA databases using the internet site SRA explorer (https://sra-explorer.info/#) (**Supplementary Table 2**).

Metagenomes 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 v1.2.9 [53] with the default parameters. The quality of the assembly has been assessed using quast v5.2.0 [64]. GeNomad software version 1.4 [15] has then been used to detect and annotate phage and plasmid contigs using the end-to-end pipeline and the default parameters. All contigs annotated as MGE were then evaluated using CheckV [12] and the ones annotated as prophages were removed for subsequent analysis.

MetaTOR v1.2.8

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 a [mge] module to the new version of MetaTOR in order 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 intra-contig contact generating a score Si,j = xi,j / sqrt(xi,i, xj,j) where xi,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 [54] to visualize the reconstructed MAGs contact matrices and to Anvi'o [66] in order 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.

MAGs binning and annotation

First, PE reads from metaHiC/3C libraries have been 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 v3.1.5 [54]) in order 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 around 300 bp length compared to a median size of 90 bp for the restriction fragments in our mock community), a sequencing read can contain multiple digestion - religation events [55]. 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 paired-end reads (PE) from 61 to 143 million, resulting in a 20% increase in the number of informative reads without increasing noise signal (1.91% compared to 1.85% initially), demonstrating the usefulness of this digestion step. The digested reads and their corresponding assembly were then used as input for MetaTOR v1.2.8 and the 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 [56] and classify in terms of quality following these criteria [57] : 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%.

vMAGs binning, host association and annotation

Contigs annotated as viral or plasmid by geNomad were used as input for the mge module of MetaTOR using the 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 has then been 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 v2.2.4.5 [58] 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 v0.8.11 [59]. Plots were displayed using different R packages [60], a linear scale with a maximum value corresponding to 99% of the maximum value contained in the contact map.

MetaTOR benchmark

In order to evaluate MetaTOR, we compared it to the existing tool ViralCC [28] and SemiBin [27]. We feed the different pipelines with viral contigs detected by geNomad for datasets Hum22, Hum48 and Hum66. We then 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 vMAGs contact maps. Indeed, contact maps of well-binned vMAG 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 2kb 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 (compared to 5 by MetaTOR, including one complete). Altogether, these analyses demonstrate that MetaTOR is able to generate $_{MGE}$MAGs with little contamination.

vMAGs proteomic tree

Proteomic tree was built using the last version of ViPtree [45] and default parameters. vMAGs proteomic tree was then annotated and visualized using iTOL v5 [61]. *Crassvirales* and associated viruses tree was computed the same way by solely adding reference Crass genomes in the input fasta file provided to ViPTree using the data from Guerin et al., [47].

Statistics & Reproducibility

No data were excluded from the analyses. Statistics test were done using R.

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 MAGs and $_{MGE}$MAGs data are provided as supplementary datasets and are available at zenodo (https://zenodo.org/records/14851637) [62].

Code availability

Open-access versions of the programs and pipelines used (Hicstuff, MetaTOR, HiContacts) are available online on github: 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.

Acknowledgements

We thank Pierrick Moreau for assistance during experimental work and the different teams of the Microbiology department from Institut Pasteur and especially Jakub Czarnecki for providing bacterial pellets. We thank Domenico d'Alelio from the Stazione Zoologica Anton Dohrm for its help with the oceanic sample. We also thank the hacienda Vergel de Gaudalupe for allowing us to sample their fermenter. This research was funded, in whole or in part, by funding Agence nationale pour la recherche ANR-20-CE92-0048 to MM and LD, and ANR-16-JPEC-0003–05 to RK, by a grant from the French government, managed by the Agence Nationale de la Recherche under the France 2030 program (ANR-23-CHBS-0002) to RK. Biomics Platform, C2RT, Institut Pasteur, Paris, France, is supported by France Génomique (ANR-10-INBS-09) and IBISA. AB was supported by an ENS fellowship from the French Ministry of Higher Education, Research and Innovation. DC is supported by a PhD grant from the PhastGut project. AB and DC belong to Ecole Doctorale Complexité du vivant ED515 of Sorbonne Université. LC, MG, MC 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).

Author contributions

Conceptualization: AB, RK, MM. Methodology: AB, RK, MM. Software: AB, MM. Validation: MM. Investigation: MM, with contributions from JS, GLT, OC, NR, DC, AT, MG, MC, JP, KL. Formal analysis: AB, MM with contributions from AP, GAM, JS. Data Curation: MM and AB. Resources: DEC, GI, NR, MM, KL, LM, AT, AB, JP, PH, DC, GLT, MGO, MC. Visualization: MM, RK. Writing - original draft: MM, RK, AB. Writing – Editing: all authors. Supervision: MM, RK, SH, LM, GI, LD, GLT, OC. Funding acquisition: MM, GI, LD, RK. Project Administration: MM, RK.

Competing interests

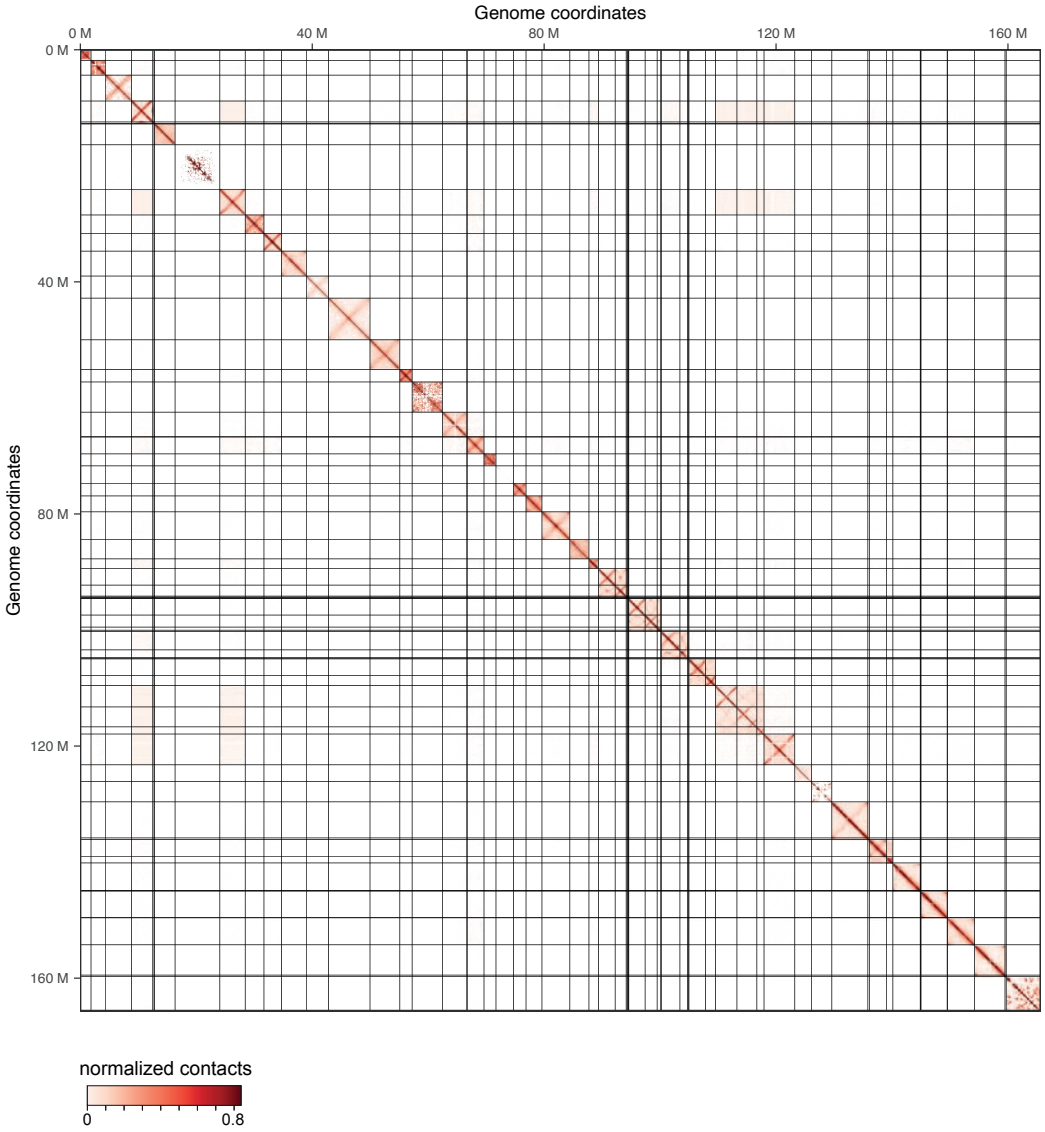
The authors declare that they do not have competing interests.

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.* fuv048 (2015) doi:10.1093/femsre/fuv048.
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.* 1–10 (2023) doi:10.1038/s41587-023-01953-y.
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. ; L., I. ;. Dunham, M. J. ;. Shendure, J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. in *G3 (Bethesda)* vol. 4 1339–46 (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.* 1 (2019) doi:10.1038/s41396-019-0446-4.
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**, (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**, (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.* 1–11 (2019) doi:10.1038/s41564-019-0625-0.
32. Press, M. O. *et al.* Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv* 198713 (2017) doi:10.1101/198713.
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.* wrae240 (2024) doi:10.1093/ismejo/wrae240.
35. DeMaere, M. Z. *et al.* Metagenomic Hi-C of a Healthy Human Fecal Microbiome Transplant Donor. *Microbiol. Resour. Announc.* **9**, (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**, e00601-23 (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**, (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).

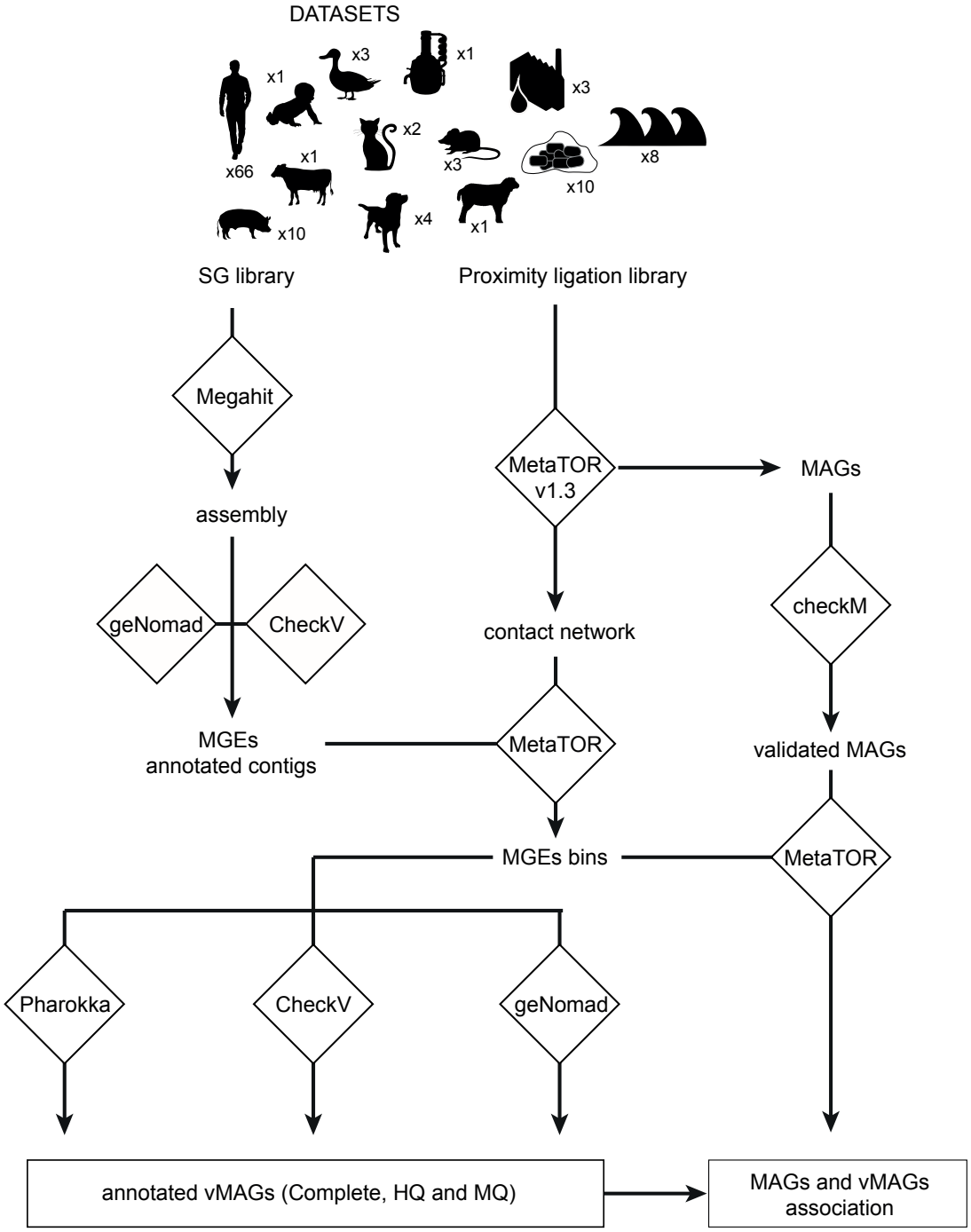
41. Pfeifer, E., Moura de Sousa, J. A., Touchon, M. & Rocha, E. P. C. Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).
42. Bouras, G. *et al.* Pharokka: a fast scalable bacteriophage 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. (2018) doi:10.1101/295642.
48. Schmidtke, D. T., Hickey, A. S., Liachko, I., Sherlock, G. & Bhatt, A. S. Analysis and culturing of the prototypic crAssphage reveals a phage-plasmid lifestyle. 2024.03.20.585998 Preprint at https://doi.org/10.1101/2024.03.20.585998 (2024).
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**, (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. *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).
54. 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).
55. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
56. 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).
57. 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).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2019).
60. Serizay, J., Matthey-Doret, C., Bignaud, A., Baudry, L. & Koszul, R. Orchestrating chromosome conformation capture analysis with Bioconductor. *Nat. Commun.* **15**, 1072 (2024).
61. 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).
62. marbouty, martial. Mobile genetic elements with a multi-host spectrum are common in the environment. (2025).

Extended Data Figures and Legends



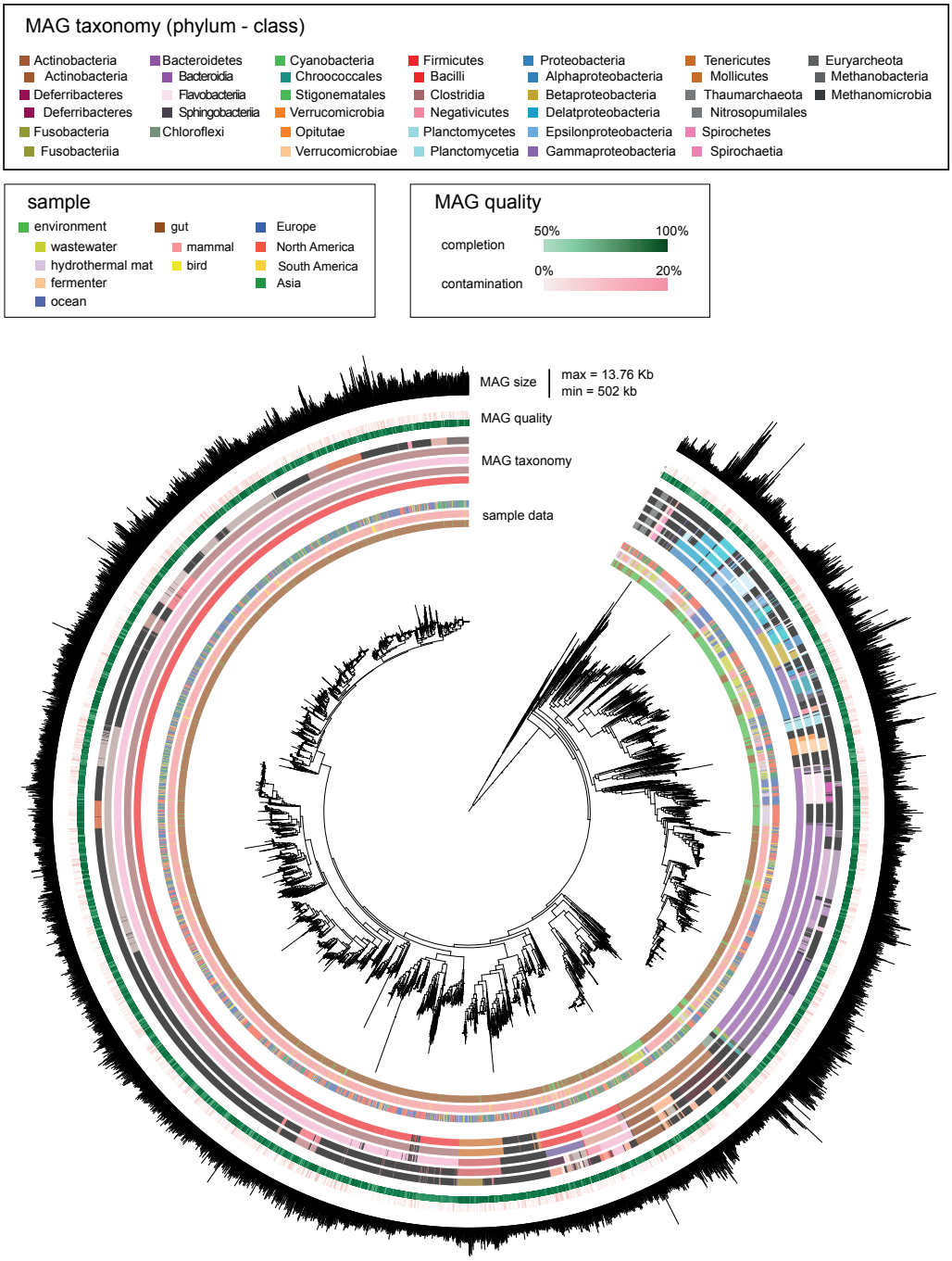
Extended Data Fig 1: contact map of the mock community.

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.



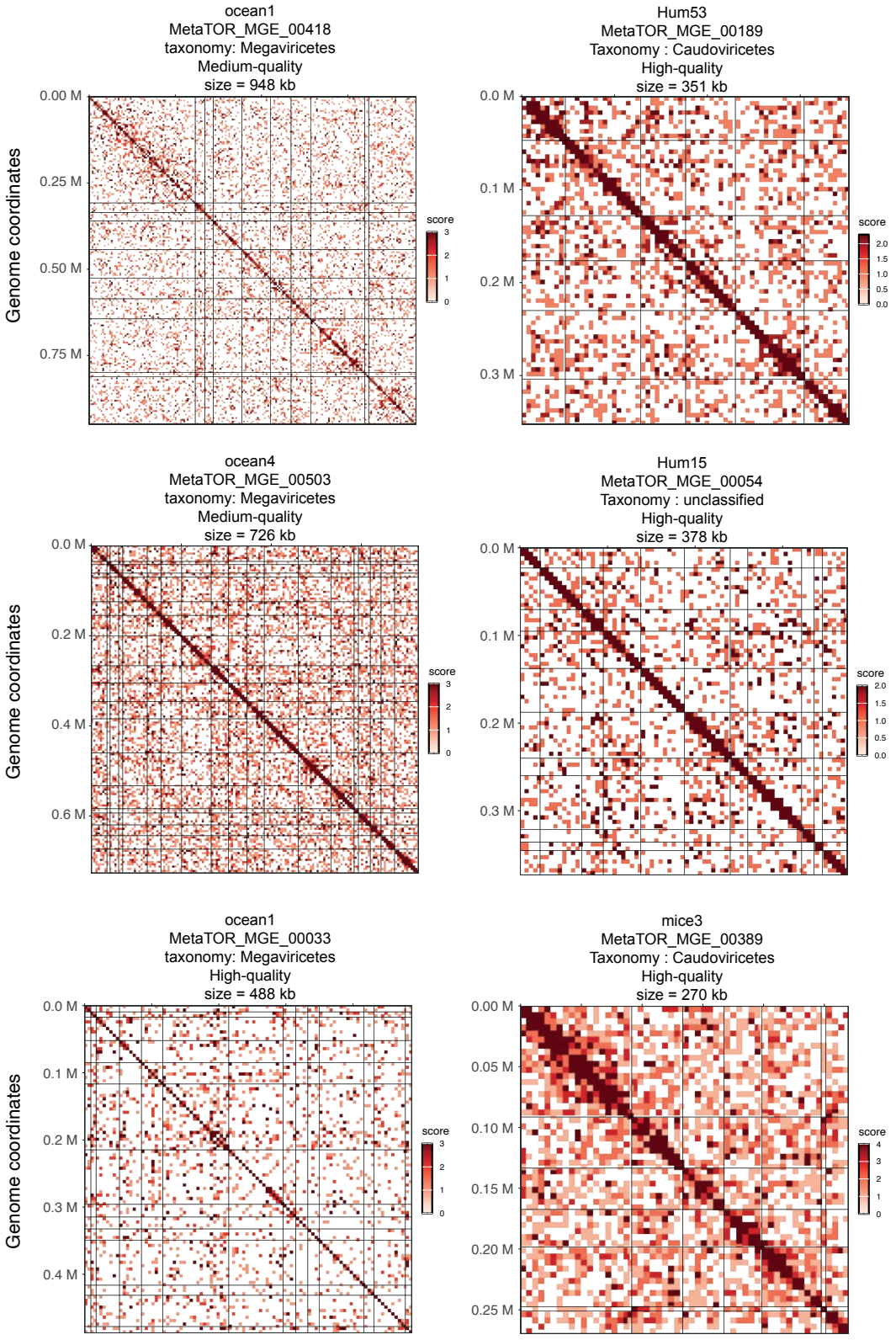
Extended Data Fig 2: datasets and pipeline used in the present study.

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).



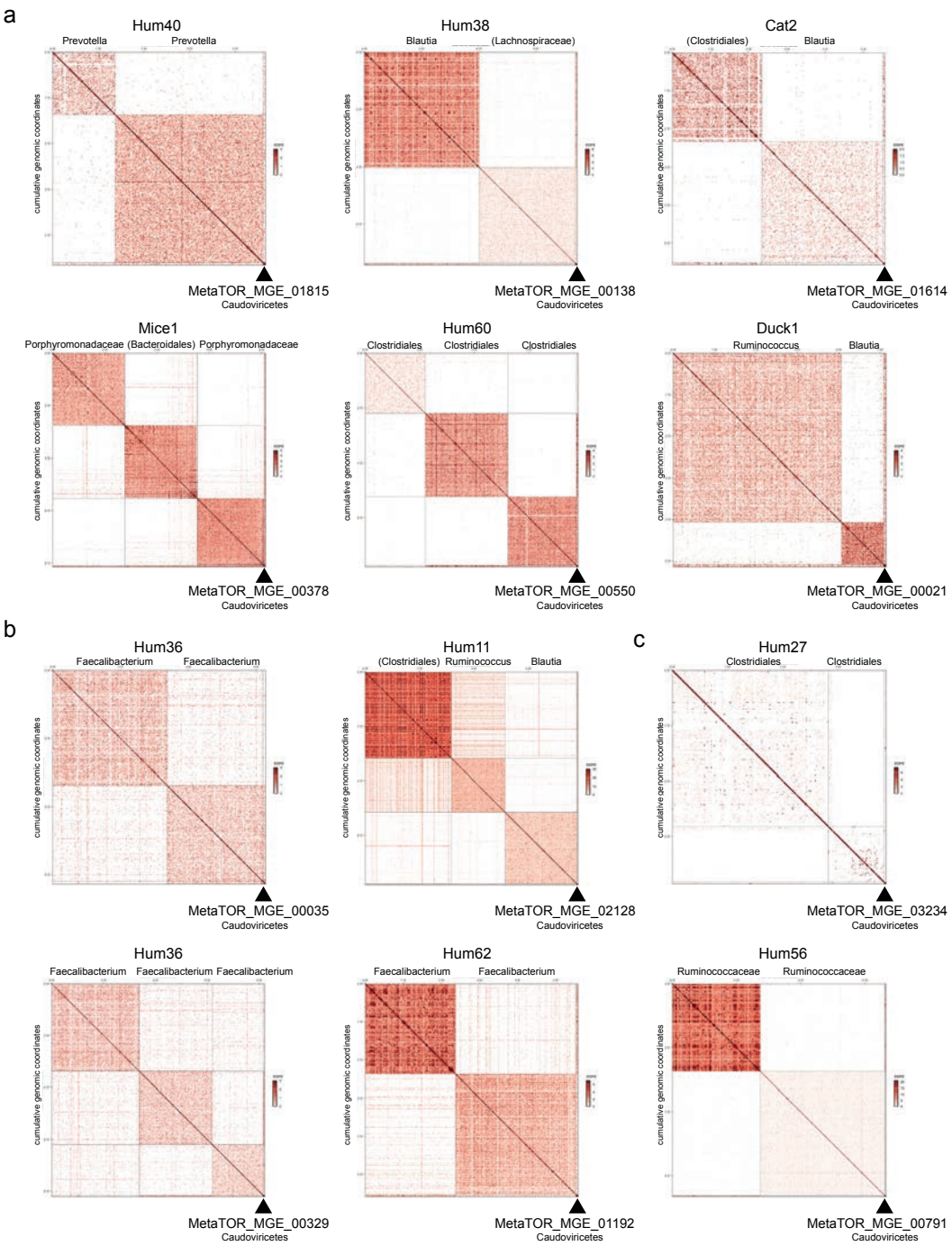
Extended Data Fig 3: Phylogenetic tree of the MAGs.

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.



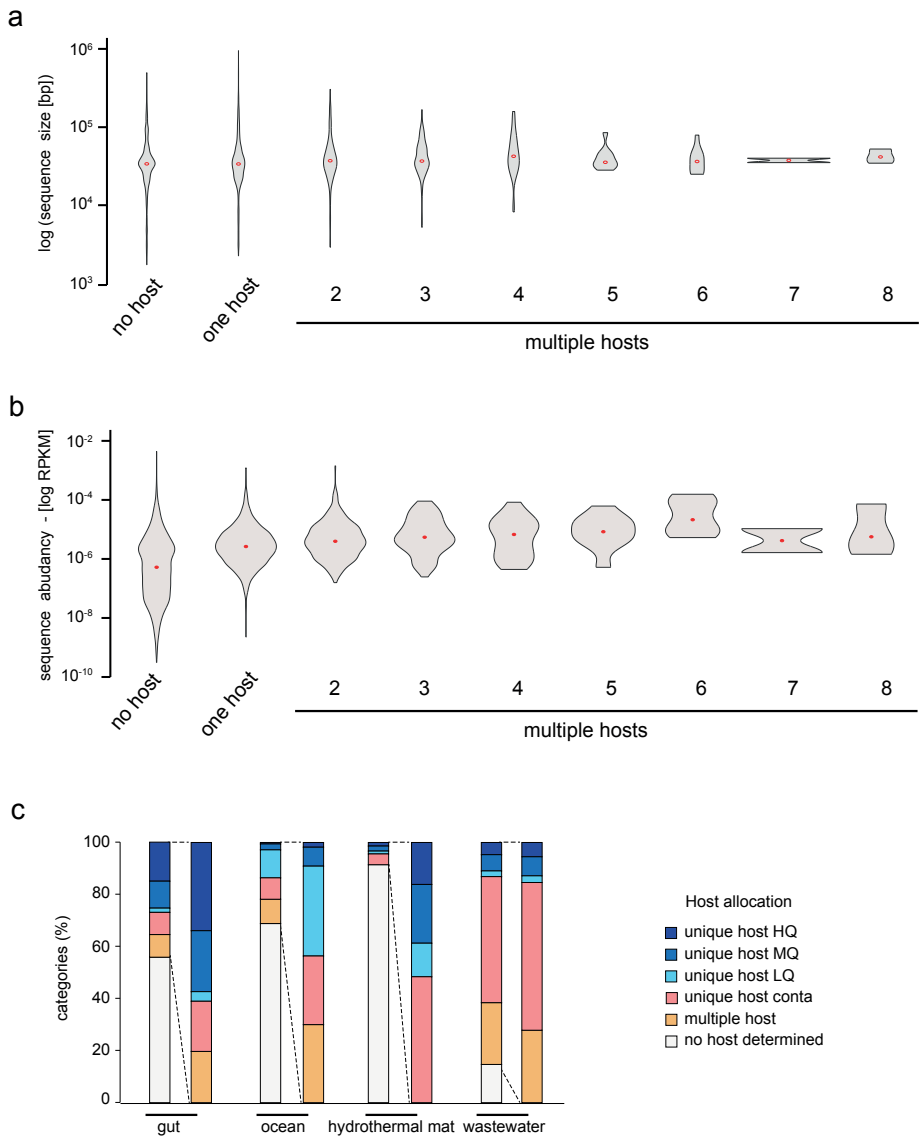
Extended Data Fig 4: Contigs contact map of different large vMAGs.

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.



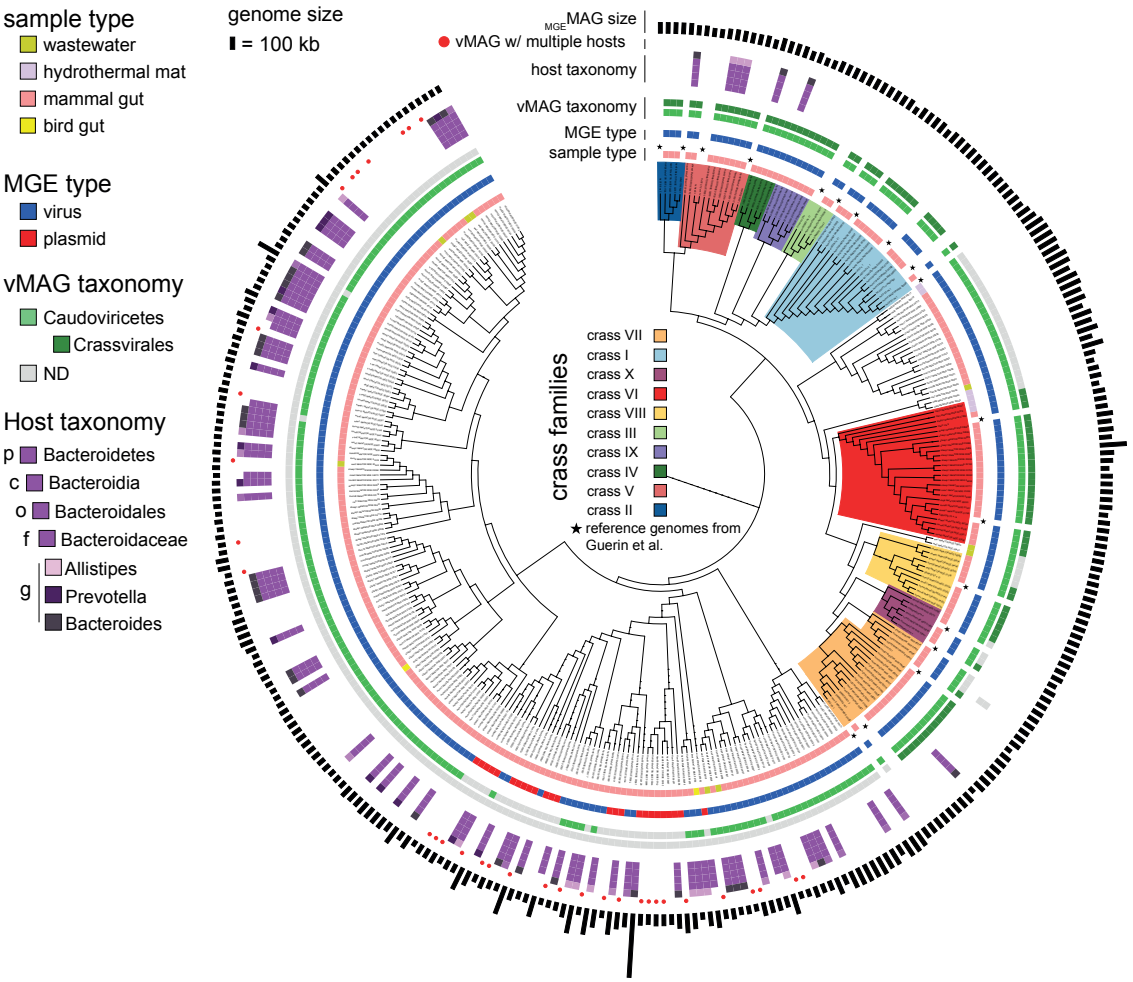
Extended Data Fig 5: contact maps of $_{MGE}$MAGs and their different hosts.

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.



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).



Extended Data Fig 7: Proteomic tree of the Crassphages family and related phages infecting *Bacteroidetes*.

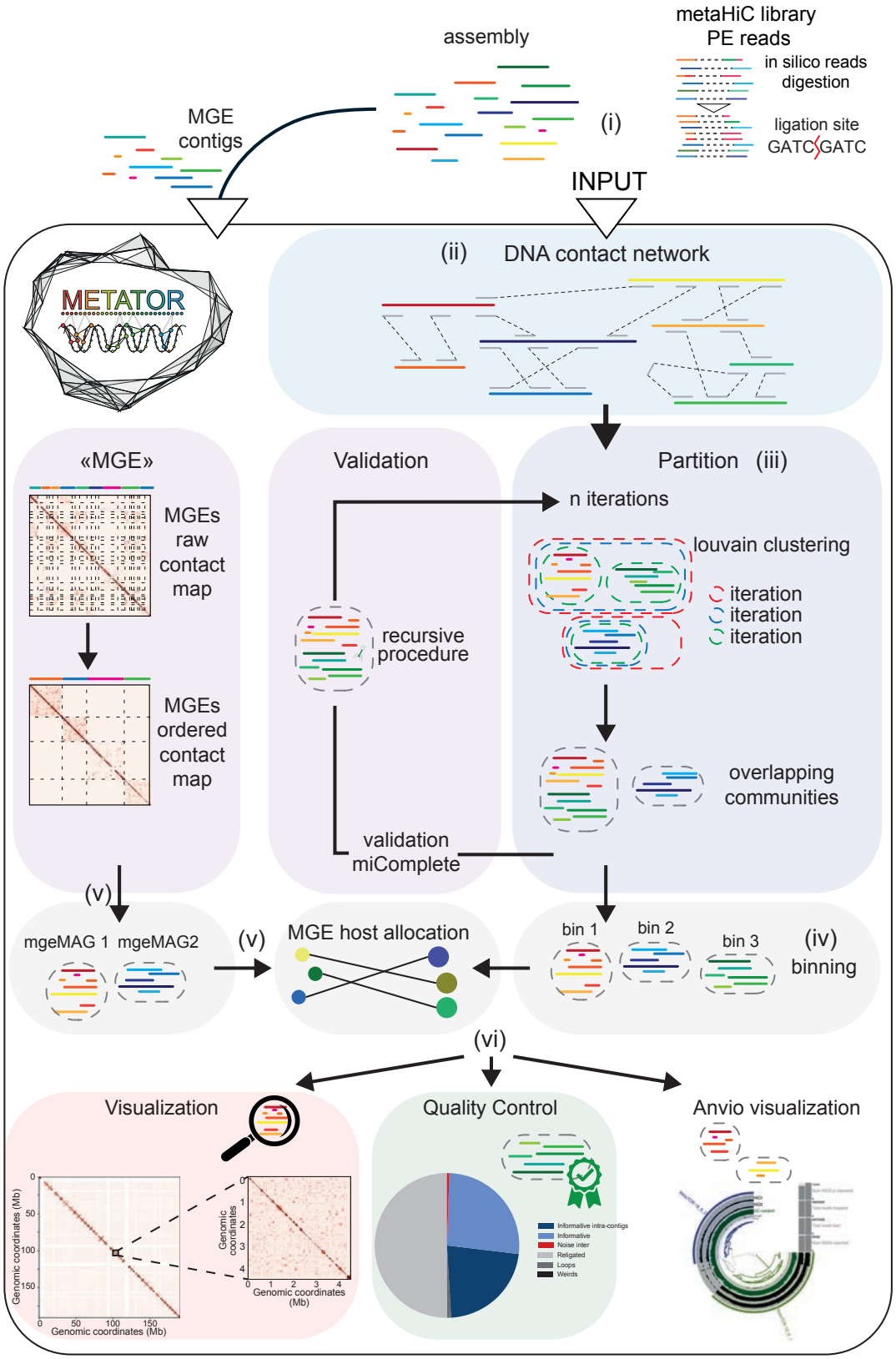
Proteomic tree of the different *Crassvirales* and related phages characterized in the present study. The Tree also encompasses different representative reference Crass genomes from Guerin et al. [56], indicated by a black star. The different crass families are indicated by colored areas over the branches. 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).

Supplementary Data

Supplementary Figures 1-3

Supplementary Tables 1 and 2

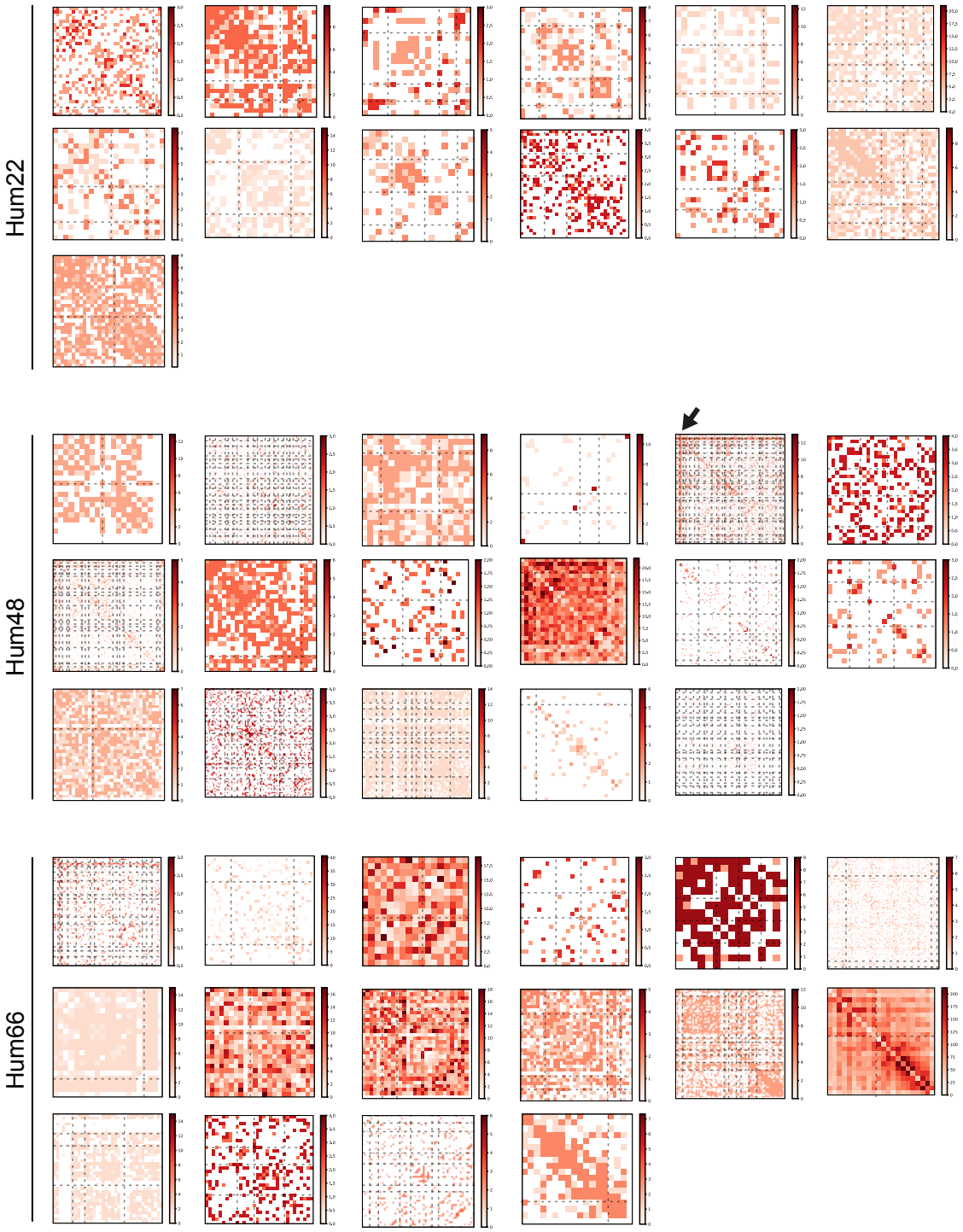
Supplementary Figure 1



Supplementary Figure 1: Overview of the new MetaTOR pipeline.

i) inputs correspond to HiC PE reads and an assembly of the studied sample. A new module allows to pre-digest the reads based on the restriction enzyme(s) used during the HiC library generation in order to increase the mapping rate and the 3D signal. *ii)* the contig network of interactions based on HiC PE reads is computed and, eventually, normalized. *iii)* the network is partitioned using an iterative procedure of the louvain algorithm and the group of contigs (overlapping communities) are recovered based on a threshold of the iterative procedure. *iv)* bins above 500 kb are evaluated using miComplete and, eventually, refined using a recursive procedure of the louvain algorithm and evaluation of the sub-communities using miComplete. Final bins with the best ratio completion / contamination are then generated. *v)* a new module (*i.e.* MetaTOR mge) allows to bin contigs (*i.e.* contigs annotated as MGE in our case) of user choice based on their intra- and inter-contact signal. MGE bins are then extracted and assigned to microbial MAGs on the base of the interactions signal. *vi)* different modules allow to 1- generate contact map of any objects (overlapping communities, final bin, MGE bin…); 2- generate different metrics of the HiC libraries quality (3D signal, intra vs. inter contigs signal…); 3- generate different files that serve as input for the Anvio platform.

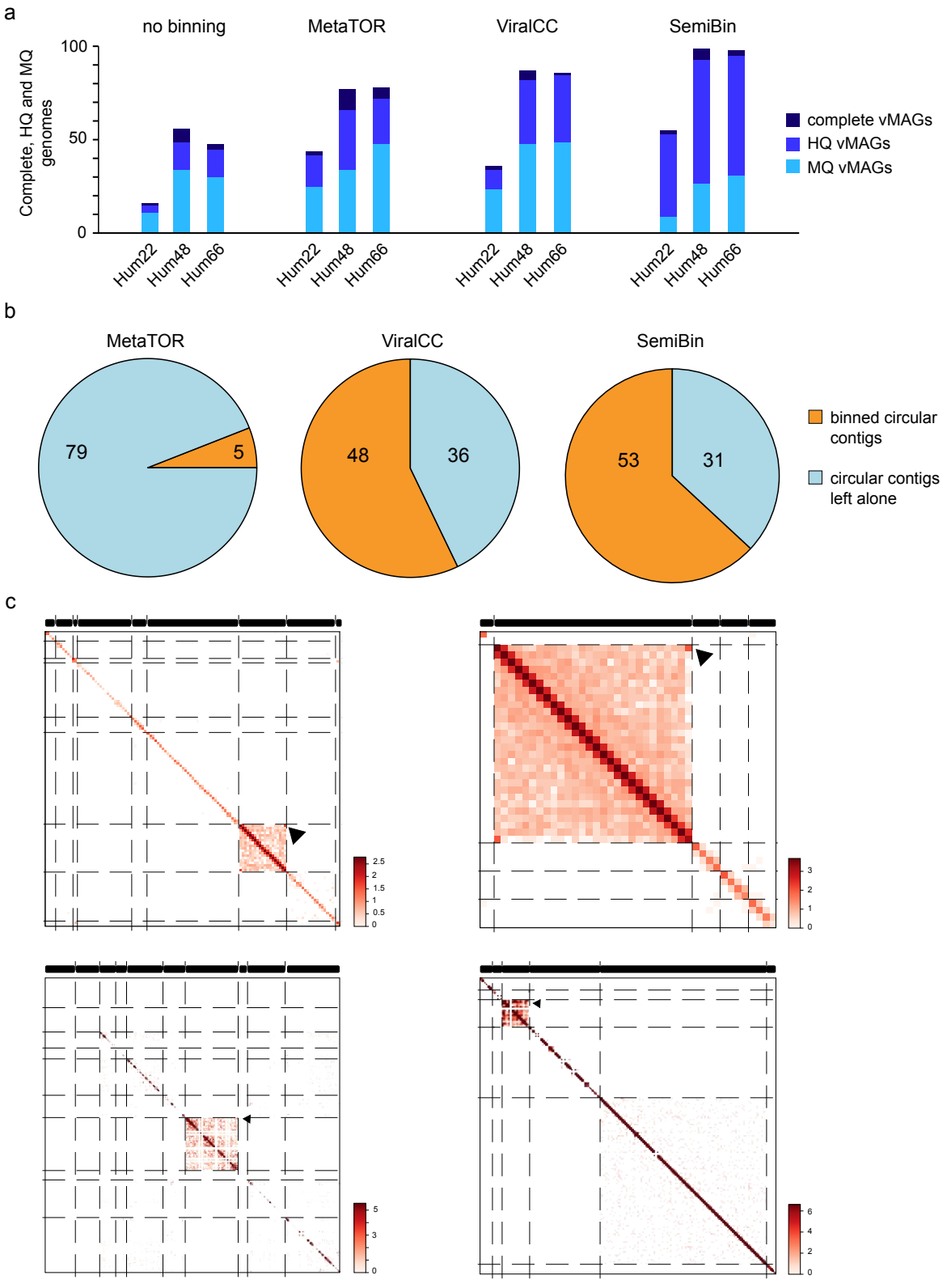
Supplementary Figure 2



Supplementary Figure 2: Contigs contact map of the complete and HQ vMAGs encompassing several contigs obtained for the Hum22, Hum48 and Hum66 samples.

Raw contact map of the 46 vMAGs obtained on samples Hum22, Hum48 and Hum66 assessed as complete or high-quality by CheckV and encompassing several contigs. Dashed lines represent borders of the different contigs. Samples are indicated on the left while scale bars are present on the right. Black arrow points to the only vMAGs exhibiting an aberrant signal. Main diagonal was set to zero prior to binning and plotting of the 3D signal.

Supplementary Figure 3



Supplementary Figure 3: Benchmark of MetaTOR pipeline.

a. Bar plot of the number of complete, HQ and MQ viral genomes assessed by CheckV for the raw contigs and the vMAGs obtained using MetaTOR, SemiBin or ViralCC for 3 human gut datasets (Hum22, Hum48 and Hum66). **b.** Pie chart of the proportion of circular viral contigs either left alone or binned with other contigs. **c.** Example of contact map of ViralCC and SemiBin bins encompassing circular viral contigs. Black arrows indicate circular signals in the contact map.

Supplementary Table 1: reference genomes of the different species present in the mock community. (NA = Not Applicable).

Genus	Species	Strain	GenBank assembly accession
akkermansia	muciniphila	YL44	GCA_016697425.1
acutalibacter	muris	KB18	GCA_016697365.1
agrobacterium	fabrum	C58	GCA_000092025.1
bifidobacterium	animalis	YL2	GCA_023278655.1
brucella	antrophi	ATCC 49188	GCA_000017405.1
bacteroides	caecimuris	I48	GCA_023277905.1
burkholderia	cepacia	BC16	GCA_009586235.1
bacillus	cereus	NC7401	GCA_000283675.1
blautia	pseudococcoides	YL58	GCA_016696745.1
bordetella	petrii	DSM 12804	GCA_000067205.1
bacillus	subtilis	BS168	GCA_000009045.1
clostridium	difficile	630	GCA_000009205.2
corynebacterium	glutamicum	ATCC 13032	GCA_000196335.1
clostridium	innocuum	I46	GCA_016697325.1
clostridium	perfringens	13	GCA_000009685.1
cereibacter	sphaeroides	KD131	GCA_000021005.1
enterocloster	clostridoformis	YL32	GCA_016696785.1
escherichia	coli	K-12 MG1655	GCA_000005845.2
enterococcus	faecalis	KB1	GCA_016696825.1
flavonifractor	plautii	YL31	GCA_023277885.1
helicobacter	pylori	G27	GCA_000021165.1
klebsiella	oxytoca	CCUG 72917	ERR2221348
klebsiella	pneumoniae	NTUH-K2044	GCA_000009885.1
leptospira	biflexa	Patoc 1 (Paris)	GCA_000017685.1
listeria	monocytogenes	EGD	GCA_000582845.1
legionella	pneumophila	Paris	GCA_000048645.1
limosilactobacillus	reuteri	I49	GCA_016697045.1
muribaculum	intestinale	YL27	GCA_016696845.1
methanobrevibacter	smithii	ACE6	GCA_000824705.1
mycobacterium	tuberculosis	H37Rv	GCA_000195955.2
pseudomonas	aeruginosa	PAO1	GCA_000006765.1

paracoccus	**denitrificans**	**ATCC 19367**	**GCA_004063735.1**
streptomyces	**coelicolor**	**DSM**	**GCA_013307045.1**
sulfolobus	**islandicus**	**REY15A**	**GCA_000189555.1**
streptococcus	**penumoniae**	**TIGR4**	**GCA_000006885.1**
synechocystis	**PCC6803**	**PCC6803**	**GCA_000009725.1**
turicimonas	**muris**	**YL45**	**GCA_016696765.1**
vibrio	**cholerae**	**O1 El Tor N16961**	**GCA_000006745.1**
veillonella	**parvula**	**SKV38**	**GCA_902810435.1**
yersinia	**enterocolitica**	**WA**	**GCA_000834195.1**
yersinia	**pseudotuberculosis**	**IP32953**	**GCA_000834295.1**
phage*	**phikz**	**NA**	**GCA_000842965.1**
phage*	**SPP1**	**NA**	**GCA_000847145.1**
phage	**Vibrio_phage_115**	**NA**	**GCA_945833845**
phage	**Vibrio_phage_191**	**NA**	**GCA_945834175**
phage	**Vibrio_phage_115E34_picmi**	**NA**	**Barcia-Cruz et al. 2023**

Supplementary Table 2: metagenomic datasets used in the present study.

(grey colors delimit the different studies)

id	ref	bioproject	enzyme	protocol	sample type
cat1	Rojas et al., 2022	PRJNA893230	Sau3AI,MluCI	proximeta	cat_gut
cat2	Rojas et al., 2022	PRJNA893230	Sau3AI,MluCI	proximeta	cat_gut
cow1	Bickhart et al., 2019	PRJNA507739	Sau3AI,MluCI	proximeta	cow_gut
dog1	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	dog_gut
dog2	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	dog_gut
dog3	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	dog_gut
dog4	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	dog_gut
duck1	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	duck_gut
duck2	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	duck_gut
duck3	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	duck_gut
mice1	Marbouty et al., Sc Adv 2017	PRJNA3021589	HpaII,MluCI	Meta3C	mice_gut
mice2	Baudry et al.,Frontiers 2019	PRJNA542645	HpaII,MluCI	Meta3C	mice_gut
mice3	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	mice_gut
pig1	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig2	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig3	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig4	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig5	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig6	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig7	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig8	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig9	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
pig10	Kalmar et al., PLOS genetics 2022	PRJEB48382	HpyCH4IV	MetaHiC	pig_gut
Hum1	Yaffe et al., 2020	PRJNA505354	DpnII,HinfI	MetaHiC	Human_gut
Hum2	Yaffe et al., 2020	PRJNA505354	DpnII,HinfI	MetaHiC	Human_gut
Hum3	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum4	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum5	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum6	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum7	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum8	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum9	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum10	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut

Hum11	Kent et al., 2020	PRJNA649316	Sau3AI	MetaHiC	Human_gut
Hum12	DeMaere et al., MRA 2020	PRJNA377403	Sau3AI,MluCI	proximeta	Human_gut
Hum13	Ivanova et al., Frontiers 2022	PRJNA718195	HpaII	Meta3C	Human_gut
Hum14	Ivanova et al., Frontiers 2022	PRJNA718195	HpaII	Meta3C	Human_gut
Hum15	Press et al., BioRXiv 2017	PRJNA413092	Sau3AI,MluCI	proximeta	Human_gut
Hum16	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum17	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum18	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum19	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum20	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum21	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum22	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum23	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum24	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum25	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum26	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum27	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum28	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum29	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum30	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum31	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum32	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum33	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum34	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum35	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut

Hum36	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum37	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum38	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum39	Gounot et al., 2022	PRJEB49168	Sau3AI,MluCI	proximeta	Human_gut
Hum40	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum41	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum42	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum43	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum44	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum45	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum46	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta3C	Human_gut
Hum47	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta-eHiC	Human_gut
Hum48	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta-eHiC	Human_gut
Hum49	Marbouty et al., Elife 2021	PRJNA627086	HpaII,MluCI	Meta-eHiC	Human_gut
Hum50	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum51	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum52	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum53	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum54	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum55	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum56	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum57	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum58	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum59	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum60	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum61	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum62	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum63	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum64	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum65	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
Hum66	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	Human_gut
mat1	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat2	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat3	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat4	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat5	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat6	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat

mat7	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat8	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat9	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
mat10	Hwang et al., 2023	PRJNA879229	Sau3AI,MluCI	proximeta	hydrothermal_mat
Mezcal	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	mezcal_fermentation
MK1	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	infant gut
ocean1	This study	PRJNA1169674	DpnII,HinfI	MetaHiC V2	oceanic
ocean2	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
oceean3	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
ocean4	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
ocean6	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
ocean7	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
ocean8	Varona et al.	PRJNA975592	Sau3AI,MluCI	proximeta	oceanic
sheep1	Bickart et al., nat Biotech 2022	PRJNA595610	Sau3AI,MluCI	proximeta	sheep gut
ww1	Stalder et al., ISME 2019	PRJNA506462	Sau3AI,MluCI	metaHiC	waste water
ww2	Stalder et al., ISME 2019	PRJNA506462	Sau3AI,MluCI	metaHiC	waste water
ww3	Chen et al., nat comm 2021	PRJNA745436	Sau3AI,MluCI	metaHiC	waste water