Supplementary Figures
for

A prevalent huge phage clade in human and animal gut microbiomes

LinXing Chen[1,*], Antonio Pedro Camargo[2,3], Yiting Qin[1], Eugene V. Koonin[4], Haoyu Wang[5], Yuanqiang Zou[5], Yi Duan[6,7], Hao Li[1]

[1] State Key Laboratory of Advanced Environmental Technology, the Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei, China

[2] Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP 05508-060, Brazil

[3] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

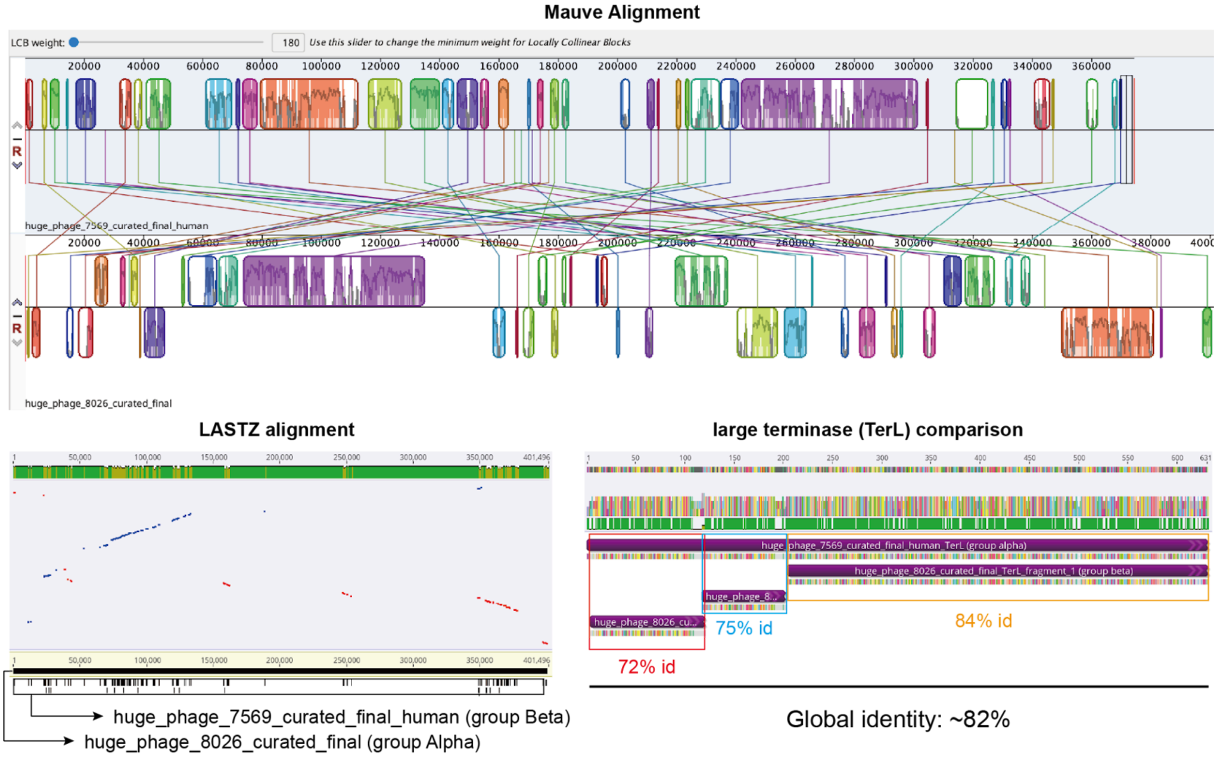
[4] Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[5] BGI-Shenzhen, Shenzhen, China

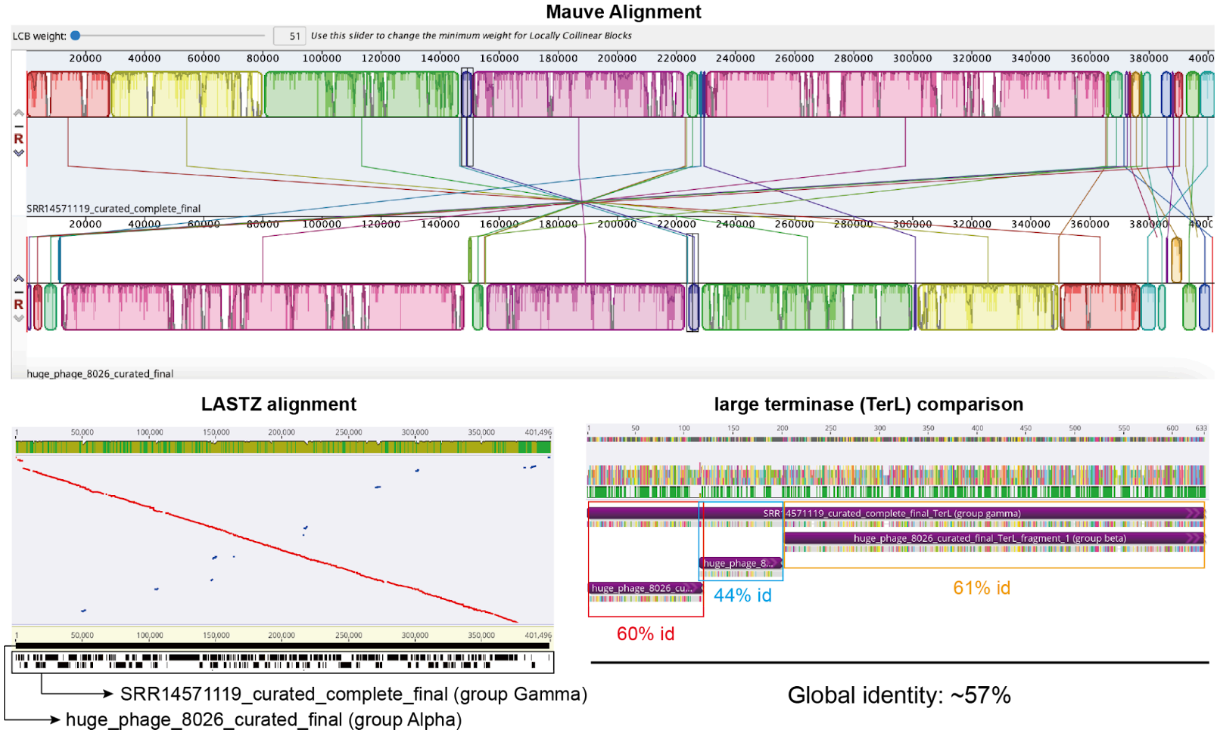
[6] National Key Laboratory of Immune Response and Immunotherapy, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[7] Center for Advanced Interdisciplinary Science and Biomedicine of IHM and Department of Infectious Diseases, The First Affiliated Hospital of USTC, University of Science and Technology of China, Hefei, Anhui, China

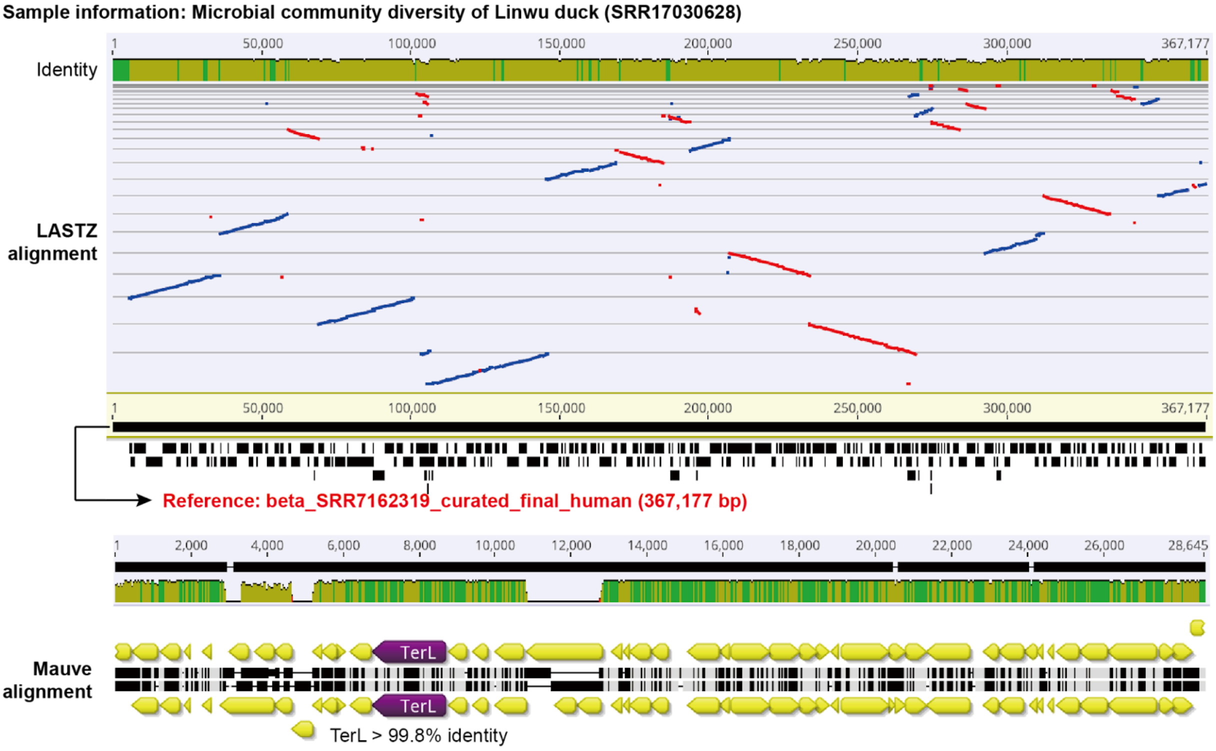
[*]Corresponding author: Linxing Chen, linxingchen@ustc.edu.cn


Mauve Alignment
LCB weight:
180
Use this slider to change the minimum weight for Locally Collinear Blocks
huge_phage_7569_curated_final_human
huge_phage_8026_curated_final
LASTZ alignment
huge_phage_7569_curated_final_human (group Beta)
huge_phage_8026_curated_final (group Alpha)
large terminase (TerL) comparison
huge_phage_7569_curated_final_human_TerL (group alpha)
huge_phage_8026_curated_final_TerL_fragment_1 (group beta)
huge_phage_8...
huge_phage_8026_cu...
72% id
75% id
84% id
Global identity: ~82%

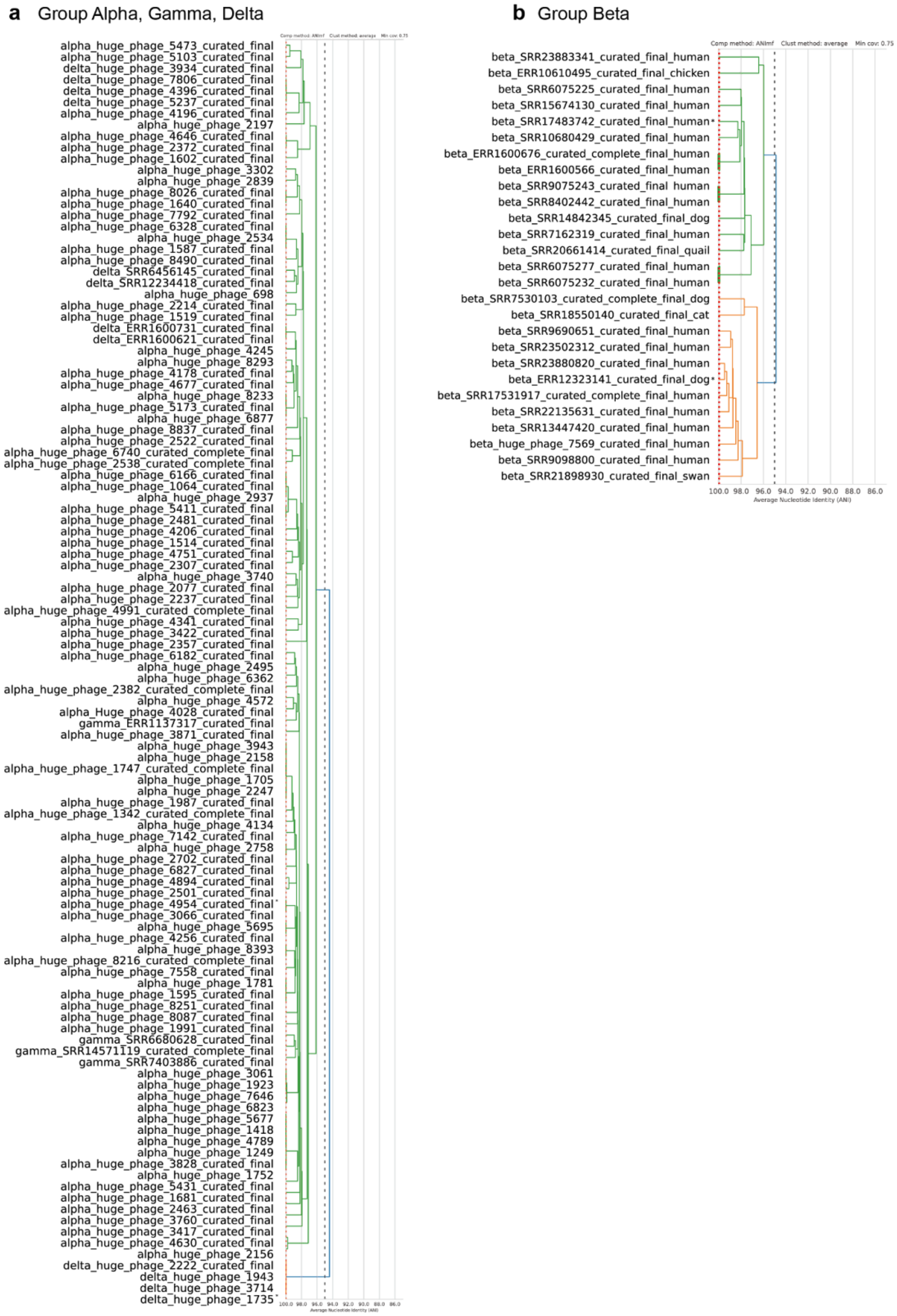
Supplementary Figure 1 | The comparison of groups Alpha and Beta Jug phage genomes. Jug phage genomes of huge_phage_7569 (group Beta) and huge_phage_8026 (group Alpha) were compared. The genome alignment was performed using Mauve (Darling *et al.*, 2004) and LASTZ version 1.04.15 (Harris, 2007) within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). The TerL alignment was conducted using the MUSCLE version 5.1 (Edgar 2004) within Geneious Prime Build, with the protein identity shown.



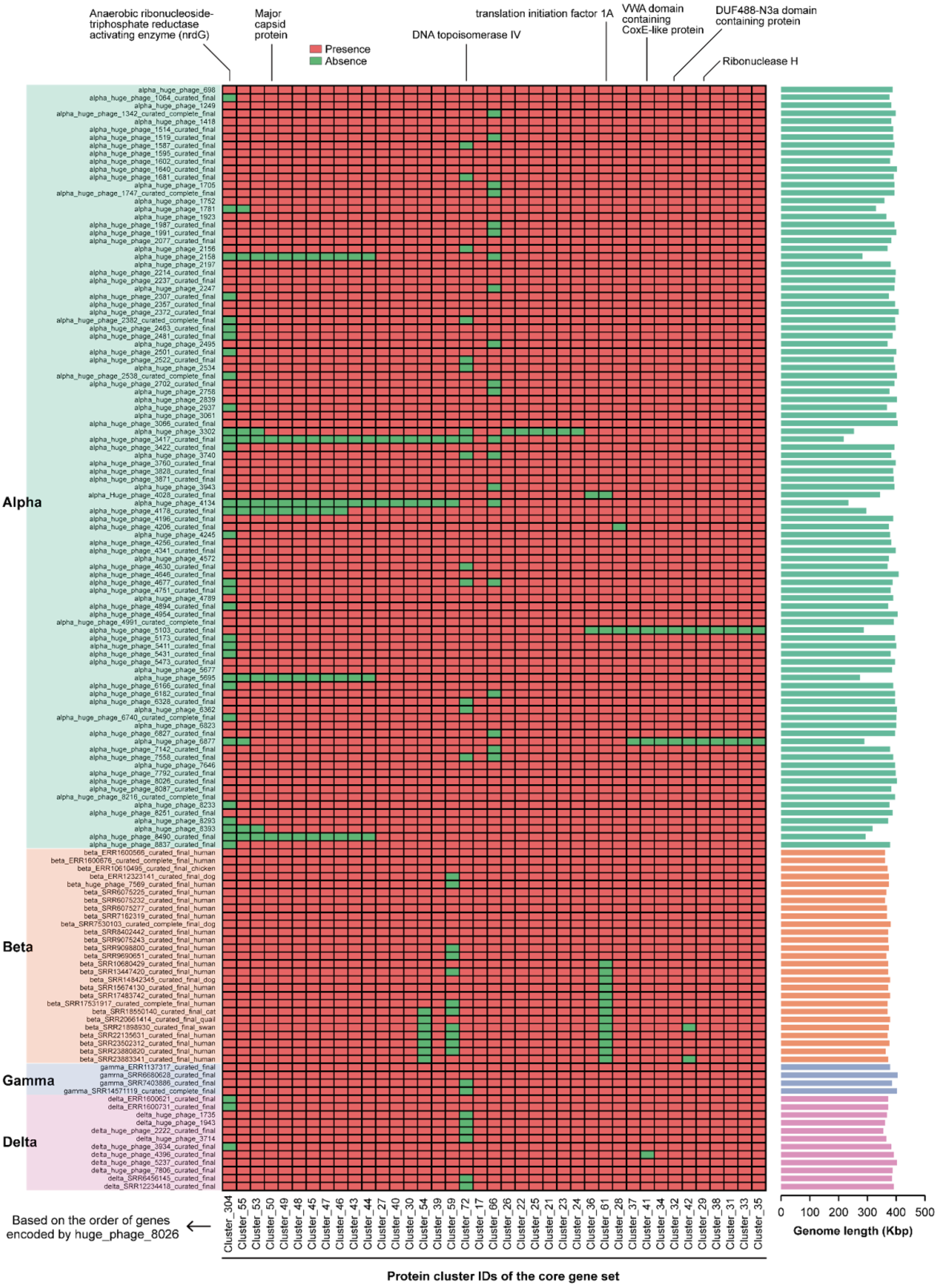
Supplementary Figure 2 | The comparison of groups Alpha and Gamma Jug phage genomes. Jug phages SRR14571119_curated_complete_final (group Gamma) and huge_phage_8026 (group Alpha) were compared. The genome alignment was performed using Mauve (Darling *et al.*, 2004) and LASTZ version 1.04.15 (Harris, 2007) within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). The TerL alignment was conducted using the MUSCLE version 5.1 (Edgar 2004) within Geneious Prime Build, with the protein identity shown.



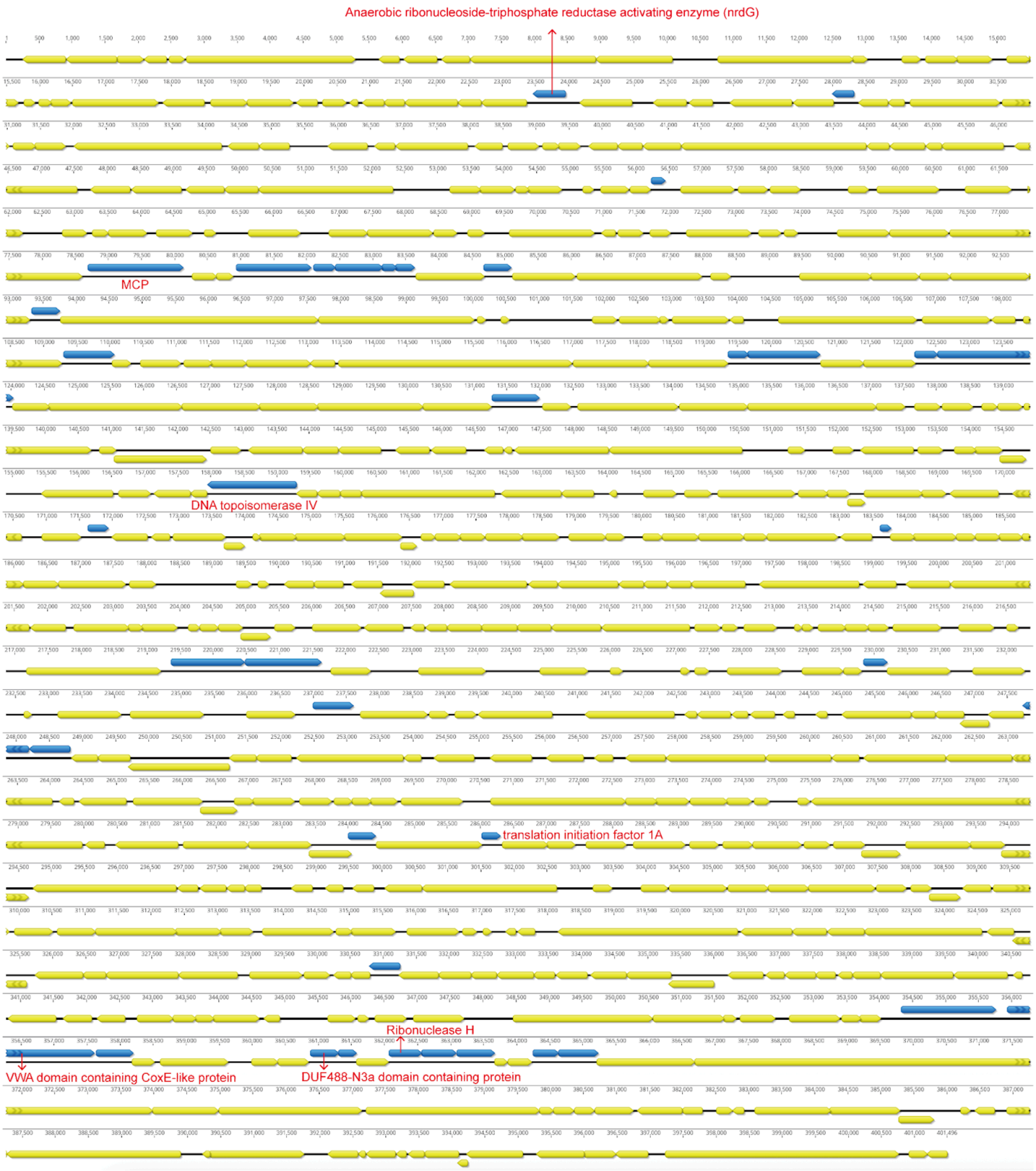
Supplementary Figure 3 | The detection of Jug phages in the gut of ducks. An example of group Beta Jug phage in the gut of a duck. The analysis was performed by aligning the Logan contigs against a curated Jug phage genome using LASTZ version 1.04.15 (Harris, 2007) within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). The Mauve alignments show the zoom-in of the large terminase (TerL) regions, with the TerL identity shown.



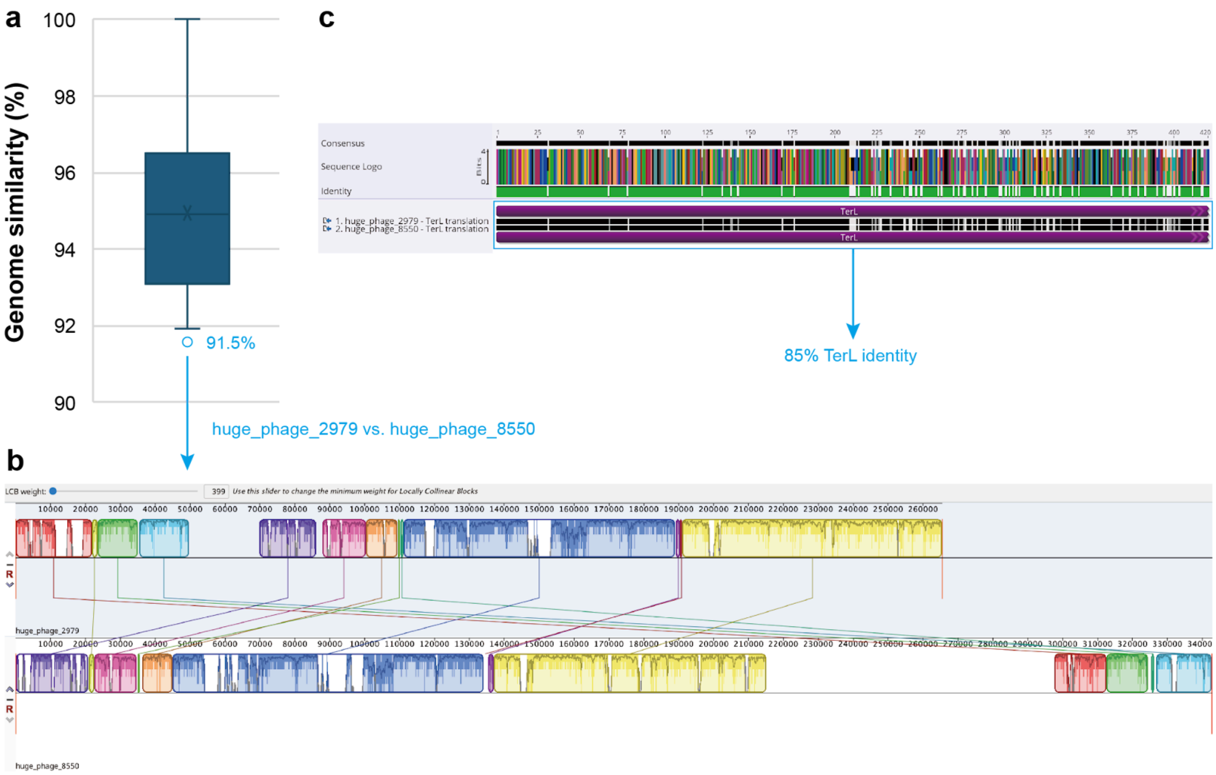
Supplementary Figure 4 | The genome-wide clustering of the 139 curated Jug phage genomes. (a) The genomes of group Alpha, Gamma, and Delta were clustered together, and (b) those of group Beta were clustered together. The clustering was performed using dRep version 3.2.2 (Olm *et al.*, 2017), with the parameters of "-sa 0.95 -nc 0.75".



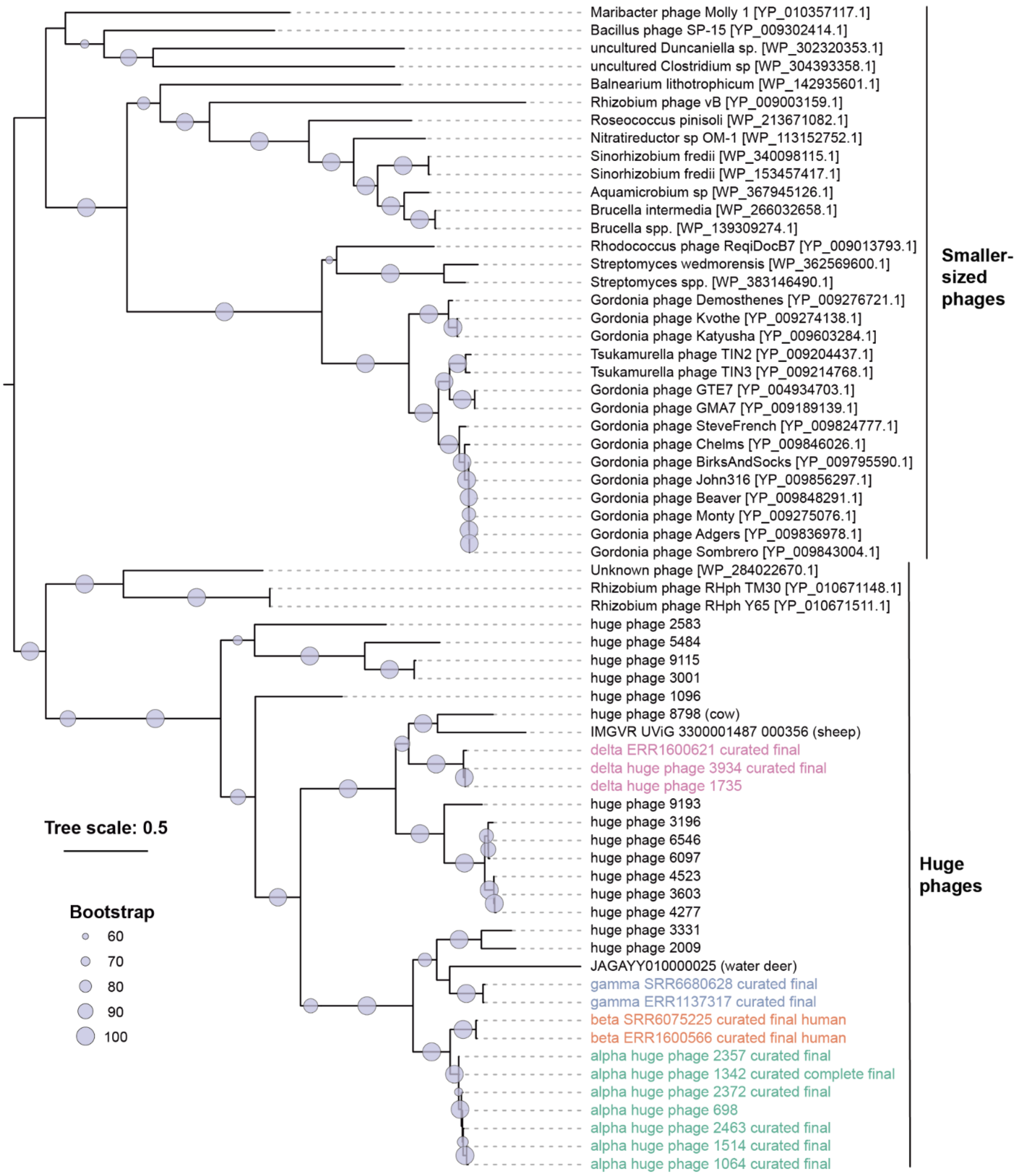
Supplementary Figure 5 | The distribution of the 39 single-copy core gene set across the 139 curated Jug phage genomes. The annotation of each gene is shown on the top if available; all those not shown are hypothetical proteins. The length of the genomes is shown to the right. Some of the genomes in the Alpha group lack many of the core gene set, which is likely due to their lower genome completeness as reflected by their genome length.


Anaerobic ribonucleoside-triphosphate reductase activating enzyme (nrdG)
MCP
DNA topoisomerase IV
translation initiation factor 1A
Ribonuclease H
VWA domain containing CoxE-like protein
DUF488-N3a domain containing protein

Supplementary Figure 6 | The distribution of the 39 single-copy genes shared by all four Jug phage groups. The protein-coding genes from all 139 curated Jug phage genomes were clustered using CD-HIT version 4.8.1 (Li and Godzik, 2006) with the parameters of "-c 0.7 -aS 0.9 -G 0". Those clusters with only one member for each genome, and including at least half of the genomes of each group, were defined as single-copy gene clusters. See Supplementary Table 7 for details of the genes. The genome of Huge_phage_8026 is used as an example to show. The genes with annotations are highlighted in red. The single-copy genes are shown in blue, and all other genes are shown in yellow. The figure was generated using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).

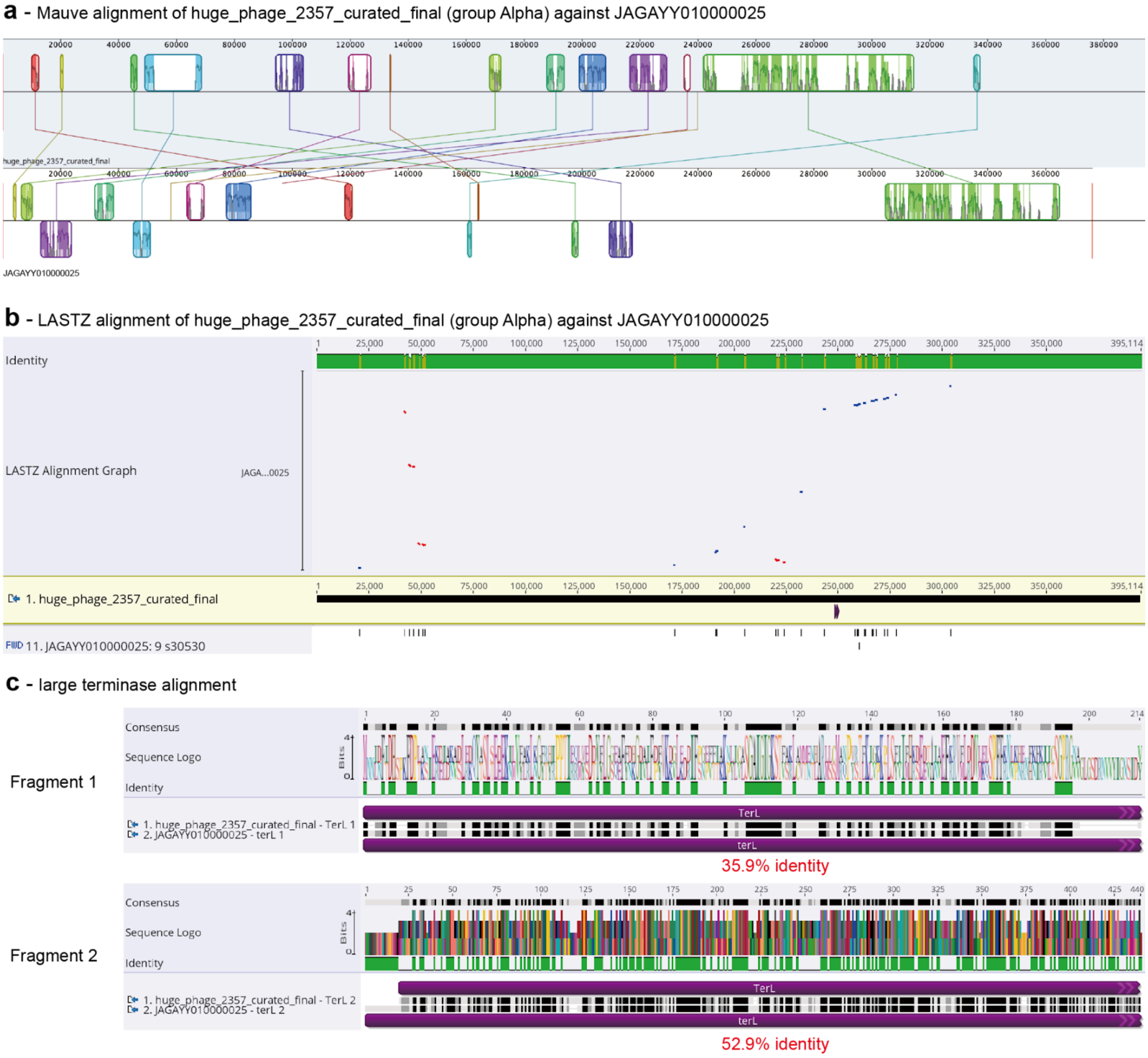


Supplementary Figure 7 | The only other HPGC cluster had a paired TerL protein identity <90%, except for the Jug phage cluster. The clusters were defined at ⩾90% genome-wide similarity. (a) The distribution of the paired genome-wide similarity within the cluster. (b) The whole-genome alignment of the two genomes (huge_phage_2979 and huge_phage_8550) in the cluster shared only 91.5% genome-wide similarity. The alignment was performed using Mauve (Darling *et al.*, 2004) within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). (c) The alignment and identity of the TerL proteins of the two genomes. The alignment was conducted using MUSCLE version 5.1 (Edgar 2004) within Geneious Prime Build 2025-03-24.



Maribacter phage Molly 1 [YP_010357117.1]
Bacillus phage SP-15 [YP_009302414.1]
uncultured Duncaniella sp. [WP_302320353.1]
uncultured Clostridium sp [WP_304393358.1]
Balnearium lithotrophicum [WP_142935601.1]
Rhizobium phage vB [YP_009003159.1]
Roseococcus pinisoli [WP_213671082.1]
Nitratireductor sp OM-1 [WP_113152752.1]
Sinorhizobium fredii [WP_340098115.1]
Sinorhizobium fredii [WP_153457417.1]
Aquamicrobium sp [WP_367945126.1]
Brucella intermedia [WP_266032658.1]
Brucella spp. [WP_139309274.1]
Rhodococcus phage ReqiDocB7 [YP_009013793.1]
Streptomyces wedmorensis [WP_362569600.1]
Streptomyces spp. [WP_383146490.1]
Gordonia phage Demosthenes [YP_009276721.1]
Gordonia phage Kvothe [YP_009274138.1]
Gordonia phage Katyusha [YP_009603284.1]
Tsukamurella phage TIN2 [YP_009204437.1]
Tsukamurella phage TIN3 [YP_009214768.1]
Gordonia phage GTE7 [YP_004934703.1]
Gordonia phage GMA7 [YP_009189139.1]
Gordonia phage SteveFrench [YP_009824777.1]
Gordonia phage Chelms [YP_009846026.1]
Gordonia phage BirksAndSocks [YP_009795590.1]
Gordonia phage John316 [YP_009856297.1]
Gordonia phage Beaver [YP_009848291.1]
Gordonia phage Monty [YP_009275076.1]
Gordonia phage Adgers [YP_009836978.1]
Gordonia phage Sombrero [YP_009843004.1]
Smaller-sized phages
Unknown phage [WP_284022670.1]
Rhizobium phage RHph TM30 [YP_010671148.1]
Rhizobium phage RHph Y65 [YP_010671511.1]
huge phage 2583
huge phage 5484
huge phage 9115
huge phage 3001
huge phage 1096
huge phage 8798 (cow)
IMGVR UViG 3300001487 000356 (sheep)
delta ERR1600621 curated final
delta huge phage 3934 curated final
delta huge phage 1735
huge phage 9193
huge phage 3196
huge phage 6546
huge phage 6097
huge phage 4523
huge phage 3603
huge phage 4277
huge phage 3331
huge phage 2009
JAGAYY010000025 (water deer)
gamma SRR6680628 curated final
gamma ERR1137317 curated final
beta SRR6075225 curated final human
beta ERR1600566 curated final human
alpha huge phage 2357 curated final
alpha huge phage 1342 curated complete final
alpha huge phage 2372 curated final
alpha huge phage 698
alpha huge phage 2463 curated final
alpha huge phage 1514 curated final
alpha huge phage 1064 curated final
Huge phages
Tree scale: 0.5
Bootstrap
60
70
80
90
100

Supplementary Figure 8 | The full phylogeny of Jug phages and related viruses based on the large terminase (TerL). The related viruses were from NCBI by searching the TerLs of Jug phages against NCBI RefSeq and IMG/VR v4 using BLASTp. Only those hit TerLs with a minimum identity of 20% to Jug phage TerL were included (all of the identities are 20-30%). The TerL protein sequences were aligned using MUSCLE version 3.8.31 (Edgar, 2004), and trimmed using trimAl version v1.4.rev22 (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) to remove columns containing ⩾90% of gaps. The phylogeny was constructed using IQ-TREE multicore version 2.4.0 (Minh *et al.*, 2020) with the parameters of "-bb 1000 -m LG+G4". The phylogenetic tree was visualized in iTOL v7 (Letunic and Bork, 2019). The TerL fragments of Jug phages were concatenated and clustered using CD-HIT version 4.8.1(Huang *et al.*, 2010) with 99% identity, and only the representative sequences were included; the numbers in the brackets indicate the count of TerLs in the corresponding clusters.

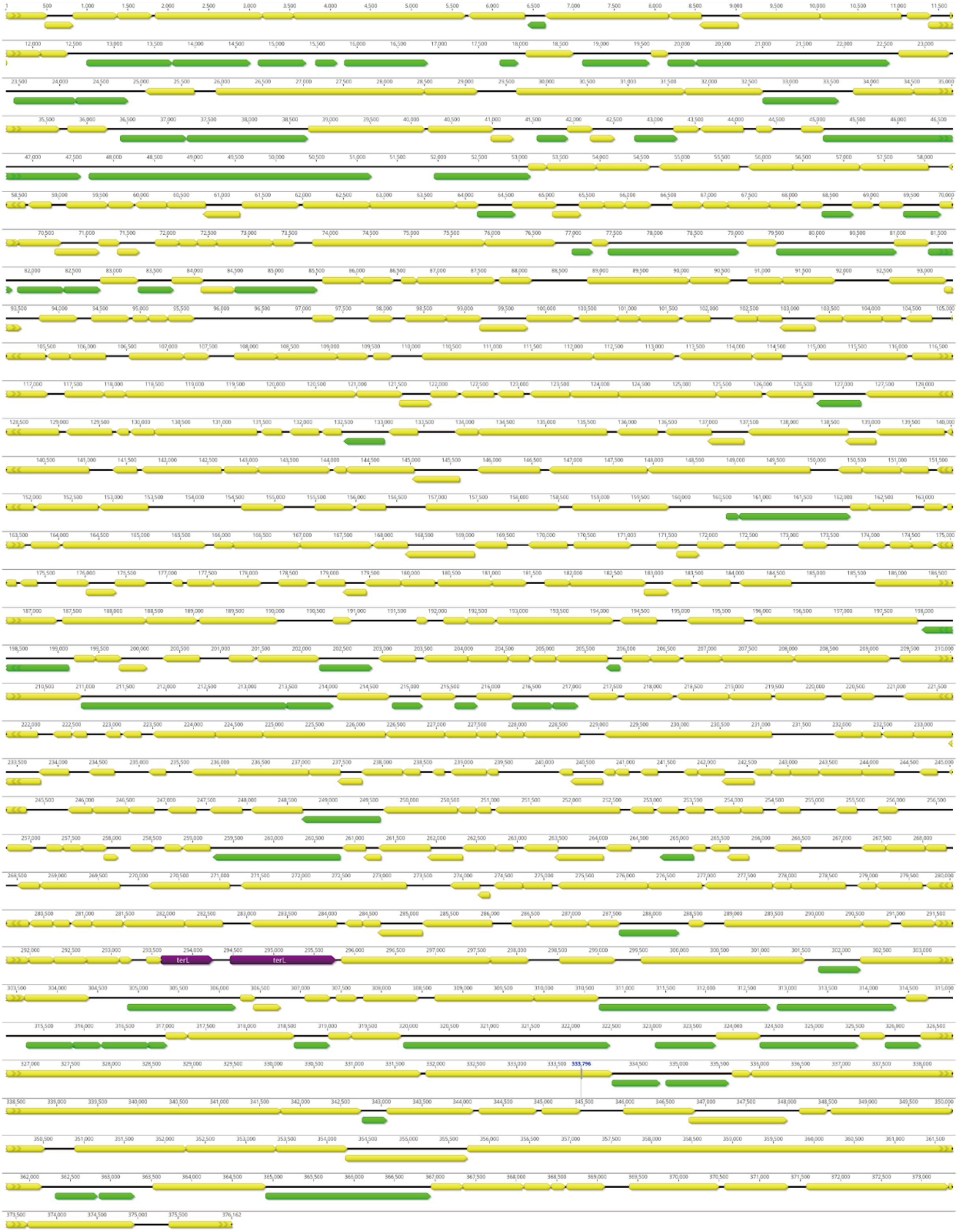


a - Mauve alignment of huge_phage_2357_curated_final (group Alpha) against JAGAYY010000025

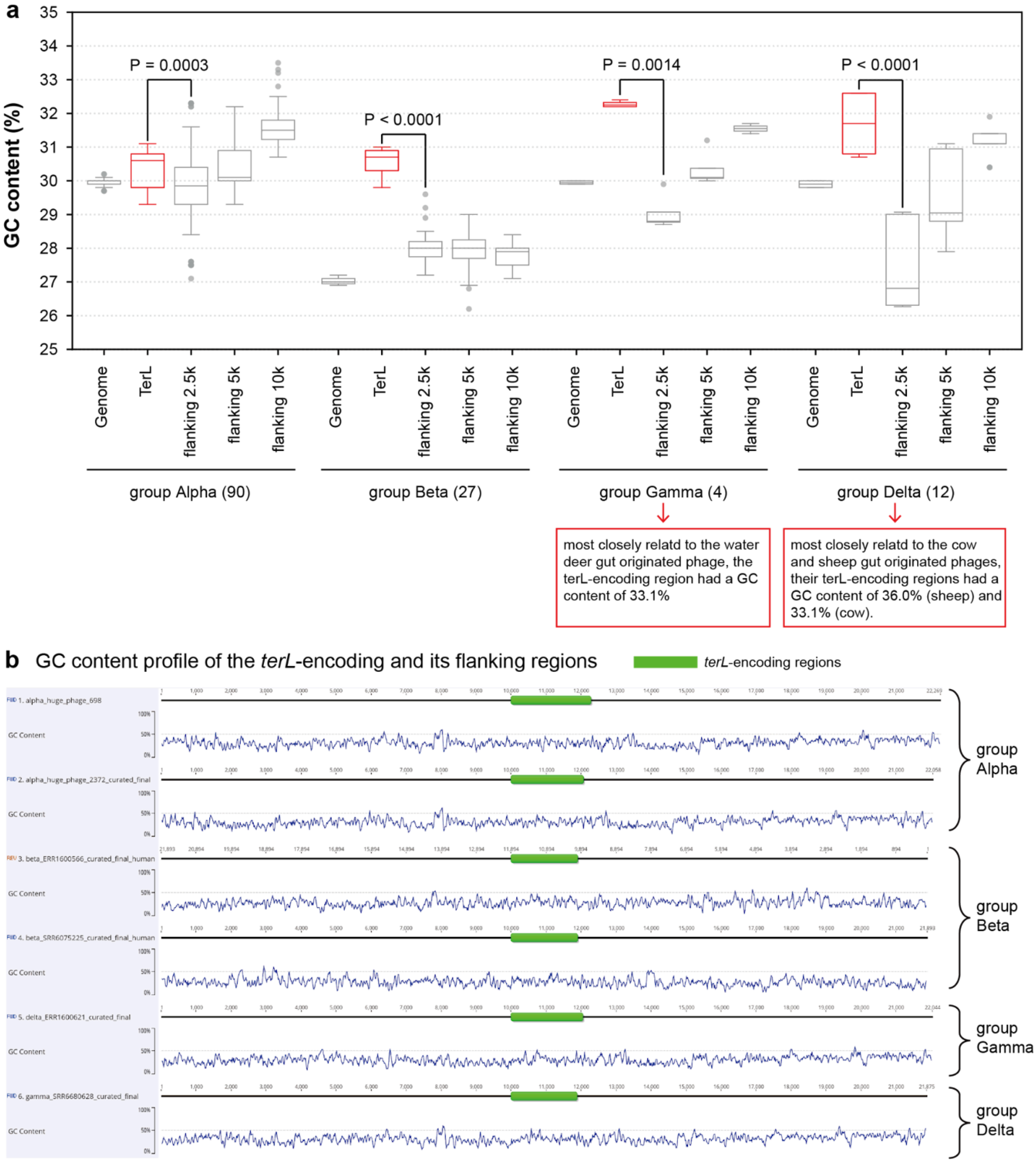
b - LASTZ alignment of huge_phage_2357_curated_final (group Alpha) against JAGAYY010000025

c - large terminase alignment

Supplementary Figure 9 | The comparison of a phage genome (NCBI id = JAGAYY010000025) from the gut of water deer against the Jug phage. When performed online BLASTp searches of Jug phage protein against NCBI-nr, we found that many of the Jug phage proteins had >50% identity to those from a metagenome-assembled genome (MAG), whose name is "Bacilli bacterium isolate RGIG8901 Water_deer_Omasum.Co__c192626, whole genome shotgun sequence". All the targeted hits were from a single contig in this MAG, JAGAYY010000025 (376,162 bp in length). We downloaded the sequence of JAGAYY010000025 and evaluated it using geNomad version 1.5.2 (Camargo *et al.*, 2023) and found that it represented a circular phage genome (with an end overlap sequence of 141 bp in length). This may indicate a misbinning of the phage contig into the MAG. The protein-coding genes of JAGAYY010000025 were predicted using Prodigal version 2.6.3 (Hyatt *et al.*, 2010) using the "single" mode, and searched against the proteins of the 139 curated Jug phages. Two large terminase (TerL) fragments were detected for JAGAYY010000025, and the longer one is most similar to that of huge_phage_2357_curated_final (group Alpha). We thus compared the genomes of huge_phage_2357_curated_final and JAGAYY010000025. (a) The Mauve alignment of the two genomes. It was analyzed using Mauve (Darling *et al.*, 2004). (b) The LASTZ alignment of the two genomes. It is analyzed using LASTZ version 1.04.15 (Harris, 2007). (c) The TerL alignments, which were conducted using MUSCLE version 5.1. All the figures were generated within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).



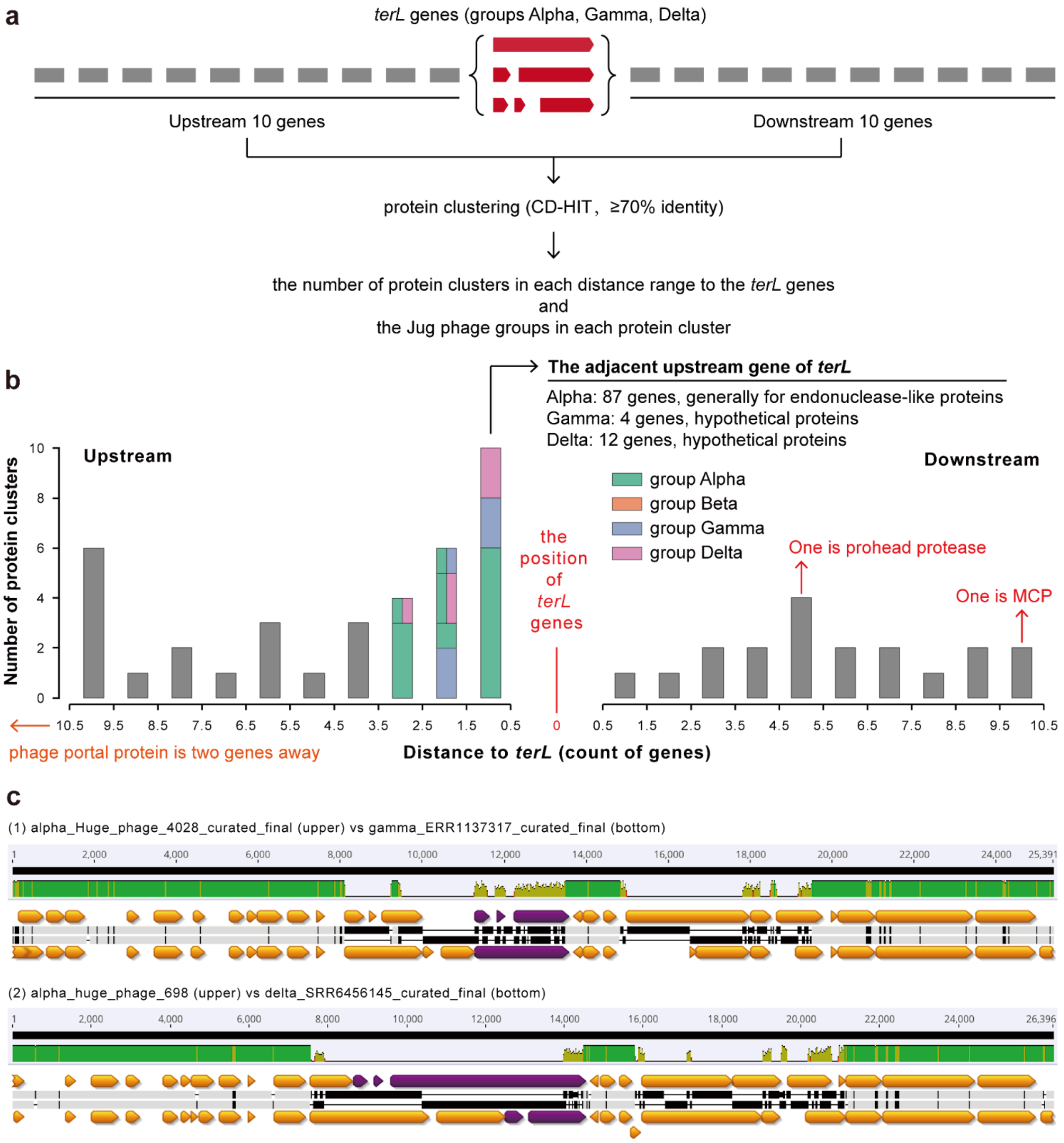
Supplementary Figure 10 | The protein-coding genes of JAGAYY010000025. The protein-coding genes shared >50% identity with those from any curated Jug phages are shown in purple (the two large terminase fragments) or green, with all other genes shown in yellow. The comparison of JAGAYY010000025 against the Jug phage protein-coding genes was performed using local BLASTp. The figure was generated within Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).



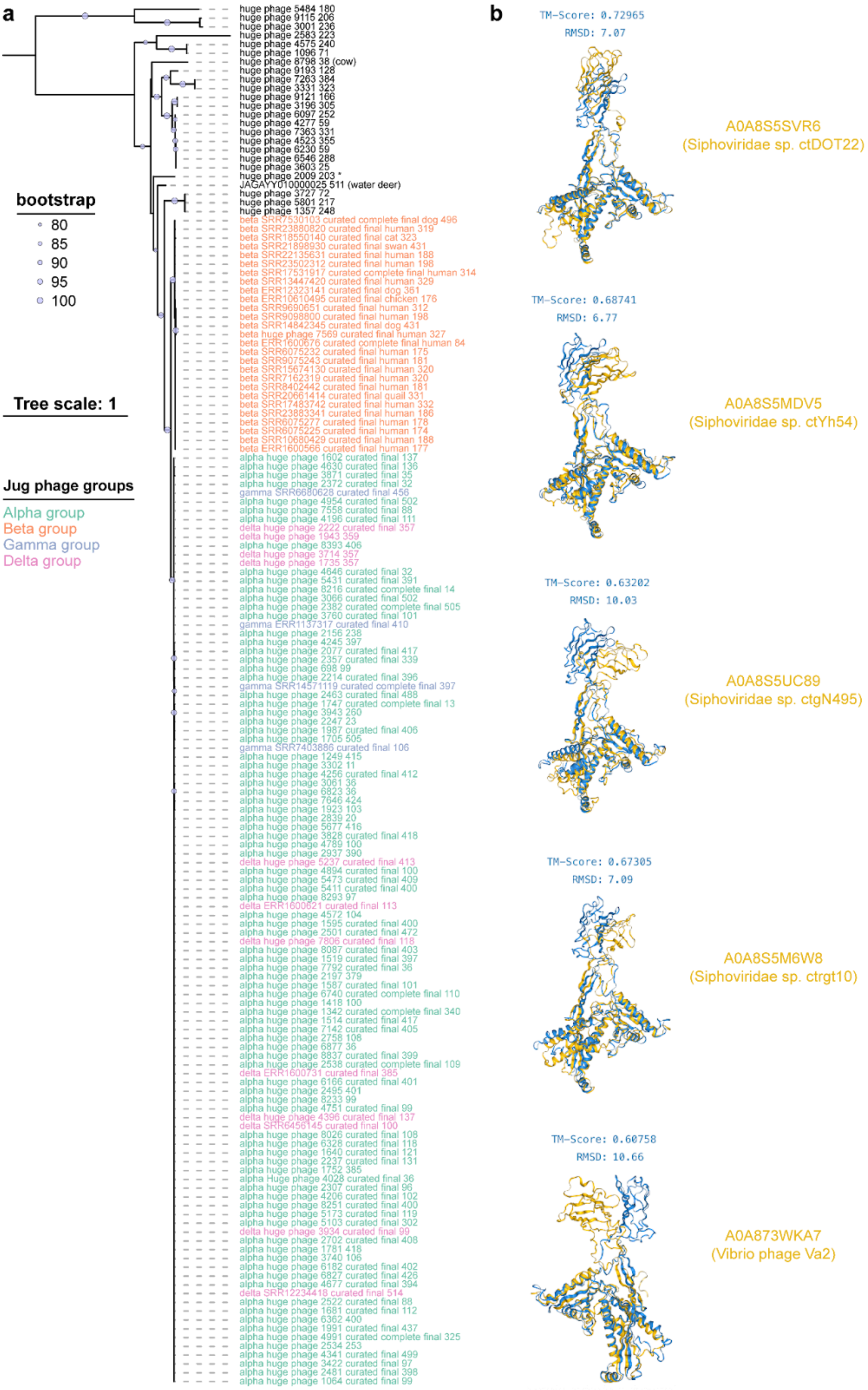
most closely relatd to the water deer gut originated phage, the terL-encoding region had a GC content of 33.1%

most closely relatd to the cow and sheep gut originated phages, their terL-encoding regions had a GC content of 36.0% (sheep) and 33.1% (cow).

Supplementary Figure 11 | The GC contents of the whole genome, TerL-encoding, and TerL-flanking regions. (a) The comparison of GC contents. The flanking regions were extracted from the genomes based on the starting and ending positions of all the *terL* fragments. The significant difference in GC contents between the *terL*-encoding region and the flanking 2.5 kbp region was tested by a paired t-test. The flanking region, for example, 2.5 kbp, is defined as the concatenated nucleotide sequence of the upstream 2.5 kbp and downstream 2.5 kbp of the *terL*-encoding region, thus 5 kbp in total. (b) The examples of genomes from all four Jug phage groups show the GC content of the *terL*-encoding and its flanking regions.

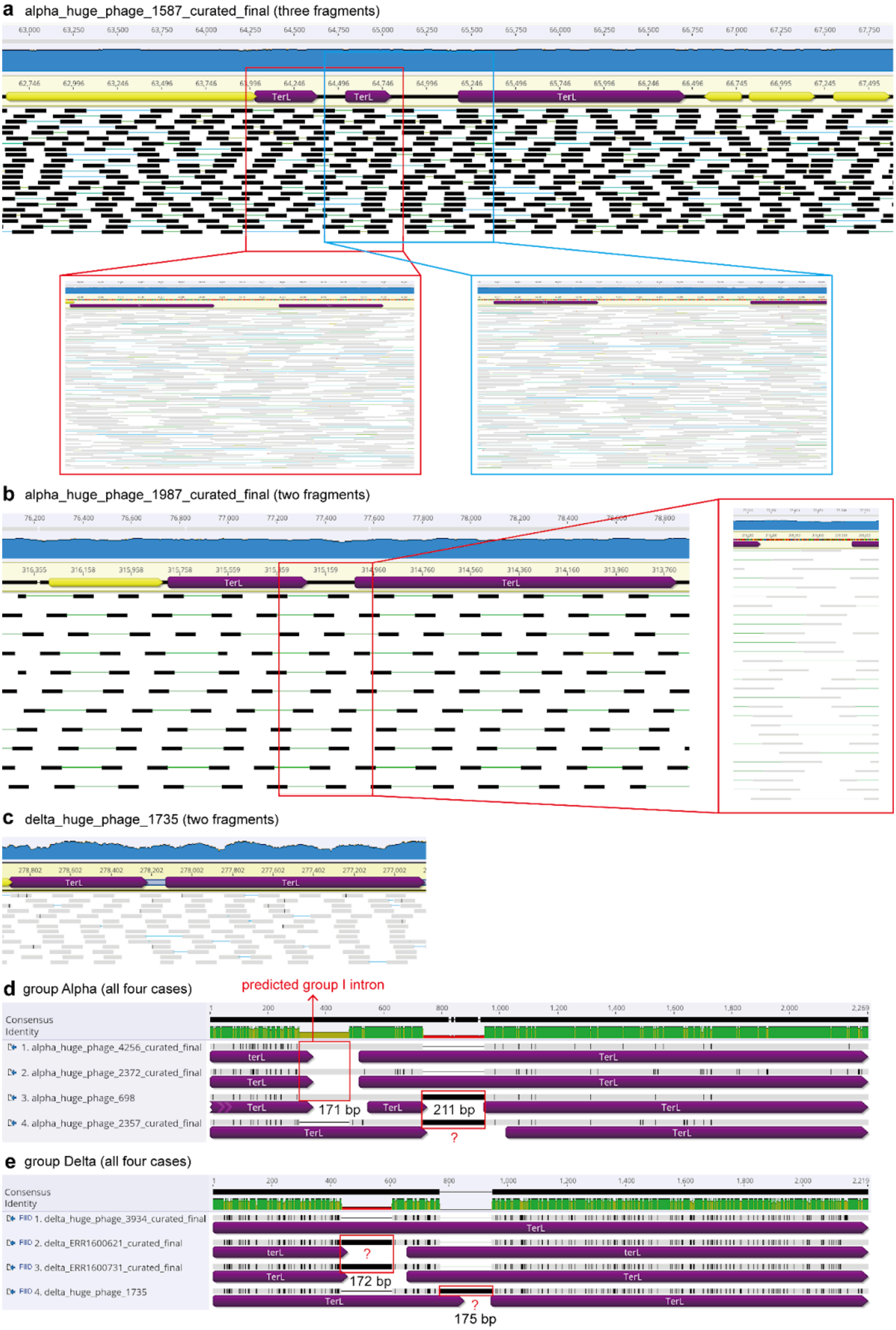


Supplementary Figure 12 | The neighbor genes of the large terminase (*terL*) genes. (a) The diagram shows the analysis pipeline of the neighborhood genes. The CD-HIT version 4.8.1 [66] was used to cluster the genes, with ⩾70% identity (-c 0.7 -aS 0.9 -G 0). Then, in each of the distance ranges, the number of protein clusters was summarized and presented in (b). (b) The number of protein clusters in each distance range to the *terL* genes. As we found that the protein clusters in the upstream distance range of 0.5-3.5 genes were very distinct, we thus profiled the corresponding Jug phage groups for each cluster. Notably, the upstream distance range of 0.5-1.5 genes contained 10 protein clusters, and those of group Alpha members generally encoded for endonuclease-like proteins, while those of groups Gamma and Delta were for hypothetical proteins. One of the protein clusters is for the major capsid protein (MCP). (c) The nucleotide alignment of the *terL*-coding regions of Jug phage genomes from different groups. Examples of (1) groups Alpha and Gamma, and (2) groups Alpha and Delta are shown. The TerL fragments are indicated in purple.


a
bootstrap
80
85
90
95
100
Tree scale: 1
Jug phage groups
Alpha group
Beta group
Gamma group
Delta group
huge phage 5484 180
huge phage 9115 206
huge phage 3001 236
huge phage 2583 223
huge phage 4575 240
huge phage 1096 71
huge phage 8798 38 (cow)
huge phage 9193 128
huge phage 7263 384
huge phage 3331 323
huge phage 9121 166
huge phage 3196 305
huge phage 6097 252
huge phage 4277 59
huge phage 7363 331
huge phage 4523 355
huge phage 6230 59
huge phage 6546 288
huge phage 3603 25
huge phage 2009 203 *
JAGAYY010000025 511 (water deer)
huge phage 3727 72
huge phage 5801 217
huge phage 1357 248
beta SRR7530103 curated complete final dog 496
beta SRR23880820 curated final human 319
beta SRR18550140 curated final cat 323
beta SRR21898930 curated final swan 431
beta SRR22135631 curated final human 188
beta SRR23502312 curated final human 198
beta SRR17531917 curated complete final human 314
beta SRR13447420 curated final human 329
beta ERR12323141 curated final dog 361
beta ERR10610495 curated final chicken 176
beta SRR9690651 curated final human 312
beta SRR9098800 curated final human 198
beta SRR14842345 curated final dog 431
beta huge phage 7569 curated final human 327
beta ERR1600676 curated complete final human 84
beta SRR6075232 curated final human 175
beta SRR9075243 curated final human 181
beta SRR15674130 curated final human 320
beta SRR7162319 curated final human 320
beta SRR8402442 curated final human 181
beta SRR20661414 curated final quail 331
beta SRR17483742 curated final human 332
beta SRR23883341 curated final human 186
beta SRR6075277 curated final human 178
beta SRR6075225 curated final human 174
beta SRR10680429 curated final human 188
beta ERR1600566 curated final human 177
alpha huge phage 1602 curated final 137
alpha huge phage 4630 curated final 136
alpha huge phage 3871 curated final 35
alpha huge phage 2372 curated final 32
gamma SRR6680628 curated final 456
alpha huge phage 4954 curated final 502
alpha huge phage 7558 curated final 88
alpha huge phage 4196 curated final 111
delta huge phage 2222 curated final 357
delta huge phage 1943 359
alpha huge phage 8393 406
delta huge phage 3714 357
delta huge phage 1735 357
alpha huge phage 4646 curated final 32
alpha huge phage 5431 curated final 391
alpha huge phage 8216 curated complete final 14
alpha huge phage 3066 curated final 502
alpha huge phage 2382 curated complete final 505
alpha huge phage 3760 curated final 101
gamma ERR1137317 curated final 410
alpha huge phage 2156 238
alpha huge phage 4245 397
alpha huge phage 2077 curated final 417
alpha huge phage 2357 curated final 339
alpha huge phage 698 99
alpha huge phage 2214 curated final 396
gamma SRR14571119 curated complete final 397
alpha huge phage 2463 curated final 488
alpha huge phage 1747 curated complete final 13
alpha huge phage 3943 260
alpha huge phage 2247 23
alpha huge phage 1987 curated final 406
alpha huge phage 1705 505
gamma SRR7403886 curated final 106
alpha huge phage 1249 415
alpha huge phage 3302 11
alpha huge phage 4256 curated final 412
alpha huge phage 3061 36
alpha huge phage 6823 36
alpha huge phage 7646 424
alpha huge phage 1923 103
alpha huge phage 2839 20
alpha huge phage 5677 416
alpha huge phage 3828 curated final 418
alpha huge phage 4789 100
alpha huge phage 2937 390
delta huge phage 5237 curated final 413
alpha huge phage 4894 curated final 100
alpha huge phage 5473 curated final 409
alpha huge phage 5411 curated final 400
alpha huge phage 8293 97
delta ERR1600621 curated final 113
alpha huge phage 4572 104
alpha huge phage 1595 curated final 400
alpha huge phage 2501 curated final 472
delta huge phage 7806 curated final 118
alpha huge phage 8087 curated final 403
alpha huge phage 1519 curated final 397
alpha huge phage 7792 curated final 36
alpha huge phage 2197 379
alpha huge phage 1587 curated final 101
alpha huge phage 6740 curated complete final 110
alpha huge phage 1418 100
alpha huge phage 1342 curated complete final 340
alpha huge phage 1514 curated final 417
alpha huge phage 7142 curated final 405
alpha huge phage 2758 108
alpha huge phage 6877 36
alpha huge phage 8837 curated final 399
alpha huge phage 2538 curated complete final 109
delta ERR1600731 curated final 385
alpha huge phage 6166 curated final 401
alpha huge phage 2495 401
alpha huge phage 8233 99
alpha huge phage 4751 curated final 99
delta huge phage 4396 curated final 137
delta SRR6456145 curated final 100
alpha huge phage 8026 curated final 108
alpha huge phage 6328 curated final 118
alpha huge phage 1640 curated final 121
alpha huge phage 2237 curated final 131
alpha huge phage 1752 385
alpha Huge phage 4028 curated final 36
alpha huge phage 2307 curated final 96
alpha huge phage 4206 curated final 102
alpha huge phage 8251 curated final 400
alpha huge phage 5173 curated final 119
alpha huge phage 5103 curated final 302
delta huge phage 3934 curated final 99
alpha huge phage 2702 curated final 408
alpha huge phage 1781 418
alpha huge phage 3740 106
alpha huge phage 6182 curated final 402
alpha huge phage 6827 curated final 426
alpha huge phage 4677 curated final 394
delta SRR12234418 curated final 514
alpha huge phage 2522 curated final 88
alpha huge phage 1681 curated final 112
alpha huge phage 6362 400
alpha huge phage 1991 curated final 437
alpha huge phage 4991 curated complete final 325
alpha huge phage 2534 253
alpha huge phage 4341 curated final 499
alpha huge phage 3422 curated final 97
alpha huge phage 2481 curated final 398
alpha huge phage 1064 curated final 99
b
TM-Score: 0.72965
RMSD: 7.07
A0A8S5SVR6
(Siphoviridae sp. ctDOT22)
TM-Score: 0.68741
RMSD: 6.77
A0A8S5MDV5
(Siphoviridae sp. ctYh54)
TM-Score: 0.63202
RMSD: 10.03
A0A8S5UC89
(Siphoviridae sp. ctgN495)
TM-Score: 0.67305
RMSD: 7.09
A0A8S5M6W8
(Siphoviridae sp. ctrgt10)
TM-Score: 0.60758
RMSD: 10.66
A0A873WKA7
(Vibrio phage Va2)

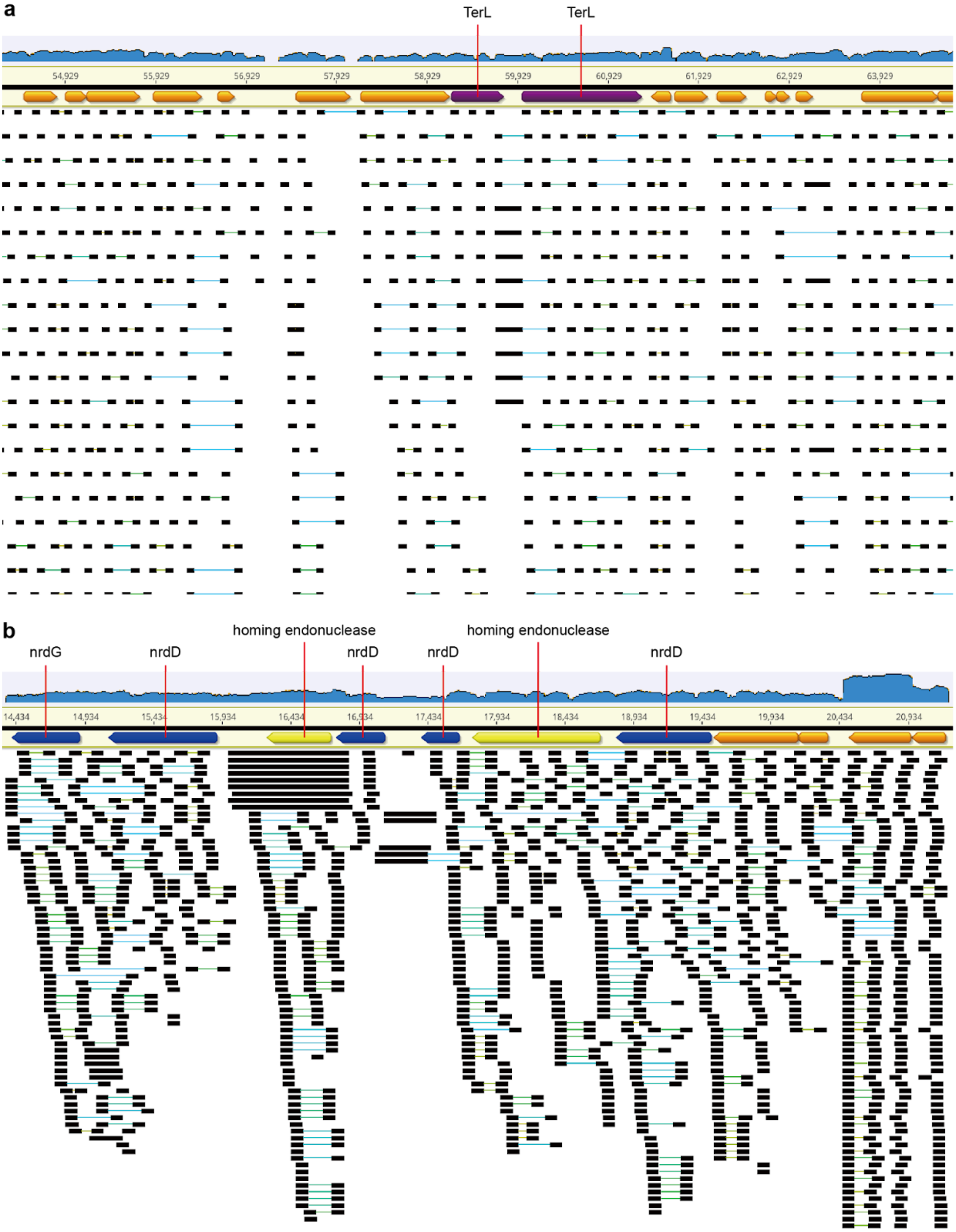
See the text legend on the next page

Supplementary Figure 13 | The major capsid proteins of Jug phages. (a) The phylogeny of the major capsid proteins. All the major capsid proteins (MCPs) from Jug phages and their relatives were included for analysis. The MCPs encoded by relatives were retrieved by searching all other HPGC proteins against the Jug phages' MCPs using BLASTp with an e-value threshold of 1e-5. The MCP sequence of the phage detected in the water deer gut microbiome was included as well. All the Jug phage relatives were from the gut microbiomes, except for the one from a wastewater sample (indicated by *). (b) The comparison of the MCP structures. The top five BFVD MCP structure hits with the highest TM-scores against Jug phages' MCP are shown. The BFVD/UniProt IDs of the reference structures and the corresponding viruses are shown on the right.

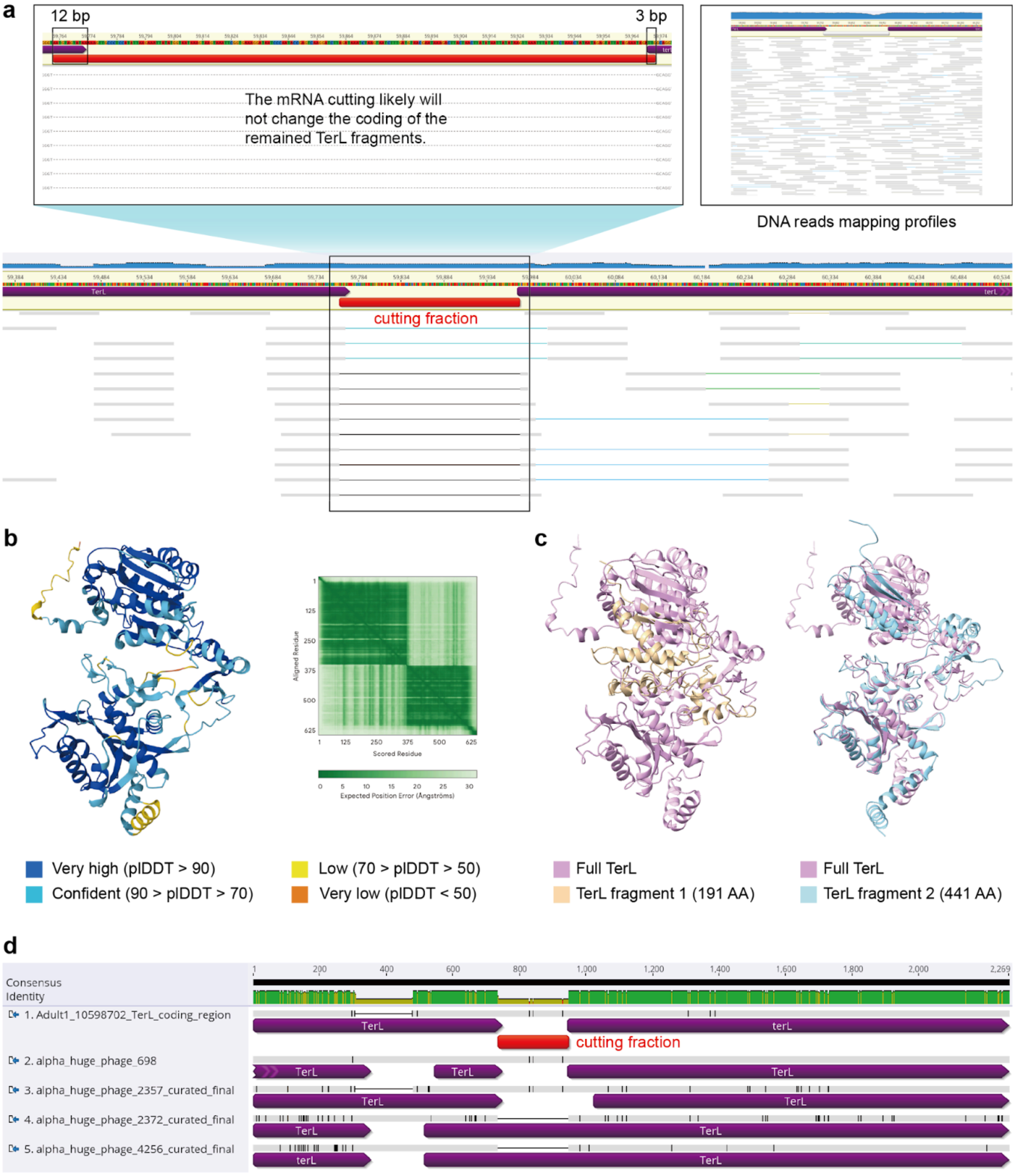


See the text legend on the next page

Supplementary Figure 14 | The fragmentation of the large terminase (*TerL*) genes encoded by Jug phages. In (a), (b), and (c), examples of paired-end reads mapping to the *TerL*-encoding regions did not show any single-nucleotide indel that led to the fragmentation of the *TerLs*. The zoom-in of the mapped paired-end reads is shown in the red and/or blue boxes. In (d) and (e), alignment of the nucleotide sequences of the *TerL*-encoding regions likely indicated that the insertion of large nucleotide sequences could be the potential reason for *TerL* fragmentation. The nucleotide length of the introns is shown. The mapped was performed using Bowtie2 version 2.4.4 (Longmead and Salzberg, 2012) with default parameters, and the figures were generated using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).

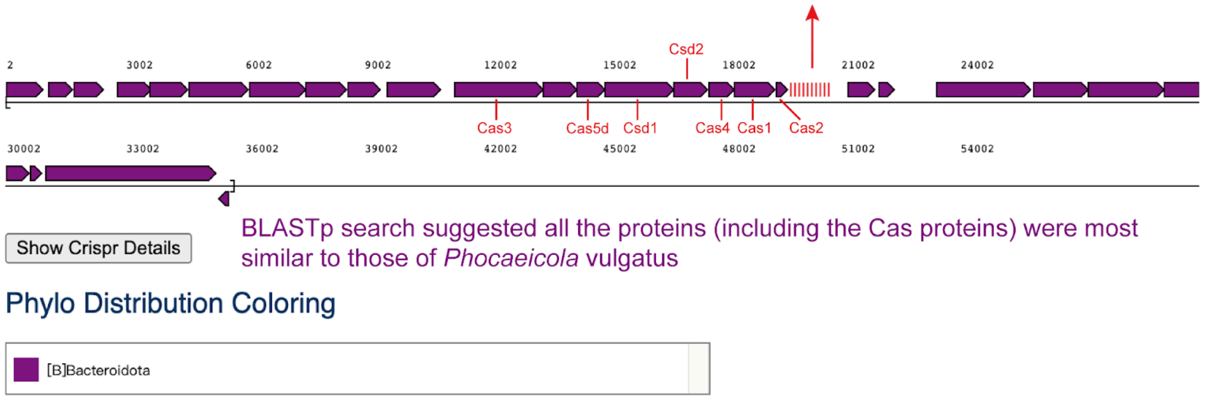


Supplementary Figure 15 | The co-transcription of fragmented genes as operons, with (a) the large terminase (TerL) and (b) the nrdD as examples. The RNA reads from the adult male 1 samples were mapped to the corresponding Jug phage genome. The RNA reads mapping bam file was imported and visualized using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).


a
12 bp
3 bp
The mRNA cutting likely will not change the coding of the remained TerL fragments.
DNA reads mapping profiles
TerL
cutting fraction
terL
b
Aligned Residue
Scored Residue
Expected Position Error (Ångströms)
🟦 Very high (pIDDT > 90)
🟦 Confident (90 > pIDDT > 70)
🟨 Low (70 > pIDDT > 50)
🟧 Very low (pIDDT < 50)
c
Full TerL
TerL fragment 1 (191 AA)
Full TerL
TerL fragment 2 (441 AA)
d
Consensus
Identity
1. Adult1_10598702_TerL_coding_region
2. alpha_huge_phage_698
3. alpha_huge_phage_2357_curated_final
4. alpha_huge_phage_2372_curated_final
5. alpha_huge_phage_4256_curated_final
TerL
terL
cutting fraction
1
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,269

Supplementary Figure 16 | An example showing the mRNA cutting and splicing to obtain a full-length TerL. The terL fragments of Jug phage from adult 1 are used as an example here. (a) The mRNA cutting won't change the coding of the remained fragments. The DNA reads mapping profiles are shown on the right, to indicate that the RNA mapping pattern was not due to a local variation in which some phage genomes lacked the "cutting fraction" region. (b) The TerL protein structure predicted by AlphaFold3 (https://alphafoldserver.com/). With the pLDDT values of each residue shown on the left, and the alignment information on the right. (c) The structure alignment of the TerL fragments to the full-length TerL. (d) The nucleotide alignment of the terL-coding regions of curated Jug phage group Alpha genomes with that of adult 1. The alignment supports our conclusion of mRNA cutting and splicing based on RNA reads mapping. See **Supplementary Figure 14** for more information.

Sample name: Human fecal microbial communities from newborn in Denmark - 128_B
NCBI SRA accession: ERR525705
Isolation: Human feces from Newborn
Country: Denmark
Scaffold ID: Ga0169188_10197
Scaffold length: 35,737 bp
CRISPR-Cas repeat sequence: GTCACACCCTGCGTGGGTGTGTGGATTGAAAC



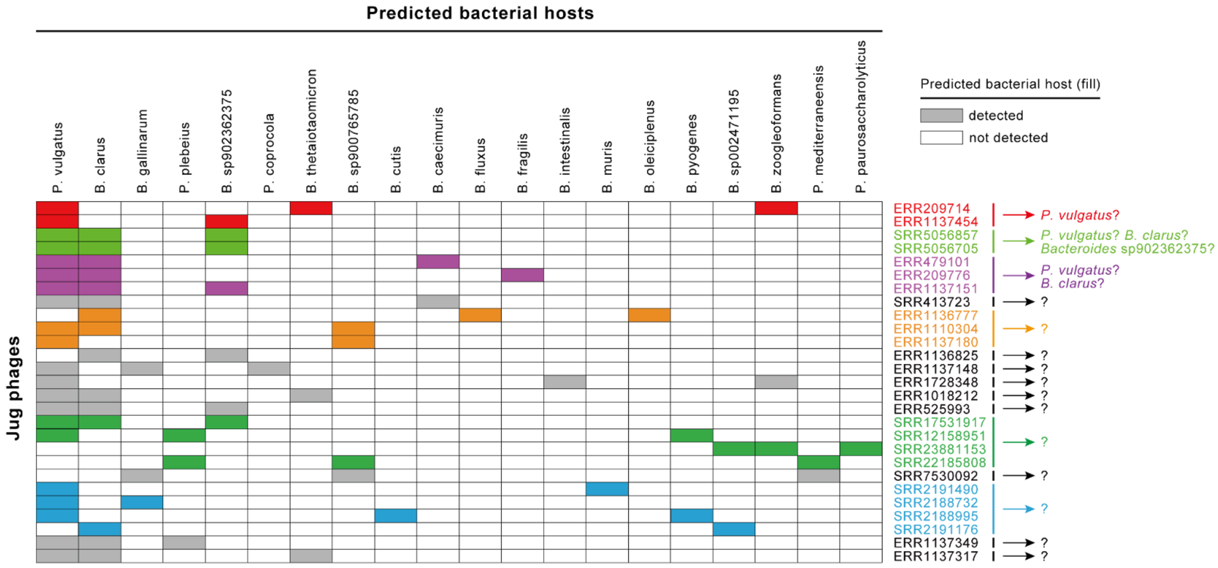
BLASTp search suggested all the proteins (including the Cas proteins) were most similar to those of *Phocaeicola* vulgatus

Spacer: AGTGATAGTAATATAGATAGTTTAGCGGTTGATAT
Jug phage: AGTGATAGTAATATAGATAGTTTAGCGGTAGATAT

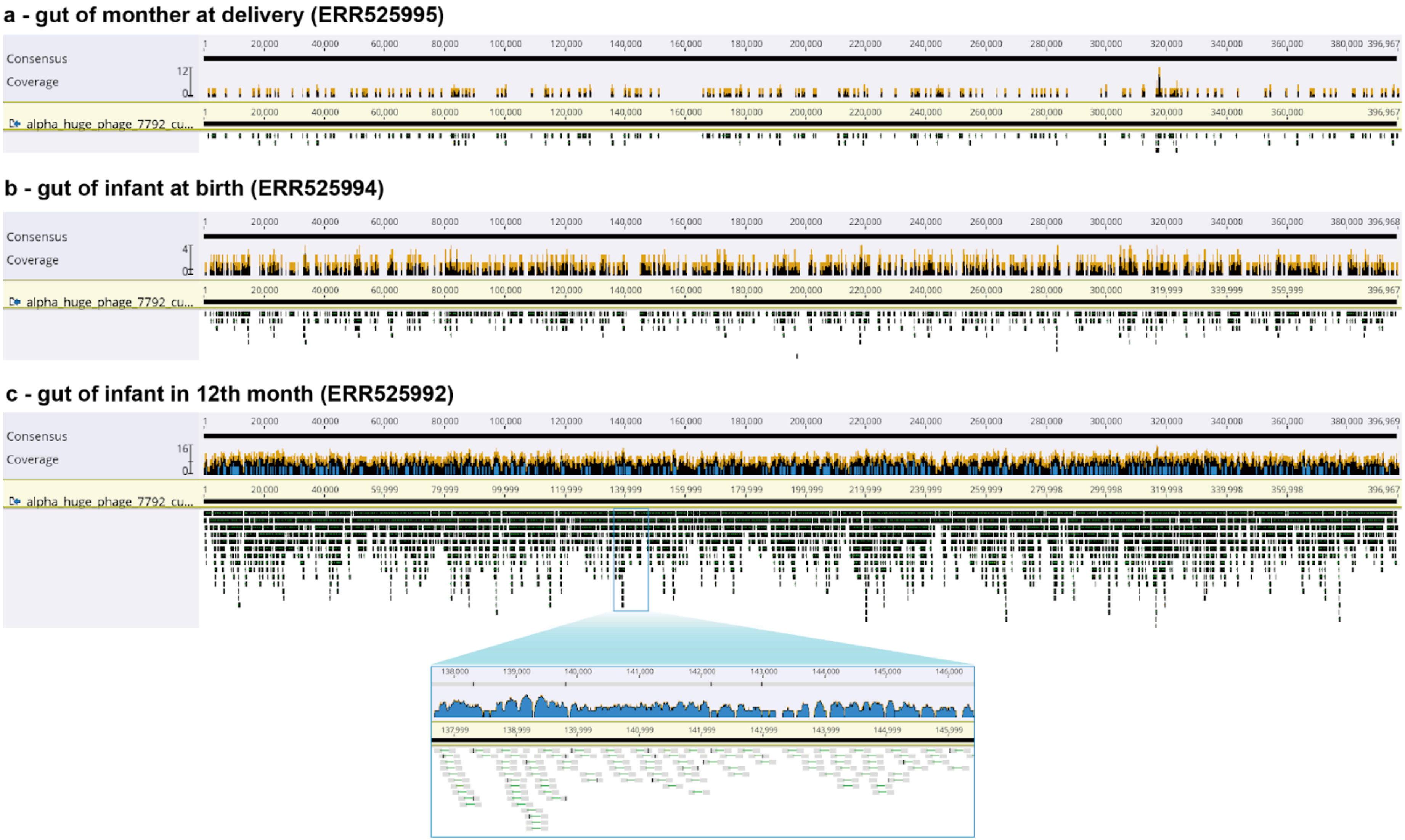
identity = 34/35 = 97.1%

part of a gene encoding a hypothetical protein

Supplementary Figure 17 | An example of CRISPR-Cas spacer targeting on the Jug phage genomes. The information about the scaffold with the CRISPR-Cas system is shown. The scaffold was assembled from the metagenomic reads of a newborn gut microbiome sample, and its taxonomic assignment is *Phocaeicola vulgatus*. The diagram in the middle shows the locations of the cas proteins and the repeat locus on the scaffold. The alignment of the targeting spacer and the Jug phage genomic fragment is shown at the bottom.



Supplementary Figure 18 | The co-occurrence of *Bacteroides* and *Phocaeicola* species and the Jug phages. Only those samples with only 2 or 3 predicted hosts (*Bacteroides* and *Phocaeicola* species) detected are shown. Each line represents a Jug phage (indicated by the sample SRA numbers, as only one Jug phage is presented in the corresponding samples), and each column represents a *Bacteroides* or *Phocaeicola* species. The SRA numbers in the same color (excepting black) indicated that the Jug phage genomes shared ⩾97% genome similarity across ⩾90% of their genome length (by dRep version 3.2.2 (Olm *et al.*, 2017)), and such phages were assumed to infect the same host(s). The colored boxes indicate the presence of the *Bacteroides* or *Phocaeicola* species in the corresponding samples. The specific host-virus relationship was speculated based on co-occurrence and is shown on the right.

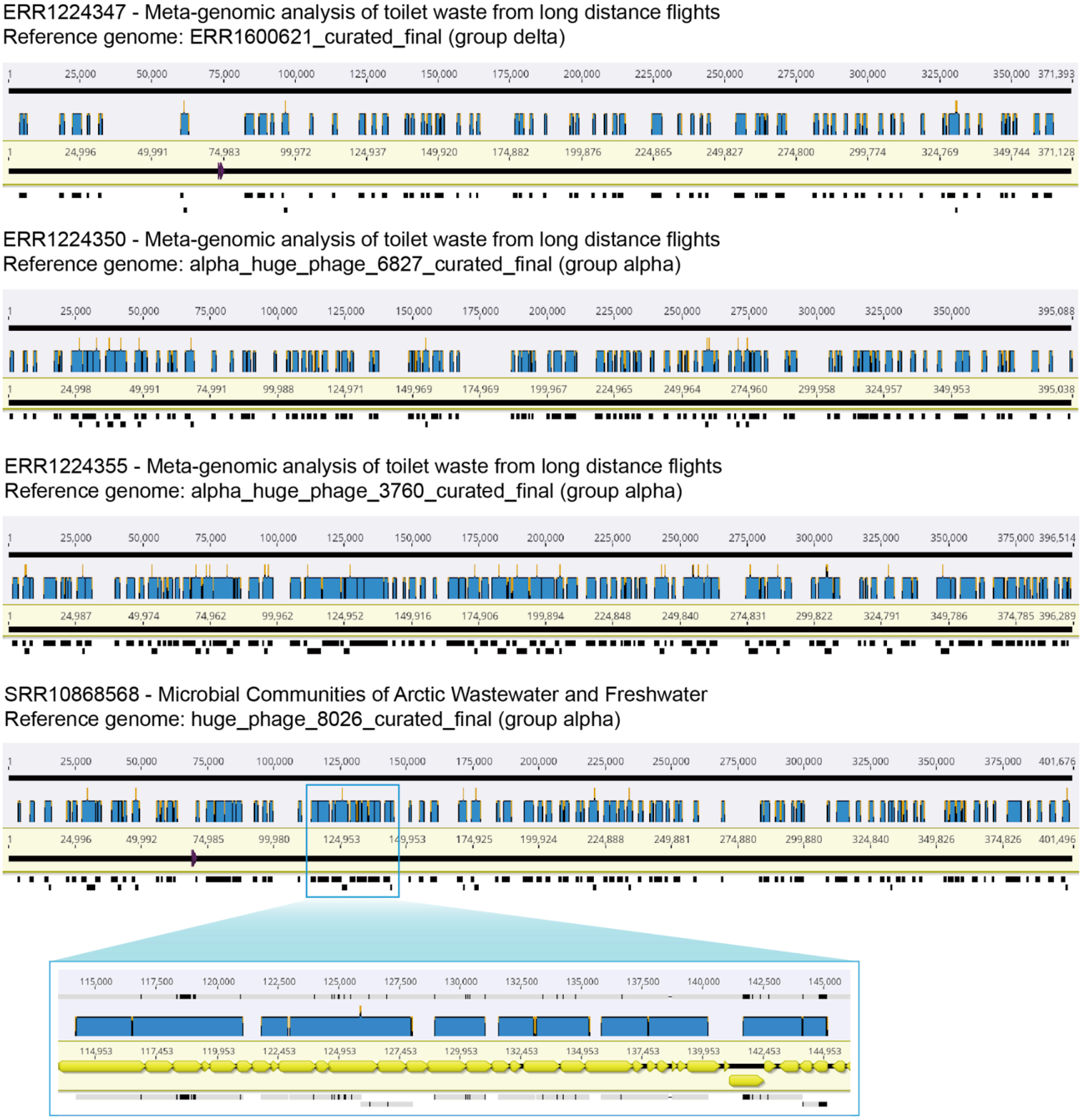

a - gut of monther at delivery (ERR525995)

b - gut of infant at birth (ERR525994)

c - gut of infant in 12th month (ERR525992)

See text legend on the next page.

Supplementary Figure 19 | The detection of the group Alpha Jug phage in the gut microbiomes of the infant and the mother. The mapping of reads from gut samples collected (a) at delivery of the mother, (b) at birth of the infant, and (c) in the 12th month of the infant. The paired-end reads were mapped to the curated Jug phage genome reconstructed from the gut samples collection in the fourth month of the infant, allowing no more than one mismatch for each read to be mapped. The mapping was performed using Bowtie2 version 2.4.4 (Longmead and Salzberg, 2012) with default parameters, and the figures were generated using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). The zoom-in indicates that each black square represents a metagenomic read. The paired-end reads are linked with a blue line.



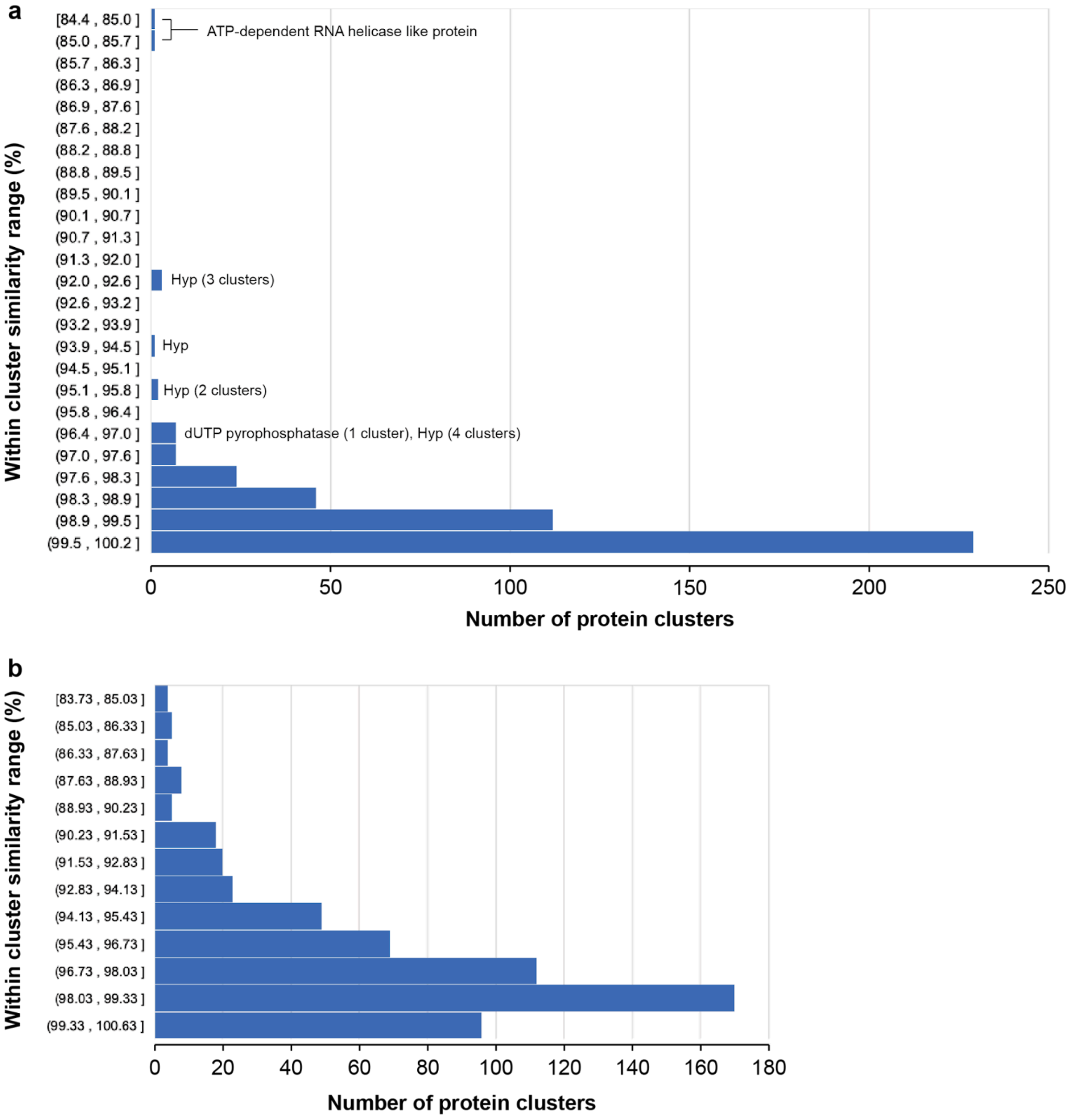
ERR1224347 - Meta-genomic analysis of toilet waste from long distance flights
Reference genome: ERR1600621_curated_final (group delta)

ERR1224350 - Meta-genomic analysis of toilet waste from long distance flights
Reference genome: alpha_huge_phage_6827_curated_final (group alpha)

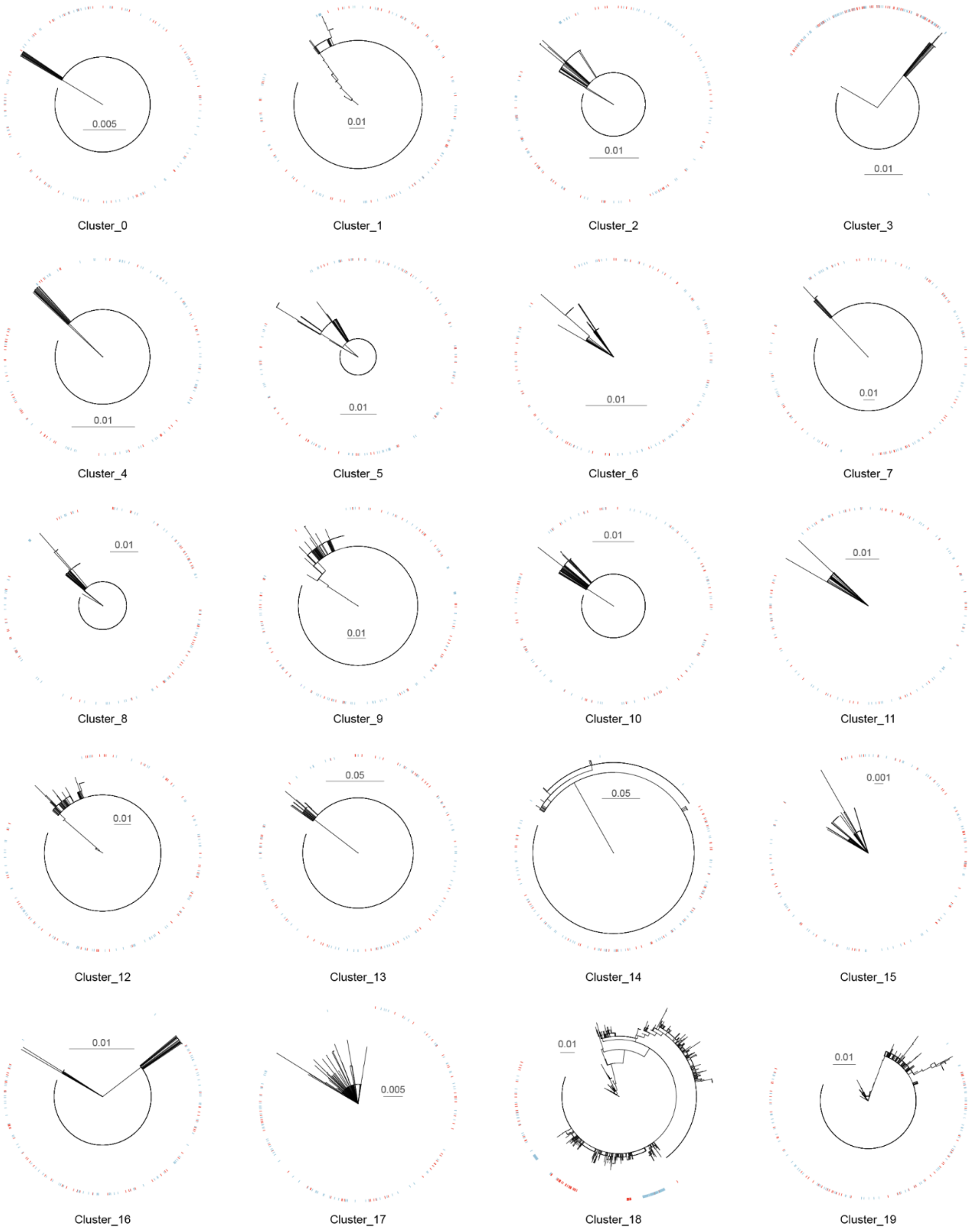
ERR1224355 - Meta-genomic analysis of toilet waste from long distance flights
Reference genome: alpha_huge_phage_3760_curated_final (group alpha)

SRR10868568 - Microbial Communities of Arctic Wastewater and Freshwater
Reference genome: huge_phage_8026_curated_final (group alpha)

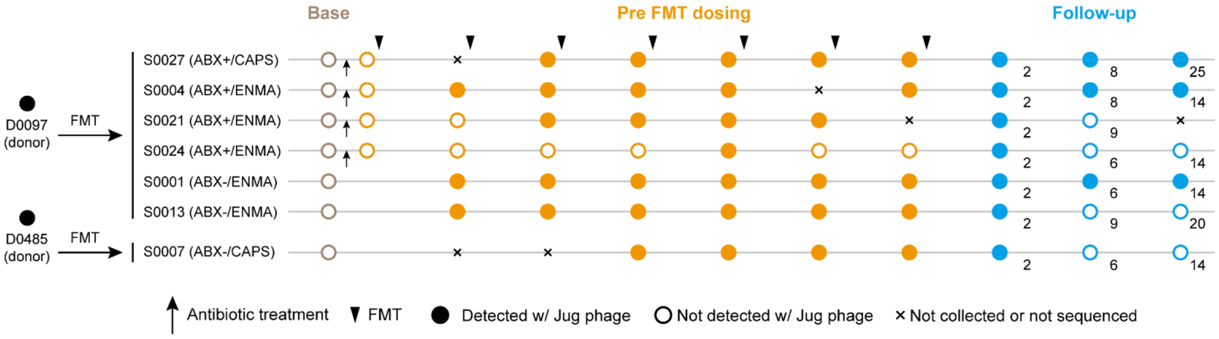
Supplementary Figure 20 | Genome alignment of Logan contigs assembled from wastewater to the curated Jug phage genomes. The alignment was performed using the "Map to Reference(s)" function of Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). Four examples (one of group Delta, and three of group Alpha) are shown with the sample information, including SRA IDs and project descriptions from NCBI. The zoom-in shows the details of the alignment as an example.



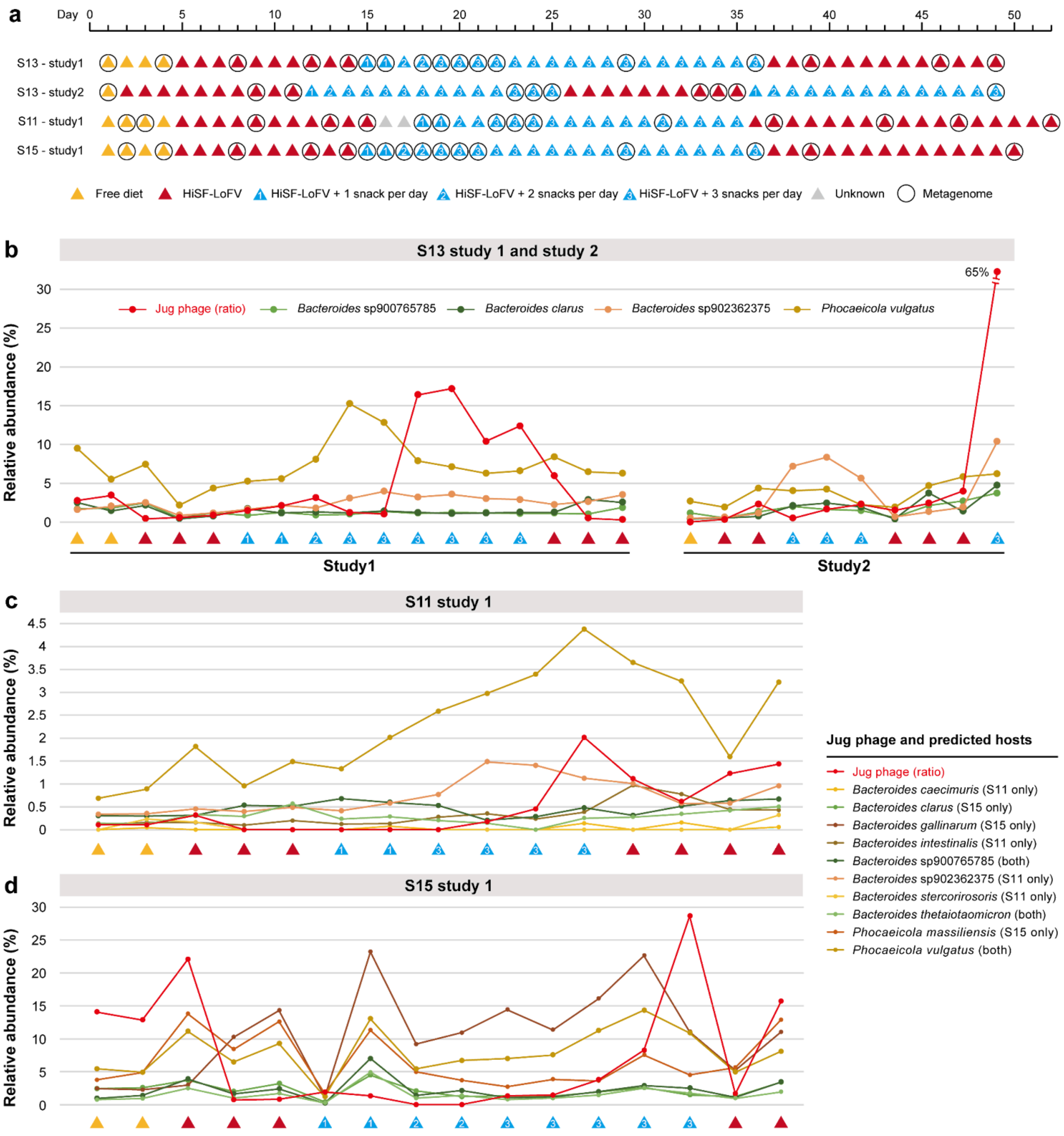
Supplementary Figure 21 | The similarity of proteins within each protein cluster of the Jug phages. The protein clusters shared by (a) the group Alpha Jug phages from the gut of humans, mice, and rats, and (b) the group Beta Jug phages from the gut of humans and dogs. The average paired protein sequence identity of each cluster was calculated first, and then the similarity ranges of all shared protein clusters were profiled. The paired protein sequence similarity was based on all-vs-all BLASTp within each cluster (e-value threshold = 1e-5).



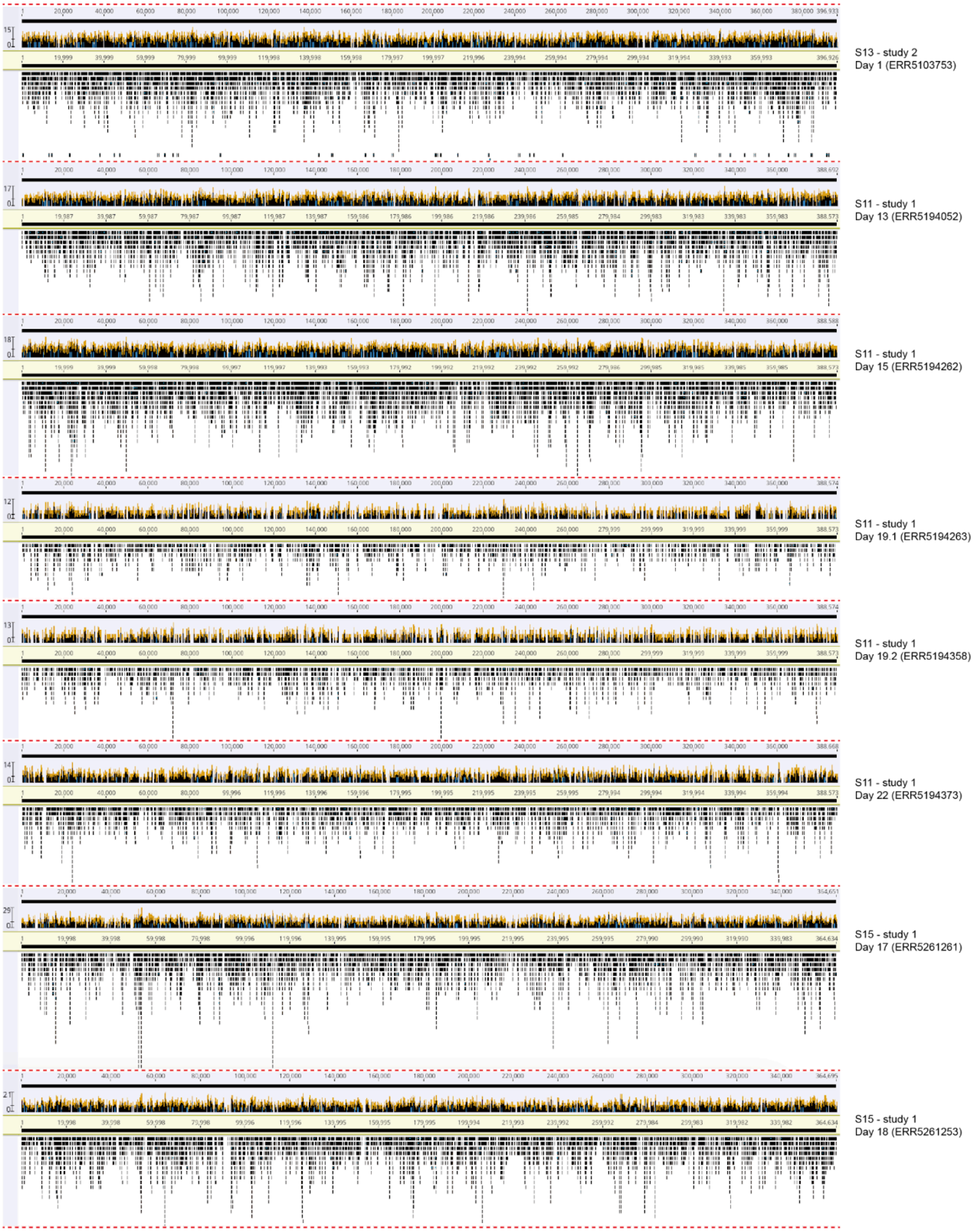
Supplementary Figure 22 | The phylogenies of protein families shared by group Apha Jug phages from humans, mice, and rats. The sequences of the protein families containing group Alpha Jug phages from the gut of humans, mice, and rats were conducted for phylogenetic analyses, and the top 20 protein clusters are shown here as examples. The colored strips indicate the animal host of the Jug phages, with blue strips for those from mice, red strips for those from rats, and all others for those from humans.



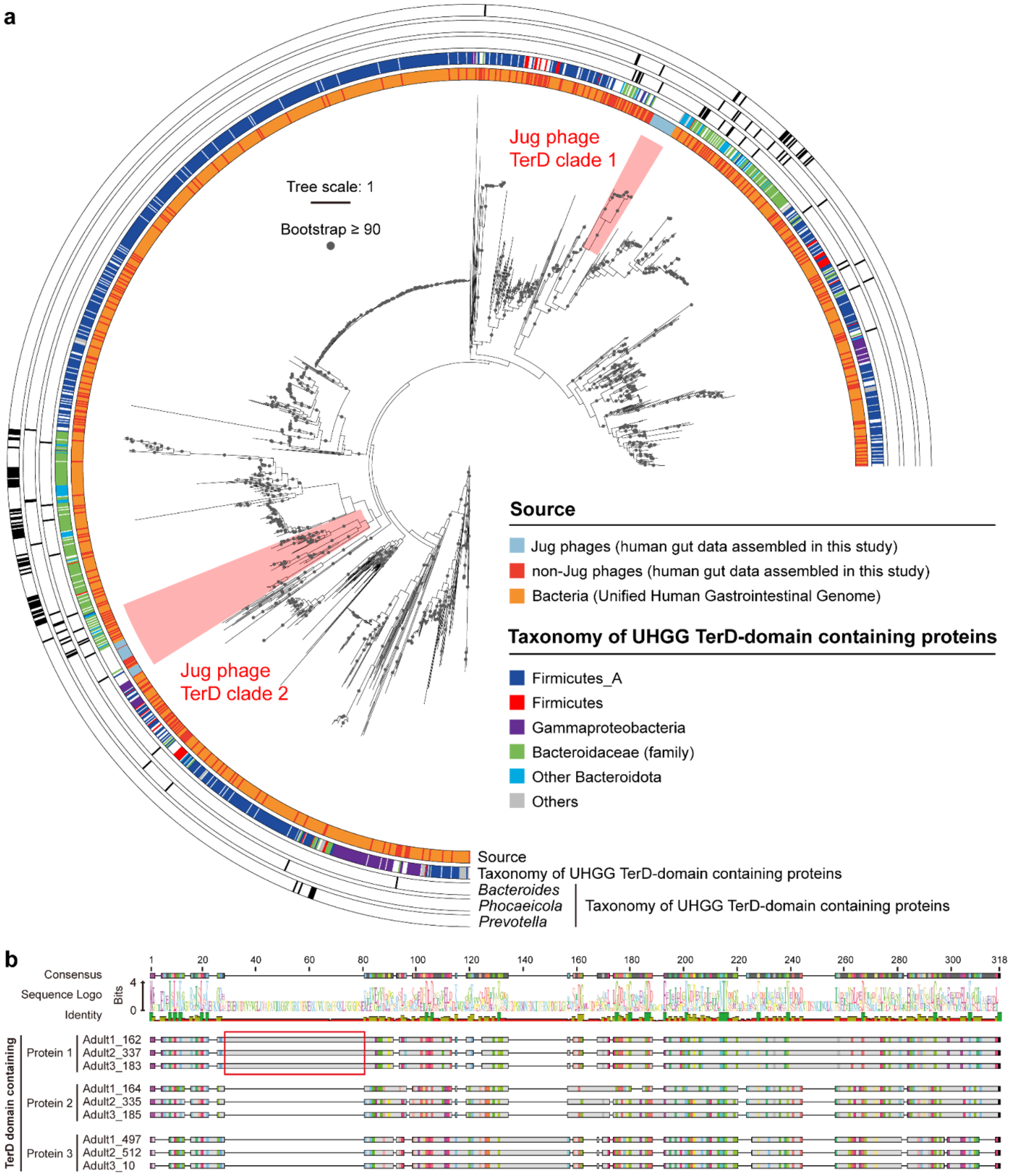
Supplementary Figure 23 | The fecal microbiota transplantation (FMT) from two donors to the recipients for the treatment of ulcerative colitis. The points of recipients that received antibiotic treatment before FMT are indicated by black arrows. A total of 7 FMT dosing and 2-3 follow-up evaluations were performed for each recipient. The colored circles indicated when the fecal samples were collected for metagenomic analyses; one more sample was collected for those recipients accepting antibiotic treatment. The follow-up time (in weeks) after the last dosing is shown. The FMT was performed in two approaches, i.e., ENMA (maintenance doses via enema) and CAPS (maintenance doses via capsules). ABX+, with antibiotic pretreatment; ABX-, without antibiotic pretreatment.



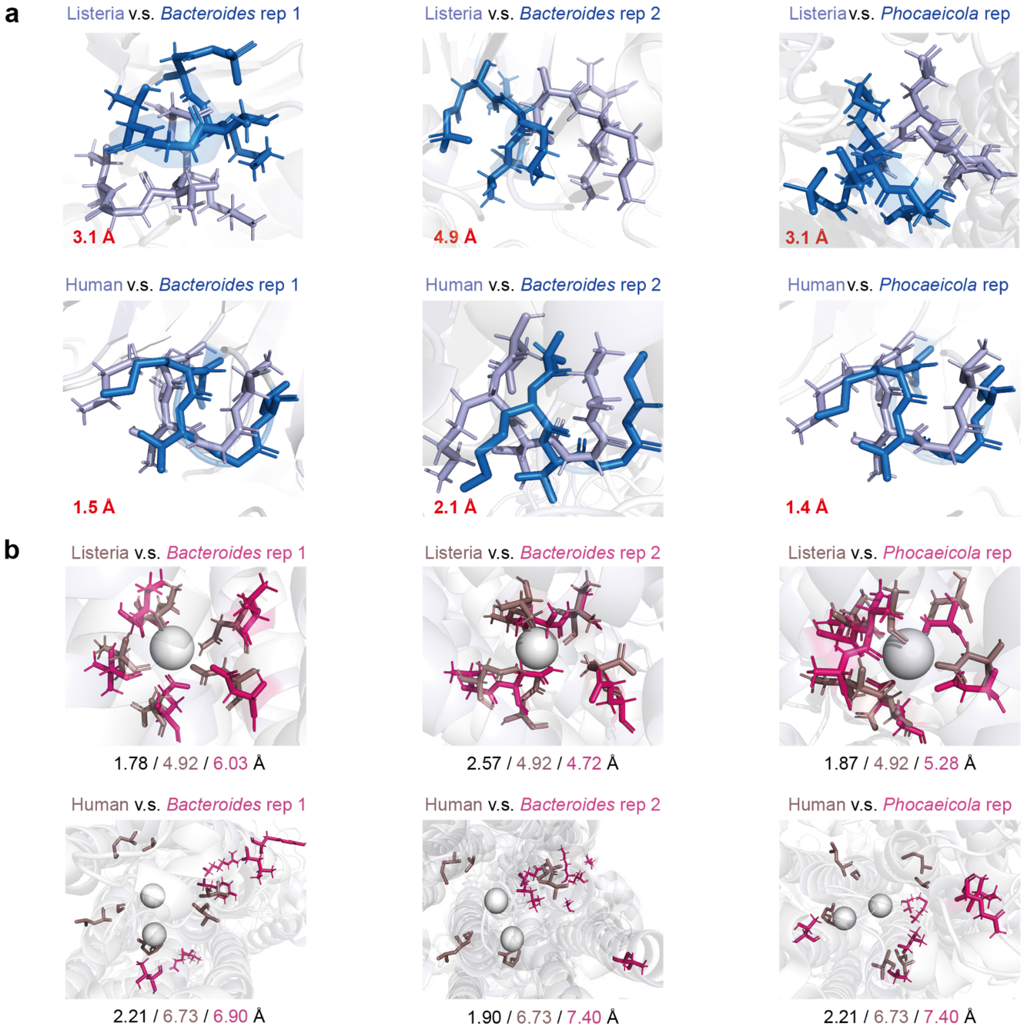
Supplementary Figure 24 | The dynamics of the Jug phage in the three participants involved in the dietary intervention studies. (a) The set-up of the dietary intervention study. The diet contents and metagenomic sampling time points (black circles) are shown. In study 1, the participants took a pea fibre-containing snack as a supplement, and had a 2-fibre (from day 12 through 25) or a 4-fibre snack prototype (from day 36 through 49). (b)-(d) The co-occurrence of Jug phages and *Bacteroides* and *Phocaeicola* species in the human gut based on sequencing coverage. The bacterial relative abundance was calculated based on the sequencing depth of the ribosomal protein S3 (rpS3) encoding scaffolds (see **Methods** in the main text). In (c) and (d), the presence of the predicted hosts in the participants is indicated in the brackets; "both" means the species was presented in both participants. For comparative analysis, we also show the presence of Jug phage by calculating the ratio of Jug phage sequencing coverage to the cumulative sequencing coverage of all bacterial and archaeal rpS3-encoding scaffolds within each respective sample.



Supplementary Figure 25 | The detection of the Jug phages in the gut microbiomes of the participants in dietary intervention studies. The quality-control metagenomic paired-end reads were mapped to the corresponding curated Jug phage genomes reconstructed for each participant, allowing no more than one mismatch for each read to be mapped. The mapping was performed using Bowtie2 version 2.4.4 (Longmead and Salzberg, 2012) with default parameters, and the figures were generated using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012).

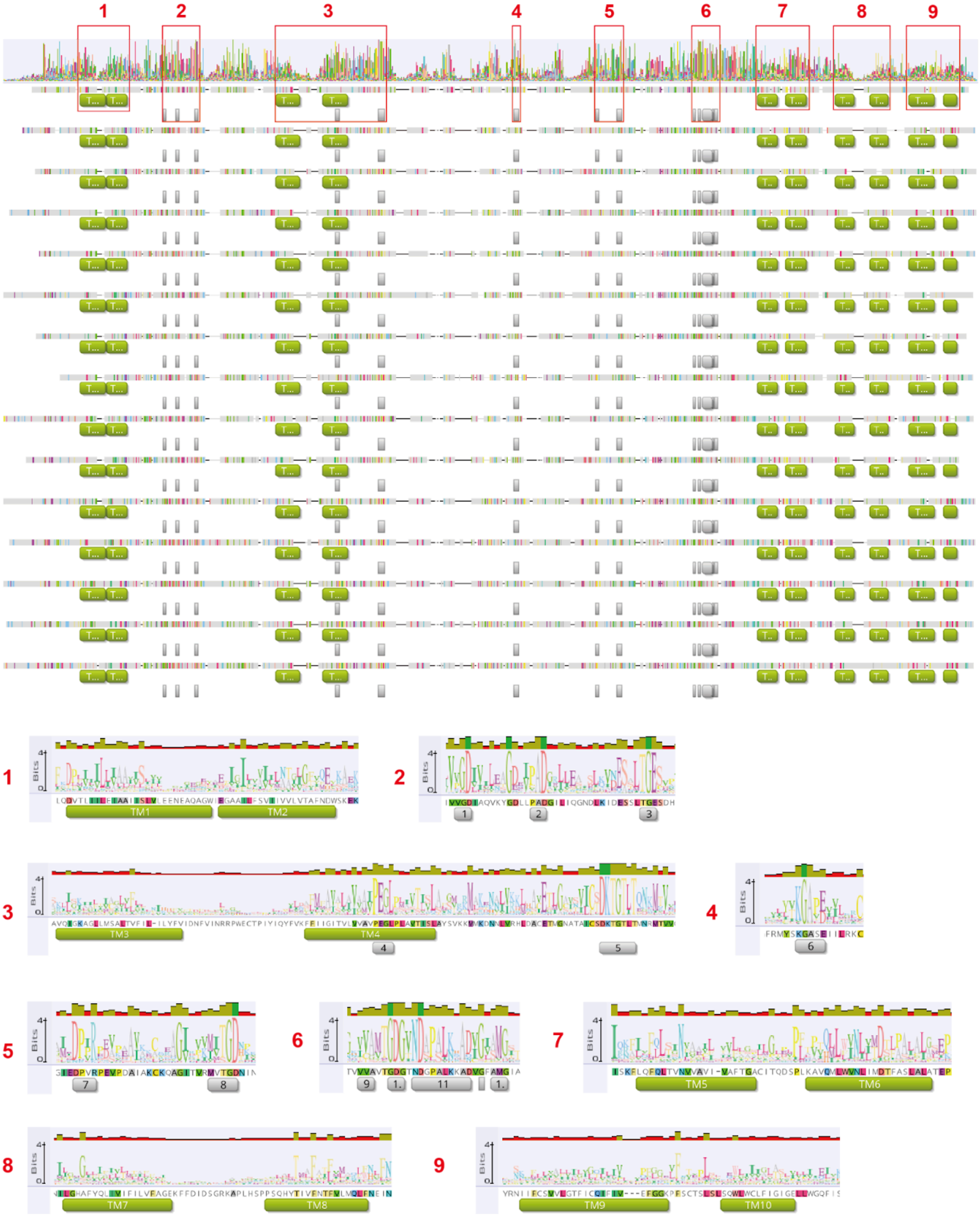


Supplementary Figure 26 | The TerD domain-containing genes encoded by Jug phages. (a) The phylogeny of the TerD domain-containing proteins. The source of the proteins (the most inner colored strip circle) includes (1) Jug phages, and (2) non-Jug phages with sequences assembled in this study, and the non-Jug phage ones from the Unified Human Gastrointestinal Genome (UHGG) unique protein database. The proteins from each source were firstly clustered with ⩾99% identity using CD-HIT, and the cluster representatives were included for phylogenetic analyses. The taxonomy of the UHGG TerD-domain containing proteins is shown in the second most inner colored strip circle; note that those of the family Bacteroidaceae (containing the genera of *Bacteroides*, *Phocaeicola*, and *Prevotella*) are shown separately. The outer three colored strip circles indicate the TerD domain-containing proteins from *Bacteroides*, *Phocaeicola*, and *Prevotella*, respectively. The two clades with TerD proteins from the Jug phages are highlighted with a colored (red) background. (b) The alignment of the TerD protein sequences. Each of the three Jug phages encoded three copies of the *terD* genes. One of the TerD proteins has an extra region (indicated by the red box).

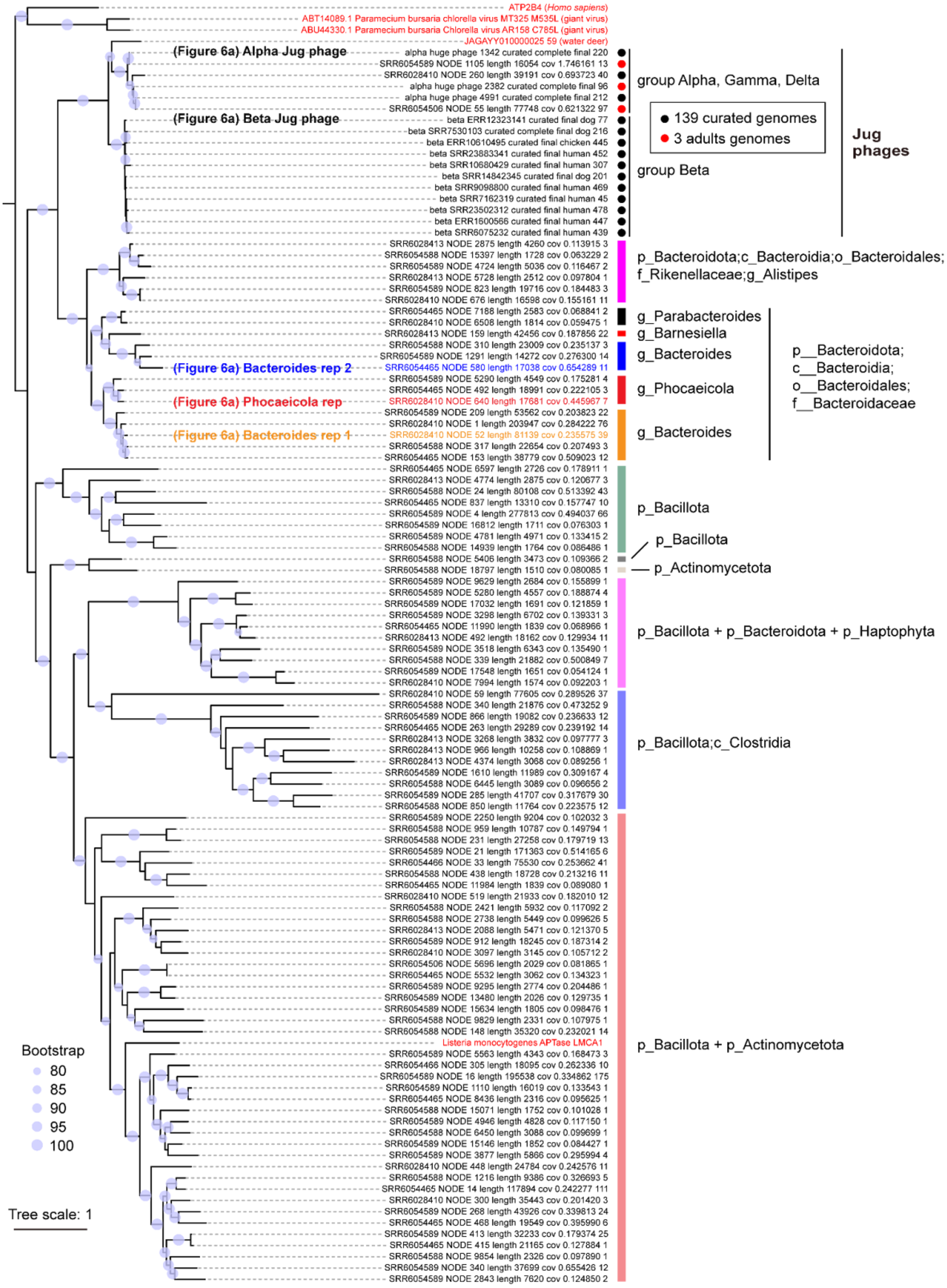


a
Listeria v.s. *Bacteroides* rep 1
3.1 Å
Listeria v.s. *Bacteroides* rep 2
4.9 Å
Listeriav.s. *Phocaeicola* rep
3.1 Å
Human v.s. *Bacteroides* rep 1
1.5 Å
Human v.s. *Bacteroides* rep 2
2.1 Å
Humanv.s. *Phocaeicola* rep
1.4 Å
b
Listeria v.s. *Bacteroides* rep 1
1.78 / 4.92 / 6.03 Å
Listeria v.s. *Bacteroides* rep 2
2.57 / 4.92 / 4.72 Å
Listeria v.s. *Phocaeicola* rep
1.87 / 4.92 / 5.28 Å
Human v.s. *Bacteroides* rep 1
2.21 / 6.73 / 6.90 Å
Human v.s. *Bacteroides* rep 2
1.90 / 6.73 / 7.40 Å
Human v.s. *Phocaeicola* rep
2.21 / 6.73 / 7.40 Å

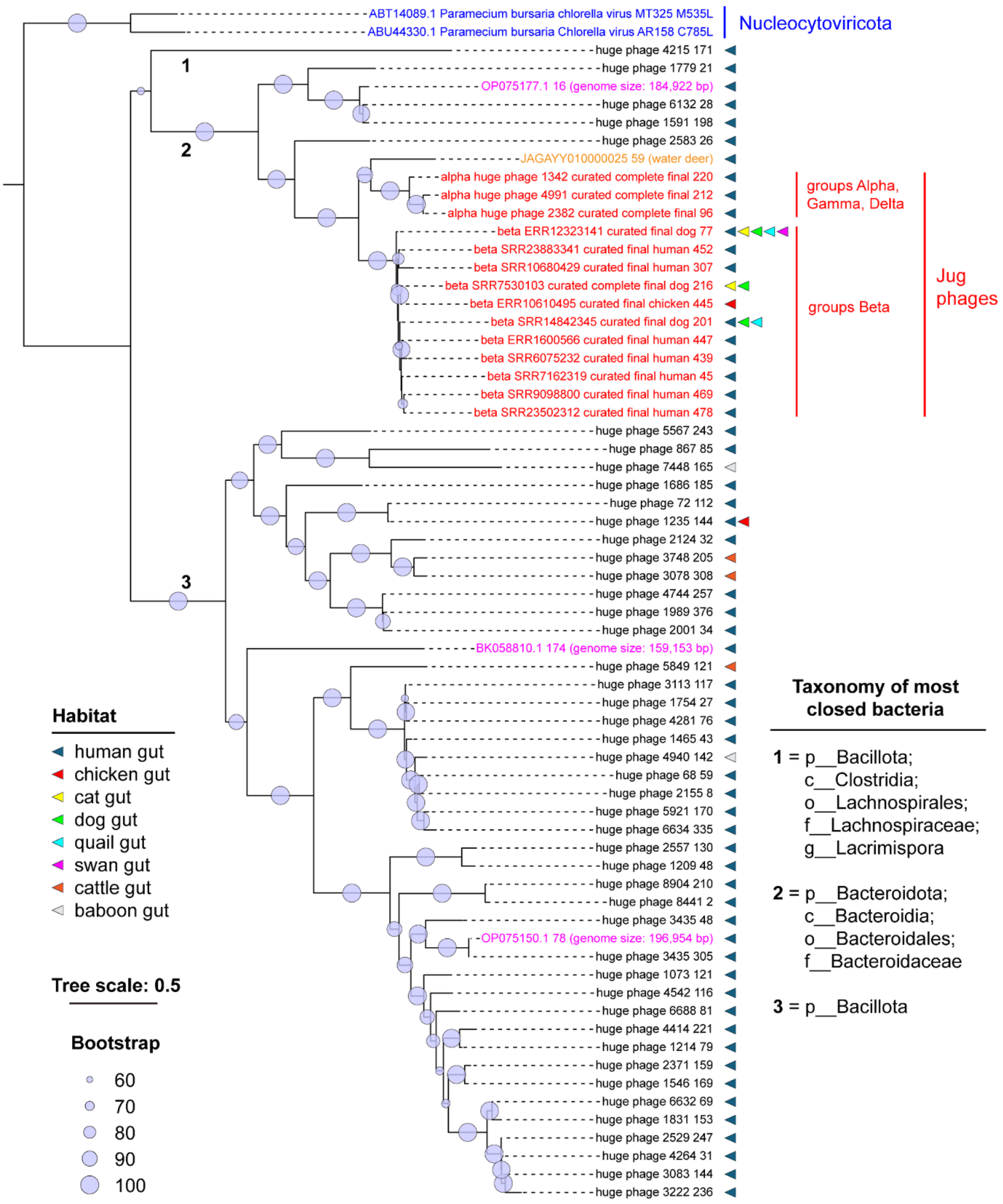
Supplementary Figure 27 | The comparison of the calcium-translocating P-type ATPases encoded by predicted bacterial hosts of Jug phages against references. (a) The comparison of the "DKTGT" motif. The distance (Å) of the "D" residue is shown in red. (b) Calcium coordination site comparison. The total distance of Ca^{2+} ion (gray spheres) binding-related residues between the ATPases of predicted Jug phages' hosts and reference, the total distance of the reference ATPase binding residues to Ca^{2+} ion, and the total distance of predicted hosts' ATPase binding residues to Ca^{2+} ion are shown at the bottom. Note that the human ATPase could combine with two Ca^{2+} ions. See Methods in the main text for more details.



Supplementary Figure 28 | The sequence alignment and key residues of calcium-tranlocating P-type ATPases. The aligned sequences of the calcium-translocating p-type ATPases in **Supplementary Figure 28**, including those from *Homo sapiens* (ATP2B4), the *Chlorella* viruses (M535L and C785L), the *Listeria monocytogenes* (LMCA1), the Jug phages, the Jug phage relative from the water deer gut, and the co-existed gut bacteria of the three adult male were visualized using Geneious Prime Build 2025-03-24 (Kearse *et al.*, 2012). Only a subset of the aligned sequences is shown here. The locations of the 10 transmembrane regions and conserved residues were manually inspected and indicated based on the information from Bonza et al. (Bonza et al. 2010). These regions and conserved residues were zoomed in below, with the sequence logo shown.

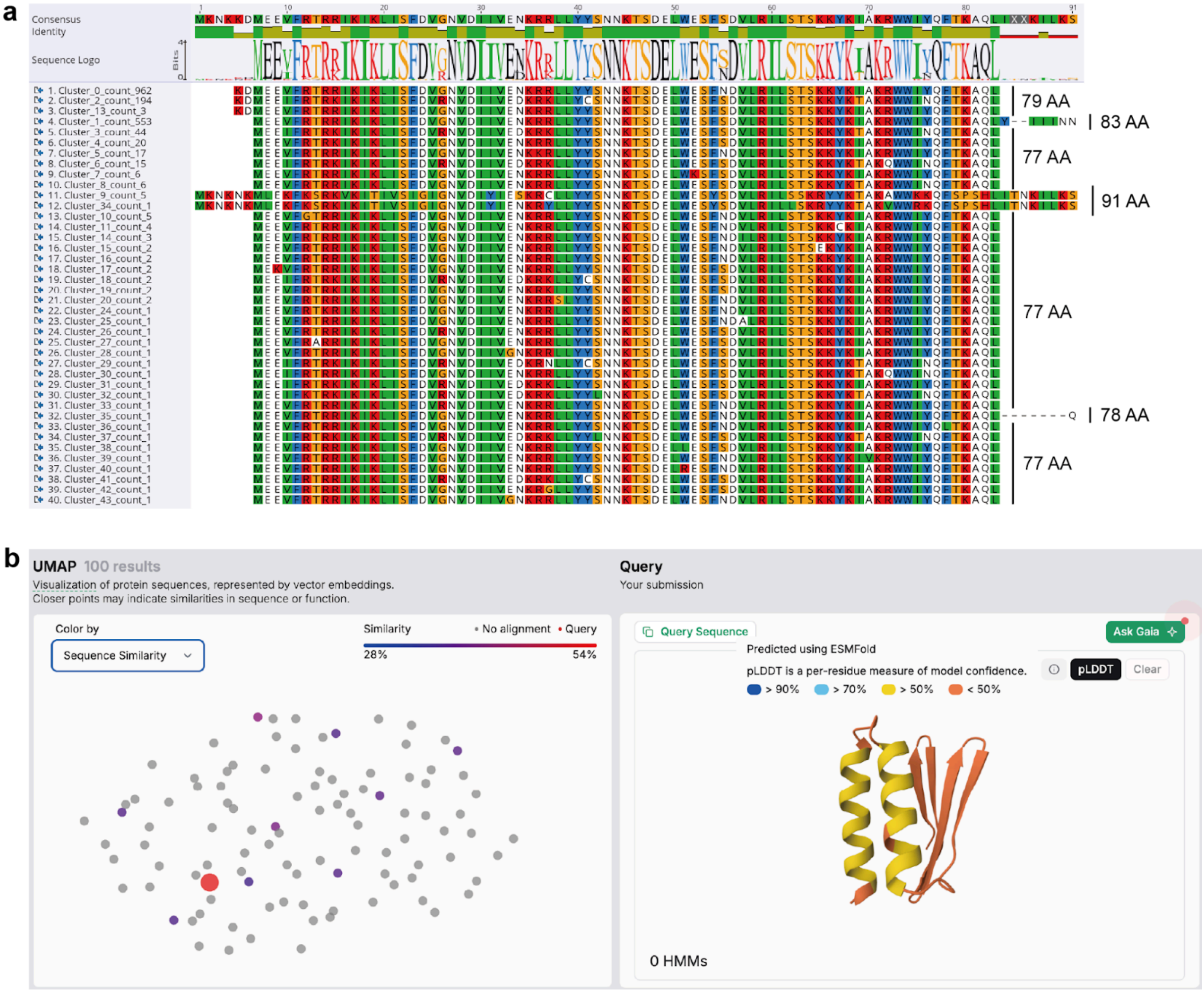

ATP2B4 (Homo sapiens)
ABT14089.1 Paramecium bursaria chlorella virus MT325 M535L (giant virus)
ABU44330.1 Paramecium bursaria Chlorella virus AR158 C785L (giant virus)
JAGAYY010000025 59 (water deer)
(Figure 6a) Alpha Jug phage
alpha huge phage 1342 curated complete final 220
SRR6054589 NODE 1105 length 16054 cov 1.746161 13
SRR6028410 NODE 260 length 39191 cov 0.693723 40
alpha huge phage 2382 curated complete final 96
alpha huge phage 4991 curated complete final 212
SRR6054506 NODE 55 length 77748 cov 0.621322 97
(Figure 6a) Beta Jug phage
beta ERR12323141 curated final dog 77
beta SRR7530103 curated complete final dog 216
beta ERR10610495 curated final chicken 445
beta SRR23883341 curated final human 452
beta SRR10680429 curated final human 307
beta SRR14842345 curated final dog 201
beta SRR9098800 curated final human 469
beta SRR7162319 curated final human 45
beta SRR23502312 curated final human 478
beta ERR1600566 curated final human 447
beta SRR6075232 curated final human 439
SRR6028413 NODE 2875 length 4260 cov 0.113915 3
SRR6054588 NODE 15397 length 1728 cov 0.063229 2
SRR6054589 NODE 4724 length 5036 cov 0.116467 2
SRR6028413 NODE 5728 length 2512 cov 0.097804 1
SRR6054589 NODE 823 length 19716 cov 0.184483 3
SRR6028410 NODE 676 length 16598 cov 0.155161 11
SRR6054465 NODE 7188 length 2583 cov 0.068841 2
SRR6028410 NODE 6508 length 1814 cov 0.059475 1
SRR6028413 NODE 159 length 42456 cov 0.187856 22
SRR6054588 NODE 310 length 23009 cov 0.235137 3
SRR6054589 NODE 1291 length 14272 cov 0.276300 14
(Figure 6a) Bacteroides rep 2
SRR6054465 NODE 580 length 17038 cov 0.654289 11
SRR6054589 NODE 5290 length 4549 cov 0.175281 4
SRR6054465 NODE 492 length 18991 cov 0.222105 3
(Figure 6a) Phocaeicola rep
SRR6028410 NODE 640 length 17681 cov 0.445967 7
SRR6054589 NODE 209 length 53562 cov 0.203823 22
SRR6028410 NODE 1 length 203947 cov 0.284222 76
(Figure 6a) Bacteroides rep 1
SRR6028410 NODE 52 length 81139 cov 0.235575 39
SRR6054588 NODE 317 length 22654 cov 0.207493 3
SRR6054465 NODE 153 length 38779 cov 0.509023 12
SRR6054465 NODE 6597 length 2726 cov 0.178911 1
SRR6028413 NODE 4774 length 2875 cov 0.120677 3
SRR6054588 NODE 24 length 80108 cov 0.513392 43
SRR6054465 NODE 837 length 13310 cov 0.157747 10
SRR6054589 NODE 4 length 277813 cov 0.494037 66
SRR6054589 NODE 16812 length 1711 cov 0.076303 1
SRR6054589 NODE 4781 length 4971 cov 0.133415 2
SRR6054588 NODE 14939 length 1764 cov 0.086486 1
SRR6054588 NODE 5406 length 3473 cov 0.109366 2
SRR6054588 NODE 18797 length 1510 cov 0.080085 1
SRR6054589 NODE 9629 length 2684 cov 0.155899 1
SRR6054589 NODE 5280 length 4557 cov 0.188874 4
SRR6054589 NODE 17032 length 1691 cov 0.121859 1
SRR6054589 NODE 3298 length 6702 cov 0.139331 3
SRR6054465 NODE 11990 length 1839 cov 0.068966 1
SRR6028413 NODE 492 length 18162 cov 0.129934 11
SRR6054589 NODE 3518 length 6343 cov 0.135490 1
SRR6054588 NODE 339 length 21882 cov 0.500849 7
SRR6054589 NODE 17548 length 1651 cov 0.054124 1
SRR6028410 NODE 7994 length 1574 cov 0.092203 1
SRR6028410 NODE 59 length 77605 cov 0.289526 37
SRR6054588 NODE 340 length 21876 cov 0.473252 9
SRR6054589 NODE 866 length 19082 cov 0.236633 12
SRR6054465 NODE 263 length 29289 cov 0.239192 14
SRR6028413 NODE 3268 length 3832 cov 0.097777 3
SRR6028413 NODE 966 length 10258 cov 0.108869 1
SRR6028413 NODE 4374 length 3068 cov 0.089256 1
SRR6054589 NODE 1610 length 11989 cov 0.309167 4
SRR6054588 NODE 6445 length 3089 cov 0.096656 2
SRR6054589 NODE 285 length 41707 cov 0.317679 30
SRR6054588 NODE 850 length 11764 cov 0.223575 12
SRR6054589 NODE 2250 length 9204 cov 0.102032 3
SRR6054588 NODE 959 length 10787 cov 0.149794 1
SRR6054588 NODE 231 length 27258 cov 0.179719 13
SRR6054589 NODE 21 length 171363 cov 0.514165 6
SRR6054466 NODE 33 length 75530 cov 0.253662 41
SRR6054588 NODE 438 length 18728 cov 0.213216 11
SRR6054465 NODE 11984 length 1839 cov 0.089080 1
SRR6028410 NODE 519 length 21933 cov 0.182010 12
SRR6054588 NODE 2421 length 5932 cov 0.117092 2
SRR6054588 NODE 2738 length 5449 cov 0.099626 5
SRR6028413 NODE 2088 length 5471 cov 0.121370 5
SRR6054589 NODE 912 length 18245 cov 0.187314 2
SRR6028410 NODE 3097 length 3145 cov 0.105712 2
SRR6054506 NODE 5696 length 2029 cov 0.081865 1
SRR6054465 NODE 5532 length 3062 cov 0.134323 1
SRR6054589 NODE 9295 length 2774 cov 0.204486 1
SRR6054589 NODE 13480 length 2026 cov 0.129735 1
SRR6054589 NODE 15634 length 1805 cov 0.098476 1
SRR6054588 NODE 9829 length 2331 cov 0.107975 1
SRR6054588 NODE 148 length 35320 cov 0.232021 14
Listeria monocytogenes APTase LMCA1
SRR6054589 NODE 5563 length 4343 cov 0.168473 3
SRR6054466 NODE 305 length 18095 cov 0.262336 10
SRR6054589 NODE 16 length 195538 cov 0.334862 175
SRR6054589 NODE 1110 length 16019 cov 0.133543 1
SRR6054465 NODE 8436 length 2316 cov 0.095625 1
SRR6054588 NODE 15071 length 1752 cov 0.101028 1
SRR6054589 NODE 4946 length 4828 cov 0.117150 1
SRR6054588 NODE 6450 length 3088 cov 0.099699 1
SRR6054589 NODE 15146 length 1852 cov 0.084427 1
SRR6054589 NODE 3877 length 5866 cov 0.295994 4
SRR6028410 NODE 448 length 24784 cov 0.242576 11
SRR6054588 NODE 1216 length 9386 cov 0.326693 5
SRR6054465 NODE 14 length 117894 cov 0.242277 111
SRR6028410 NODE 300 length 35443 cov 0.201420 3
SRR6054589 NODE 268 length 43926 cov 0.339813 24
SRR6054465 NODE 468 length 19549 cov 0.395990 6
SRR6054589 NODE 413 length 32233 cov 0.179374 25
SRR6054465 NODE 415 length 21165 cov 0.127884 1
SRR6054588 NODE 9854 length 2326 cov 0.097890 1
SRR6054589 NODE 340 length 37699 cov 0.655426 12
SRR6054589 NODE 2843 length 7620 cov 0.124850 2
group Alpha, Gamma, Delta
● 139 curated genomes
● 3 adults genomes
group Beta
Jug phages
p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;
f_Rikenellaceae;g_Alistipes
g_Parabacteroides
g_Barnesiella
g_Bacteroides
g_Phocaeicola
g_Bacteroides
p__Bacteroidota;
c__Bacteroidia;
o__Bacteroidales;
f__Bacteroidaceae
p_Bacillota
p_Bacillota
p_Actinomycetota
p_Bacillota + p_Bacteroidota + p_Haptophyta
p_Bacillota;c_Clostridia
p_Bacillota + p_Actinomycetota
Bootstrap
80
85
90
95
100
Tree scale: 1

Supplementary Figure 29 | The phylogeny of calcium-translocating P-type ATPases of Jug phages and human gut bacteria. The calcium-translocating P-type ATPases included those from *Homo sapiens* (ATP2B4), the *Chlorella* viruses (M535L and C785L), the *Listeria monocytogenes* (LMCA1), the Jug phages and their relative from the water deer gut, and the co-existed gut bacteria of the three adult males. Those sequences from the curated Jug phage genomes were dereplicated using CD-HIT by clustering them with 99% identity, and only the cluster representatives were included. The taxonomy of scaffolds encoding the bacterial ATPases was determined based on the taxonomic assignment of all protein-coding genes on the corresponding scaffolds. The ones predicted with protein structures and shown in Figure 6a are indicated.



ABT14089.1 Paramecium bursaria chlorella virus MT325 M535L
ABU44330.1 Paramecium bursaria Chlorella virus AR158 C785L
Nucleocytoviricota
huge phage 4215 171
1
huge phage 1779 21
OP075177.1 16 (genome size: 184,922 bp)
huge phage 6132 28
huge phage 1591 198
2
huge phage 2583 26
JAGAYY010000025 59 (water deer)
alpha huge phage 1342 curated complete final 220
alpha huge phage 4991 curated complete final 212
alpha huge phage 2382 curated complete final 96
groups Alpha, Gamma, Delta
beta ERR12323141 curated final dog 77
beta SRR23883341 curated final human 452
beta SRR10680429 curated final human 307
beta SRR7530103 curated complete final dog 216
beta ERR10610495 curated final chicken 445
beta SRR14842345 curated final dog 201
beta ERR1600566 curated final human 447
beta SRR6075232 curated final human 439
beta SRR7162319 curated final human 45
beta SRR9098800 curated final human 469
beta SRR23502312 curated final human 478
groups Beta
Jug phages
huge phage 5567 243
huge phage 867 85
huge phage 7448 165
huge phage 1686 185
huge phage 72 112
huge phage 1235 144
huge phage 2124 32
huge phage 3748 205
huge phage 3078 308
3
huge phage 4744 257
huge phage 1989 376
huge phage 2001 34
BK058810.1 174 (genome size: 159,153 bp)
huge phage 5849 121
huge phage 3113 117
huge phage 1754 27
huge phage 4281 76
huge phage 1465 43
huge phage 4940 142
huge phage 68 59
huge phage 2155 8
huge phage 5921 170
huge phage 6634 335
huge phage 2557 130
huge phage 1209 48
huge phage 8904 210
huge phage 8441 2
huge phage 3435 48
OP075150.1 78 (genome size: 196,954 bp)
huge phage 3435 305
huge phage 1073 121
huge phage 4542 116
huge phage 6688 81
huge phage 4414 221
huge phage 1214 79
huge phage 2371 159
huge phage 1546 169
huge phage 6632 69
huge phage 1831 153
huge phage 2529 247
huge phage 4264 31
huge phage 3083 144
huge phage 3222 236
Habitat
◀ human gut
◀ chicken gut
◀ cat gut
◀ dog gut
◀ quail gut
◀ swan gut
◀ cattle gut
◁ baboon gut
Tree scale: 0.5
Bootstrap
● 60
● 70
● 80
● 90
● 100
Taxonomy of most closed bacteria
1 = p__Bacillota; c__Clostridia; o__Lachnospirales; f__Lachnospiraceae; g__Lacrimispora
2 = p__Bacteroidota; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae
3 = p__Bacillota

Supplementary Figure 30 | The phylogeny of virus-encoded calcium-translocating p-type ATPases. The ATPase protein sequences from the *Chlorella* viruses (M535L and C785L; in blue), the Jug phages (in red), the Jug phage relative from the water deer gut (in orange), the smaller-sized phages (NCBI Genbank; in pink), and other HPGC phages (in black) were included. Those sequences from Jug phages, NCBI Genbank, and other HPGC phages were respectively dereplicated using CD-HIT at 99% identity, and only the cluster representatives were included. The tree was rerooted using the two sequences from *Chlorella* viruses. A colored triangle indicated the animal host of the phages of each cluster. The most closed related taxonomy of the ATPases from smaller-sized phages (NCBI Genbank) and other HPGC phages was determined by searching them against the UniProt database.



UMAP 100 results

Visualization of protein sequences, represented by vector embeddings.
Closer points may indicate similarities in sequence or function.

Query

Your submission

Predicted using ESMFold

pLDDT is a per-residue measure of model confidence.

0 HMMs

Supplementary Figure 31 | The co-transcribed adjacent upstream gene of the calcium-translocating P-type ATPase gene encoded by the Jug phages. (a) The protein sequence alignment of the adjacent upstream genes. The genes were from all curated and Logan-retrieved Jug phage genomes by BLASTp, and were dereplicated using CD-HIT (100% identity). Only the CD-HIT clusters are shown here (4 clusters were only partial genes and thus excluded). The cluster IDs and the total count of members in the cluster are shown on the left. The length of each representative is shown on the right. (b) The analyses of the adjacent upstream gene (the 77 AA version) were conducted using the recently published Gaia (Genomic AI Annotator) sequence annotation platform. The left panel shows the sequence similarity of the Ca^{2+} ATPase adjacent gene (the big red circle) to those in the Gaia database; the right panel shows the predicted protein structure of the adjacent gene, with the pLDDT shown for each residue.

References

Camargo, A.P. *et al.* (2023) 'Identification of mobile genetic elements with geNomad', *Nature biotechnology* [Preprint]. Available at: https://doi.org/10.1038/s41587-023-01953-y.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics* , 25(15), pp. 1972–1973.

Darling, A.C.E. *et al.* (2004) 'Mauve: multiple alignment of conserved genomic sequence with rearrangements', *Genome research*, 14(7), pp. 1394–1403.

Edgar, R.C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic acids research*, 32(5), pp. 1792–1797.

Harris, R.S. (2007) *Improved Pairwise Alignment of Genomic DNA*.

Huang, Y. *et al.* (2010) 'CD-HIT Suite: a web server for clustering and comparing biological sequences', *Bioinformatics* , 26(5), pp. 680–682.

Hyatt, D. *et al.* (2010) 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC bioinformatics*, 11, p. 119.

Jain, C. *et al.* (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature communications*, 9(1), p. 5114.

Kearse, M. *et al.* (2012) 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics* , 28(12), pp. 1647–1649.

Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic acids research*, 47(W1), pp. W256–W259.

Li, W. and Godzik, A. (2006) 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics* , 22(13), pp. 1658–1659.

Longmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie2'. Available at: https://www.sid.ir/en/journal/ViewPaper.aspx?ID=436196.

Minh, B.Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular biology and evolution*, 37(5), pp. 1530–1534.

Olm, M.R. *et al.* (2017) 'dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication', *The ISME journal*, 11(12), pp. 2864–2868.

Steinegger, M. and Söding, J. (2017) 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature biotechnology*, 35(11), pp. 1026–1028.