Ultra-conserved bacteriophage genome sequence identified in 1300-year-old human paleofaeces

Authors:
Piotr Rozwalak[1], Jakub Barylski[2], Yasas Wijesekara[3], Bas E. Dutilh[4,5*], Andrzej Zielezinski[1*]

Affiliations:

1. Department of Computational Biology, Faculty of Biology, Adam Mickiewicz University, Poznan 61-614, Poland
2. Department of Molecular Virology, Faculty of Biology, Adam Mickiewicz University, Poznan 61-614, Poland
3. Institute for Bioinformatics, University of Medicine Greifswald, Felix-Hausdorff-Str. 8, 17475 Greifswald, Germany
4. Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, 07743 Jena, Germany
5. Theoretical Biology and Bioinformatics, Science4Life, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands

* Correspondence to: andrzej.zielezinski@amu.edu.pl and bedutilh@gmail.com

ORCIDs:

PR	0000-0002-4014-6624	piotr.rozwalak@amu.edu.pl
JB	0000-0001-6630-6932	jakub.barylski@amu.edu.pl
YW	0000-0002-9955-7735	yasas.wijesekara@uni-greifswald.de
BD	0000-0003-2329-7890	bedutilh@gmail.com
AZ	0000-0002-8096-3776	andrzej.zielezinski@amu.edu.pl

Keywords: bacteriophage, ancient DNA, ancient metagenomics

Running title: Bacteriophages in the ancient human gut

Abstract

Bacteriophages are widely recognised as rapidly evolving biological entities. However, in our exploration of ancient metagenomes, we discovered a genome nearly identical to present-day *Mushuvirus mushu*, a phage infecting commensal microorganisms in the human gut ecosystem. The DNA damage patterns of this genome have confirmed its ancient origin, and despite 1300 years of evolution, the ancient *Mushuvirus* genome shares 97.7% nucleotide identity with its modern counterpart indicating a long-term relationship between the prophage and its host. We also reconstructed and authenticated other 297 phage genomes from the last 5300 years, including those belonging to unknown families. In sum, our findings demonstrate the feasibility of reconstructing ancient phage genomes, expand the known virosphere, and offer new insights into phage-bacteria interactions covering several millennia.

Significance statement

Bacteriophages are viruses infecting bacteria with a huge impact on microbial communities. However, due to challenges in studying ancient DNA, little is known about their evolution from a long-time perspective. Here, we show the possibility of overcoming current limitations by reconstructing hundreds of fragments and complete bacteriophage genomes from the last few thousand years. We carefully confirmed their ancient origin and unexpectedly found an ultra-conserved genome of a virus living in at least a 1300-year-old relationship with its bacterial host in the human gut.

Introduction

Bacteriophages diversified over billions of years through a co-evolutionary arms race with their microbial hosts (1). Due to recent advancements in metagenomic sequencing (2) and computational analysis (3), the vast genomic diversity of phages can now be explored, catalogued (4) or even tracked through space (5) and time (6). However, most studies only sample present-day phages, thus lacking a long-term perspective on their evolution. The longest study of substitutions and recombinations in phage genomes spans almost three decades (7), which represents only a partial glimpse of the complex evolutionary history of bacteriophages. To fully understand the phylogeny of these viruses and their impact on microbial ecosystems, it is essential to go deeper in time to get a more comprehensive understanding of bacteriophage evolution.

Previous research on ancient virology has focused mainly on the reference-based reconstruction of human pathogens (8), retroviral elements incorporated in animal genomes (9) or eukaryotic viruses that have remained dormant since prehistory (10). However, knowledge about ancient bacteriophages is limited (11). The first reports on ancient phages described viruses from 14th-century faecal material in Belgium (12) and pre-Columbian Andean mummy's gut contents (13). These studies, however, lacked authentication based on terminal deamination patterns observed in ancient DNA (aDNA) (14), which can be used to distinguish ancient sequences from modern contaminants (15). In addition, the small amount of sequenced DNA in these investigations resulted in the recovery of only fragmented phage genomes. To our knowledge, only one complete oral phage genome has been published that meets the criteria of authenticated ancient DNA reconstruction (16).

While the recovery of hundreds of high-quality ancient bacterial genomes is now possible (17, 18), their phage counterparts remain largely unexplored. However, recent developments in computational tools for viral metagenomics (3, 4), combined with the extraction of exceptionally well-preserved ancient DNA samples (17, 19–23) offer a new opportunity to unlock the potential of phage genome diversity from the past.

Here, we analyse draft or complete genomes assembled from 150–5300 years old paleofaeces and human gut-content samples (17, 19–23). We use this collection to (i) identify 298 ancient phage genomes, (ii) authenticate their ancient origin based on DNA damage patterns, (iii) determine their taxonomic assignments and relationships to modern viruses, (iv) predict hosts using state-of-the-art bioinformatic tools, and finally (v) characterise the particularly stable genome of one encountered virus, *Mushuvirus mushu* that is 97.7% identical to its present-day reference. Together, these results demonstrate that the large-scale recovery of virus genomes from ancient samples is possible and may provide unexpected insights into the evolutionary history of the human virome.

Results

Identification and validation of ancient phage genomes

We selected aDNA sequence datasets from 30 samples previously published in studies about the ancestral human gut microbiome (17, 19–23). These samples were derived from eight sites in Europe and North America, dating back between 150 and 5300 years ago using the C14 method (Table S1). *De novo* assembly of selected libraries resulted in 72,693 high-quality contigs (mean length: 12,592 nucleotides ± 59.3 standard error; Fig. 1A). However, we deemed less than half of assembled contigs as ancient based on deamination patterns observed at the 5' ends of sequencing reads (see Methods). A total of 2,375 sequences were classified as viral by at least two methods (VIBRANT (24), VirSorter2 (25) Jaeger; Table S2) and were selected as *bona fide* virus contigs. Among those, 383 were at least 20 kb long or assessed by CheckV as medium or better quality (>50% completeness; Fig. 1B and Table S3) and were selected for further analysis. Finally, we clustered selected sequences into 298 species-level viral operational taxonomic units (hereafter referred to as aMGVs - ancient metagenomic gut viruses) on the basis of 95% average nucleotide identity (ANI) over 85% alignment fraction (26). Despite fragmentation and degradation of aDNA, our aMGV collection has a size range from 5 kbp to 292 kbp, an estimated mean completeness of 50.3%, and comprises 49 high-quality or complete genomes (Table S3). Although the mean damage at the first position of reads mapping to collected aMGVs was relatively low (0.025 frequency), it was 12 times higher than the control modern viral genomes from the human gut (0.002 frequency) (Table S3; Fig. 1C, D).

Human gut origin of ancient phage genomes

The use of deamination damage patterns is a well-established technique for ancient DNA validation, but the complexity of ancient metagenomic data may necessitate the implementation of additional authentication steps. In our case, we decided to rule out post-depositional contamination with DNA from the environment (e.g., cave sediments), which may also exhibit damage due to the same degradational processes as the aDNA of interest (27). To estimate the risk of such contamination we constructed a gene-sharing network using

vContact2 (28) to phylogenetically relate ancient and modern phages (Fig. 2A), including reference viruses from different environments in the IMG/VR database (29) and prokaryotic viruses classified by the International Committee of Virus Taxonomy (ICTV) (30). In the network (Table S4), 151 ancient phages were distributed across 122 viral clusters (VCs). These VCs were dominated by mammalian-gut-associated viruses from IMG/VR (80%, see Fig. 2B) with a high representation (68%) of phages found in humans. We recognized that 49% of aMGVs (n = 147) did not cluster with any modern viruses, suggesting they represent potential novel viruses or incomplete genomic fragments with insufficient phylogenetic information. The estimated contamination is relatively low; only 5% of aMGVs (n = 15) clustered with viruses from environmental or engineered ecosystems.

Further evidence supporting the ancient gut or paleofecal origin of 298 analysed aMGVs comes from the distribution of their predicted hosts (see Methods, Fig. S1 and Table S5). Over half of the ancient viruses were predicted to infect *Clostridia* and *Bacteroidia* hosts, two dominant classes in the gut microbiome (31). The distribution of host bacteria assigned to ancient viruses is more similar to that of modern viruses from the digestive systems of different animals than viruses of other environments. For example, we observed a strong correlation (Pearson's $r = 0.95$, $P = 3.29 \times 10^{-6}$) between the proportion of host classes of ancient phages and hosts of modern viruses in human large intestines (Table S6). To sum up, both the results of the host prediction and the gene-sharing network analysis suggest that the ancient phage sequences primarily originate from the human gut rather than the surrounding environment.

Catalogue of the human gut virome expands into the past

We performed geNomad's (32) marker-based taxonomic assignment for all 298 ancient viral sequences, 293 of which (~98%) were assigned to the *Caudoviricetes* class, three to the *Megaviricetes* class (nucleo-cytoplasmic large DNA viruses - NCLDV), and two single-strand viruses to the *Cressdnaviricota* and *Phixviricota* phyla (Table S7). Most aMGVs from the *Caudoviricetes* class were not classified at lower taxonomic ranks, except nine that were assigned to the order *Crassvirales* and the families *Autographiviridae*, *Straboviridae*, *Herelleviridae*, and *Rountreeviridae*. To improve the taxonomic resolution, we selected 49 high-quality or complete ancient genomes and performed a clustering with RefSeq and IMG/VR virus genomes based on pairwise average amino acid identity and gene sharing, followed by manual assignment to the ICTV taxonomy (see Methods). Additionally, we constructed a proteomic tree to support the clustering results and illustrate the relationships between modern and ancient viruses (Fig. 3). This analysis revealed that high-quality or complete aMGVs were distributed across 39 putative families and 46 putative genera (Table S8).

Most of the putative families (90%, 35 out of 39) and genera (72%, 33 out of 46) contained both aMGVs and modern representatives from IMG/VR and/or RefSeq. Only a small proportion of those groups have been classified by ICTV, including six families (15%, *Straboviridae*, *Peduoviridae*, *Casjensviridae*, *Microviridae*, *Chaseviridae*, and *Guelinviridae*), and eight genera (17%, *Tequatrovirus*, *Plaisancevirus*, *Pbunavirus, Astrithrvirus*, *Brucesealvirus*, *Loughboroughvirus*, and *Mushuvirus*). Most of these putative taxa included a single ancient phage grouped with multiple contemporary viruses. An interesting example of such a grouping is a clade of *Escherichia*-infecting members of the *Tequatrovirus* genus (including bacteriophage T4, Fig. S2) with an ancient virus located on the ancestral branch.

Only a few putative taxa contain multiple ancient representatives. For example, two genus-like groups ("13", "14" see Table S8) form a family-like unit that encompasses three ancient phages, 36 IMG/VR viruses and *Salmonella*-infecting phage *Astrithrvirus astrithr*.

In contrast to groups with multiple representatives, there are four putative families and 13 genera represented by only a single aMGV. While these phages might hypothetically represent extinct lineages, it is likely that modern representatives of those groups have not yet been discovered. Although it is beyond the scope of the current study, one way to address this question could be to identify conserved marker regions in the ancient genomes and attempt to amplify them in locations where many modern human gut phages come together, such as sewage systems (5). Notably, our taxonomic analysis of high-quality ancient genomes was limited by the database-dependent tool CheckV, which has a bias for phages that are similar to known reference genomes. Consequently, the other 249 aMGVs can still include viral genomes that are unlike any sequence characterised up to date. Despite this limitation, we showed that our approach is promising for discovering previously unknown viruses and shedding light on "dark matter" lingering in ancient metagenomes.

Persistence of *Mushuvirus mushu*

Among the collection of diverse ancient phages, we found the genome sequence of a bacteriophage from the *Mushuvirus mushu* species that was recently discovered in a prophage analysis of its host *Faecalibacterium prausnitzii* (33). In that study, the prophage could not be induced after DNA damage, but virions were observed in the gastrointestinal tract in a mouse model (33). We are convinced that our aMGV represents an ancient sequence because the frequency of C → T substitutions at the first position of *Mushuvirus mushu* reads was high (0.042) compared to other sequences in our collection of ancient viruses (mean: 0.025; see Figure S3 and Table S3). Moreover, the *F. prausnitzii* host of *Mushuvirus mushu* is a strict anaerobe that typically inhabits the mucosal surface in the gastrointestinal tract (34), so we do not expect that it could have contaminated the original paleofaeces sample (21).

Recovered from 1300-year-old faecal material from Zape cave in Mexico, the ancient *Mushuvirus mushu* genome was 97.7% identical to the modern reference genome FP_Mushu (RefSeq accession: NC_047913.1) that was extracted from wastewater in France (33) (Fig. 4A). The mean sequence identity to ten IMG/VR genomes from the same species that were derived from the large intestine was 97.1% ± 0.6% (Fig. 4c; Table S9). The ancient and modern genome sequences had similar lengths (36,623 bp and 36,636 bp, respectively) and perfect colinearity between 52 protein-coding genes (Fig. 4B). Phage mutation rates reported in the literature range from 1.976×10^{-4} to 4.690×10^{-3} nucleotide alterations per site per year due to substitutions and recombinations for virulent phages (7, 35), and 1.154×10^{-4} substitutions per site per year for temperate phages (6). As can be observed in Fig. 4D, the probability of finding a 36 kbp long genome sequence with at least 97.7% identity approaches zero after a little over 200 years, even at the lowest mutation rates reported in the literature (6, 7, 35).

A total of 22,059 sequencing reads were mapped to the ancient *Mushuvirus mushu* genome, resulting in 28x genome coverage. Our comparison of the ancient and modern genomes revealed 869 single nucleotide variants (SNVs). The distribution of these SNVs was not homogeneous but concentrated in the specific section of the gene encoding the Hoc-like

capsid decoration protein (Fig. 4B). This fragment is the target of diversity-generating retroelement (DGR) that produces a large number of localized mutations through error-prone reverse transcription and is involved in host switching in some phages (36–38). Indeed, Hoc-like gene had a lower level of amino acid identity to the reference FP_Mushu (91.8%) than other protein-coding genes (mean identity: 98.22%) and accounted for 16% of all SNVs in the genome. Notably, according to recently published RNA-seq data from cultured strains of *Faecalibacterium prausnitzii* the Hoc-like gene is expressed at levels similar to the average bacterial gene under most studied conditions (39) (Fig. 4E). This finding suggests that this protein might have a role in lysogenized bacterial cells. Despite the potential significance of the Hoc-like gene in the evolution of *Mushuvirus mushu*, its function remains poorly understood, with only homologs of the gene being characterized as enhancing binding to mucus on the mucosal surface of the intestine and potentially increasing encounters with mucosa-dwelling hosts (40, 41).

To analyse whether our ancient *Mushuvirus mushu* sequence represented an integrated prophage, we investigated the flanking regions of the contig (see Methods). These regions differ not only between the ancient virus and modern ones but also among different modern phages, indicating a potential bacterial origin. The sequence similarity of the flanking regions of ancient *Mushuvirus mushu* points to *Oscillospiraceae* genus ER4, suggesting this may be a host for the ancient *Mushuvirus mushu* (Table S11). *Mushuvirus mushu* seems to have a broad host range (33). Indeed, when comparing the ancient genome to bacterial sequences from the Genome Taxonomy Database (GTDB) (42) and the Unified Human Gastrointestinal Genome (UHGG) collection (43), we found prophage genomes with 97% sequence identity in seven different *Oscillospiraceae* genera and *Faecalibacterium* genomes from the *Ruminococcaceae* family (see Fig. 4F; Table S10).

Discussion

In this study, we demonstrated the reconstruction of high-quality ancient phage genomes on a large scale using state-of-the-art bioinformatic methods. To validate the authenticity of the reconstructed ancient genomes, we analysed their DNA damage patterns, relationships with contemporary viruses, and host associations. We could not find any significant similarity to known viruses for approximately half of the reconstructed genomes. This indicates that our understanding of the human gut virome history is still limited. Nevertheless, advancements in viral metagenomics and access to well-preserved ancient DNA samples hold promise for expanding our view on virosphere evolution.

Bacteriophage genomes are often thought to be variable, shaped by rampant mutations, recombinations, and horizontal gene transfers (1, 7). Therefore, recent reports of near-identical phage genomes found over vast geographical distances (44) and spanning several years in the same location (45) were striking. Here, we reconstructed a high-quality ancient genome sequence of *Mushuvirus mushu* that was highly conserved despite at least 1300 years of evolution and presence on different continents. Integrated as a prophage in its bacterial host, *Mushuvirus mushu* is able to generate nearly identical virions over vast periods of time, with very low mutation rates. The observed conservation of the sequence might be related to the relatively stable environment of the human gastrointestinal tract, temperate lifestyle, and potential mutual relationships with a broad range of hosts like *Faecalibacterium*

prausnitzii which is widely present in the human gut across populations ranging from hunter-gatherers to industrialized societies (46).

In summary, our study highlights the utility of publicly accessible ancient metagenomes in the investigation of viruses associated with microorganisms. Our focus was on sequences from the well-preserved gut and paleofeces samples, however, much of the currently available data comes from ancient teeth or dental plaque and these samples await further exploration. In the future, we anticipate that similar studies of virome in ancient metagenomes will contribute to elucidating the complex history of viruses and their role in co-evolution with bacterial, animal and plant hosts.

Methods

Sample selection

Based on community-curated metadata from AncientMetagenomeDir (47), we selected 72 ancient DNA metagenomic libraries from 30 well-preserved human faeces or digestive gut contents. Data from palaeofeces originated mainly from archaeological excavations in caves located in Boomerang Shelter, Arid West Cave or Zape in the USA and Mexico (*n* = 38) as well as the underground salt mines of the Hallstatt in Austria (*n* = 19). Moreover, we used the digestive contents material (*n* = 15) from a biopsy of the Tyrolean Iceman mummy discovered more than 30 years ago in a melting glacier (19). The studied samples represented material from 150 to 5300 years ago. Details about original publications, localizations (coordinates), and samples (sequencing depth, instruments) are described in supplementary materials (Table S1).

Sequence data preprocessing

Pair-ended Illumina reads were trimmed using Cutadapt v.4.1 (48) with quality cutoff = 25, minimum read-length = 30, and a minimum overlap with adapter sequence to reads = 1. A total of 2,628,045,312 clean reads from all 72 libraries were mapped to *Homo sapiens* reference genome (hg37) using KneadData v.0.12.0 (with --bypass-trim option) in order to filter out human DNA (https://github.com/biobakery/kneaddata). Finally, the quality of 2,352,455,887 remaining reads after preprocessing was controlled in Fastqc.

De novo contig assembly

Filtered reads from each library were assembled into contigs using Metaspades v.3.15.5 (49) with default settings. Only contigs longer than 4000 nucleotides with a minimum coverage of 20 were considered in downstream analyses (see Fig. 1).

aDNA authentication

After preprocessing, clean reads were mapped to the assembled contigs using Bowtie2 v2.4.4 (50) with default settings. The resulting alignment was sorted and indexed with SAMtools v.1.14 (51). DNA damage patterns observed in reads were used to label corresponding contigs as ancient or modern. The analysis was run using Pydamage v.0.70 (15) - a program that calculates the frequency of C to T transitions at the first 20 positions of mapped reads

compared to a reference sequence. The filtering threshold for predicted accuracy was determined by the Kneedle method (52) and we imposed an additional cut-off of transition frequency (minimum 0.01 at the first position) to filter out contigs with weakly damaged reads that could confuse the program. Additionally, the same analytical process of authentication was performed for 943 randomly selected contigs from Metagenomic Gut Viruses (MGV) database (31) to compare deamination patterns between modern and ancient viral genomes (see Fig. 1C, D).

Viral contigs identification

Three machine learning tools were used to identify viral contigs. The first one was Jaeger 1.1.0, a deep-learning model that identifies phage genome sequences in metagenomes (https://github.com/Yasas1994/Jaeger) based on automatic compositional feature extraction. The second and third - were VIBRANT v.1.2.1 (24) and VirSorter2 v.2.2.3 (25), Both these tools rely on the analysis of HMM profiles representing conserved families and/or domains similar to predicted proteins but apply different classifiers and reference databases. Jaeger and Vibrant were run with default settings, for VirSorter2 we used positional arguments '--include-groups dsDNAphage,NCLDV,ssDNA,lavidaviridae all'. Contigs classified as viral by at least two tools were passed to further analysis steps.

Genomes quality assessment

CheckV v.1.0.1 (53) was applied to assess the genome quality of ancient viral contigs using the end_to_end command. Ancient viral genomes classified as complete, high-quality, medium-quality, or fragments longer than 20kb and with at least one viral gene were considered in the next steps.

Viral contigs clustering

Ancient MGVs (*n* = 383) were clustered into 298 species-level vOTUs using scripts, published in the CheckV repository (https://bitbucket.org/berkeleylab/checkv/src/master/). Briefly, sequences were grouped on the basis of 95% average nucleotide identity (ANI) and 85% alignment fraction of the shorter sequence, as recommended in MIUViG (Minimum information about an uncultivated virus genome) (26).

Gene-sharing network

To compare ancient phages with their modern counterparts from different environments, we selected from the IMG/VR (v.4 - high-confidence genomes only) 10 modern phage genomes for each aMGV. Specifically, we selected genomes with the highest number of shared proteins determined by Diamond v2.0.15 (54) search in blastp mode (--very-sensitive) with a minimum of 50% query coverage and 50% sequence identity. We then visualized this collection of aMGVs, selected modern viral genomes, and all prokaryotic DNA viruses with assigned ICTV taxonomy (VMR_20-190822_MSL37.2, created 08/31/2022) using vContact2 v.0.11.3 (28). The network was displayed in Cytoscape v.3.9.0 and refined in Inkscape v.1.2.2.

Host prediction

Four computational tools (BLASTn (55), PHIST (56), VirHostMatcher-Net (57), RaFAH (58)) were used to assign hosts to ancient phages and sequences from IMG/VR v.4 (only PHIST). PHIST v.1.1.0 and BLASTn v.2.13.0+ predictions were run against representative genomes of GTDB (42) database v.07-RS207 (62,291 bacterial species + 3,412 archaeal species). Prokaryotic species whose genomes obtained the highest similarity score to the virus genome, and have *e*-value < 10^{-5} (BLASTn) or *p*-value < 10^{-5} (PHIST), were assigned as a putative host. In methods integrating machine learning approaches like VirHostMatcher-Net or RaFAH, we selected host predictions with at minimum 0.5 and 0.14 scores, respectively.

Taxonomy assignment, clustering, and phylogenetic analysis

We performed the taxonomic assignment of viral genomes using 'annotate' function from the geNomad 1.3.3 tool (32). To classify aMGVs on genus- and family-level, we clustered viral genomes using a combination of gene sharing and amino acid identity (AAI). At first, we selected aMGVs assessed as high-quality or complete by CheckV, and we then added prokaryotic viruses from RefSeq (*n* = 4703; access: 30.01.2023) and IMG/VR sequences (*n* = 265) forming clusters (VC) with aMGVs (see: Methods, Gene-sharing-network). In this collection, pairwise protein sequence alignments were performed using DIAMOND v0.9.24.125 (54) with the options '-e-value 1 × 10−5–max-target-seqs 10000'. Then we calculated the percentage of shared genes and AAI between each pair of viruses. Following the criteria from the previous studies (31, 59), we kept connections between viruses with >20% AAI and >10% genes shared for the family level, and >50% AAI and >20% genes shared for the genus level. Finally, clustering was performed based on the connections between viral genomes using MCL with the option '-I 1.2' for the family level or '-I 2' for the genus level. All scripts to perform analyses at this step are available here (https://github.com/snayfach/MGV/blob/master/aai_cluster/). To visualise phylogenetic relationships of genus- and family-level groups, we generated a proteomic tree of 5017 viral sequences using ViPTreeGen v1.1.3 (60) and GraPhlAn v1.1.3 (61).

Analyses of *Mushuvirus mushu* genome

We performed a BLASTn search of 298 ancient metagenomic viral sequences against the nr/nt NCBI database. This search revealed that one identified aMGV (NODE_310_length_36983_cov_28.516681) is remarkably similar to present-day *Mushuvirus mushu* (NC_047913.1). This genome was present in a vContact2 cluster, along with one reference from NCBI (contemporary genome of *Mushuvirus mushu*) and 10 viral operational taxonomic units (vOTUs) from IMG/VR v.4 (high-confidence genomes only). All 12 sequences from the vContac2 cluster were aligned using MAFFT v7.308 to differentiate the core of the phage genome (~36,623 bp) from the flanking regions coming from the host integration sites (~241 bp and ~119 bp in the ancient contig). Only those core sequences were used in further analyses. To assess the nucleotide variability of modern and ancient *M. mushu* sequences, we calculated the intergenomic similarity within the cluster using the VIRIDIC (62) with the following parameters: '-word_size 7 -reward 2 -penalty -3 -gapopen 5 -gapextend 2'. The obtained similarity matrix, we used to construct a phylogenetic tree (Fig. 4D) using the bioNJ algorithm with default parameters (63). To annotate protein-coding genes in the analyzed genomes, we used end-to-end script (https://github.com/Yasas1994/phage_contig_annotator) that annotates phage genes based on HMMER 3.3.2 search against Prokaryotic Virus Remote Homologous Groups (PHROGs) (64). Visualisation and manual curation of the genome were

conducted in Geneious Prime 2023.04 (Fig. 4C). MAFFT plugin in the same tool was used to generate multiple sequence alignment of entire viral genomes. The identity of protein-coding genes at the amino acid level was calculated based on local alignment performed using EMBOSS v6.6.0.0 (65). To calculate SNVs in the ancient genome, a Python script was used (https://github.com/pinbo/msa2snp). We detected template and variable regions of diversity generating retroelements using myDGR (36). Finally, we assigned potential hosts by searching for sequences similar to prophage core and flanking regions (blastn -task megablast) in the Unified Human Gastrointestinal Genome (UHGG) collection (43) and Genome Taxonomy Database (GTDB) (42).

Data and Code Availability

Access numbers to raw data are available in Table S1. Bioinformatic scripts and a guide to data analysis performed in the article are provided in the GitHub repository (https://github.com/rozwalak/Ancient_gut_phages). Ancient phage genome sequences are available in Zenodo (https://doi.org/10.5281/zenodo.7919433).

Acknowledgements
The computations were performed at PLGrid Infrastructure and the Poznan Supercomputing and Networking Center (grant 312). This work was funded by the Polish Ministry of Science and Higher Education under the program “Perły Nauki”, project number PN/01/0063/2022, the total value of the project: 228,448PLN awarded to P.R., A.Z is supported by Polish National Science Centre [2018/31/D/NZ2/00108], B.E.D. is supported by the European Research Council (ERC) Consolidator grant 865694: DiversiPHI, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2051 – Project-ID 390713860, the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by German Federal Ministry of Education and Research, and the European Union’s Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF).

Author contributions:
P.R. conceived the idea of studying ancient phages; P.R., A.Z., J.B., B.E.D. designed research; P.R. and A.Z. selected samples and performed preprocessing, assembly and aDNA authentication; J.B., Y.W. performed identification of viral contigs; P.R., A.Z. created a gene-sharing network and performed host prediction; P.R., A.Z., J.B. performed taxonomy assignment, clustering, phylogenetic studies and analyses of *Mushuvirus mushu* genome; Y.W. contributed to phage genome annotation; P.R., A.Z., J.B., and B.E.D. analysed the results. P.R. wrote the manuscript with substantial contributions to the writing from A.Z., J.B. and B.E.D. All authors reviewed and approved the manuscript.

Competing interests: The authors declare no competing interests.

Figures

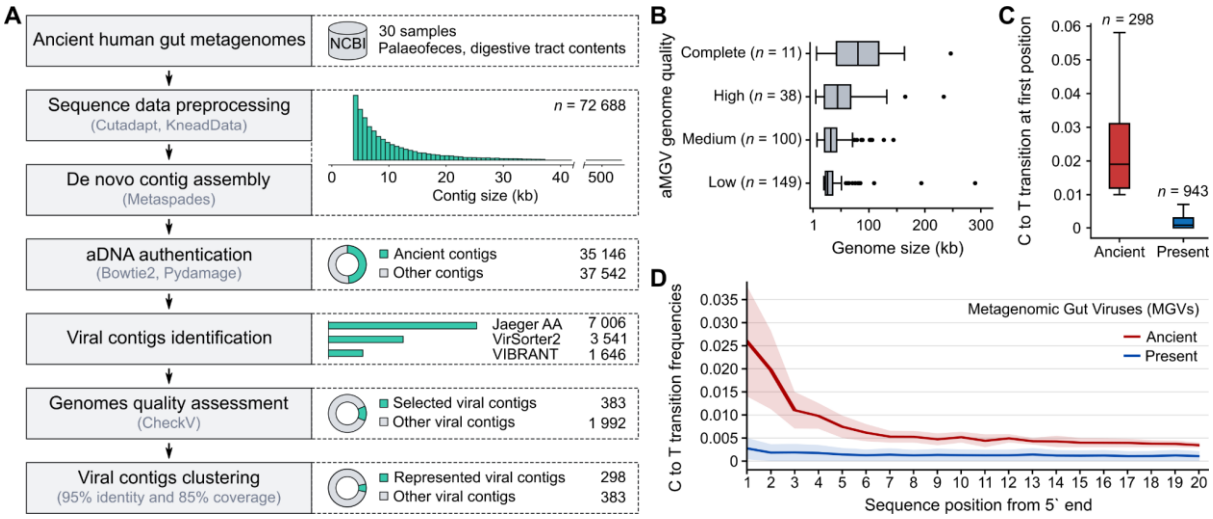


Fig. 1. Identification of phage genomes in ancient human gut metagenomes
(A) Overview of workflow to identify ancient phage genomes. **(B)** Distribution of genome sizes of 298 ancient metagenomic viruses (aMGVs) stratified by CheckV (53) genome quality. Box and whisker plots show median (center line), upper and lower quartiles (represented by box), and highest (upper whisker) or lowest (lower whisker) value that is within 1.5 inter-quartile range (IQR) while black dots indicate values outside of the IQR. **(C)** Distribution of C → T substitutions frequency at the first position of 5' end of sequencing reads. **(D)** Comparison of damage patterns between selected modern viral genomes from MGV (31) and aMGVs. The solid line shows the mean frequency of C → T substitutions and the shade indicates the standard deviation.

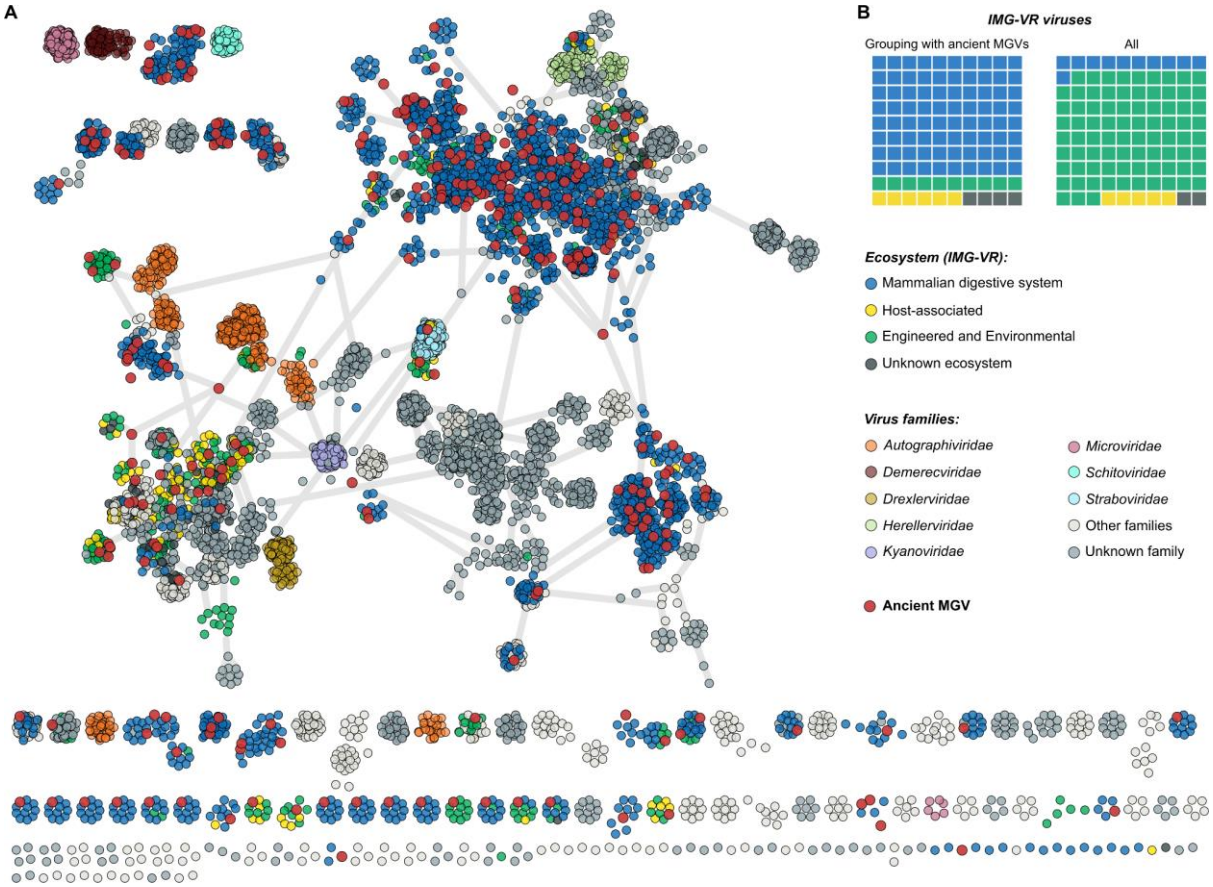


Fig. 2. Ancient phage genomes from paleofeces are related to the currently known mammalian-gut-associated viruses.

(A) vContact2 gene-sharing network (28) of 151 ancient metagenomic gut virus genomes (aMGVs), 2198 selected close relatives from the IMG/VR database, and 3655 prokaryotic viruses classified by the International Committee of Virus Taxonomy (ICTV). Distantly related aMGVs (*n* = 147) are outliers, not included in the gene-sharing network. **(B)** Waffle charts representing the proportion of contemporary viruses categorized by ecosystems in two data sets: clusters comprising aMGVs (left) and the entire IMG/VR database.

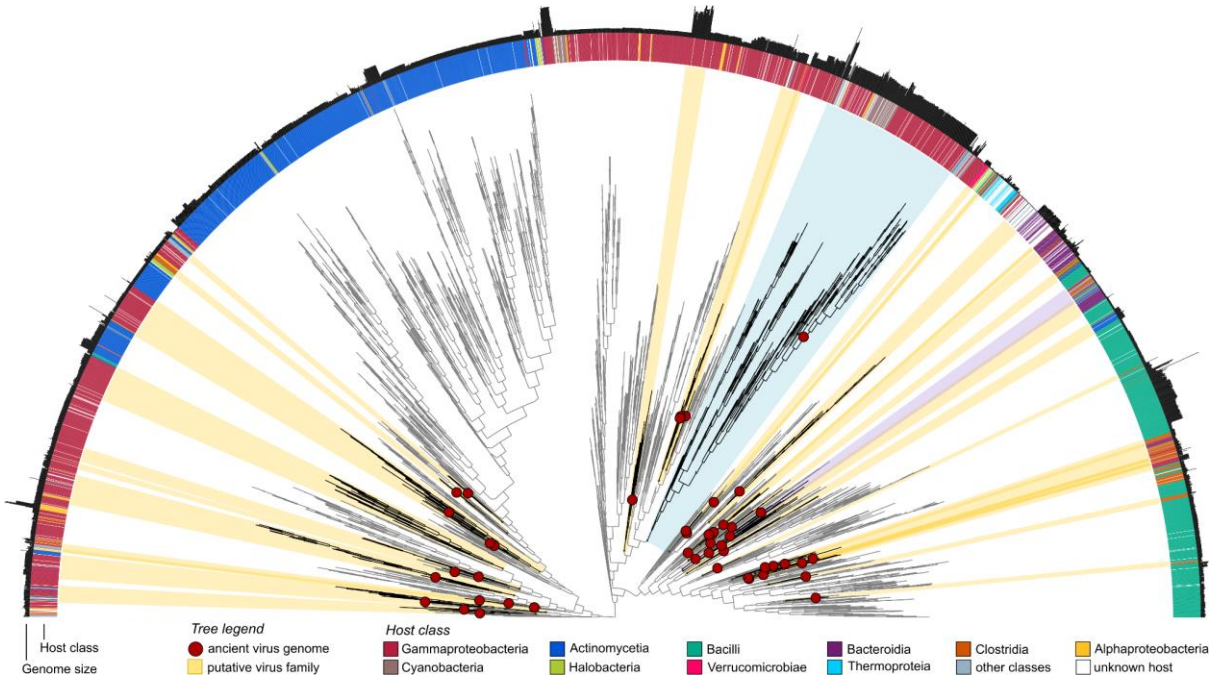
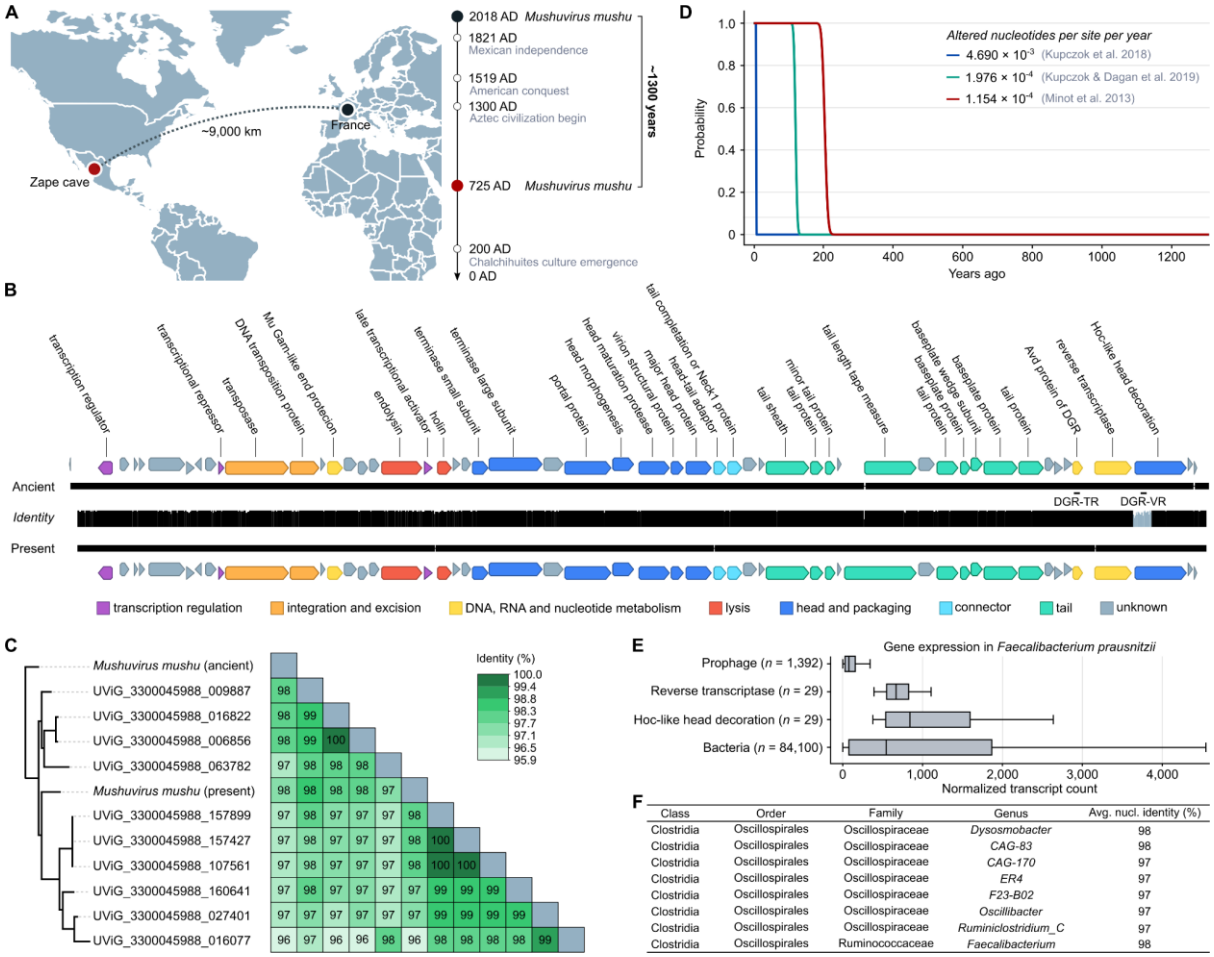


Fig. 3. Proteomic tree of ancient and present phage genomes.
The tree was generated in ViPTreeGen (60) and encompasses 49 high-quality and complete ancient phages, their 265 closest relatives from IMG/VR (29), and all 4703 prokaryote-infecting viruses from RefSeq (60, 66). The ancient phages are identified in the tree with red circles. Putative families were defined using a percentage of shared genes and AAI between each pair of viruses. Clusters containing at least one ancient phage are highlighted in yellow and listed in Table S8. The purple and blue highlighted clades represent putative families of *Salmonella*-infecting phages and T4 bacteriophage-related genomes, respectively. Outer rings provide additional information for each phage, such as its host assignment and genome size.



Class	Order	Family	Genus	Avg. nucl. identity (%)
Clostridia	Oscillospirales	Oscillospiraceae	*Dysosmobacter*	98
Clostridia	Oscillospirales	Oscillospiraceae	*CAG-83*	98
Clostridia	Oscillospirales	Oscillospiraceae	*CAG-170*	97
Clostridia	Oscillospirales	Oscillospiraceae	*ER4*	97
Clostridia	Oscillospirales	Oscillospiraceae	*F23-B02*	97
Clostridia	Oscillospirales	Oscillospiraceae	*Oscillibacter*	97
Clostridia	Oscillospirales	Oscillospiraceae	*Ruminiclostridium_C*	97
Clostridia	Oscillospirales	Ruminococcaceae	*Faecalibacterium*	98

Fig. 4. Stability of the *Mushuvirus mushu* prophage genome

(A) Geographic and historical distance between modern reference (France) and ancient (Zape cave in Mexico) *Mushuvirus mushu* genomes. **(B)** Comparison of sequence conservation and gene organization between ancient and modern reference (NC_047913.1) genomes of *Mushuvirus mushu*. Black lines indicate alignment with small gaps. The height of the middle plot depends on the identity between genomes. The lowest similarities (grey color) are concentrated in a ~600-nucleotide variability region of the Hoc-like gene impacted by the diversity-generating retroelement (DGR) mechanism. Two major structural differences in the ancient and modern genomes constitute additional codon stop in the gene of unknown function (genome position: 9518) and 17 nucleotide insertion/deletion changing open reading frame in tail length tape measure protein **(C)** Intergenomic similarity matrix calculated by VIRIDIC (62) reveals high identity between ancient and modern representatives of species *Mushuvirus mushu* deposited in RefSeq and IMG/VR v.4. **(D)** Probability of a 36,630 nucleotide phage genome having at least 97.7% unaltered nucleotides over a period of 1300 years. The probability was calculated using a binomial cumulative distribution function assuming different mutation rates reported in the literature (6, 7, 35). The blue and green lines represent the mutation rates due to substitutions and recombinations in lytic phages, with rates of 4.690×10^{-3} (7) and 1.976×10^{-4} (35) altered nucleotides per site per year, respectively. The green line includes a rate of 1.154×10^{-4} (6) altered nucleotides per site per year due to substitutions in temperate phage genomes. **(E)** Gene expression in cultured *Faecalibacterium prausnitzi* from 29 RNA-seq samples (39) shows that Hoc-like head decoration and Reverse transcriptase genes are expressed at a comparable level to bacterial genes and higher than

other phage genes. **(F)** Table with *Mushuvirus mushu* hosts found using BLASTn search against Unified Human Gastrointestinal Genome (UHGG) collection (43) and Genome Taxonomy Database (GTDB) (42).

Supplementary Figures

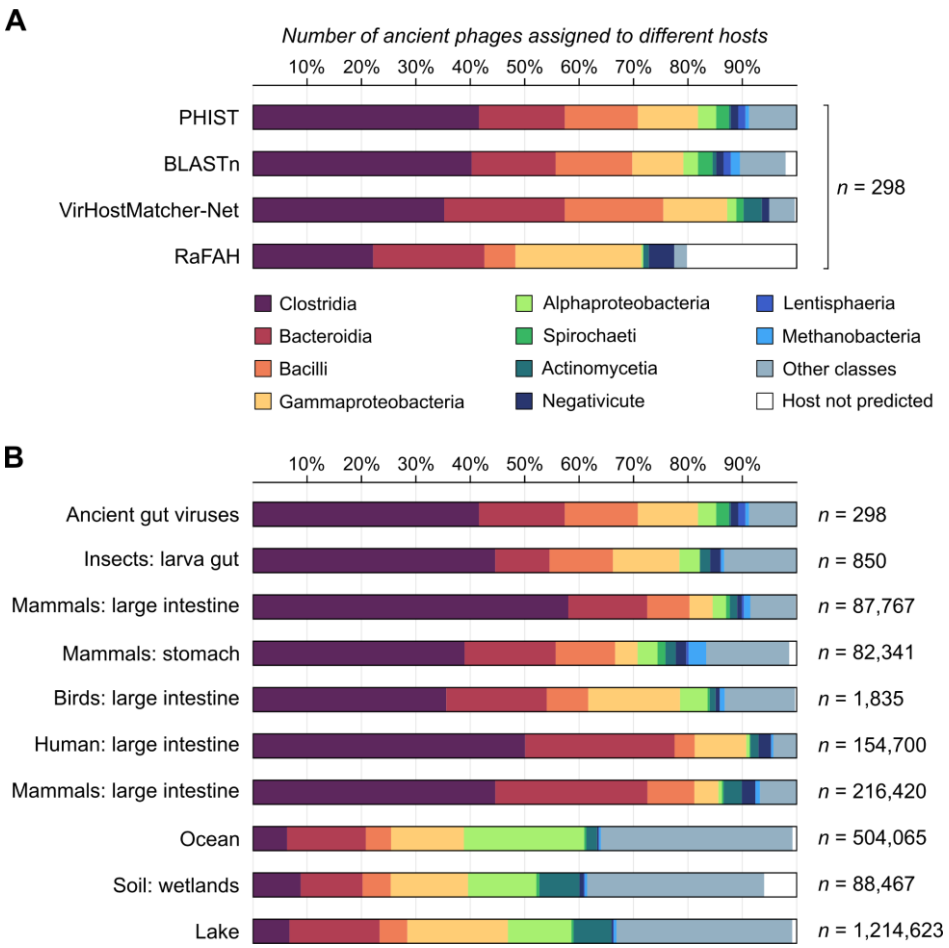


Fig. S1. Bacteria classes linked to ancient and contemporary viruses from IMG/VR. **(A)** Comparison of host predictions for 298 aMGVs using four tools BLASTn (55), PHIST (56), VirHostMatcher-Net (57), and RaFAH (58). **(B)** The distribution of different classes of host bacteria assigned to ancient viruses and hosts of viruses from different ecosystems in IMG/VR (29).

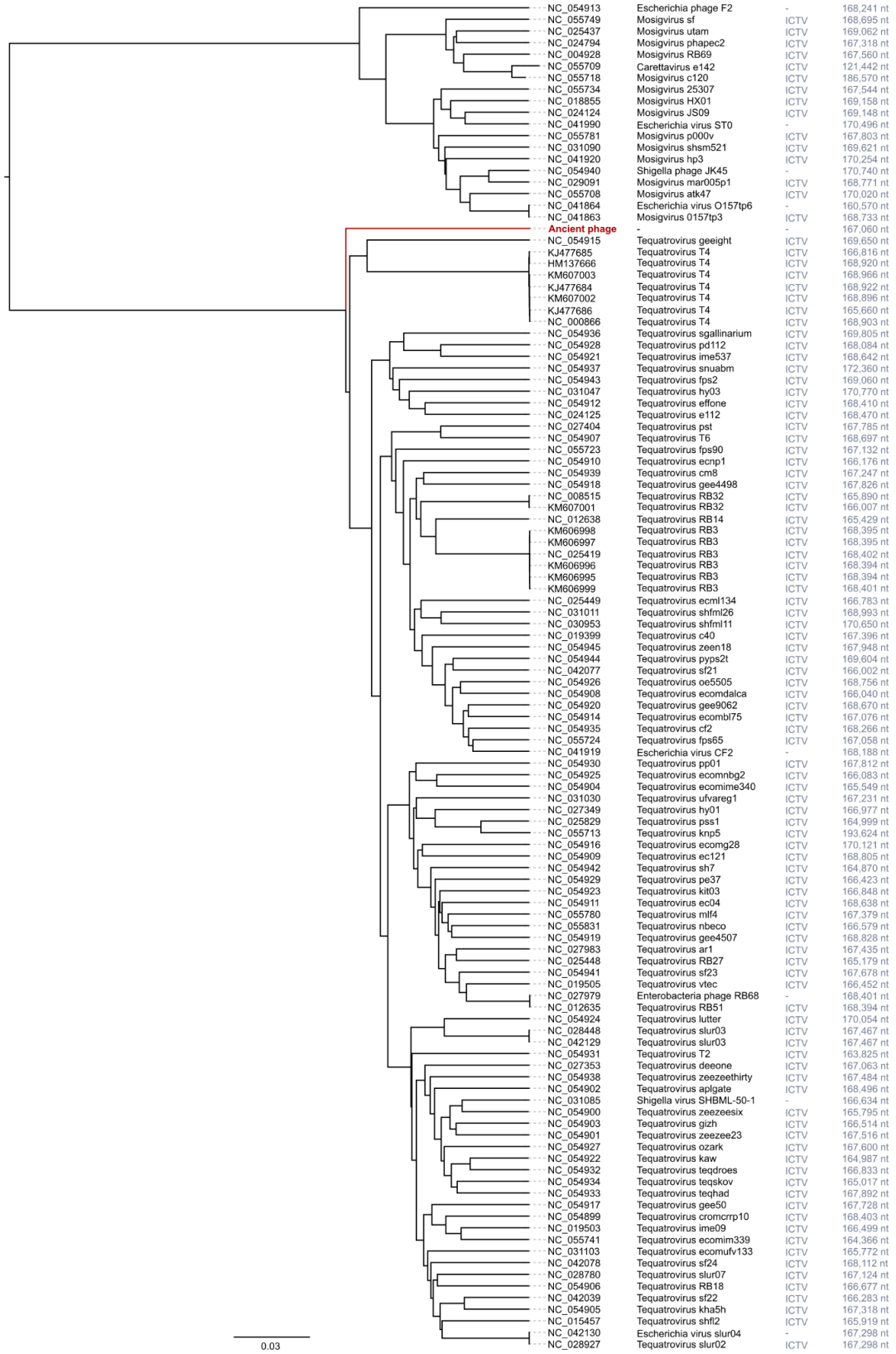


Fig. S2. Proteomic tree of T4-related ancient bacteriophage

Fragment of Proteomic tree created by VipTree (60) server for ancient phage genome NODE_1_length_167060_cov_116.590443.

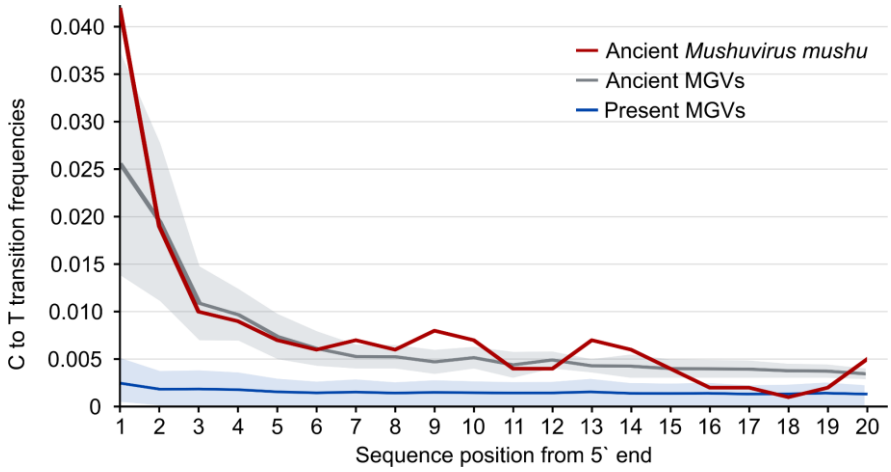


Fig. S3. Damage patterns in ancient and contemporary virus genomes. Comparison of damage patterns between ancient *Mushuvirus mushu*, other aMGVs and modern viral genomes. The solid line shows the mean frequency of C → T substitutions and the shade indicates the standard deviation.

References:

1. G. F. Hatfull, R. W. Hendrix, Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**, 298–303 (2011).

2. S. Roux, J. Matthijnssens, B. E. Dutilh, 'Metagenomics in Virology' in *Reference Module in Life Sciences*, (Elsevier, 2019) https:/doi.org/10.1016/b978-0-12-809633-8.20957-6.

3. N. Pappas, *et al.*, 'Virus Bioinformatics' in *Encyclopedia of Virology*, (Elsevier, 2021), pp. 124–132.

4. L. Call, S. Nayfach, N. C. Kyrpides, Illuminating the Virosphere Through Global Metagenomics. *Annual Review of Biomedical Data Science* **4**, 369–391 (2021).

5. R. A. Edwards, *et al.*, Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology* **4**, 1727–1736 (2019).

6. S. Minot, *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).

7. A. Kupczok, *et al.*, Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades. *Mol. Biol. Evol.* **35**, 1147–1159 (2018).

8. M. A. Spyrou, K. I. Bos, A. Herbig, J. Krause, Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* **20**, 323–340 (2019).

9. J. G. N. Barreat, A. Katzourakis, Paleovirology of the DNA viruses of eukaryotes. *Trends Microbiol.* **30**, 281–292 (2022).

10. J.-M. Alempic, *et al.*, An Update on Eukaryotic Viruses Revived from Ancient Permafrost. *Viruses* **15** (2023).

11. L. Nishimura, N. Fujito, R. Sugimoto, I. Inoue, Detection of Ancient Viruses and Long-Term Viral Evolution. *Viruses* **14** (2022).

12. S. Appelt, *et al.*, Viruses in a 14th-Century Coprolite. *Applied and Environmental Microbiology* **80**, 2648–2655 (2014).

13. T. M. Santiago-Rodriguez, *et al.*, Natural mummification of the human gut preserves bacteriophage DNA. *FEMS Microbiol. Lett.* **363**, fnv219 (2016).

14. R. Eisenhofer, L. S. Weyrich, Proper Authentication of Ancient DNA Is Still Essential. *Genes* **9** (2018).

15. M. Borry, A. Hübner, A. B. Rohrlach, C. Warinner, PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA assembly. *PeerJ* **9**, e11845 (2021).

16. L. Nishimura, *et al.*, Identification of ancient viruses from metagenomic data of the Jomon people. *J. Hum. Genet.* **66**, 287–296 (2020).

17. M. C. Wibowo, *et al.*, Reconstruction of ancient microbial genomes from the human gut. *Nature* **594**, 234–239 (2021).

18. M. Klapper, *et al.*, Natural products from reconstructed bacterial genomes of the Middle and Upper Paleolithic. *Science*, eadf5300 (2023).

19. F. Maixner, *et al.*, The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**, 162–165 (2016).

20. A. Tett, *et al.*, The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* **26**, 666–679.e7 (2019).

21. R. W. Hagan, *et al.*, Comparison of extraction methods for recovering ancient microbial DNA from paleofeces. *Am. J. Phys. Anthropol.* **171**, 275–284 (2020).

22. M. Borry, *et al.*, CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content. *PeerJ* **8**, e9001 (2020).

23. F. Maixner, *et al.*, Hallstatt miners consumed blue cheese and beer during the Iron Age and retained a non-Westernized gut microbiome until the Baroque period. *Curr. Biol.* **31**, 5149–5162.e6 (2021).

24. K. Kieft, Z. Zhou, K. Anantharaman, VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 1–23 (2020).

25. J. Guo, *et al.*, VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 1–13 (2021).

26. S. Roux, *et al.*, Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).

27. C. Der Sarkissian, *et al.*, Ancient Metagenomic Studies: Considerations for the Wider Scientific Community. *mSystems* **6**, e0131521 (2021).

28. H. Bin Jang, *et al.*, Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

29. A. P. Camargo, *et al.*, IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* (2022) https:/doi.org/10.1093/nar/gkac1037.

30. P. J. Walker, *et al.*, Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch. Virol.* **167**, 2429–2440 (2022).

31. S. Nayfach, *et al.*, Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology* **6**, 960–970 (2021).

32. A. P. Camargo, *et al.*, You can move, but you can’t hide: identification of mobile genetic elements with geNomad. *bioRxiv*, 2023.03.05.531206 (2023).

33. J. K. Cornuault, *et al.*, Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes. *Microbiome* **6**, 65 (2018).

34. N. Juge, Relationship between mucosa-associated gut microbiota and human diseases. *Biochem. Soc. Trans.* **50**, 1225–1236 (2022).

35. A. Kupczok, T. Dagan, Rates of Molecular Evolution in a Marine Phage Lineage. *Viruses* **11** (2019).

36. F. Sharifi, Y. Ye, MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res.* **47**, W289–W294 (2019).

37. S. Doulatov, *et al.*, Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004).

38. S. Roux, *et al.*, Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).

39. S. Auger, *et al.*, Gene co-expression network analysis of the human gut commensal bacterium Faecalibacterium prausnitzii in R-Shiny. *PLoS One* **17**, e0271847 (2022).

40. J. J. Barr, *et al.*, Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 10771–10776 (2013).

41. J. J. Barr, *et al.*, Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13675–13680 (2015).

42. D. H. Parks, *et al.*, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785–D794 (2022).

43. A. Almeida, *et al.*, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2020).

44. C. M. Bellas, D. C. Schroeder, A. Edwards, G. Barker, A. M. Anesio, Flexible genes establish widespread bacteriophage pan-genomes in cryoconite hole ecosystems. *Nat. Commun.* **11**, 1–10 (2020).

45. M. F. Marston, J. B. H. Martiny, Genomic diversification of marine cyanophages into stable ecotypes. *Environ. Microbiol.* **18**, 4240–4253 (2016).

46. B. D. Merrill, *et al.*, Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers Vanishing Gut Microbes. *bioRxiv* (2022) https:/doi.org/10.1101/2022.03.30.486478.

47. J. A. Fellows Yates, *et al.*, Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci Data* **8**, 31 (2021).

48. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).

49. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

50. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

51. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

52. V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan, Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior in *2011 31st International Conference on Distributed Computing Systems Workshops*, (IEEE, 2011) https:/doi.org/10.1109/icdcsw.2011.20.

53. S. Nayfach, *et al.*, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2020).

54. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale

using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

55. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

56. A. Zielezinski, S. Deorowicz, A. Gudyś, PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* **38**, 1447–1449 (2021).

57. W. Wang, *et al.*, A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics* **2** (2020).

58. F. H. Coutinho, *et al.*, RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns (N Y)* **2**, 100274 (2021).

59. S. Li, *et al.*, A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience* **25** (2022).

60. Y. Nishimura, *et al.*, ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).

61. F. Asnicar, G. Weingart, T. L. Tickle, C. Huttenhower, N. Segata, Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).

62. C. Moraru, A. Varsani, A. M. Kropinski, VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* **12**, 1268 (2020).

63. O. Gascuel, BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).

64. P. Terzian, *et al.*, PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**, lqab067 (2021).

65. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

66. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).