Ultraconserved bacteriophage genome sequence identified in 1300-year-old human palaeofaeces

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 of Bioinformatics, University 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, we discovered an ancient genome nearly identical to present-day *Mushuvirus mushu*, a phage that infects 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 297 other phage genomes from the last 5300 years, including those belonging to unknown families. Our findings demonstrate the feasibility of reconstructing ancient phage genomes, expanding the known virosphere, and offering new insights into phage-bacteria interactions that cover several millennia.

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 an evolutionary perspective. The longest study of substitutions and recombinations in phage genomes spans nearly three decades[7], representing only a partial glimpse into the complex evolutionary history of bacteriophages. To fully understand the phylogeny of these viruses and their effect on microbial ecosystems, it is essential to go further back in time.

Previous research on ancient virology has mainly focused 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 the gut contents of pre-Columbian Andean mummies[13]. However, these studies lacked authentication based on terminal deamination patterns observed in ancient DNA (aDNA)[14], which is used to distinguish ancient sequences from modern contaminants[15]. Furthermore, 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 aDNA 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 new opportunities to unlock the mystery of past phage genome diversity.

Here, we analyse draft and complete genomes assembled from 150–5300 years old palaeofaeces 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 on 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 (**Supplementary Table S1**). The *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 the assembled contigs as ancient based on deamination patterns observed at the 5` ends of the sequencing reads (see Methods). A total of 2,375 sequences were classified as viral by at least two methods (VIBRANT[24], VirSorter2[25] and Jaeger; **Supplementary Table 2**) and were selected as *bona fide* virus contigs. Among these, 383 were at least 20 kb long or assessed by CheckV as either medium or better quality (>50% completeness; **Fig. 1b** and **Supplementary Table S3**) and selected for further analysis. We clustered the selected sequences into 298 species-level viral operational taxonomic units (hereafter referred to as aMGVs - ancient metagenomic gut viruses) based on 95% average nucleotide identity (ANI) with over 85% alignment fraction[26]. Despite fragmentation and degradation of the aDNA, our aMGV collection had size range from 5 kbp to 292 kbp, an estimated mean completeness of 50.3%, and comprised of 49 high-quality or complete genomes (**Supplementary 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) (**Fig. 1c,d** and **Supplementary Table S3**).

Human gut origin of ancient phage genomes

Deamination damage patterns is a well-established technique for ancient DNA validation, but the complexity of ancient metagenomic data can 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 degradation 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 (**Supplementary Table 4**), 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 represented potential novel viruses or incomplete genomic fragments with insufficient phylogenetic information. The estimated contamination was relatively low; only 5% of aMGVs (n = 15) clustered with viruses from environmental or engineered ecosystems.

Further evidence supporting the ancient gut origin of the 298 analysed aMGVs came from the distribution of their predicted hosts (see Methods; **Supplementary Fig. 1** and **Supplementary Table 5**). 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 was 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 (**Supplementary Table 6**). Hence, both the results of host prediction and the gene-sharing network analysis suggest that the ancient phage sequences primarily originated 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 (**Supplementary Table 7**). Most aMGVs from the *Caudoviricetes* class were not classified at lower taxonomic ranks, except for 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 clustered them with RefSeq and IMG/VR virus genomes based on pairwise average amino acid identity (AAI) 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 the 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 (**Supplementary Table 8**).

Most putative families (90%, 35 out of 39) and genera (72%, 33 out of 46) contained 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, **Supplementary Fig. 2**) with an ancient virus located on the ancestral branch. Only a few putative taxa contained multiple ancient representatives. For example, two genus-like groups ("13" and "14" see **Supplementary Table 8**) formed a family-like unit that encompassed three ancient phages, 36 IMG/VR viruses and the *Salmonella*-infecting phage, *Astrithrvirus astrithr*.

In contrast to groups with multiple representatives, there were four putative families and 13 genera represented by only a single aMGV. While these phages could hypothetically represent extinct lineages, it is likely that modern representatives of those groups have yet to be 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 similar to known reference genomes. Consequently, the other 249 aMGVs could still include viral genomes unlike any sequence characterised to date. Despite this limitation, we have shown 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 ancient phages, we found the genome sequence of a bacteriophage from the *Mushuvirus mushu* species 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 believe 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 **Supplementary Fig. 3** and **Supplementary Table 3**). Moreover, the *F. prausnitzii* host of *Mushuvirus mushu* is a strict anaerobe that typically inhabits the mucosal surface in the gastrointestinal tract[34], so it is unlikely it could have contaminated the original palaeofaeces sample[21].

The ancient *Mushuvirus mushu* genome that was recovered from 1300-year-old faecal material from the Zape cave in Mexico 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,b**). 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** and **Supplementary Table 9**). 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 that are 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 shown 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 concentrated in the specific section of the gene encoding the Hoc-like capsid decoration protein (**Fig. 4b**). This fragment is the target of a 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, the Hoc-like gene had a lower level of AAI 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 *F. prausnitzii*, the Hoc-like gene is expressed at levels similar to a typical bacterial gene under most studied conditions[39] (**Fig. 4e**). This finding suggests that the Hoc-like protein might play 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 the ancient *Mushuvirus mushu* sequence represented an integrated prophage, we investigated the flanking regions of the contig (see Methods). These regions differed between the ancient and modern viruses 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* (**Supplementary Table 10**). *Mushuvirus mushu* appears to have a broad host range[33]; 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** and **Supplementary Table 11**).

Discussion

In this study, we demonstrated the large-scale reconstruction of high-quality ancient phage genomes 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 their host associations. There were no significant similarities to known viruses for approximately half of the reconstructed genomes. This indicates that our understanding of human gut virome history is limited. Nevertheless, advancements in viral metagenomics and access to well-preserved aDNA samples hold promise for expanding our understanding of virosphere evolution.

Bacteriophage genomes are often 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 a presence on different continents. Integrated as a prophage in its bacterial host, *Mushuvirus mushu* could have generated nearly identical virions over vast periods, 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 such as *F. prausnitzii* which is widely present in the human gut across populations ranging from hunter-gatherers to industrialized societies[46].

This study highlights the utility of publicly accessible ancient metagenomes in investigating viruses associated with microorganisms. Our focus was on sequences from well-preserved gut and palaeofaeces 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-evolving with bacterial, animal and plant hosts.

Methods

Sample selection

Based on community-curated metadata from the AncientMetagenomeDir[47], we selected 72 aDNA metagenomic libraries from 30 well-preserved human faeces or digestive gut contents. Data from palaeofaeces primarily originated 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 (n = 15) from a biopsy of the Tyrolean Iceman mummy discovered more than 30 years ago in a melting glacier[19]. The samples represented material from 150 to 5300 years ago. Details about original publications, localisations (coordinates), and samples (sequencing depth and instruments) are described in the supplementary materials (**Supplementary Table 1**).

Sequence data preprocessing

Pair-ended Illumina reads were trimmed using Cutadapt v.4.1[48] with a 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 the *Homo sapiens* reference genome (hg37) using KneadData v.0.12.0 (with --bypass-trim option) to filter out human DNA (https://github.com/biobakery/kneaddata). The quality of the 2,352,455,887 remaining reads after preprocessing was controlled in Fastqc.

294 **De novo contig assembly**
295
296 Filtered reads from each library were assembled into contigs using Metaspades v.3.15.5[49] with
297 default settings. Only contigs longer than 4000 nucleotides with a minimum coverage of 20
298 were considered in the downstream analyses (see **Fig. 1**).
299
300 **aDNA authentication**
301
302 After preprocessing, clean reads were mapped to the assembled contigs using Bowtie2 v.2.4.4[50]
303 with default settings. The resulting alignment was sorted and indexed with SAMtools v.1.14[51].
304 DNA damage patterns observed in reads were used to label corresponding contigs as ancient
305 or modern. The analysis was run using Pydamage v.0.70[15] - a program that calculates the
306 frequency of C to T transitions at the first 20 positions of mapped reads compared to a reference
307 sequence. The filtering threshold for predicted accuracy was determined by the Kneedle
308 method[52] and we imposed an additional cut-off of transition frequency (minimum 0.01 at the
309 first position) to filter out contigs with weakly damaged reads that could introduce a random
310 noise generated by the inherent error of the Illumina method. Additionally, the same analytical
311 process of authentication was performed for 943 randomly selected contigs from the
312 Metagenomic Gut Viruses (MGV) database[31] to compare deamination patterns between
313 modern and ancient viral genomes (see **Fig. 1c, d**).
314
315 **Viral contigs identification**
316
317 Three machine learning tools were used to identify viral contigs. The first was Jaeger v.1.1.0,
318 a deep-learning model that identifies phage genome sequences in metagenomes
319 (https://github.com/Yasas1994/Jaeger) based on automatic compositional feature extraction.
320 The second and third were VIBRANT v.1.2.1[24] and VirSorter2 v.2.2.3[25], which rely on
321 analysing HMM profiles representing conserved families and/or domains similar to predicted
322 proteins but applying different classifiers and reference databases. Jaeger and VIBRANT were
323 run with default settings. For VirSorter2, we used the positional arguments '--include-groups
324 dsDNAphage,NCLDV,ssDNA,lavidaviridae all'. Contigs classified as viral by at least two
325 tools were further analysed.
326
327 **Genomes quality assessment**
328
329 CheckV v.1.0.1[53] was applied to assess the genome quality of ancient viral contigs using the
330 'end_to_end' command. Ancient viral genomes classified as complete, high-quality, medium-
331 quality, or fragments longer than 20 kb and with at least one viral gene were considered in the
332 next steps.
333
334 **Viral contigs clustering**
335
336 The ancient MGVs (n = 383) were clustered into 298 species-level viral operational taxonomic
337 units (vOTUs) using scripts, published in the CheckV repository

338 (https://bitbucket.org/berkeleylab/checkv/src/master/). Accordingly, sequences were grouped
339 based on 95% ANI and 85% alignment fraction of the shorter sequence, as recommended in
340 MIUViG (Minimum information about an uncultivated virus genome)[26].
341
342 **Gene-sharing network**

343 We selected 10 modern phage genomes for each aMGV from the IMG/VR (v.4 - high-
344 confidence genomes only) to compare ancient phages with their modern counterparts from
345 different environments. Specifically, we selected genomes with the highest number of shared
346 proteins determined by a DIAMOND v.2.0.15[54] search in the 'blastp' mode (--very-sensitive)
347 with a minimum of 50% query coverage and 50% sequence identity. We then visualized this
348 collection of aMGVs, selected modern viral genomes, and all prokaryotic DNA viruses with
349 assigned ICTV taxonomy (VMR_20-190822_MSL37.2, created 08/31/2022) using vContact2
350 v.0.11.3[28]. The network was displayed in Cytoscape v.3.9.0 and refined in Inkscape v.1.2.2.
351
352 **Host prediction**
353
354 Four computational tools (BLASTn[55], PHIST[56], VirHostMatcher-Net[57], and RaFAH[58]) were
355 used to assign hosts to ancient phages and sequences from IMG/VR v.4 (only PHIST). PHIST
356 v.1.1.0 and BLASTn v.2.13.0+ predictions were run against representative genomes of GTDB
357 [42] database v.07-RS207 (62,291 bacterial species + 3,412 archaeal species). Prokaryotic species
358 whose genomes obtained the highest similarity score to the virus genome and had an *e*-value <
359 10^{-5} (BLASTn) or *p*-value < 10^{-5} (PHIST) were assigned as a putative host. For methods
360 integrating machine learning approaches such as VirHostMatcher-Net or RaFAH, we selected
361 host predictions with, at minimum, 0.5 and 0.14 scores, respectively.
362
363 **Taxonomy assignment, clustering, and phylogenetic analysis**
364
365 Taxonomic assignment of viral genomes was performed in the geNomad 1.3.3 tool[32] using the
366 'annotate' function. To classify aMGVs at the genus and family level, viral genomes were
367 clustered using a combination of gene sharing and AAI. Initially, we selected aMGVs assessed
368 as high-quality or complete by CheckV and then added prokaryotic viruses from RefSeq (*n* =
369 4703; access: 30.01.2023) and IMG/VR sequences (*n* = 265) to form clusters (VC) with
370 aMGVs (see: Methods, Gene-sharing-network). In this collection, pairwise protein sequence
371 alignments were performed using DIAMOND v.0.9.24.125[54] with the options '-e-value
372 1 × 10−5–max-target-seqs 10000'. Next, we calculated the percentage of shared genes and AAI
373 between each pair of viruses. Following the criteria from previous studies[31,59], we kept
374 connections between viruses with >20% AAI and >10% genes shared for the family level and
375 >50% AAI and >20% genes shared for the genus level. Finally, clustering was performed based
376 on the connections between viral genomes using MCL with the option '-I 1.2' for the family
377 level or '-I 2' for the genus level. All scripts used to perform analyses at this step are available
378 at (https://github.com/snayfach/MGV/blob/master/aai_cluster/README.md). To visualise the
379 phylogenetic relationships of genus- and family-level groups, we generated a proteomic tree of
380 5017 viral sequences using ViPTreeGen v.1.1.3[60] and GraPhlAn v.1.1.3[61].

Analyses of the *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) was 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 vOTUs from IMG/VR v.4 (high-confidence genomes only). All 12 sequences from the vContact2 cluster were aligned using MAFFT v.7.308 to differentiate the core of the phage genome (~36,623 bp) from the flanking regions coming from 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 *Mushuvirus 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'. A phylogenetic tree (**Fig. 4d**) was constructed with the obtained similarity matrix using the bioNJ algorithm with default parameters[63]. To annotate protein-coding genes in the analysed genomes, we used an end-to-end script (https://github.com/Yasas1994/phage_contig_annotator) that annotates phage genes based on the HMMER v.3.3.2 search against Prokaryotic Virus Remote Homologous Groups[64]. Visualisation and manual curation of the genome were conducted in Geneious Prime v.2023.04 (**Fig. 4c**). The MAFFT plugin for the same tool was used to generate multiple sequence alignments of entire viral genomes. The identity of protein-coding genes at the amino acid level was calculated based on local alignment performed using EMBOSS v.6.6.0.0[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 DGRs using myDGR[36]. Finally, we assigned potential hosts by searching for sequences similar to the prophage core and flanking regions (blastn -task megablast) in the UHGG collection[43] and GTDB[42].

Data and Code Availability

Access numbers to raw data are available in **Supplementary Table 1**. Ancient phage genome sequences are available in Zenodo (https://doi.org/10.5281/zenodo.7919433). The reconstructed ancient genome sequence of *Mushuvirus mushu* is available from NCBI GenBank under accession BK063464. Bioinformatic scripts and a guide to the data analysis performed are provided in the GitHub repository (https://github.com/rozwalak/Ancient_gut_phages).

Acknowledgements

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 was awarded to P.R. A.Z is supported by Polish National Science Centre [2018/31/D/NZ2/00108], and 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 the German Federal Ministry of Education and Research. B.E.D and Y.W are supported by the European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF). The computations were performed at the PLGrid Infrastructure and the Poznan Supercomputing and Networking Center (grant 312 and grant 647).

Author contributions

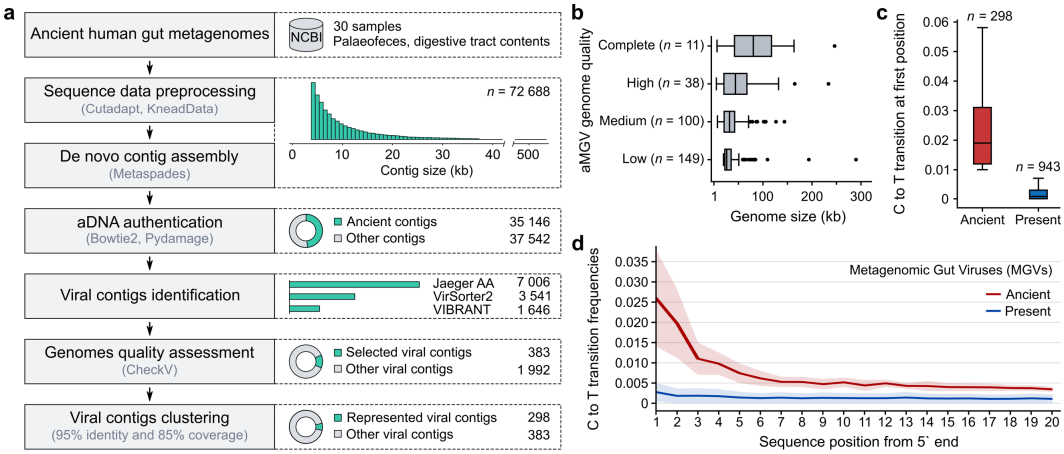
P.R. conceived the idea of studying ancient phages; P.R., A.Z., J.B. and B.E.D. designed the research; P.R. and A.Z. selected samples and performed preprocessing, assembly and aDNA authentication; J.B. and Y.W. performed identification of viral contigs; P.R. and A.Z. created a gene-sharing network and performed host prediction; P.R., A.Z. and J.B. performed taxonomy assignment, clustering, phylogenetic studies, and the analyses of the *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 from A.Z., J.B. and B.E.D. All authors reviewed and approved the manuscript.

Competing interests: The authors declare no competing interests.

Supplementary information: Supplementary Tables S1-S11.

449 **Figures**

450



451

452

453 **Fig. 1. Identification of phage genomes in ancient human gut metagenomes.**
454 **a** Overview of workflow to identify ancient phage genomes. **b** Distribution of the genome sizes
455 of 298 ancient metagenomic viruses (aMGVs) stratified by CheckV[53] genome quality. Box and
456 whisker plots show median (center line), upper and lower quartiles (represented by boxes), and
457 highest (upper whisker) or lowest (lower whisker) value within a 1.5 inter-quartile range (IQR)
458 while black dots indicate values outside of the IQR. **c** Distribution of C → T substitution
459 frequencies at the first position of the 5' end of sequencing reads. **d** Comparison of damage
460 patterns between selected modern viral genomes from MGV[31] and aMGVs. The solid line
461 shows the mean frequency of C → T substitutions and the shade indicates the standard
462 deviation.

463

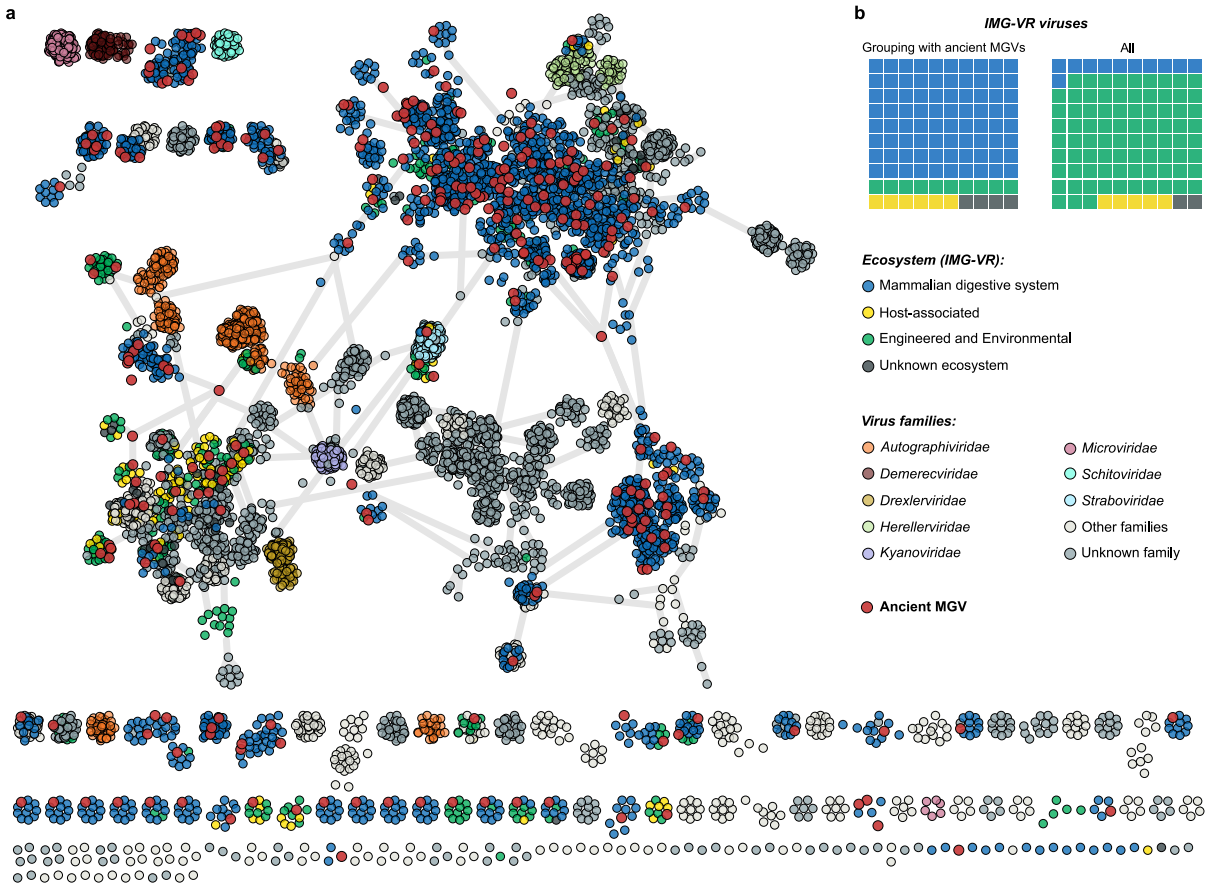


Fig. 2. Ancient phage genomes from palaeofaeces are related to currently known mammalian-gut-associated viruses.

a vContact2 gene-sharing network[28] of 151 ancient metagenomic gut virus (aMGV) genomes, 2,198 selected close relatives from the IMG/VR database, and 3,655 prokaryotic viruses classified by the International Committee of Virus Taxonomy. Distantly related aMGVs (n = 147) were outliers and not included in the gene-sharing network. **b** Waffle charts represent the proportion of contemporary viruses categorized by ecosystems in two data sets: clusters comprising aMGVs (left) and the entire IMG/VR database.

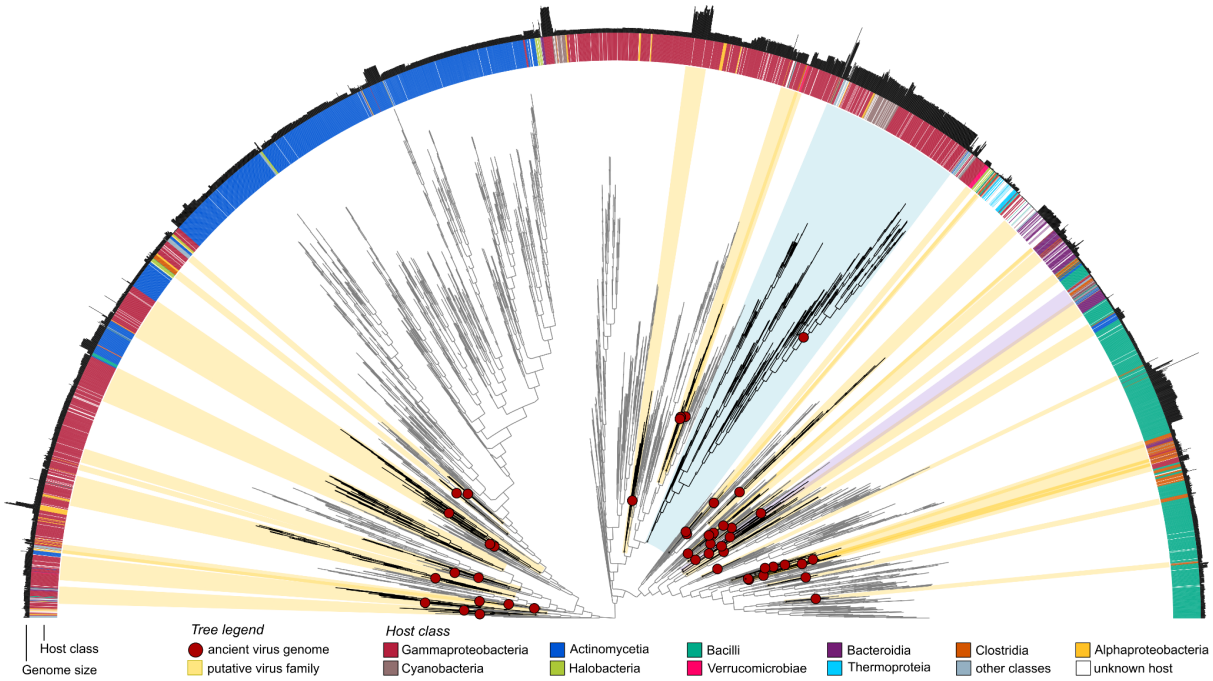
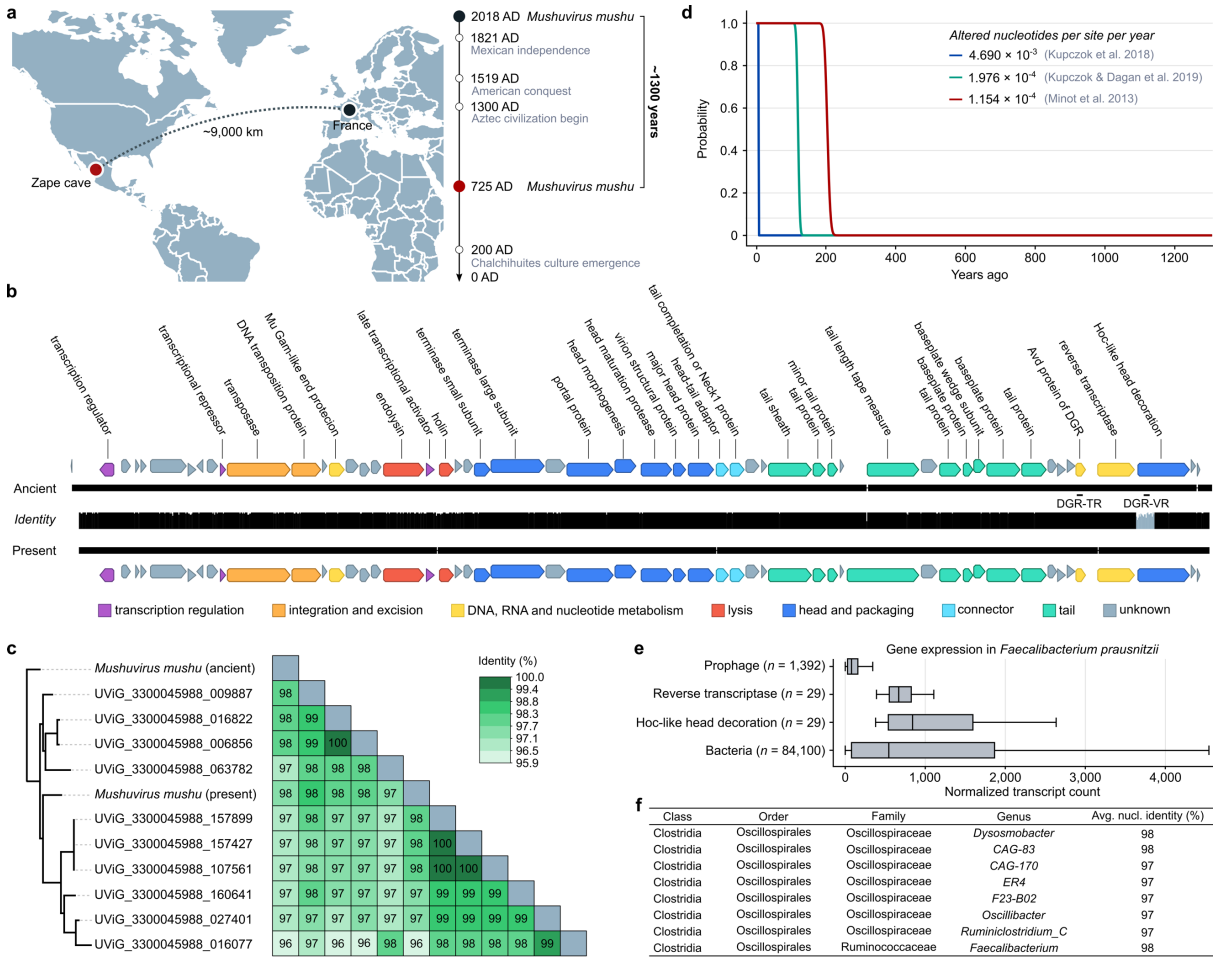


Fig. 3. Proteomic tree of ancient and contemporary phage genomes.
A tree generated in ViPTreeGen[60] encompassing 49 high-quality and complete ancient phages, their 265 closest relatives from IMG/VR[29], and all 4,703 prokaryote-infecting viruses from RefSeq[66]. The ancient phages are identified with red circles. Putative families were defined using a percentage of shared genes and amino acid identity between each pair of viruses. Clusters containing at least one ancient phage are highlighted in yellow and listed in **Supplementary Table 8**. 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 distances between the modern reference (France) and the ancient *Mushuvirus mushu* (from the Zape cave in Mexico) genomes. **b** Comparison of sequence conservation and gene organization between the ancient and modern reference (NC_047913.1) genomes of *Mushuvirus mushu*. Black lines indicate alignments with small gaps. The height of the middle plot depends on the identity between genomes. The lowest similarities (grey color) were concentrated in a ~600-nucleotide variability region of the Hoc-like gene affected by the diversity-generating retroelement (DGR) mechanism. Two major structural differences in the ancient and modern genomes constituted an additional codon stop in a gene of unknown function (genome position: 9518) and 17 nucleotide insertion/deletions, which changed the open reading frame in the tail length tape measure protein. **c** Intergenomic similarity matrix calculated by VIRIDIC[62] revealed high identity between ancient and modern representatives of the 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} and 1.976×10^{-4} altered nucleotides per site per year, respectively. The green line includes a rate of 1.154×10^{-4} 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 more than other phage genes. **f** Table with *Mushuvirus mushu* hosts found using the BLASTn search against the Unified Human Gastrointestinal Genome collection[43] and the Genome Taxonomy Database[42].

Supplementary Figures

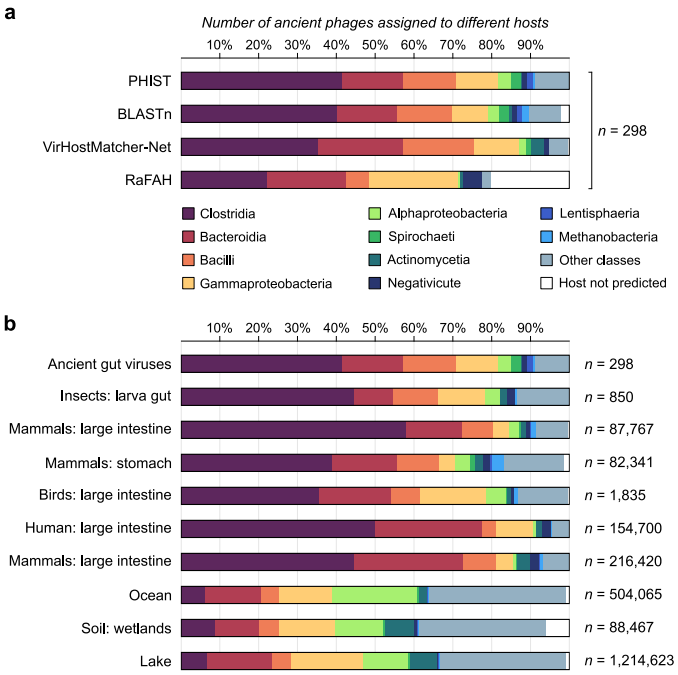


Fig. S1. Bacteria classes linked to ancient and contemporary viruses from IMG/VR. **a** Comparison of host predictions for 298 ancient metagenomic gut viruses (aMGVs) using four tools BLASTn[55], PHIST[56], VirHostMatcher-Net[57], and RaFAH[58]. **b** The distribution of different classes of host bacteria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

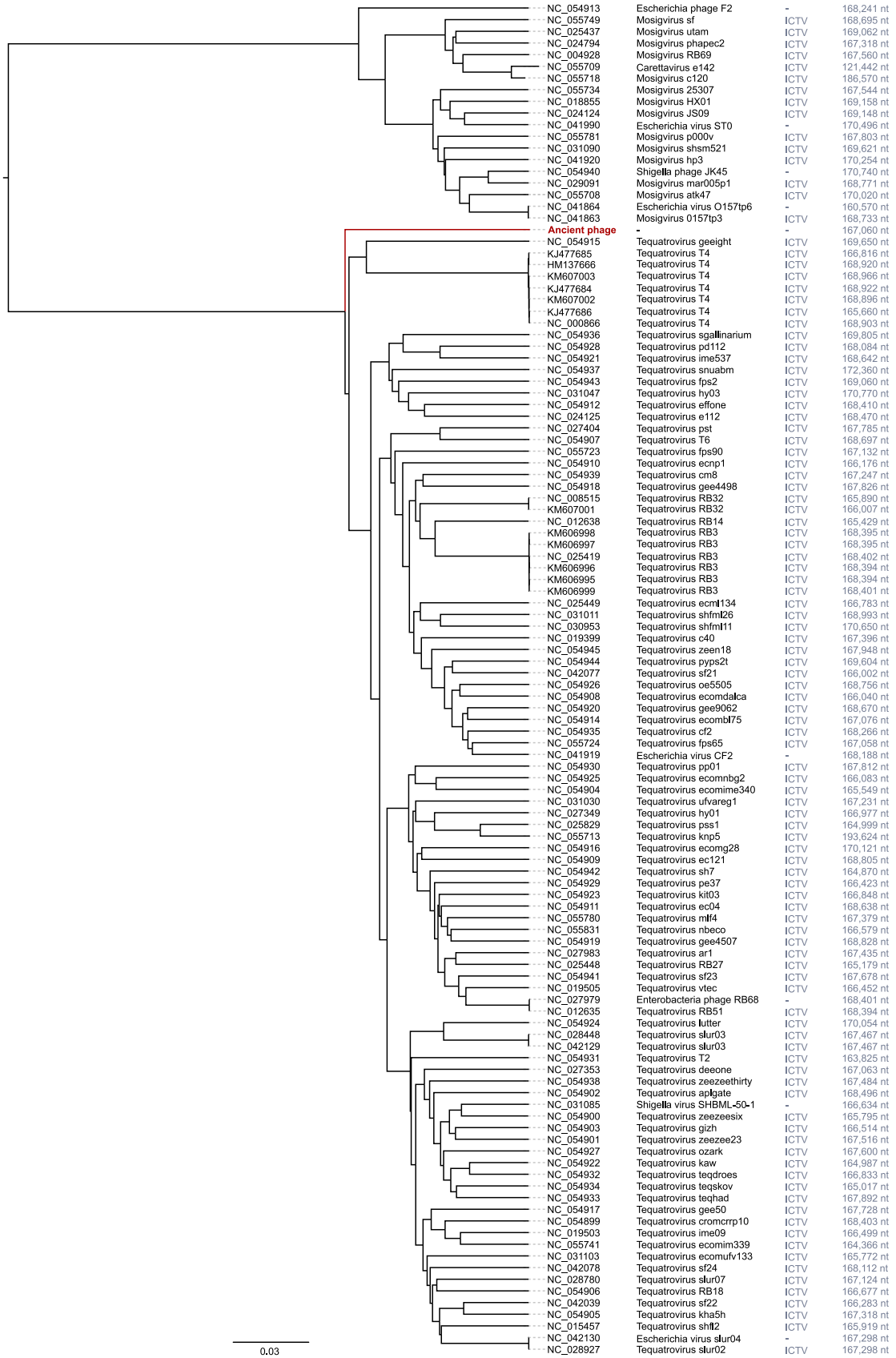


Fig. S2. Proteomic tree of T4-related ancient bacteriophage.

Fragment of the proteomic tree created by the VipTree[60] server for ancient phage genome NODE_1_length_167060_cov_116.590443.

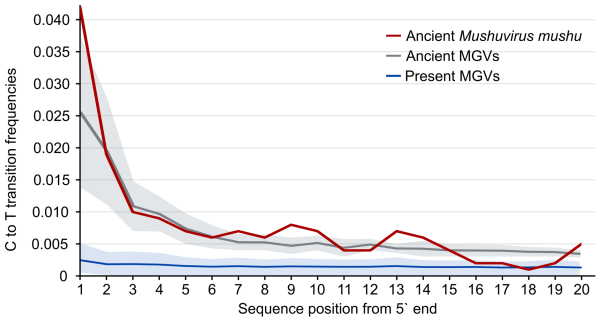


Fig. S3. Damage patterns in ancient and contemporary virus genomes.
Comparison of damage patterns between ancient *Mushuvirus mushu*, other ancient metagenomic gut viruses (aMGVs), and modern viral genomes. The solid line shows the mean frequency of C → T substitutions, and the shaded areas indicate the standard deviation.

536 **References:**
537

538 1. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**,
539 298–303 (2011).
540 2. Roux, S., Matthijnssens, J. & Dutilh, B. E. Metagenomics in Virology. in *Reference*
541 *Module in Life Sciences* (Elsevier, 2019). doi:10.1016/b978-0-12-809633-8.20957-6.
542 3. Pappas, N. *et al.* Virus Bioinformatics. in *Encyclopedia of Virology* 124–132 (Elsevier,
543 2021). doi:10.1016/b978-0-12-814515-9.00034-5.
544 4. Call, L., Nayfach, S. & Kyrpides, N. C. Illuminating the Virosphere Through Global
545 Metagenomics. *Annual Review of Biomedical Data Science* vol. 4 369–391 Preprint at
546 https://doi.org/10.1146/annurev-biodatasci-012221-095114 (2021).
547 5. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread
548 human gut virus crAssphage. *Nature Microbiology* **4**, 1727–1736 (2019).
549 6. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.*
550 **110**, 12450–12455 (2013).
551 7. Kupczok, A. *et al.* Rates of Mutation and Recombination in Siphoviridae Phage Genome
552 Evolution over Three Decades. *Mol. Biol. Evol.* **35**, 1147–1159 (2018).
553 8. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an
554 emerging tool for infectious disease research. *Nat. Rev. Genet.* **20**, 323–340 (2019).
555 9. Barreat, J. G. N. & Katzourakis, A. Paleovirology of the DNA viruses of eukaryotes.
556 *Trends Microbiol.* **30**, 281–292 (2022).
557 10. Alempic, J.-M. *et al.* An Update on Eukaryotic Viruses Revived from Ancient
558 Permafrost. *Viruses* **15**, (2023).
559 11. Nishimura, L., Fujito, N., Sugimoto, R. & Inoue, I. Detection of Ancient Viruses and
560 Long-Term Viral Evolution. *Viruses* **14**, (2022).
561 12. Appelt, S. *et al.* Viruses in a 14th-Century Coprolite. *Applied and Environmental*
562 *Microbiology* vol. 80 2648–2655 Preprint at https://doi.org/10.1128/aem.03242-13
563 (2014).
564 13. Santiago-Rodriguez, T. M. *et al.* Natural mummification of the human gut preserves
565 bacteriophage DNA. *FEMS Microbiol. Lett.* **363**, fnv219 (2016).
566 14. Eisenhofer, R. & Weyrich, L. S. Proper Authentication of Ancient DNA Is Still
567 Essential. *Genes* **9**, (2018).
568 15. Borry, M., Hübner, A., Rohrlach, A. B. & Warinner, C. PyDamage: automated ancient
569 damage identification and estimation for contigs in ancient DNA assembly. *PeerJ* **9**,
570 e11845 (2021).
571 16. Nishimura, L. *et al.* Identification of ancient viruses from metagenomic data of the
572 Jomon people. *J. Hum. Genet.* **66**, 287–296 (2020).
573 17. Wibowo, M. C. *et al.* Reconstruction of ancient microbial genomes from the human gut.
574 *Nature* **594**, 234–239 (2021).
575 18. Klapper, M. *et al.* Natural products from reconstructed bacterial genomes of the Middle
576 and Upper Paleolithic. *Science* eadf5300 (2023) doi:10.1126/science.adf5300.
577 19. Maixner, F. *et al.* The 5300-year-old Helicobacter pylori genome of the Iceman. *Science*
578 **351**, 162–165 (2016).
579 20. Tett, A. *et al.* The Prevotella copri Complex Comprises Four Distinct Clades

580 Underrepresented in Westernized Populations. *Cell Host Microbe* **26**, 666–679.e7
581 (2019).
582 21. Hagan, R. W. *et al.* Comparison of extraction methods for recovering ancient microbial
583 DNA from paleofeces. *Am. J. Phys. Anthropol.* **171**, 275–284 (2020).
584 22. Borry, M. *et al.* CoproID predicts the source of coprolites and paleofeces using
585 microbiome composition and host DNA content. *PeerJ* **8**, e9001 (2020).
586 23. Maixner, F. *et al.* Hallstatt miners consumed blue cheese and beer during the Iron Age
587 and retained a non-Westernized gut microbiome until the Baroque period. *Curr. Biol.* **31**,
588 5149–5162.e6 (2021).
589 24. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and
590 curation of microbial viruses, and evaluation of viral community function from genomic
591 sequences. *Microbiome* **8**, 1–23 (2020).
592 25. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse
593 DNA and RNA viruses. *Microbiome* **9**, 1–13 (2021).
594 26. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG).
595 *Nat. Biotechnol.* **37**, 29–37 (2019).
596 27. Der Sarkissian, C. *et al.* Ancient Metagenomic Studies: Considerations for the Wider
597 Scientific Community. *mSystems* **6**, e0131521 (2021).
598 28. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is
599 enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
600 29. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes
601 within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic*
602 *Acids Res.* (2022) doi:10.1093/nar/gkac1037.
603 30. Walker, P. J. *et al.* Recent changes to virus taxonomy ratified by the International
604 Committee on Taxonomy of Viruses (2022). *Arch. Virol.* **167**, 2429–2440 (2022).
605 31. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human
606 gut microbiome. *Nature Microbiology* **6**, 960–970 (2021).
607 32. Camargo, A. P. *et al.* You can move, but you can't hide: identification of mobile genetic
608 elements with geNomad. *bioRxiv* 2023.03.05.531206 (2023)
609 doi:10.1101/2023.03.05.531206.
610 33. Cornuault, J. K. *et al.* Phages infecting Faecalibacterium prausnitzii belong to novel viral
611 genera that help to decipher intestinal viromes. *Microbiome* **6**, 65 (2018).
612 34. Juge, N. Relationship between mucosa-associated gut microbiota and human diseases.
613 *Biochem. Soc. Trans.* **50**, 1225–1236 (2022).
614 35. Kupczok, A. & Dagan, T. Rates of Molecular Evolution in a Marine Phage Lineage.
615 *Viruses* **11**, (2019).
616 36. Sharifi, F. & Ye, Y. MyDGR: a server for identification and characterization of
617 diversity-generating retroelements. *Nucleic Acids Res.* **47**, W289–W294 (2019).
618 37. Doulatov, S. *et al.* Tropism switching in Bordetella bacteriophage defines a family of
619 diversity-generating retroelements. *Nature* **431**, 476–481 (2004).
620 38. Roux, S. *et al.* Ecology and molecular targets of hypermutation in the global
621 microbiome. *Nat. Commun.* **12**, 3076 (2021).
622 39. Auger, S. *et al.* Gene co-expression network analysis of the human gut commensal
623 bacterium Faecalibacterium prausnitzii in R-Shiny. *PLoS One* **17**, e0271847 (2022).

40. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 10771–10776 (2013).
41. Barr, J. J. *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. Parks, D. H. *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* vol. 50 D785–D794 Preprint at https://doi.org/10.1093/nar/gkab776 (2022).
43. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2020).
44. Bellas, C. M., Schroeder, D. C., Edwards, A., Barker, G. & Anesio, A. M. Flexible genes establish widespread bacteriophage pan-genomes in cryoconite hole ecosystems. *Nat. Commun.* **11**, 1–10 (2020).
45. Marston, M. F. & Martiny, J. B. H. Genomic diversification of marine cyanophages into stable ecotypes. *Environ. Microbiol.* **18**, 4240–4253 (2016).
46. Merrill, B. D. *et al.* Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers Vanishing Gut Microbes. *bioRxiv* (2022) doi:10.1101/2022.03.30.486478.
47. Fellows Yates, J. A. *et al.* Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci Data* **8**, 31 (2021).
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
49. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'kneedle' in a haystack: Detecting knee points in system behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* (IEEE, 2011). doi:10.1109/icdcsw.2011.20.
53. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2020).
54. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* vol. 215 403–410 Preprint at https://doi.org/10.1016/s0022-2836(05)80360-2 (1990).
56. Zielezinski, A., Deorowicz, S. & Gudyś, A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* **38**, 1447–1449 (2021).
57. Wang, W. *et al.* A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics* vol. 2 Preprint at https://doi.org/10.1093/nargab/lqaa044 (2020).

58. Coutinho, F. H. *et al.* RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns (N Y)* **2**, 100274 (2021).
59. Li, S. *et al.* A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience* **25**, (2022).
60. Nishimura, Y. *et al.* ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
61. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
62. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* vol. 12 1268 Preprint at https://doi.org/10.3390/v12111268 (2020).
63. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
64. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**, lqab067 (2021).
65. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
66. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).