Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut

Ran Li
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Yongming Wang
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Han Hu
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Yan Tan
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Yingfei Ma (yingfei.ma@siat.ac.cn)
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences https://orcid.org/0000-0002-2563-5390

Article

Keywords:

Posted Date: May 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1667356/v1

Version of Record: A version of this preprint was published at Nature Communications on December 29th, 2022. See the published version at https://doi.org/10.1038/s41467-022-35735-y.

Abstract

The human gut microbiome has been extensively explored, while the archaeal viruses remain largely unknown. Here, we conducted a comprehensive analysis to identify human gut archaeal virus sequences in 3,971 assembled metagenomes and the existing human gut virus collections using the CRISPR spacer and viral signature-based approach. This resulted in 1,279 archaeal viral species, of which, 95.2% infected *Methanobrevibacteria_A*, 85.1% were not found in existing virus collections, 43.5% shared identity < 95% with the proviruses in the UHGG archaeal genomes, 37.2% had a host range spanning > 2 archaeal species, and 55.7% were highly prevalent in > 1% of the human population. Besides viral hallmark genes, the archaeal viral genetic repertoire includes a methanogenic archaeal virus-specific gene *peiW* that frequently occurred in the viral sequences (n = 150). The analysis of 33 complete caudoviral genomes often discovered the genes (*integrase*,n = 29; *MazE*, n = 10) regulating the viral lysogenic-lytic cycle, demonstrating that the temperate viruses likely predominate in the human gut archaeal virome. Together, our work uncovers the considerable unexplored diversity of the human gut archaeal viruses, revealing the novel and unexpected facet of the human gut microbiome.

Introduction

The human gut microbiome is closely linked with human health[1]. In addition to the predominant bacterial component, non-bacterial members of the gut microbiota (archaea, fungi, and viruses) are known to play important roles in microbiome dynamics and human physiology, immunity, disease, *etc*.[2]. The archaeal GTDB (04-RS89) comprises 2,392 quality-filtered genomes and they were clustered into 1,248 species[3]. Archaea were originally discovered and isolated from extreme ecosystems[4]. Moreover, over the past decades, cultivation-independent studies have revealed that archaea are universally distributed and could be among the most abundant and active microorganisms in moderate environments such as the human body sites[5]. Methanogenic archaea contribute to substantial methane production and have been linked to various conditions of health and disease in humans, as they can contribute up to 12% of total anaerobes in the human gut[6]. Methanogenic archaea are mainly represented by the orders Methanobacteriales and Methanomassiliicoccales in the human gastrointestinal tract, with the species *Methanobrevibacter smithii* being an almost ubiquitous inhabitant of the intestinal microbiome[7]. Several recent studies have reported that *Methanobrevibacter smithii* co-occurred with a specific bacterial community, which is specialized to degrade dietary fibers[6,8]. Also, some archaeal species of the order Methanomassiliicoccales in the human gut microbiota can reduce the adverse effect of trimethylamine through their specific activity and higher occurrence among subjects with high trimethylamine production potential, a potential factor for arteriosclerosis[9]. In addition, non-methanogenic archaea in the human gut including the members of the order Thermoplasmatales, and the family Halobacteriaceae, were recently detected in human feces as well[5]. Archaea are also among the commensal microorganisms inhabiting other organ systems of the human body, such archaea are regularly detected in the respiratory tract, the oral cavity, and the skin[5]. Nevertheless, human-associated archaea are often overlooked and remain unconsidered, since archaea are relatively low in abundance as compared to bacteria and mostly are

unculturable. As such, culture-independent methods, such as next-generation sequencing, can help capture their identity and allow a broad assessment of the human archaeome as well as the archaeal virome.

Microbial viruses exert control over the composition and metabolism of microbial communities. The dynamics of bacterial viruses in the human gut have been studied in detail so far[10,11], while few studies report the detection of the human gut archaeal viruses in the human gut[12,13]. Viruses infecting archaea are notoriously diverse both in terms of their genome sequences and virion structures[14,15]. Most archaeal viruses have been thus far isolated from hyperthermophilic or halophilic hosts, with only a handful of virus species described for methanogenic and ammonia-oxidizing archaea[16]. Recent exhaustive metagenomic surveys aided the discovery of novel archaeal viruses from multiple ecosystems, including the ocean, fresh water, hot spring, and soil habitats[14]. In human feces, smacoviruses were once thought to infect eukaryotes. Recently, they were found to infect the methanogenic archaeon *Candidatus Methanomassiliicoccus intestinalis* using a CRISPR spacer-based host prediction method[17]. Archaeal viruses in the human gut remain highly enigmatic. Analysis of the CRISPR-Cas systems encoded by archaea revealed that 90% of all sequenced archaeal genomes hold CRISPR loci, implying a rich archaeal virome in this ecosystem[18].

The knowledge gap on archaeal viruses is fostered by the lack of their genome entries in public databases, missing conserved marker genes for viruses[13]. Only 250 archaeal viruses that infect 23 host genera have been described and publicly available to date[19]. These archaeal viruses are greatly diverse and the encoding proteins display very low levels of sequence homology to those in the public database[20]. Prokaryotes harbor CRISPRs to foster immunity against viruses and other invasive genetic elements, making it possible to uncover the associations between viruses and their hosts[21]. Indeed, the approach of matching the CRISPR spacers from a known organism to viruses for assigning a virus discovered by metagenomics to a host is highly reliable[22]. When viral genomic data can be linked to a specific host organism, it becomes possible to uncover novel viruses and study how they interact with their hosts within various ecosystems.

Here, we harness spacer sequences from the archaeal CRISPR-Cas systems and viral signatures to search for archaeal viruses in the human gut. First, we performed large-scale identification of archaeal genomic contigs from 2,971 metagenomes derived from previously published studies. Then, we obtained the spacers from the identified archaeal genomic contigs and the 1,162 archaeal genomes of UHGG[23]. Based on the archaeal spacer collection and the signatures of protein homology present in the archaeal viruses, we established a pipeline for archaeal viral detection and obtained 1,279 archaeal viral species in the human gut. This effort will contribute to a better characterization of archaeal viruses and their archaeal hosts in the human gut and provide a complementary view of the human gut microbiome.

Result

Identification of archaeal genomic contigs from the metagenomes expands the archaeal diversity in the human gut

We identified the human-associated archaeal genome contigs from 12 human microbial metagenomic datasets consisting of 3,971 samples from rural and urban human populations across 13 countries (Table S1), resulting in 17,830 archaeal genomic contigs from the human gut samples, but only 33 from the samples of other body sites (detailed in Methods, Fig. 1a, Fig. S1, and Fig. S2). Thus, we focused on the archaea inhabiting the human gut. These contigs were taxonomically assigned based on the taxonomic information of the encoding proteins using the GTDB taxonomy system[24] (detailed in Methods). The result revealed a remarkably high taxonomic diversity of as-yet undescribed archaea in the human gut across 4 phyla including Methanobacteriota (72.76%), Thermophlasmatota (27.10%), Halobacteriota (0.10%) and Altarchaeota (0.03%), 8 families, 22 genera and 56 species (Table S2a and Table S2b). To further validate this result, these 17830 archaeal contigs were mapped to the 1,162 classified human gut archaeal genomes collected in UHGG[23] with BLASTn (E-value $\leq 10-5$ and coverage ≥ 0.5, Table S3a). The result showed that 15,732 contigs were matched to 833 gut archaeal genomes (see Table S3b, Fig. 1b), while 2,098 (11.8%) did not yield a significant match in the 1,162 reference genomes. These 833 genomes were classified into 3 families (with 7 genera and 13 species) (Fig. 1b). Most genomes were taxonomically affiliated with the genus Methanobrevibacter_A (735 genomes; 88.24%), in agreement with earlier reports[12]. Other genomes were affiliated to Methanomethylophilaceae UBA71 (44; 5%), Methanomethylophilus (19; 2.3%), Methanosphaera (17; 2%), Methanomassiliicoccus_A (11; 1.3%) and Methanocorpusculum MX-02 (6; 0.7%). The discrepancy in the number of the archaeal taxa assigned by these two methods suggested that more novel archaea with extremely low abundance likely are present in the human gut.

Then, these 17,830 archaeal genomic contigs were de-replicated by clustering according to 95% average nucleotide identity (ANI) and only the contigs with length $\geq$3 kbp were kept. The longest contig within each cluster was chosen as the representative sequence, resulting in 2,948 nonredundant archaeal genomic contigs. We assessed the prevalence of these archaea in the human populations based on the number of metagenomic sequencing reads mapped to the representative sequences (Fig. 1c). Due to raw reads of some HMP samples are not available, the total reads we used for the following analysis were derived from 1,904 metagenomic samples (Table S4). As a result, a total of 1,770 (92.26%) samples had at least one read that was mapped to the archaeal contigs. It turned out that the most prevalent archaeal genera in the human gut were *Methanobrevibacter_A* (82.14%), followed by *Methanomethylophilaceae ISO4-G1* (74.32%), *Methanomethylophilaceae UBA71*(58.56%), *Methanomethylophilus* (28.2%) and *Methanosphaera* (20.27%), indicating the most common archaea in the human intestine.

The human gut carries a complex, previously unexplored virome

To perform a comprehensive search for human gut archaeal viruses, first, we constructed a Human Gut Associated Archaeal Spacer Database (HGASDB) including 13,021 nonredundant CRISPR spacers recruited from the identified archaeal genomic contigs and the 1,162 archaeal genomes of UHGG[23] (detailed in Method). These spacers were derived from the contigs and genomes of different archaea lineages, with the genus *Methanobrevibacter_A* contributing to the largest number of spacers (89.82%). In particular, 8,962 spacers specifically were derived from *M. smithii*, 2,549 spacers from *M. smithii_A*, and 185 spacers from other three species (*M. woesei*, *M. orals*, and *M. millerae*) (Fig. 1d, Table S5). A small number (n = 1,325; 10.18%) of spacers were derived from other archaeal genera. We then developed a pipeline based on the spacers and viral signatures (*i.e.* hallmark genes for the known archaeal viruses) to recruit the archaeal viral sequences in the 2,271 assembled metagenomic datasets and the publicly available human gut virus collections (see Methods and Fig. 2a for further details), resulting in 16,234 sequences. After we filtered out archaeal and bacterial genomic contamination and the sequences not encoding the viral signatures (see in Methods), these sequences were ultimately clustered ($>$95% Average Nucleotide Identity) into 1,279 nonredundant viral species, and the longest sequences within each species were selected as the representative. These representative sequences were considered as final archaeal virus sequences, namely Human Gut Archaeal Virome Database (HGAVD), for further analysis. In particular, 1,080 archaeal viral representative sequences in HGAVD were detected from the assembled metagenomic datasets, 89 from IMG/VR[25], 92 from GPD[26], 14 from GVD[13], 2 from HGV[10], 1 from EVP[27], and 1 from GL-UVAB[28].

Subsequently, genome completeness for the representative sequences of these viral species was estimated using CheckV[29], giving rise to four different quality tiers: complete genomes (3%), high-quality (9%), medium-quality (7%), low-quality (17%), and the remainders (67%) being undetermined (Fig. 2b and Table S6). In addition, we applied VirSorter [30] (categories 1–6) and VirFinder [31] (score $\geq$0.7 and p $<$0.05) on the sequences in HGAVD, and in total 442 HGAVD sequences (Table S6) were classified as viral sequences by these tools. We did not detect plasmid signatures using PlasForest[32] in these sequences. Two sequences encoded both transposase genes and viral signatures. We then aligned the HGAVD species with the 85 nonredundant proviruses derived from 557 (50–100% completeness) of the 1,162 gut archaeal genomes in UHGG[23], resulting in 56.5% (n=723) of the 1,279 species sharing identity > 95% with those proviruses. These results demonstrated that the identification workflow we constructed here is reliable. Considering that the HGAVD sequences encode viral signatures and were matched to the spacers, a number of the HGAVD sequences likely are novel archaeal viruses and were ignored by these well-developed software tools.

To further explore the extent to which the HGAVD viral species were homologous to the known archaeal viruses in the RefSeq database (v201) (built-in database of vConTACT2) and thereby taxonomically classify these viruses, we constructed the gene sharing networks generated by vConTACT2, where viral clusters (VCs) approximate genus level taxonomy[33]. With the sequences from the archaeal viral genomes in the database RefSeq and the 1,279 archaeal viral species, this analysis clustered 735 HGAVD species into 61 VCs, 391 viral species into outliers (where contigs were assigned to a VC but shared fewer similar

proteins than the bulk of the cluster), and 153 viral species into singletons (sequences that did not cluster with any other sequences). Only 2 VCs included one known reference viral sequence, respectively. This suggests that the majority of the VCs derived from the human gut likely represent viral genera that were novel to the archaeal viruses in RefSeq (Table S7). Moreover, in agreement with the previous gut virome studies[26,34], the majority (67.9%) of the HGAVD viral species can't be taxonomically classified into any known viral order. Less than half of the species (32.1%) were taxonomically classified into, specifically, the Caudovirales order (n = 388) (virus characterized by having tails and icosahedral capsids), the Cremevirales order (n = 13), and the Haloruvirales order (n = 2) (Fig. 2c).

We further compared the HGAVD viruses to those of the publicly available virus collections (detailed in Method) including 11,827 sequences (only the sequences encoding at least one protein sequence with hit to those of HGAVD) from MGV (Metagenomic Gut Virus) catalog[34], 3,502 from Prokaryotic Viral Refseq (supplied by vConTACT2), and 37 provirus sequences (sharing identity$\leq$95% with the HGAVD sequences) derived from the archaeal genomes in UHGG (Table S8, Fig. 3a and Fig. S4). The MGV catalog is the newest human gut viral database and contains extensive viral genomic diversity. The vConTACT2 network analysis resulted in the generation of 68 VCs with at least 1 HGAVD prediction. However, the 102 MGV archaeal virus sequences were clustered into 15 VCs, and 37 proviruses were only clustered into 9 VCs, reflecting the great diversity of the gut archaeal virus taxa represented by HGAVD at the genus level. Strikingly, we found that 1,097 of the 1,279 HGAVD viral species (86%) were not grouped with any viral genomes from other collections (Fig. 3a), while a majority of 37 archaeal proviruses (78.4%) and the MGV archaeal viral sequences (83.3%) were grouped with the HGAVD viruses, indicating that HGAVD can represent most of the archaeal viruses in other gut virus collections, and the number of HGAVD viruses is much higher than that of other databases. Taken together, HGAVD considerably expanded the previously unknown archaeal viral diversity in the human gut.

Archaeal viruses are highly prevalent in the human gut

We estimated the abundance of the HGAVD viral species in the human gut samples by metagenomic read recruitment and accordingly performed the principal coordinate analysis (PCoA). No significant difference in the human gut archaeal viral composition was observed between male and female sex (ANOSIM, r = 0.002, p = 0.306) or according to BMI distribution (ANOSIM, r = 0.011, p = 0.201) (Fig. S7). Nevertheless, when the analysis was stratified by country, we observed that the diversity of these archaeal viruses was distinct in the samples of different locations. In particular, the archaeal viral communities between the Tanzanian and the populations from China, America, and the UK displayed significant differences, respectively (ANOSIM, R > 0.7, p < 0.001; Table S9 and Fig. 3b).

Based on the abundance determined by the reads mapping, we further investigated the prevalence of these viruses among the human populations. The result indicated that the prevalence of 7 archaeal viral species was > 10% across the human populations. These viruses belonged to 7 different VCs (Table S7 and Fig. 3c). These 7 viral species all were predicted to infect *M. smithii* and had a higher prevalence in Asian, European, and American populations than in the African population. Moreover, 712 archaeal viral

species were prevalent in 1% of the human population. Remarkably, the virus IMG|UGV-GENOME-0271153, one putative medium-quality viral genome (40.51 kb, CheckV), had the highest prevalence (72.16%) among the human populations and was predicted to infect *Methanobrevibacter smithii*. This virus genome encodes 46 genes and 8 of them were predicted for the caudoviral functional proteins (Fig. 3d and Table S10a). Furthermore, all the viral sequences (23kbp-55kbp in length) in the same VC with this virus had the host of *M. smithii* (Fig. 3a) and were derived from the samples of United Kindom, Sweden, Austria, United States, China, Spain, and Madagascar, respectively, further suggesting the wide distribution of this virus among the global population. In particular, another highly prevalent caudovirus (10.7%) IMG|UGV-GENOME-0263128 encoding 51 genes was detected more frequently in the African population than IMG|UGV-GENOME-0271153 (see Fig. 3c). The viral sequences in the IMG|UGV-GENOME-0263128-contained VC were from 19kbp to 56 kbp in size and were predicted to infect the hosts of *M. smithii* and *M. smithii_A* (Fig. 3a). These two highly prevalent viruses likely are temperate because integrase gene was detected on the genome of the virus (IMG|UGV-GENOME-0263128) or the genomes of other viruses within the same VC (IMG|UGV-GENOME-0271153) (Fig. 3d and Table S10b).

It is worth mentioning that 13 smacoviruse species were identified and were clustered into 3 VCs with lengths ranging from 2.0 ~ 2.5kbp in HGAVD, reflecting the diversity of these small viruses in the human gut. Smacoviruse in the order of Cremevirales has a small circular single-stranded DNA genome and had been identified in fecal samples (both feces and rectal swabs) of various animals[35]. These HGAVD smacoviruses were targeted by 7 spacers derived from the archaeal genomes in UHGG and they were predicted to infect *Methanomassiliicoccus intestinalis* or *Methanomassiliicoccus_A intestinalis.* Compared with the cohort of Asia and America, the prevalence of smacovirus was higher in African and European populations (Fig. 3e).

Viruses infecting M. smithii are a major component of the archaeal virome in the human gut

To accurately investigate the diverse virus-host interactions, we particularly screened for the CRISPR spacers present in the 1162 archaeal genomes in UHGG to target the HGAVD viral sequences (see Methods). As expected, the majority (n = 1217; 95.2%) of the viral species connected to the genus *Methanobrevibacteria_A*, which is dominant in the human gut archaeaome (Fig. 4a and Fig. 1c). We then measured viral diversity by determining the number of VCs for each archaeal genera, revealing that the genus *Methanobrevibacter_A* harbored a significantly higher viral diversity than those of other archaeal genera (Fig. 4b), with 51 VCs assigned to this genus. Among the 51 VCs, 47 VCs were specific to *M. smithii*, only 17 VCs specific to *M. smithii_A*, and 13 VCs were linked to both these two archaeal species, reflecting archaeal viruses can infect their hosts cross-species. To show this in detail, we constructed the network of host-virus by matching the HGAVD viruses with the CRISPR spacers derived from the 1,162 gut archaeal genomes. Surprisingly, we revealed that approximately one-third of HGAVD viral species had a broad host range (Fig. 4c). Namely, 434 viral species had a host range that spanned 2 archaeal species (*M. smithii* and *M. smithii_A*) and 12 viral species had a host range across 3 archaeal species (*M. smithii*, *M. smithii_A*, and *M. woesei*). These analyses provide a comprehensive blueprint of archaeal virus-mediated gene flow networks in the human gut microbiome.

Most of (305/388 = 78.6%) the caudoviral species in HGAVD connected to the host of *M. smithii*. Thus, phylogenetic trees were constructed based on the predicted large subunit terminase (LST) to estimate the diversity of these archaeal viruses. In total, the LSTs were identified in 80 of 305 viral species, with 1 belonging to the Terminase_1(PF03354) domain, 33 to Terminase_3 (PF04466) and 46 to Terminase_6 (PF03237). Therefore, phylogenetic trees of the HGAVD viruses and closely related viruses in the RefSeq database (v201) were constructed using the proteins encoding the domains Terminase_3 and Terminase_6, respectively (Fig. 4d and Fig. S5). As shown on the trees, all gut archaeal viral clades detected in our study did not include any known viruses, likely representing novel viral types. In consequence, the LST phylogeny expanded the diversity of the viruses that infect *M. smithii* and defined new lineages.

Archaeal virus genomes encode an extensive functional repertoire

The functional potential of human gut archaea has been extensively studied[12]. HGAVD can enable us to explore the functional potential of the archaeal virome in the human gut. To do this, we identified 97,208 protein-coding genes on the representative sequences for these 1,279 viral species and compared these genes with the Pfam (v32) database. Overall, 40% (n = 39,268) of the viral genes did not have significant matches (cutoff: e-value$<$1e-5, score $>$ 50) in this database and were not assigned to any biological functions, indicating that remarkably little is known about the functional potential of human gut archaeal viruses (see Fig. S6 and Fig. 5a).

The viruses of *M. smithii* contained the most functional diversity with proteins homologous to 1,034 different kinds of caudovirus-specific proteins in the Pfam database (only the proteins assigned biological function were taken into consideration), such as prohead protein, baseplate J, portal protein, tail fibers, and terminase large subunit, whereas other archaeal viruses lacked some caudovirus-specific functional proteins (Fig. 5b). For example, except for the viruses infecting *M. smithii*, the remainder had no proteins annotated for lysis. In particular, the genes encoding HNH endonuclease were observed on the viral genomes of both *M. smithii* and *M. woesei*. This protein potentially cleaves DNA into genome-length units during packaging and may operate in concert with their terminase large subunit and portal proteins[36].

The representative sequences of 36 archaeal viral species in HGAVD were measured as complete genomes by CheckV. They were clustered into 7 different VCs and taxonomically classified to Caudovirales (n = 33, 6 VCs) and Cremevirales (n = 3, 1 VC). Analysis of these whole viral genomes in the order Caudovirales (Fig S10 and Table S11) resulted in an interesting finding that a gene encoding the protein homologous to PeiW frequently occurred on many viral genomes (n = 23). The prototype PeiW found in the archaeal prophage psiM100 was identified as the autolytic enzyme pseudomurein endoisopeptidase produced by the thermophilic methanoarchaeon *Methanothermobacter wolfeii* to cleave pseudomurein cell-wall sacculi of archaeal methanogens[37]. The phylogenetic analysis of PeiW

revealed that except for the viruses of *Methanothermobacter wolfeii*, other archaeal viruses could also be the carrier of PeiW, such as the viruses of *M. smithii* and *Methanobrevibacter olleyae* (Fig. 5d). When extending this analysis to all HGAVD viruses, 150 viruses encoded the genes of PeiW (Fig. S9), suggesting the importance of this gene for the archaeal viruses in infecting methanogenic archaea.

In the analysis of these complete caudoviral genomes, 29 of 33 contained the genes encoding phage integrase protein. However, only 9 genomes were predicted as proviruses, and 20 were not flanked by host DNA by CheckV. In particular, we observed that 10 genomes infecting *M. smithii* or *M. olleyae* also encoded proteins belonging to the antitoxin MazE superfamily. MazE-MazF Toxin–Antitoxin system (TA) has been found in the genomes of *E. coli* and other bacteria[38]. TA systems can function as anti-virus mechanisms and viruses have evolutionally obtained defense mechanisms to avoid these systems. For example, T4 phage harbors a MazE antitoxin against MazF encoded by *E.coli* for efficient growth[39]. In addition, a small protein belonging to the antitoxin MazE superfamily protein was found in the temperate haloarchaeal virus SNJ1 to control the lysis-lysogeny switch[40]. Accordingly, the antitoxin MazE protein encoded by the HGAVD archaeal viruses might highlight an arms race between the gut archaea and their viruses over TA systems by regulating the state of viruses (lysis and lysogenic). Further, we performed a phylogenetic analysis based on the MazE antitoxin protein sequences detected in these viral genomes. The phylogenetic tree shows that (Fig. 5e), the viruses predicted to infect *M. smithii* and *M. olleyae* were separated into different clades. We performed comparative genomic analysis on the representative caudoviral sequences selected for each VC of the complete HGAVD caudoviruses (Fig. 5f). Although these viral sequences encoded the genes for the core viral functional proteins (such as portal and terminase proteins), and the accessory genes for PeiW, MazE, or integrase, they were shown divergent in genomic sequence. Overall, the analysis on these complete HGAVD viral genomes indicated that temperate archaeal viruses were dominant in the human gut, similar to the human gut bacterial phages[41], while most of the archaeal viruses detected in this study likely were in the lytic status.

Discussion

In this study, taking advantage of the metagenomic sequencing data, we conducted a comprehensive analysis of the human-associated archaeal viruses recovered from the human gut metagenomes collected worldwide, showing that the archaeal viruses were widespread in the human gut ecosystem. The results obtained in this study based on the metagenomic sequencing datasets were well-complemented with the previous study of 1,167 nonredundant archaeal genomes[12]. Based on the Minimum Information about an Uncultivated Virus Genome (MIUViG) standards[43], we report the archaeal viruses related to virus origin, genome quality, functional annotation, taxonomic classification, biogeographic distribution, and host prediction. Leveraging this comprehensive archaeal viral sequence collection, we provided unprecedented glimpses into the human archaeal virome, thus leading to a better understanding of the human gut ecosystem.

Mining metagenomic sequencing data can recover the genomic fragments of extremely low abundant microbes that were often overlooked by the amplicon-based method or the metagenomic binning method

[44]. In this study, we collected the archaeal genomic contigs from the human metagenome datasets and the results expanded the number of the archaeal taxa in the human gut to 4 phyla, 8 families, 22 genera, and 56 species. We recognized that due to the limitations of the fragment nature of the assembled metagenomic sequencing data and the Genome Taxonomy Database R95 (GTDB), the taxonomic assignment of the contigs was likely not perfectly accurate. We validated this taxonomic assignment by mapping the identified archaeal contigs against the 1,162 archaeal genomes in UHGG, yielding 98.5% contigs assigned consistently at the genus level (Table S3b, Fig. 1b). Meanwhile, 2,098 (11.8%) archaeal contigs did not yield a significant match in the 1,162 reference genomes, likely representing the novel and extremely low abundant species in the human gut. With the development of sequencing technology and the updating methods for archaeal detection, more archaea will be unearthed from the human gut in the future.

To date, compared to the bacterial phages, fewer archaeal viral genomes derived from the human gut were available. In the database GVD, 24 viral populations (equal to species in this study) were predicted as archaeal viruses[13]; The study related to gut archaeome reported 94 proviruses derived from the archaeal genomes[12]. These large-scale gut virus collections were conducted using several popular bioinformatic tools, such as VirSorter[30], VirFinder[31], etc. These tools are heavily dependent on the reference phage sequences for viral identification. This limitation caused the failure to recognize the majority of the archaeal viruses in the human gut due to the lack of references. CRISPR spacer-based method may overcome this limitation and has been widely used for linking viral and host genomes in various studies[22,45,46]. In particular, analysis of previous studies indicated that more than 90% of archaea genomes harbor the CRISPR system as compared to 50% of the human gut bacterial genomes[18]. In this study, CRISPR loci were identified in 53% of the human gut archaeal genomes (including MAGs) and 80% of the isolated human gut archaeal genomes. We collected the spacer sequences of the CRISPR system from the gut archaeal genomes in UHGG and the archaeal contigs that we identified from the human gut microbial metagenomic data and developed a stringent workflow based on the spacers and viral signatures (identified in this study, see details in Methods, Fig. 2b) to recruit the archaeal viral sequences from the assembled metagenomic data and the previous human-associated virus collections[10,13,25–28,34]. After filtering out potential contamination of bacterial genomes, archaeal genomes, and plasmids through the stringent criteria, the workflow generated 2,126 nonredundant contigs that can perfectly match the spacers. Our stringent workflow showed a high sensitivity in identifying genome fragments for diverse gut viruses. This was evident by the detection of smacoviruses which is very small (2.5kbp) and low abundant in the human gut microbiome.

Nevertheless, some non-viral mobile elements that were perfectly matched to the spacers might also be mixed in the collection of candidate II (Fig. 2b), such as transposons and plasmids. In total, 847 sequences were not detected encoding genes homologous to the viral hallmark genes. These sequences likely were derived from transponsons or plasmids and were excluded from the collection of candidate II (Fig. 2b), and we further curated the viral sequences (n = 1,279) encoding the genes homologous to those of well-described archaeal viruses and bacterial phages for further analysis. Notwithstanding this, some

of these excluded sequences that matched the spacers also likely represent additional families of as-yet-unidentified viruses. These novel viruses could not be identified by metagenomic approaches due to the lack of knowledge and must be determined by establishing a culture-dependent method. The isolated archaeal viruses may in turn improve the bioinformatic methods for identifying archaeal viruses to recover more novel archaea viruses.

Taking together, in this study, we conducted a comprehensive metagenomic data mining of the archaea and the archaeal viruses in the human gut. The result revealed the diversity of the archaeal viruses and the archaea in the human gut. Considerable diversity of the unexplored archaeal viruses in the human gut and the novel viral species in HGAVD can exactly fill in the gaps in this field and serve as an expansion of the human gut archaeal viruses. Our data, together with the bacteria and bacterial phages, will provide a complementary view of the human gut virome and thus help us better understand the human gut ecosystem.

Materials And Methods

Collection of metagenomic sequencing data sets used for this study

Here, we collected and curated 12 human microbial metagenomic datasets consisting of 3,971 human metagenomes from 1,904 individuals across rural and urban populations from 13 countries (Table S1, publicly available as of January 2021), of which, 30 fecal shotgun metagenomic samples were obtained from Chinese patients with enteritis (https://db.cngb.org/search/project/CNP0002257/). Sequencing reads of the human gut metagenomes and the associated metadata were obtained from their respective hosting databases (eg. SRA, iVirus, or MG-RAST). Reads were then assembled using SPAdes v3.10.0[47] with option ‘–meta&rsequo;. The assembled contig sequences of five body sites (including the gastrointestinal tract, mouth, airways, skin, and vagina) were directly downloaded from the HMP Data Portal (https://portal.hmpdacc.org/)[48].

Detection of Archaeal genome contigs in the metagenomic sequencing datasets

The genes were predicted on the assembled contig sequences using Prodigal v. 2.6.3 (-p meta option)[49]. The resulting protein sequences were aligned to the Genome Taxonomy Database R95 (GTDB, R95)[24] using DIAMOND (options: --e-value 1e-3 –min-score 50)[50]. According to the GTDB taxonomy system, the taxonomy of each protein was assigned based on the top hit in the database at each taxonomic rank (Phylum, order, family, genus, and species). Subsequently, Archaeal contigs were screened based on the following criteria[51]: (i) the number of encoding proteins with hit derived from archaeal genomes $>$ the number of encoding proteins with hit derived from bacterial genomes; and (ii) the number of encoding proteins with the hits from archaeal genomes $\geq$5 (Fig. 1a). In summary, we detected 17,830 archaeal

contigs from the whole gut metagenomes and 33 archaeal contigs from other body sites (23 from the oral, 5 from the skin, and 5 from the vagina) (taxonomic information of these 33 archaeal contigs are listed in Table S12). Meanwhile, the taxonomy of an identified archaeal contig was assigned if the number of the proteins on the contig assigned to this taxonomy was higher than others. Then all curated gut archaeal contigs sharing identity ≥ 95% and coverage ≥ 90% were dereplicated by CD-HIT v4.6[52]. Using this clustering strategy, we finally obtained 2,948 nonredundant archaeal genome fragments with length $>$ 3kbp for subsequent analysis.

Construction of phylogenetic tree for archaeal genomes

To compare these archaeal contig sequences to the known archaeal genomes derived from the human gut, these 17,830 archaeal contigs were mapped to 1,162 species-level gut archaeal genomes derived from the UHGG[23] using BLASTn (e-value ≤ 10 − 5, coverage ≥ 0.5)[53]. UHGG contains 286,997 genomes, representing 4,644 species of Bacteria and Archaea from the human gut that are taxonomically annotated using GTDB-tk v.0.3.1 (GTDB R89). Taxonomy of these genomes was assigned using GTDB-Tk v0.3.3[54] based on the Genome Taxonomy Database R202 (GTDB, http://gtdb.ecogenomic.org) taxonomy. The results were further refined using maximum-likelihood phylogeny inferred from a concatenation of 122 archaeal marker genes produced by GTDB-Tk. The archaeal tree was built using RAxML v8[55] called as follows: raxmlHPCHYBRID -f a -n result -s ge input -c 25 -N 100 -p 12345 -m ROTCATLG -x 12345 and Newick tree output files were visualized with iTOL v.6[56] (https://itol.embl.de/).

Establishment of Human Gut Associated Archaeal Spacer Database (HGASDB)

The CRISPR spacer sequences were derived from two databases: (i) 17,830 gut archaeal contigs detected from the gut metagenomes, (ii) 1,162 species-level archaeal genomes from the UHGG catalogue. Spacer sequences were predicted using the CRISPR Recognition Tool (CRT)[57] with default parameters. In total, 19,055 and 6,553 CRISPR spacer sequences were predicted from 1,162 UHGG archaeal genomes and the 17,830 gut archaeal contigs, respectively. Redundant spacer sequences were dereplicated using CD-HIT (parameters: -c = 1, -aS = 1, -aL = 1, -g = 1), resulting in 13,021 nonredundant CRISPR spacers sequences.

Collection reference of Archaeal Viral Genomes

We collected a database for 202 Archaeal Viral Genomes as a reference from 3 sources:

1. 97 reference archaeal viral genomes available in NCBI RefSeq as of December 2020.
2. 102 archaeal virus genomes provided in the studies of Iranzo et.al[58]. The 59 duplicated genomes compared to the genomes in (i) were removed. What's more, there were 16 genomes were labeled as "Proviruses" by Iranzo et.al. However, sequences of these proviruses have not been provided by the authors, for which reason, we used VirSorter[30] to predict the provirus from the 16 genomes. By this means, 14 proviruses have been extracted from 14 genomes. Taken together, we got 41 archaeal virus genomes from this source.

3. To complete the archaeal viral dataset, we included genomes of *Methanobacterium virus Drs3*[59], 43 new putative archaeal virus genomes identified from two depth profiles in the Eastern Tropical North Pacific (ETNP) oxygen minimum zone[60], 24 unknown archaeal viral populations detected by GVD[13] and 8 genomes of smacoviruses that were found to infect Archaea[17].

In total, the final archaeal virus database consisted of 202 archaeal viral genomes or fragments.

Selection of hallmark Genes for Archaeal Viruses

Firstly, we predicted genes from the 202 archaeal virus genomes using Prodigal v. 2.6.3 (default parameters) and obtained 21,985 proteins encoded by these genes. Subsequently, functional annotations were assigned to the proteins using the hmmsearch command in HMMER3 (e-value cutoff set to 1e-5)[61] against the Pfam. v. 32 database[62], a custom comprehensive viral HMM database including viral protein families (VPF)[27] from JGI Earth's virome project and the Virus Orthologous Groups (VOG) (release 202, http://vogdb.org) containing orthologous groups of numerous viruses. Then the database of archaeal viral hallmark genes was composed of the following four parts (Fig. S3):

(1) Exclusive archaeal viral proteins based on the annotations in the Pfam database

1. We collected 35 genomes of archaeal isolates from UHGG catalog and each protein encoded by the genomes was annotated in the Pfam database. We selected the proteins (n = 1,523) with the Pfam homologs only occurring on the 202 archaeal viral genomes as hallmark genes.
2. If any proteins encoded by the archaeal virus genomes and the 35 isolated archaeal genomes were annotated in the Pfam database with the keywords including portal, terminase, spike, capsid, sheath, tail, coat, virion, lysin, holin, base plate, lysozyme, head, fiber, whisker, neck, lysis, tapemeasure or structural, then these (n = 164) were added to the collection of hallmark genes for archaeal viruses.

(2) To include the proviruses in the archaeal genomes, we collected 11 proviruses predicted from the 35 isolated archaeal genomes in UHGG by CheckV[29], and then the 249 proteins predicted from the provirus were added to the collection of the hallmark genes for archaeal viruses.

(3) The 5,907 archaeal virus proteins with the best hit to the members of the VOG database were selected.

(4) The 3,368 archaeal virus proteins with the best hit to the members of the VPF database were selected.

After combining and de-replicating the proteins from these four sources, in total, 8,485 proteins were selected as the hallmark genes for archaeal viruses.

Development of archaeal viral detection workflow

To perform a comprehensive search for human gut archaeal viruses, sequences for archaeal virus detection were derived from two sources: 1) the assembled contigs of the metagenomic sequencing data we described above; 2) viral genomes identified in the published viral databases (see Fig. 2b), including 125,842 partial DNA viral genomes obtained from the Earth's Virome (hereafter 'EVP')[27], 57,721 viral

contigs from the Human Gut Virome database (HGV)[10], 195,698 viral contigs from Uncultured Viral Database of Archaeal and Bacteria (hereafter 'GL-UVAB')[28], 33,243 viral sequences obtained from GVD[13], 142,809 nonredundant phage genomes from GPD[26] and 2,332,702 viral genomes from IMG/VR v3[25]. To identify archaeal viral sequences from these data, we developed a viral detection workflow as follows:

1. All the assembled metagenomic contigs were searched against HGASDB using blastn from the blast + package v.2.2.31 (e-value 1e-5), and 16,234 contigs that matched to the spacers were assigned as archaeal virus candidate I. These contigs were further dereplicated using the CD-HIT (v4.6) with the parameters "-aS 0.9 -c 0.95". Multiple reports[13,43] have revealed that $>$95% ANI (Average Nucleotide Identity) was a suitable threshold for defining a set of closely related discrete 'viral group'; follow-on studies suggest that this cut-off establishes populations that are largely concordant with a biologically relevant "viral species" definition[63]. Thus, this clustering strategy resulted in 2238 viral species (represented by the longest contig within each viral species) in archaeal virus candidate I.
2. To remove potential bacterial genome contamination, sequences of archaeal virus candidate I were queried against 16,234 isolated bacterial genomes from UHGG collection using blastn. The cutoffs defined for these searches were the minimum identity of 50%, and minimum query coverage of 80%, with a maximum e-value of 10^{-5}. Thus 10 contigs were filtered out from candidate I and 2,228 viral species remained for candidate II.
3. To remove the contamination of archaeal genomes, sequences of archaeal virus candidate II were performed blastn against 35 isolated archaeal genomes from the UHGG collection. The cutoffs defined for these searches were the minimum identity of 50%, minimum query coverage of 100%, with maximum e-value of 10^{-5}, Thus 102 contigs were removed from candidate II and 2,126 viral species remained for candidate III.

Protein sequences derived from the contigs in candidate III were compared with the protein sequences of the archaeal viral hallmark genes (identified in *Selection of hallmark Genes for Archaeal Viruses*) using DIAMOND. Any contigs containing best hits with a maximum e-value of 10^{-5} were picked. For these contigs, CheckV[29] was used to detect proviruses boundaries and remove contamination from host-derived sequences.

Finally, 1,279 viral species were retained for the Human Gut Archaeal Virome Database (HGAVD).

Taxonomic classification of gut archaeal viruses

Two complementary approaches were used for the taxonomic classification of the 1,279 archaeal viral species. First, for 1,279 representative contigs of these archaeal viral species, genes were predicted using Prodigal v2.6.3 with the -p meta option. Then these predicted genes were used to cluster the 1,279 archaeal viral contigs with the prokaryotic viral Refseq (v201) using vConTACT (v.2.0)[33] with default parameters (The Refseq were supplied by the built-in database of vConTACT2). Thus, we leveraged the taxonomic information provided by the viral Refseq to taxonomically classify the contigs in these VCs.

For example, if one contig in a VC is classified to the Caudovirales order, the rest contigs in this VC will also be assigned to the Caudovirales order.

Second, we used taxonomical informative profiles from the VOG database (http://vogdb.org) and eggNOG (v5.0) database[64] to find out viruses likely to be the members of the Caudovirales order. Specifically, we first picked out the VOGs with annotation containing the keywords (portal, terminase, spike, capsid, sheath, tail, base plate, fiber, and tape measure) and named them as Hallmark VOGs. Then the predicted proteins from the archaeal viral contigs were compared to the VOG HMM profiles and the eggNOG database using hmmsearch v3.2.1 and eggNOG-mapper v.2.0.0[65] respectively. During this process, the minimum score and maximum E-value were set to 40 and 1e-5. If the viral contig encoding genes with hits against the Hallmark VOGs or eggNOGs whose annotation contains the keywords mentioned above, then this contig will be classified into the Caudovirales order (Table S6 and Fig. 2c for Order-level taxonomy).

Comparison of the viral species to other gut viral databases

A comparison between HGAVD and the viruses in publicly available databases derived from the gut microbiome was performed based on the following databases:

1. Metagenomic Gut Virus (MGV) catalogue[34], the newest gut virus collection, contains 189,680 viral draft genomes estimated to be $>$50% complete and representing 54,118 candidate viral species. The protein sequences of the representative archaeal viral contigs were used as queries in a BLAST search in the MGV database with a threshold of e-value $\leq$1e-3. Only the sequences in the MGV database encoding at least one protein sequence with the hits to those of the archaeal viral contigs were retained for network analysis. (11,827/189,680 = 0.06).
2. Proviruses detected from 1,162 gut archaeal genomes. 118 proviruses were predicted by CheckV from the 557 archaeal genome contigs in UHGG with a quality assignment of medium quality (50–90% completeness) and high quality (> 90% completeness) or were complete. These proviruses were then clustered at 95% identity and 80% coverage, resulting in 85 nonredundant viral species. We future clustered the 85 proviruses with the viruses in HGAVD. Only the 37 proviruses sharing identity $\leq$95% with the 1279 viral contigs in HGAVD were considered for further analysis.
3. The Prokaryotic Viral Refseq (V201) Database supplied by vConTACT2.

Estimation of the relative abundance of viruses and hosts.

First, we mapped all reads of the metagenomic sequencing data to the identified archaeal contigs and archaeal viral contigs by the software Soap2[66], only the contigs with > 30% breadth of coverage were counted. Second, the number of the reads corresponding to each of the identified archaeal genome contigs and archaeal viral contigs was normalized by the total number of the reads of each sample; the normalized value thereby represents the relative abundance of the contig in the sample.

Statistical analyses

All statistical analyses were performed in R version 4.0.5. Based on Bray–Curtis dissimilarity matrices, which were calculated using the VEGAN function vegdist, principal coordinate analysis (PCoA) was performed using the pcoa function in the APE package, and ANOSIM using the VEGAN function anosim was performed to test the significance of dissimilarity between groups.

Virus-host prediction

Host-virus interactions were resolved by searching CRISPR spacer sequences in the hosts and the viral contigs. To accurately investigate the gut archaeal viruses that have a broad host range, we particularly predicted CRISPR spacers from the 1,162 archaeal genomes in the UHGG database[23] based on the following criteria: (i) CRISPR arrays were identified on the archaeal genomes longer than 10kb using CRT[57]; (ii) To minimize spurious predictions, we dropped arrays with fewer than three spacers; (iii) CRISPR spacers were longer than 25 bp. The retained CRISPR spacers were aligned with the archaeal viral contigs using BLASTn to identify spacers present in the viral contigs, and matches satisfying the thresholds of 100% identity were selected (settings: -task blastn-short, - gapopen 10, -gapextend 2, -penalty 1, -word_size 7 -perc_identity 100).

Phylogenetic tree analysis of genes

To construct the phylogenetic trees for large terminase subunit, TerI, PeiW and MazE-antitoxin, amino acid sequences were aligned using the MUSCLE algorithm[67] included in MEGA X[68]. The maximum-likelihood phylogenetic tree was constructed using Fasttree v2.1.11[69] with the substitution model WAG. The final consensus tree was visualized and beautified in iTOL[56].

Declarations

Acknowledgments

This work received support from Guangdong Provincial Key Laboratory of Synthetic Genomics (2019B030301006); Shenzhen Key Laboratory of Synthetic Genomics (ZDSYS201802061806209); the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB29050501); Shenzhen Institute of Synthetic Biology Scientific Research Program (Grant no. JCHZ20200001)

Contributions

Y.M., Y.W., R.L. designed the study. R.L. and Y.W. performed the metagenomic analysis. H.H and Y.T. provided suggestions. Y.M. R.L and Y.W. contributed to the scientific discussion and preparation of the manuscript.

Corresponding authors

Correspondence to Yingfei Ma.

Competing interests

The authors declare no competing interests.

Data availability

https://github.com/SIAT-MaLab/Archaea-/releases

References

1. Borrel, G., Brugère, J. F., Gribaldo, S., Schmitz, R. A. & Moissl-Eichinger, C. The host-associated archaeome. Nature Reviews Microbiology (2020) doi:10.1038/s41579-020-0407-y.
2. Coker, O. O., Kai Wu, W. K., Wong, S. H., Sung, J. JY. & Yu, J. Altered Gut Archaea Composition and Interaction with Bacteria are Associated with Colorectal Cancer. Gastroenterology (2020) doi:10.1053/j.gastro.2020.06.042.
3. Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. Nature Microbiology **6**, (2021).
4. DeLong, E. F. Exploring Marine Planktonic Archaea: Then and Now. Frontiers in Microbiology **11**, (2021).
5. Koskinen, K. *et al.* First insights into the diverse human archaeome: Specific detection of Archaea in the gastrointestinal tract, lung, and nose and on skin. mBio (2017) doi:10.1128/mBio.00824-17.
6. Kumpitsch, C. *et al.* Methane emission of humans is explained by dietary habits, host genetics, local formate availability and a uniform archaeome. *bioRxiv* (2020) doi:10.1101/2020.12.21.423794.
7. Dridi, B., Henry, M., el Khéchine, A., Raoult, D. & Drancourt, M. High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS ONE **4**, (2009).
8. Ruaud, A. *et al.* Syntrophy via interspecies H2 transfer between christensenella and methanobrevibacter underlies their global cooccurrence in the human gut. mBio **11**, (2020).
9. Borrel, G. *et al.* Genomics and metagenomics of trimethylamine-utilizing Archaea in the human gut microbiome. ISME Journal (2017) doi:10.1038/ismej.2017.72.
10. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host and Microbe **26**, (2019).
11. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. Cell Host and Microbe **26**, (2019).
12. Chibani, C. M. *et al.* A catalogue of 1,167 genomes from the human gut archaeome. Nature Microbiology **7**, (2022).
13. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. Cell Host and Microbe (2020) doi:10.1016/j.chom.2020.08.003.
14. Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. v. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Research **244**, 181–193 (2018).

15. Wirth, J. & Young, M. The intriguing world of archaeal viruses. PLoS Pathogens **16**, 5–9 (2020).
16. Prangishvili, D. *et al.* The enigmatic archaeal virosphere. Nature Reviews Microbiology (2017) doi:10.1038/nrmicro.2017.125.
17. Díez-Villaseñor, C. & Rodriguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nature Communications (2019) doi:10.1038/s41467-018-08167-w.
18. Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annual Review of Biochemistry vol. 82 (2013).
19. Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. Nucleic Acids Research **49**, (2021).
20. Dellas, N., Snyder, J. C., Bolduc, B. & Young, M. J. Archaeal viruses: Diversity, replication, and structure. Annual Review of Virology **1**, (2014).
21. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews **40**, (2016).
22. Rahlff, J. *et al.* Lytic archaeal viruses infect abundant primary producers in Earth's crust. Nature Communications **12**, (2021).
23. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology (2021) doi:10.1038/s41587-020-0603-3.
24. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology **38**, (2020).
25. Roux, S. *et al.* IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Research (2021) doi:10.1093/nar/gkaa946.
26. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. Cell (2021) doi:10.1016/j.cell.2021.01.029.
27. Paez-Espino, D. *et al.* Uncovering Earth's virome. Nature (2016) doi:10.1038/nature19094.
28. Coutinho, F. H., Edwards, R. A. & Rodríguez-Valera, F. Charting the diversity of uncultured viruses of Archaea and Bacteria. BMC Biology **17**, (2019).
29. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology (2021) doi:10.1038/s41587-020-00774-7.
30. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: Mining viral signal from microbial genomic data. PeerJ **2015**, 1–20 (2015).
31. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome **5**, 69 (2017).
32. Pradier, L., Tissot, T., Fiston-Lavier, A. S. & Bedhomme, S. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. BMC Bioinformatics **22**, (2021).
33. bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nature Biotechnology (2019) doi:10.1038/s41587-019-0100-8.

34. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nature Microbiology (2021) doi:10.1038/s41564-021-00928-6.
35. Varsani, A. & Krupovic, M. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Archives of Virology **163**, (2018).
36. Kala, S. *et al.* HNH proteins are a widespread component of phage DNA packaging machines. Proc Natl Acad Sci U S A **111**, (2014).
37. Luo, Y., Pfister, P., Leisinger, T. & Wasserfallen, A. Pseudomurein endoisopeptidases PeiW and PeiP, two moderately related members of a novel family of proteases produced in Methanothermobacter strains. FEMS Microbiology Letters **208**, (2002).
38. Hazan, R. & Engelberg-Kulka, H. Escherichia coli mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. Molecular Genetics and Genomics **272**, (2004).
39. Otsuka, Y. Prokaryotic toxin–antitoxin systems: novel regulations of the toxins. Current Genetics vol. 62 (2016).
40. Chen, B. *et al.* ORF4 of the Temperate Archaeal Virus SNJ1 Governs the Lysis-Lysogeny Switch and Superinfection Immunity. Journal of Virology **94**, (2020).
41. Canchaya, C., Fournous, G. & Brüssow, H. The impact of prophages on bacterial chromosomes. Molecular Microbiology vol. 53 (2004).
42. Rambo, I. M., Anda, V. de, Langwig, M. v & Baker, B. J. Unique viruses that infect Archaea related to eukaryotes. bioRxiv (2021).
43. Roux, S. *et al.* Minimum information about an uncultivated virus genome (MIUVIG). Nature Biotechnology (2019) doi:10.1038/nbt.4306.
44. Wooley, J. C. & Ye, Y. Metagenomics: Facts and artifacts, and computational challenges. Journal of Computer Science and Technology **25**, (2010).
45. Jian, H. *et al.* Diversity and distribution of viruses inhabiting the deepest ocean on Earth. ISME Journal **15**, (2021).
46. Li, Z. *et al.* Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. ISME Journal **15**, (2021).
47. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology (2012) doi:10.1089/cmb.2012.0021.
48. Turnbaugh, P. J. *et al.* The Human Microbiome Project. Nature **449**, 804–810 (2007).
49. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics (2010) doi:10.1186/1471-2105-11-119.
50. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nature Methods (2014) doi:10.1038/nmeth.3176.
51. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology **2**, 1533–1542 (2017).

52. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics (2012) doi:10.1093/bioinformatics/bts565.
53. Johnson, M. *et al.* NCBI BLAST: a better web interface. Nucleic Acids Res (2008) doi:10.1093/nar/gkn201.
54. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. Bioinformatics (2020) doi:10.1093/bioinformatics/btz848.
55. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**, (2014).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Research **49**, (2021).
57. Bland, C. *et al.* CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics (2007) doi:10.1186/1471-2105-8-209.
58. Iranzo, J., Koonin, E. v., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. Journal of Virology (2016) doi:10.1128/jvi.01622-16.
59. Wolf, S. *et al.* Characterization of the lytic archaeal virus Drs3 infecting Methanobacterium formicicum. Archives of Virology (2019) doi:10.1007/s00705-018-04120-w.
60. Vik, D. R. *et al.* Putative archaeal viruses from the mesopelagic ocean. PeerJ **2017**, 1–26 (2017).
61. Eddy, S. R. Accelerated profile HMM searches. PLoS Computational Biology **7**, (2011).
62. El-Gebali, S. *et al.* The Pfam protein families database in 2019. Nucleic Acids Research (2019) doi:10.1093/nar/gky995.
63. Bobay, L. M. & Ochman, H. Biological species in the viral world. Proc Natl Acad Sci U S A (2018) doi:10.1073/pnas.1717593115.
64. Huerta-Cepas, J. *et al.* EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research (2019) doi:10.1093/nar/gky1085.
65. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular Biology and Evolution (2017) doi:10.1093/molbev/msx148.
66. Li, R. *et al.* SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics (2009) doi:10.1093/bioinformatics/btp336.
67. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**, (2004).
68. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution (2018) doi:10.1093/molbev/msy096.
69. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE (2010) doi:10.1371/journal.pone.0009490.

Figure S10

Figure S10 is not available with this version.

Figures

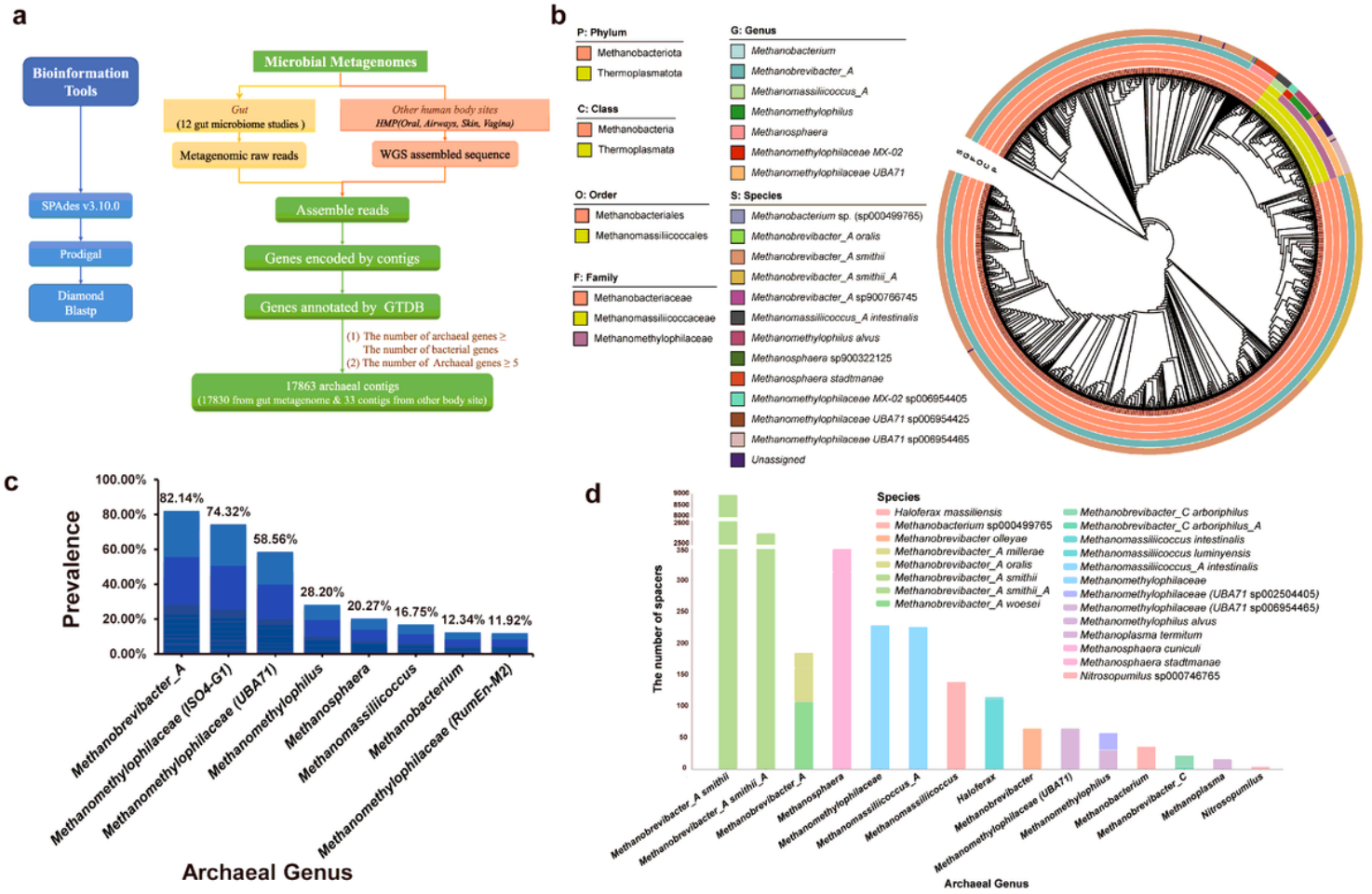

Figure 1

Identification of the human gut archaeal genomic contigs and recruitment of the CRISPR spacers.

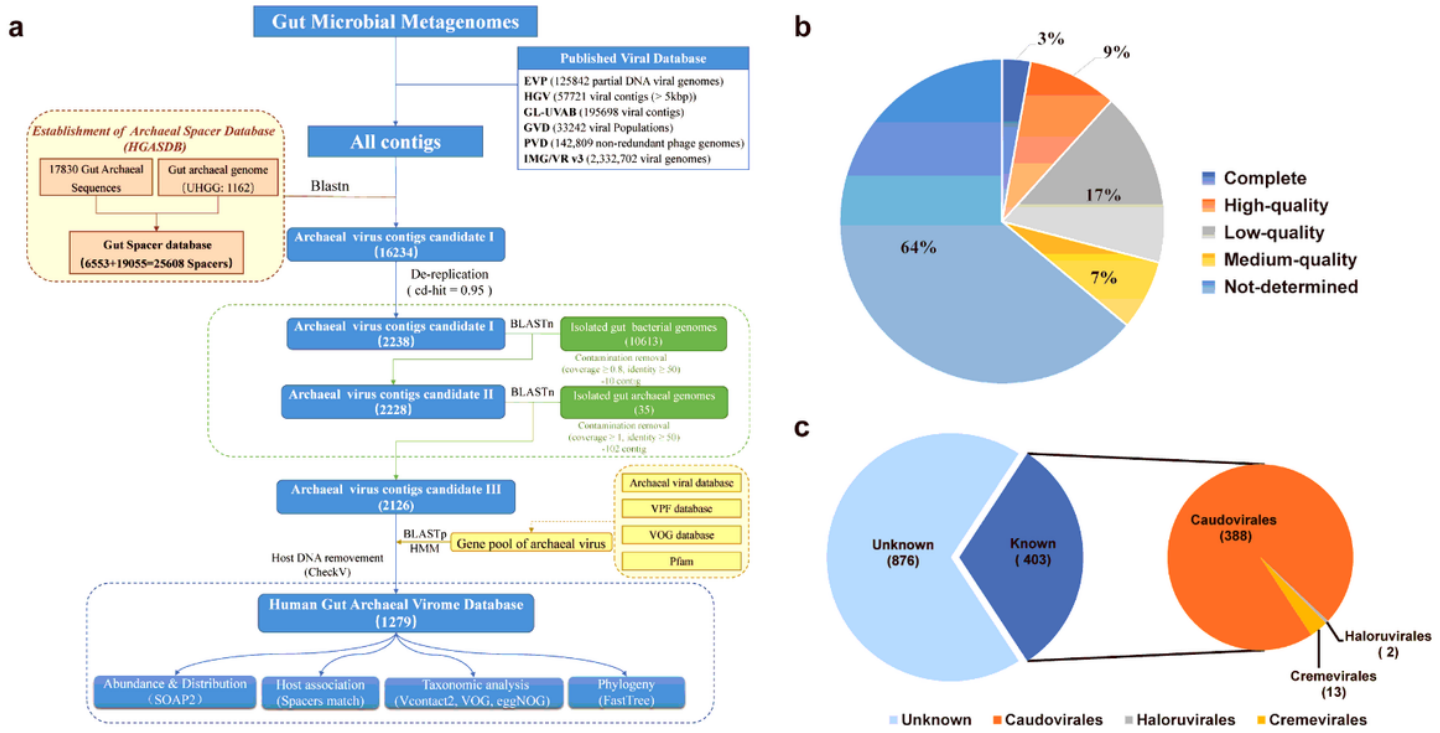

Figure 2

Identification of Archaeal viruses from the human gut.

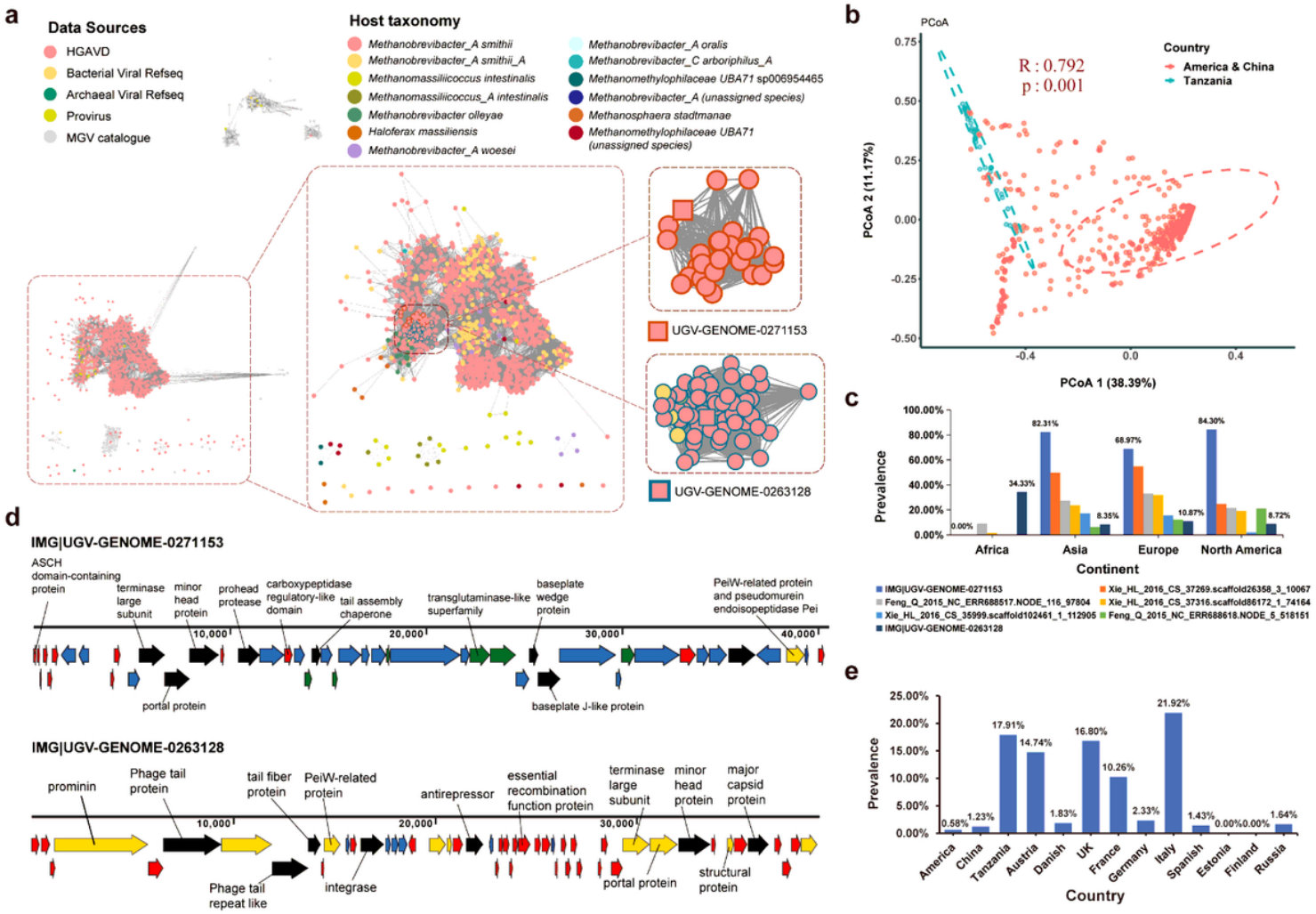

Figure 3

Protein clustering network and global distribution of the HGAVD viruses in the human gut.

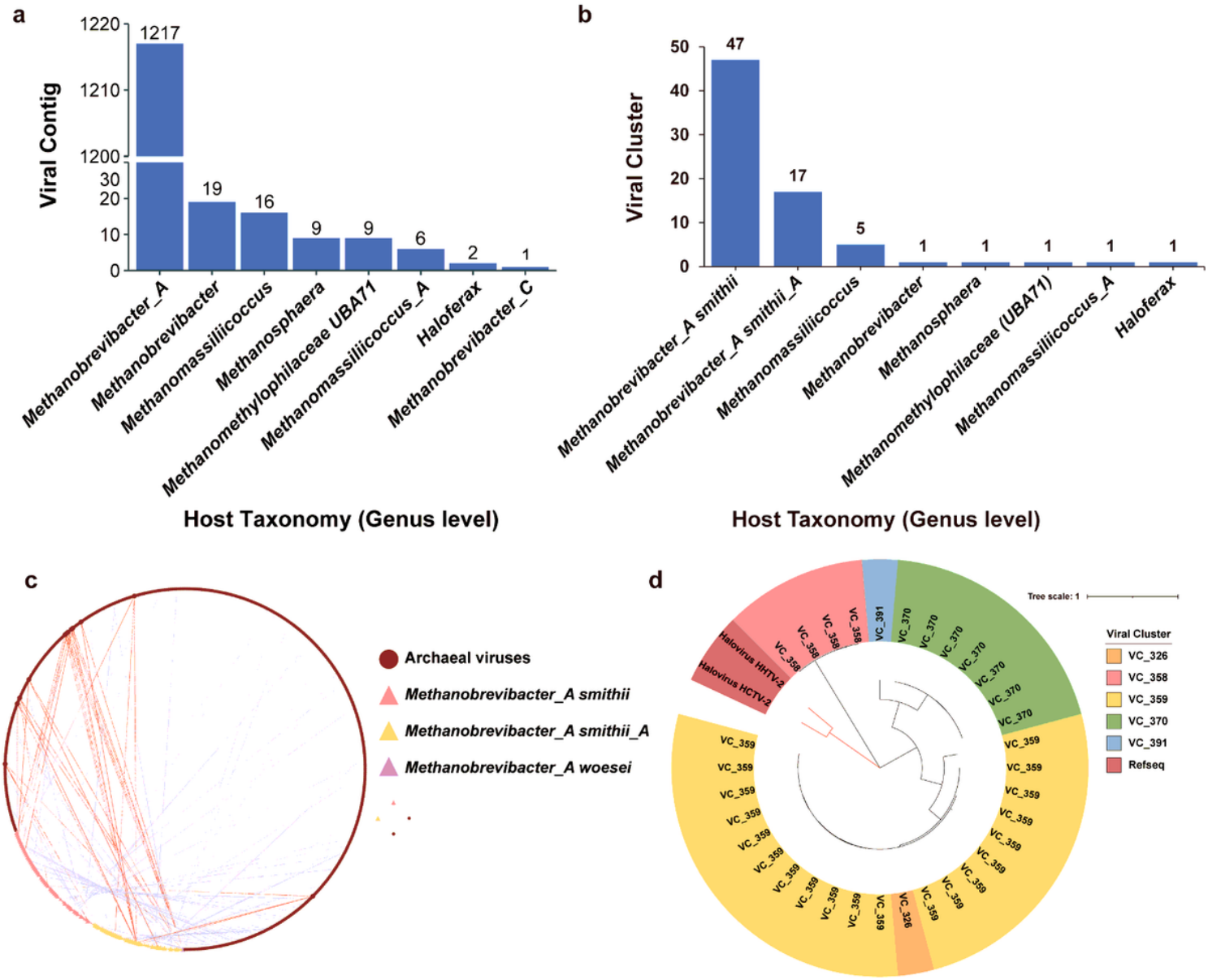

Figure 4

Archaeal viral host assignment and host range determination.

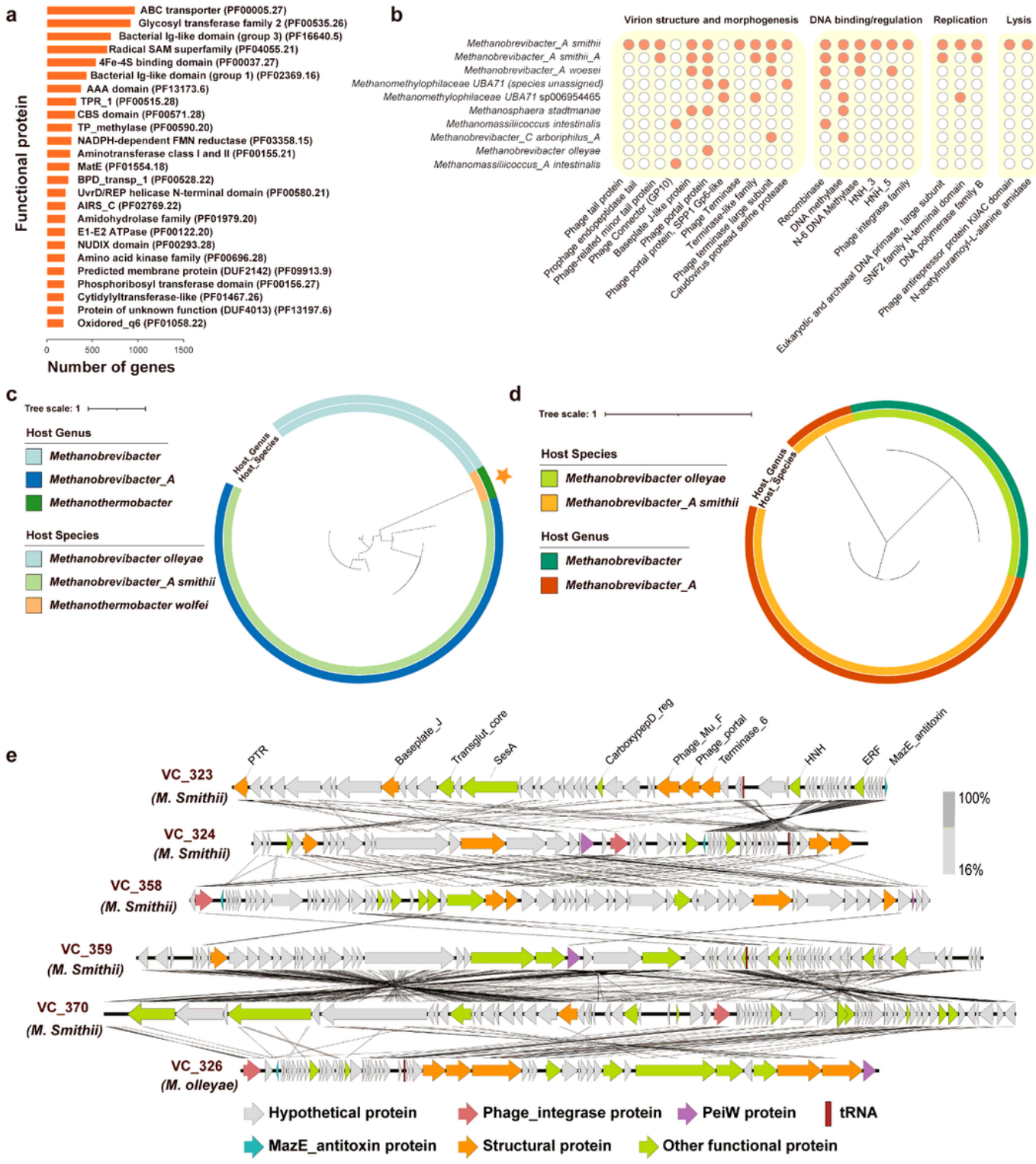

Figure 5

Functional landscape of the HGAVD viruses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx
- TableS11.xlsx
- TableS12.xlsx