

## A distinct class of conjugative megaplasמידs includes potential vehicles for prophage dissemination

Ling Yuan (袁凌)<sup>1</sup>, Yiting Qin (秦祎婷)<sup>1</sup>, Jacob West-Roberts<sup>2</sup>,

Karthik Anantharaman<sup>3</sup>, Haoyu Wang (王浩宇)<sup>4,5</sup>, Yuanqiang Zou (邹远强)<sup>4</sup>, Yi Duan (段屹)<sup>6,7</sup>,

Antonio Pedro Camargo<sup>8,\*</sup>, Eugene V. Koonin<sup>9,\*</sup>, LinXing Chen (陈林兴)<sup>1,\*</sup>

<sup>1</sup> State Key Laboratory of Advanced Environmental Technology, the Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei, China

<sup>2</sup> Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA 94720, USA

<sup>3</sup> Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA

<sup>4</sup> State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen 518083, China

<sup>5</sup> College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

<sup>6</sup> State Key Laboratory of Immune Response and Immunotherapy, Department of Infectious Diseases, The First Affiliated Hospital of USTC, Center for Advanced Interdisciplinary Science and Biomedicine of IHM, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China

<sup>7</sup> Key Laboratory of Anhui Province for Emerging and Reemerging Infectious Diseases, Hefei, Anhui 230027, China

<sup>8</sup> Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP 05508-060, Brazil

<sup>9</sup> Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

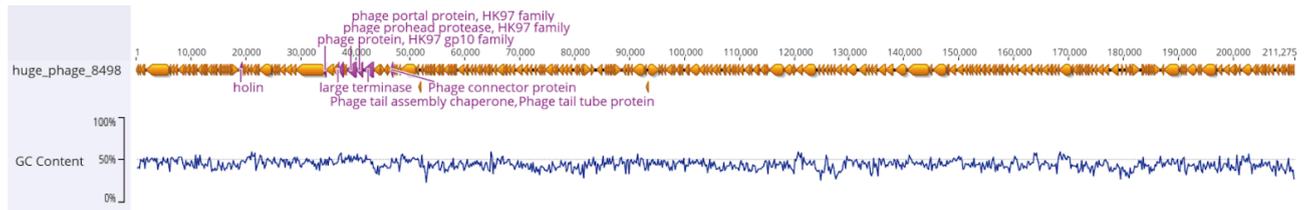
\*Corresponding author:

LinXing Chen, [linxingchen@ustc.edu.cn](mailto:linxingchen@ustc.edu.cn)

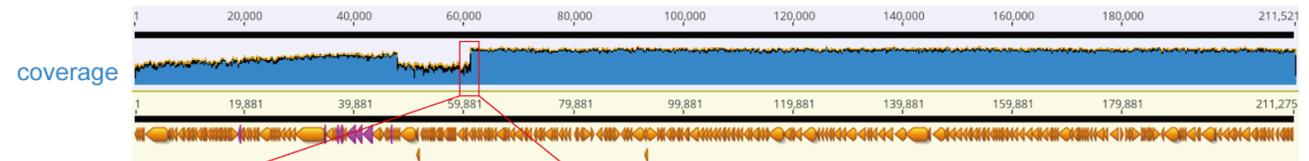
Eugene V. Koonin, [koonin@ncbi.nlm.nih.gov](mailto:koonin@ncbi.nlm.nih.gov)

Antonio Pedro Camargo, [antonio@iq.usp.br](mailto:antonio@iq.usp.br)

**a** The original genome (“huge\_phage\_8498”, 211,275 bp) with phage genes

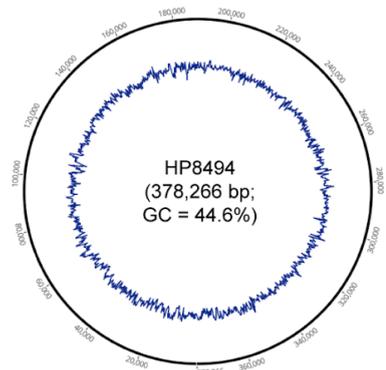


**b** Reads mapping to the original genome



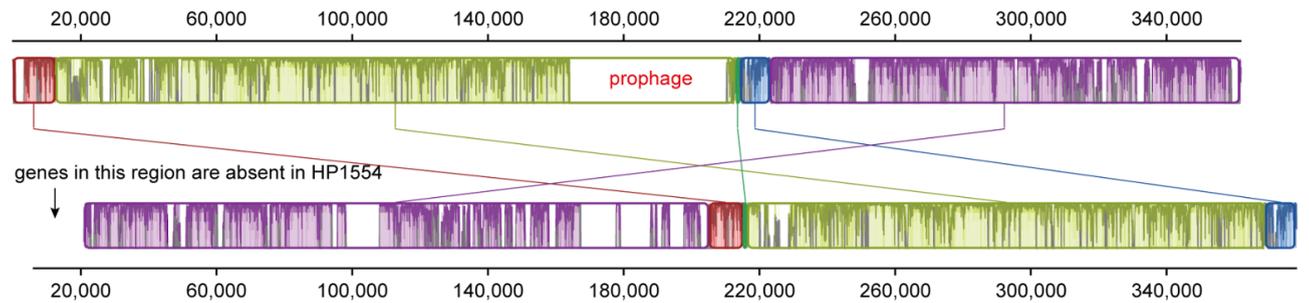
no paired-end reads spanning this region, suggesting potential chimeric assembly

**c**



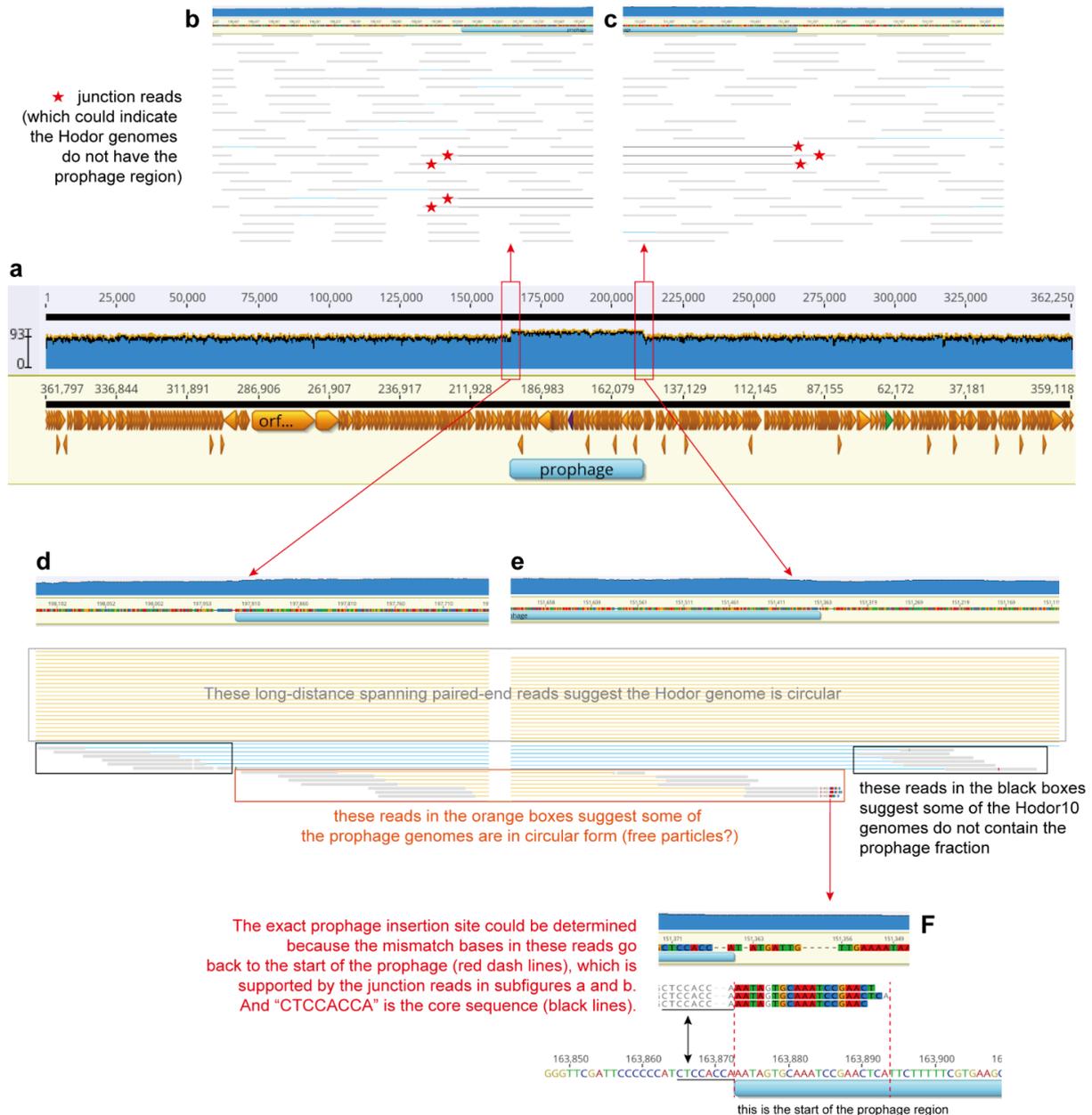
*de novo* assembly of the reads with manual curation obtained a circular genome without the phage genes encoding region

**d** HP1554 (top) vs HP8494 (bottom)

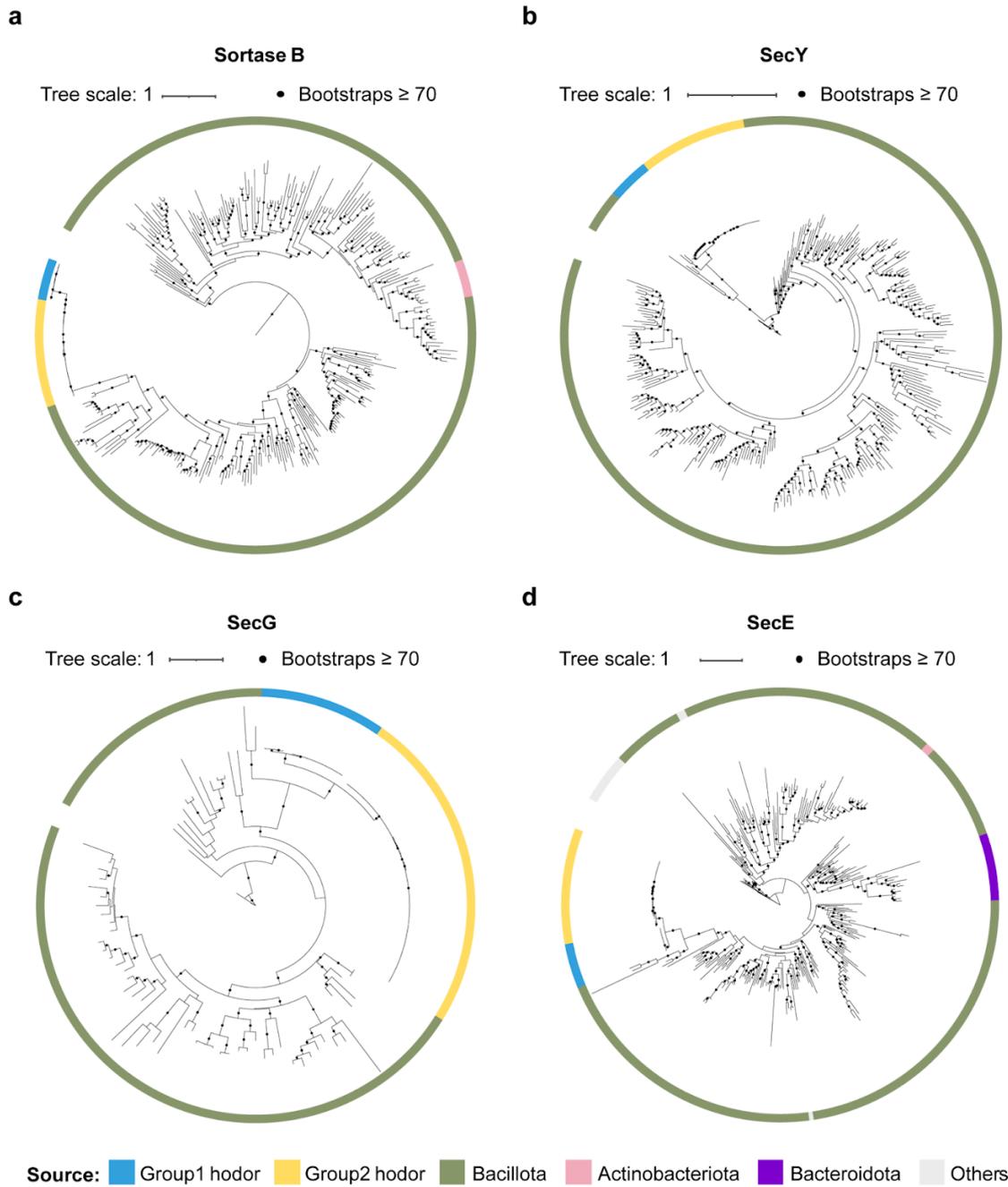


**Supplementary Figure S1 | Genome curation of “huge\_phage\_8498”.** (a) The genome encoded some phage genes. (b) Paired-end reads mapping to the genome suggested potential chimeric assembly issues. (c) The complete and circular genome. (d) The genome-wide comparison of HP8494 (Hodor15) and HP1554 (Hodor10). The genome comparison was performed using Mauve<sup>1</sup> within Geneious Prime<sup>2</sup>.

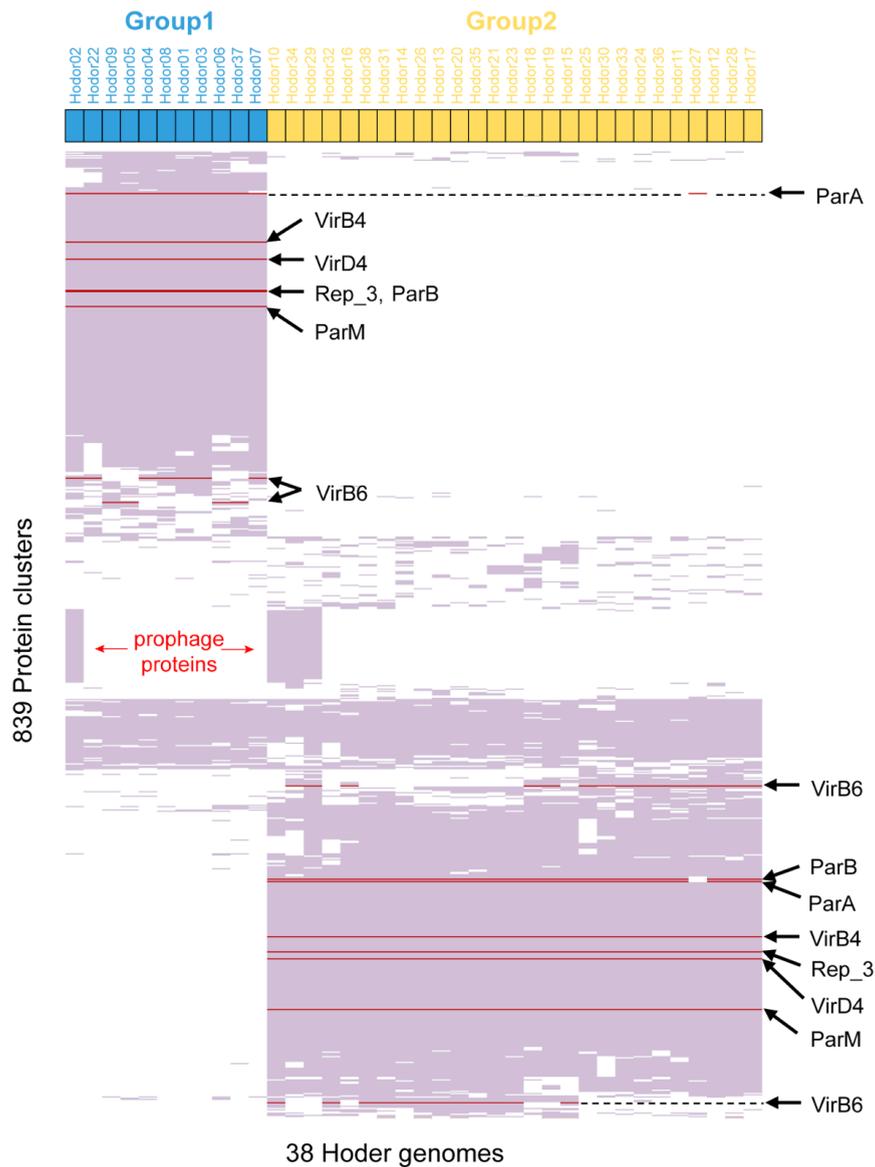
**Hodor10 as an example (prophage region: 163,873-210,433 bp)**



**Supplementary Figure S2 | Some Hodor genomes contain complete prophage genomes.** Hodor10 is used as an example for illustration. (a) Genome-wide view of a representative Hodor genome carrying an integrated prophage. The upper panels show paired-end read mapping across the entire circularized Hodor genome. The coverage track (blue area with black/yellow line) indicates sequencing depth across the replicon. The prophage region is highlighted. The junction reads indicating Hodor genomes without the prophage are enlarged in panels (b) and (c). Long-distance spanning paired-end reads supporting circularization of the Hodor genome are enlarged in panels (d) and (e). (f) Evidence for circularized prophage genomes. Read pairs spanning the prophage termini (orange boxes) suggest that a subset of prophage genomes exists in a circular form, consistent with excision from the Hodor backbone and potential presence in extrachromosomal form and/or phage particles. A high-resolution view of the prophage insertion site is shown. The junction reads (including those highlighted in subpanels a and b) defined the precise prophage integration boundary. Mismatched bases at the read ends map to the start of the prophage region (indicated by red dashed lines), allowing accurate identification of the attachment site. The conserved core sequence "CTCCACCA" (black lines) is present at the insertion junction and represents the recombination core motif.



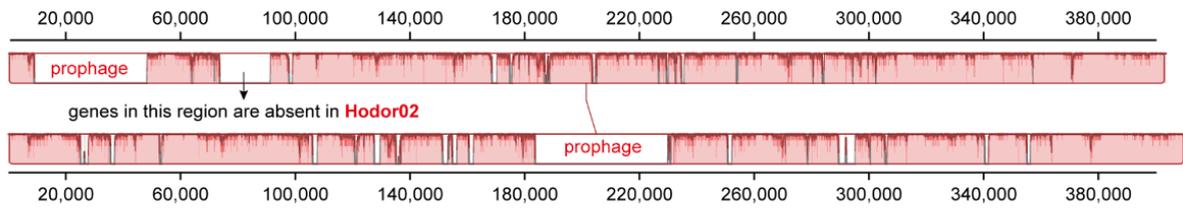
**Supplementary Figure S3 | Hodor genes similar to those in Bacillota.** Maximum-likelihood phylogenetic trees of Hodor-encoding Sortase B (a), SecY (b), SecG (c), and SecE (d), and reference sequences from the HRGMv2\_100 protein database. The outer color strip denotes the Hodor group or the taxonomic origin of each sequence.



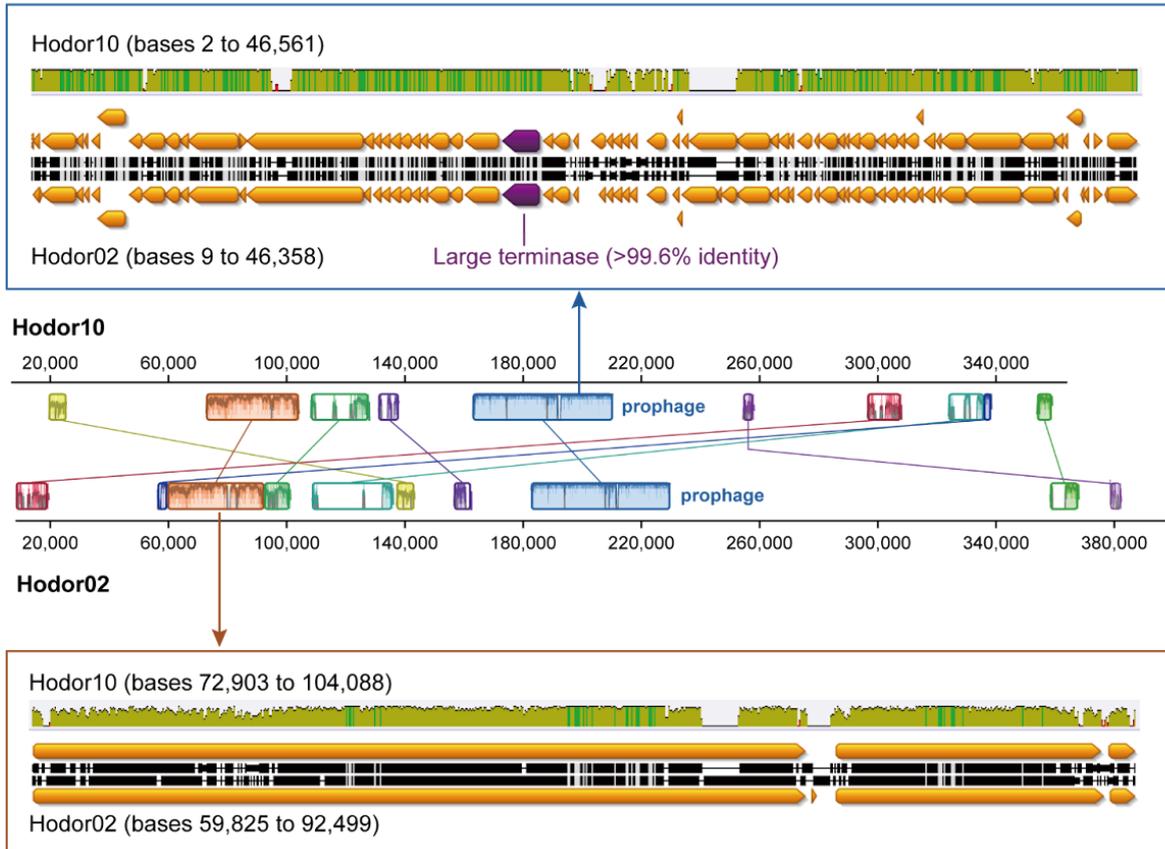
**Supplementary Figure S4 | Protein family clustering analysis of the 38 curated Hodor genomes.**

Protein-family distribution across 38 Hodor genomes. Heatmap showing the presence/absence profiles of protein families across 38 Hodor genomes. Rows represent protein families and columns represent genomes. Hodor genomes are colored by the indicated groups (Group1, blue; Group2, yellow). Plasmid marker genes are highlighted in red. Protein families were defined by clustering predicted proteins with mmSeqs2 (--min-seq-id 0.7 -c 0.8 --cov-mode 1 --cluster-mode 2), and only those detected in at least 3 genomes are shown here. The proteins of the prophage identified in one G1 and three G2 Hodor genomes are highlighted. Rows and columns were hierarchically clustered based on Euclidean distance.

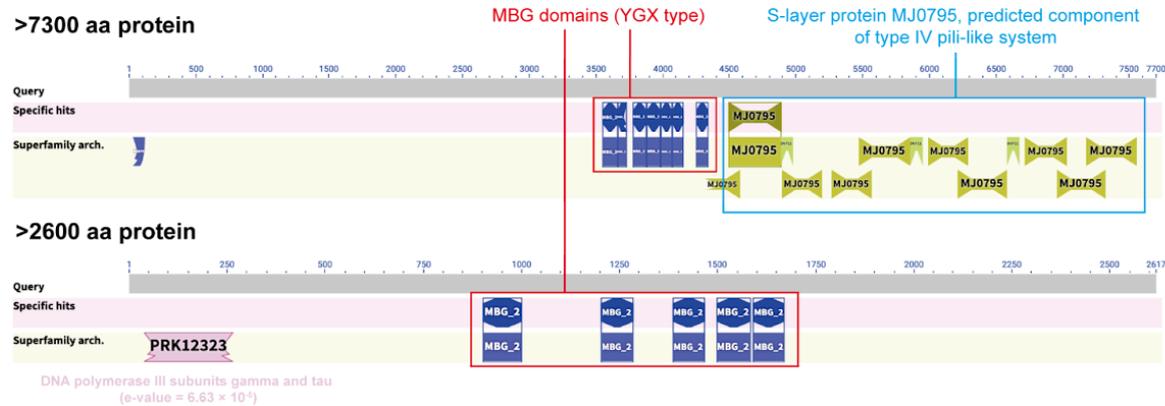
**a Hodor01 (G1, top) vs Hodor02 (G1, bottom)**



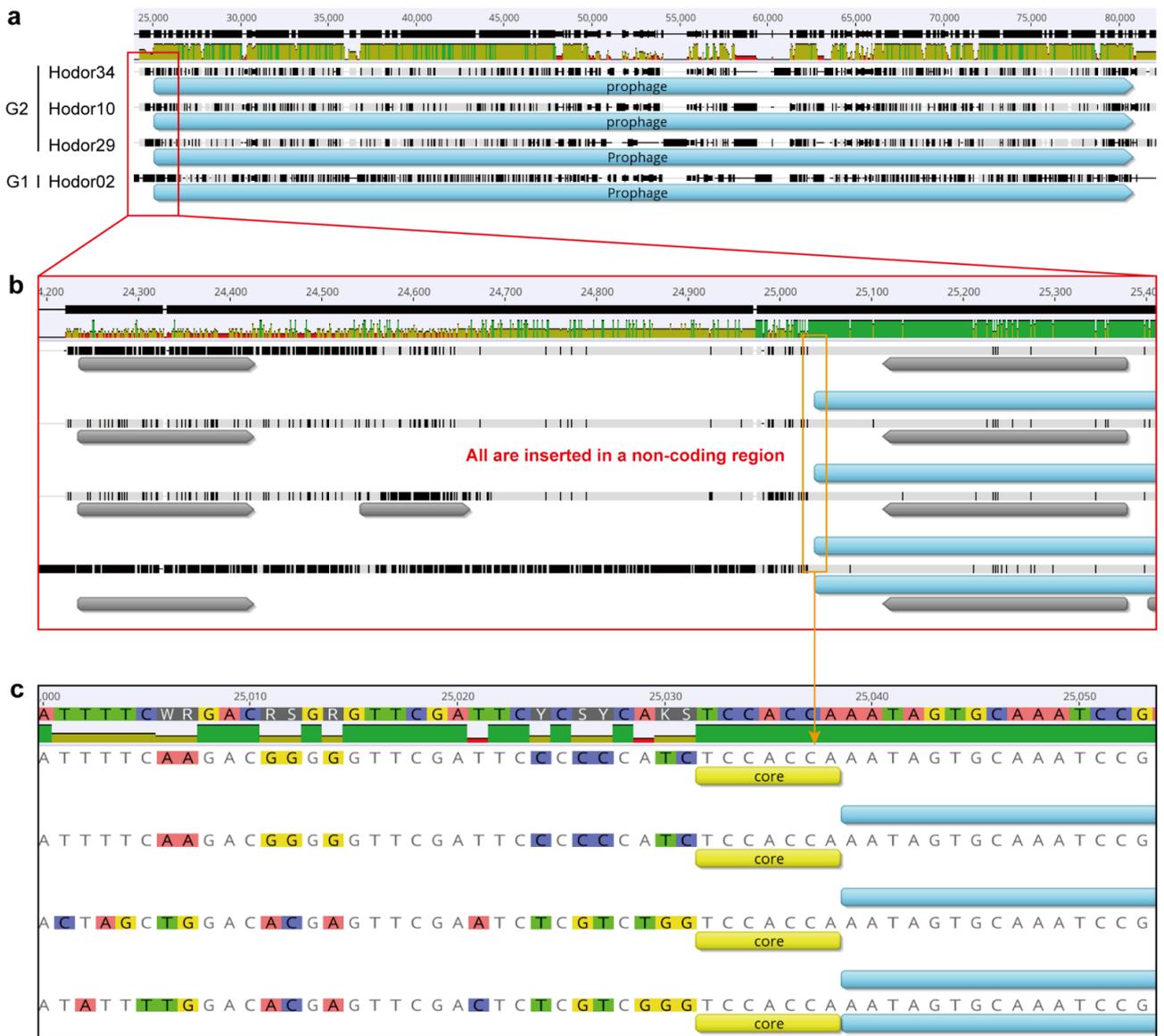
**b**



**c**

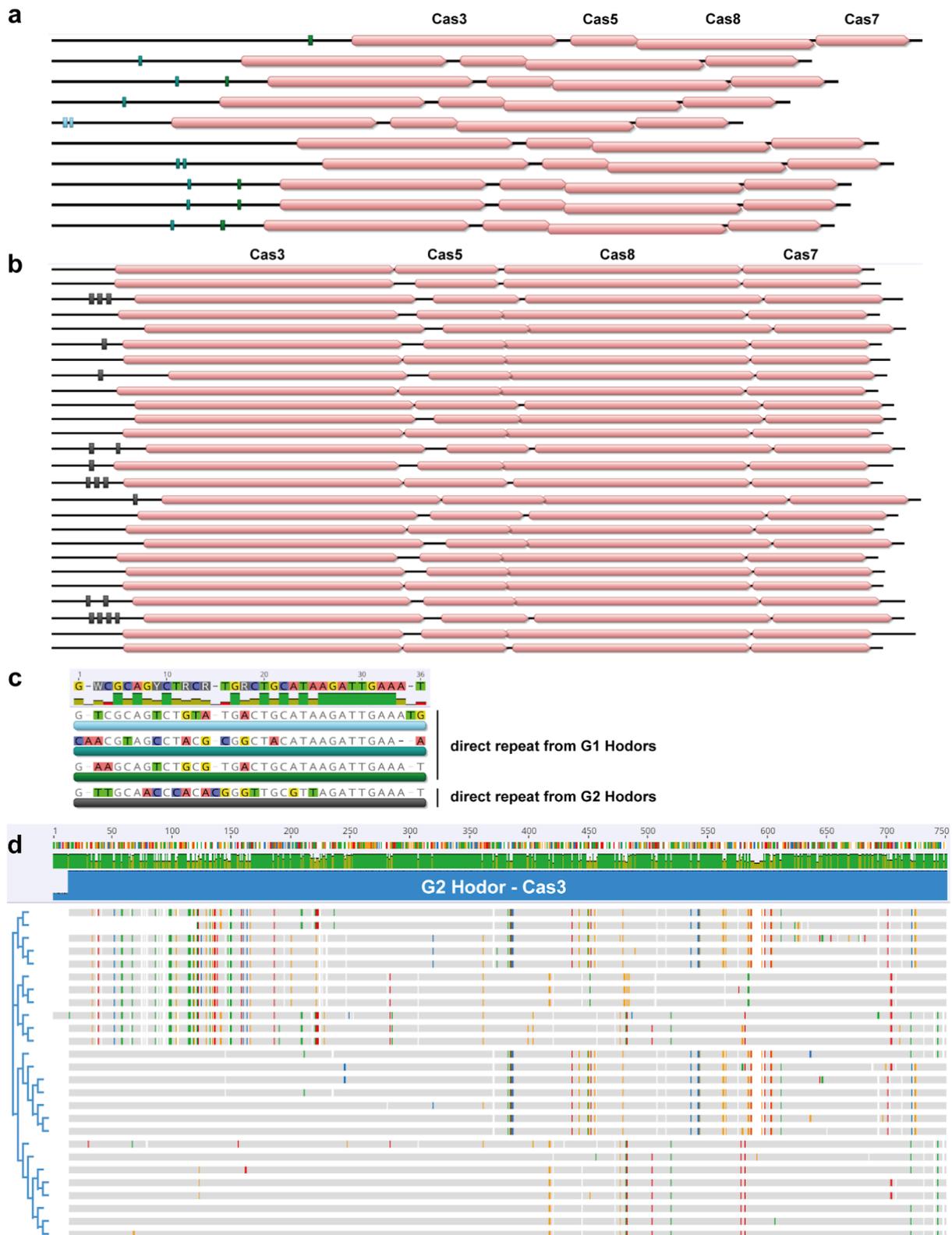


**Supplementary Figure S5 | Detailed information on high similarity regions between Hodor10 and Hodor02.** (a) Hodors from the same group contained different prophages. The genome comparison of Hodor01 and Hodor02 is shown as an example. (b) Hodors with divergent genomes share a similar prophage. The genomes of Hodor10 and Hodor02 belong to different Hodor groups, but the embedded prophage genomes are highly similar. (c) Information on the conserved gene block encoding surface proteins. The genome alignment was performed using Mauve<sup>1</sup> within Geneious Prime Build 2025-03-24<sup>2</sup>.

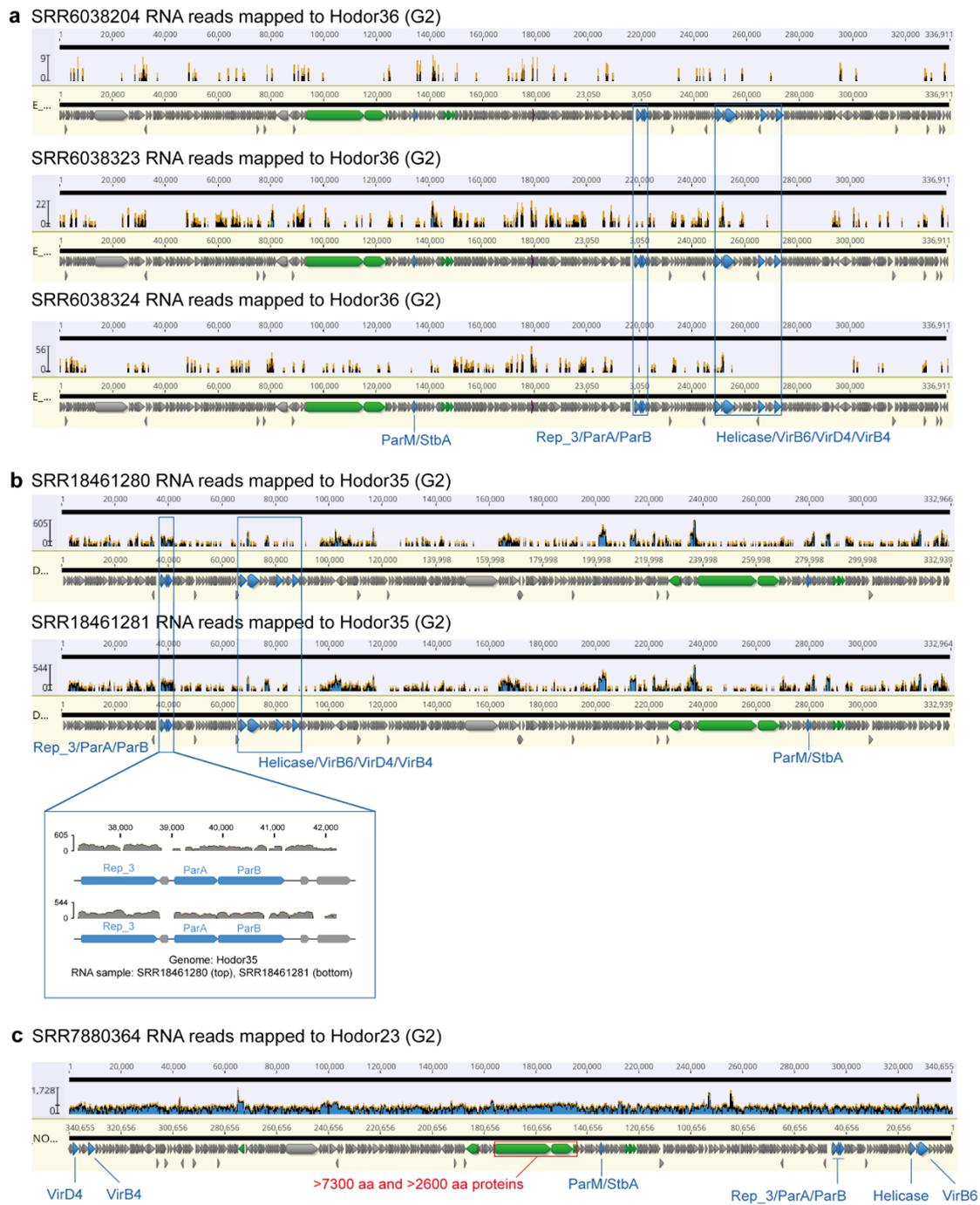


All have the same core sequence  
“TCCACCA” for integration recognition

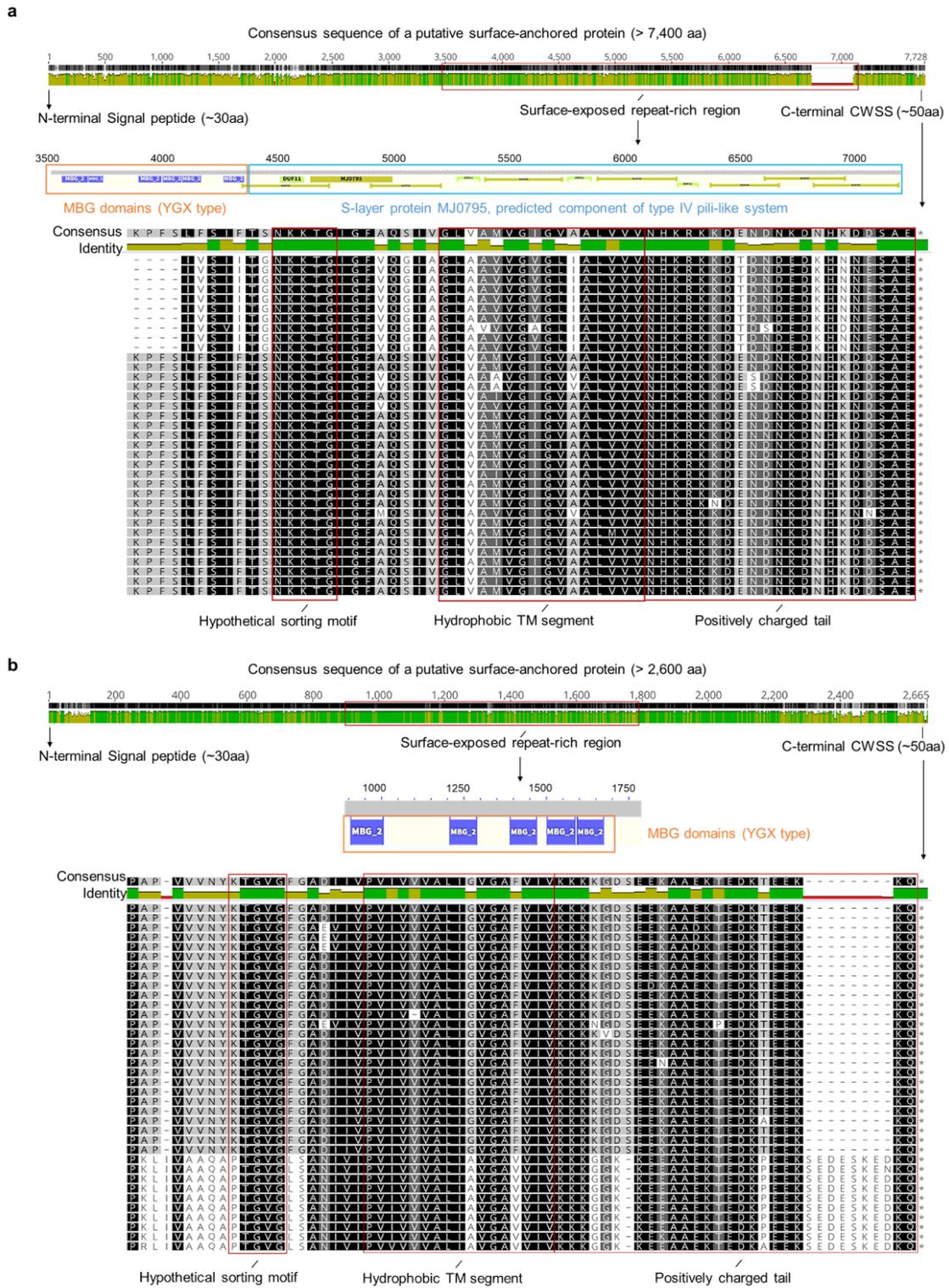
**Supplementary Figure S6 | The conserved integrase location of prophages in Hodor02, 10, 29, and 34.** (a) The prophage regions extracted from the whole Hodor genome alignments. (b) Zoom-in to show that the prophages are all inserted into a non-coding region in the Hodor genomes. (c) Zoom in to show that the core sequences for integration recognition are the same for all prophages.



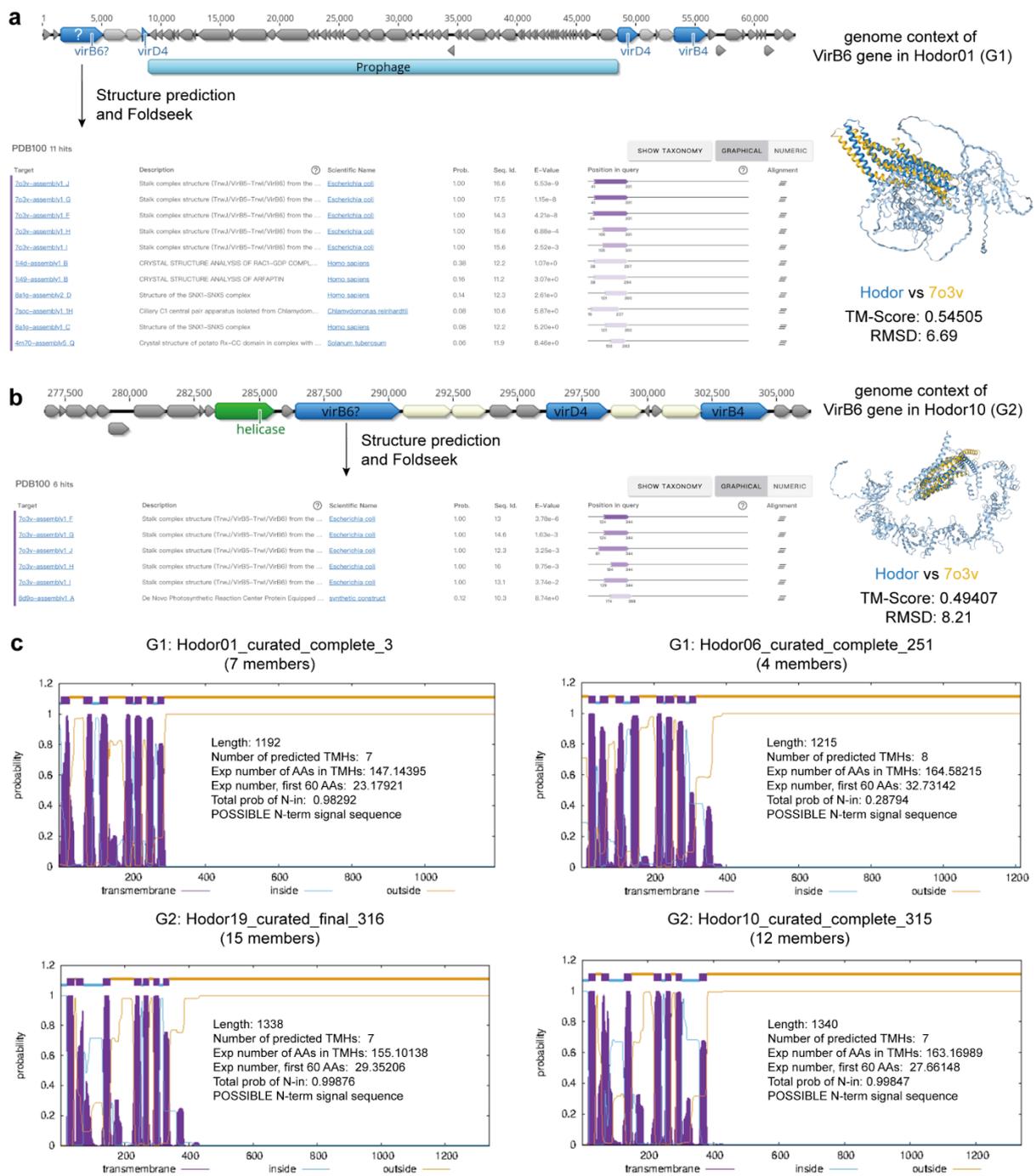
**Supplementary Figure S7 | CRISPR-repeat analyses of Logan retrieved G2 Hodor contigs.** (a) The CRISPR-Cas system in G1 Hodors. (b) The CRISPR-Cas system was in G2 Hodors. In (a) and (b), the squares indicate the predicted direct repeat sequences. (c) Comparison of direct repeat sequences predicted from G1 and G2 Hodors. (d) The phylogeny and alignment of the Cas3 proteins in the CRISPR-Cas systems of G2 Hodors. The Cas3 proteins in G1 CRISPR-Cas systems were almost identical and thus not shown here.



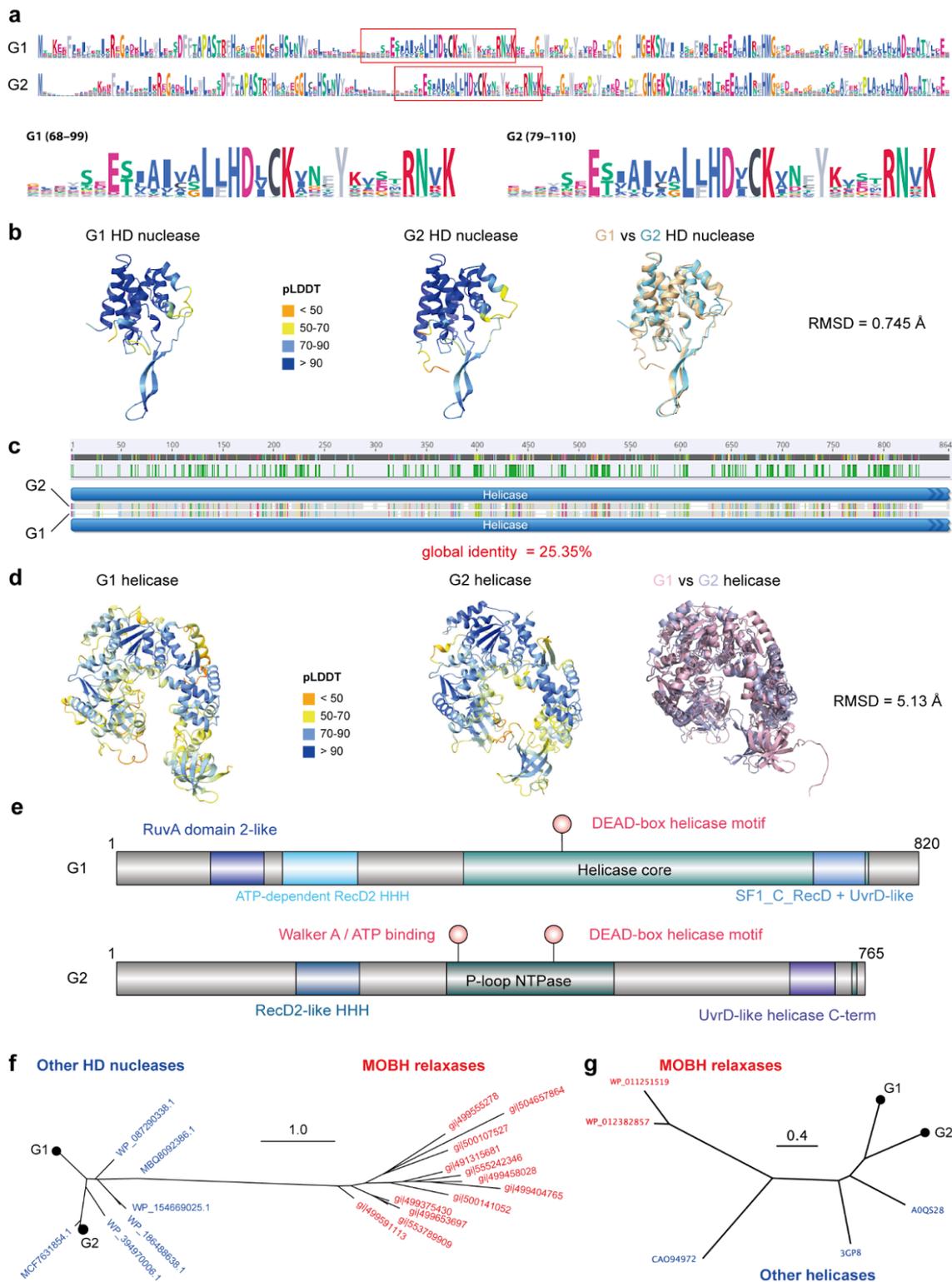
**Supplementary Figure S8 | Transcriptional activities of Hodors.** (a) The RNA reads from three time points were mapped to the Hodor36 genome of G2. (b) The RNA reads mapped to the Hodor35 genome of G2. The Hodor35 genome was reconstructed from the stool sample of a pediatric patient, which is used as an example, along with two corresponding RNA samples, one from a distal colon sample (SRR18461280), the other from a proximal colon sample (SRR18461281)<sup>3</sup>. (c) The RNA reads mapping profile of the Hodor23 genome of G2. The >7300 aa protein and its neighborhood genes are highlighted with a red box. The NCBI SRA id of the RNA dataset is listed above the figure.



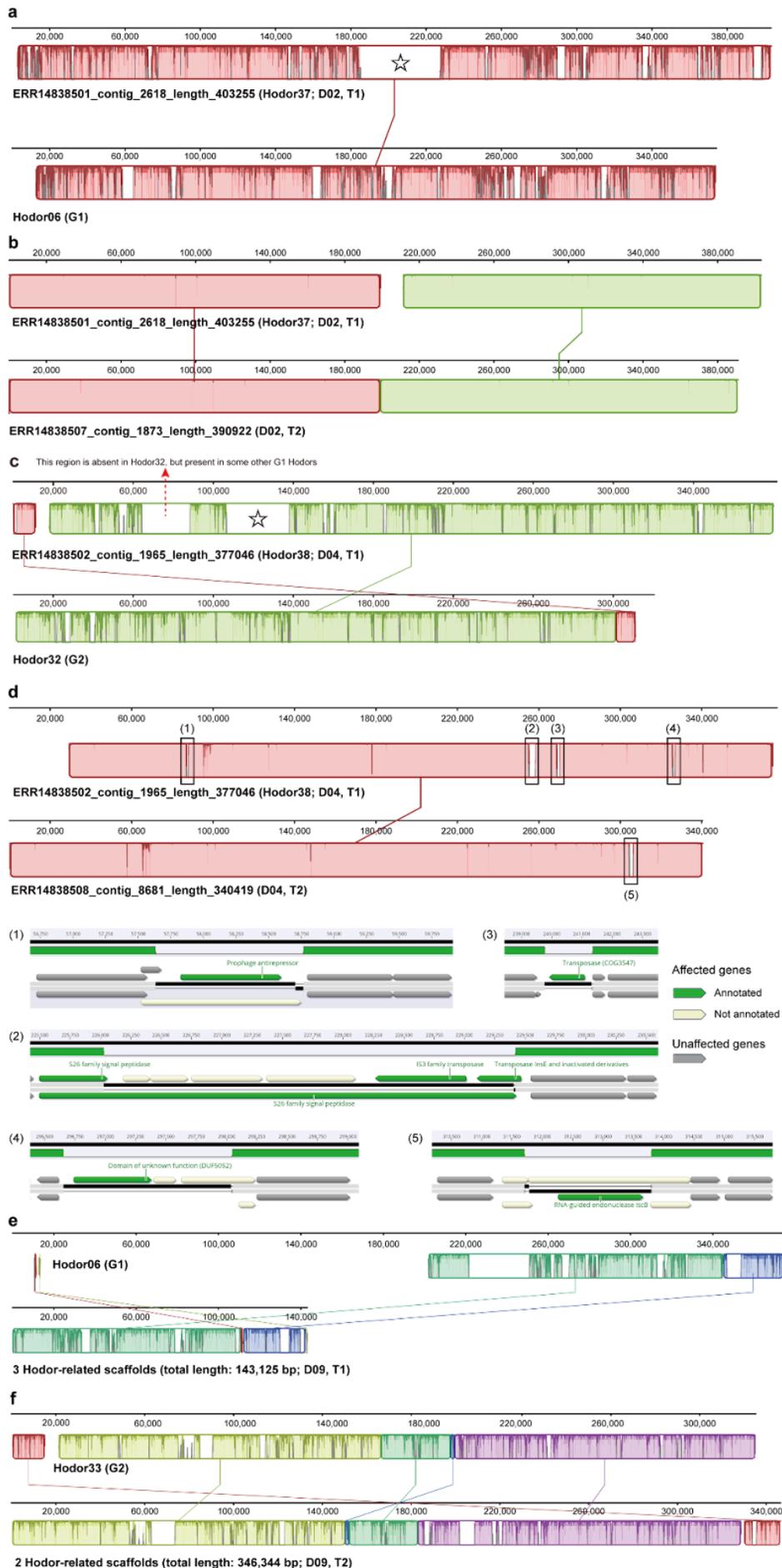
**Supplementary Figure S9 | Architectural features of two putative sortase-anchored surface proteins.** Two large proteins (a and b) with an N-terminal signal peptide, a central surface-exposed repeat-rich region, and a C-terminal cell wall sorting signal (CWSS), consistent with the canonical architecture of sortase-anchored surface proteins. For each protein, multiple-sequence alignments of the C-terminal CWSS region highlight the putative sorting motif, the hydrophobic transmembrane (TM) helix, and the positively charged cytosolic tail. The surface-exposed repeat-rich region was annotated using online BLASTp searches of the consensus sequences.



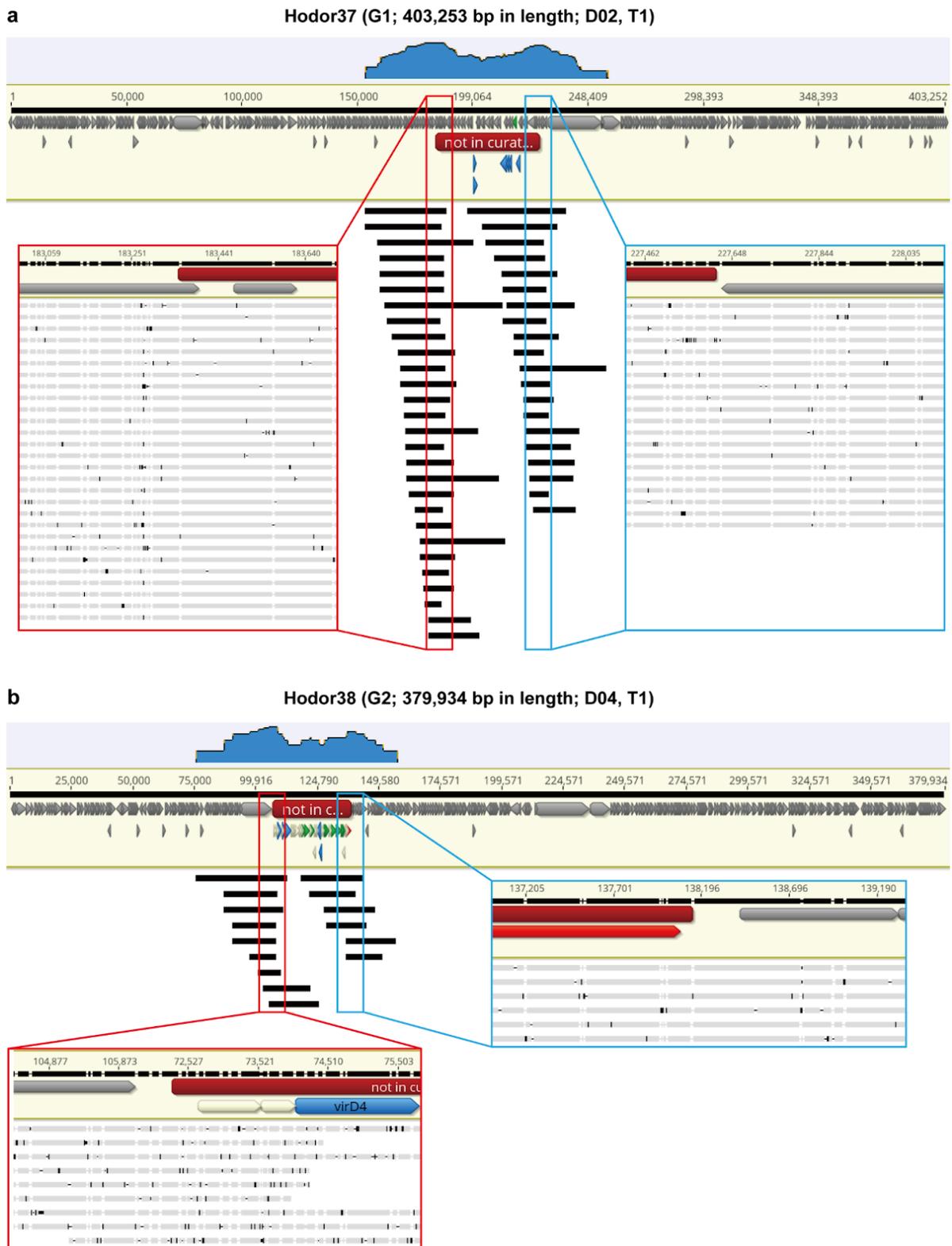
**Supplementary Figure S10 | VirB6 prediction.** The genetic context of the potential VirB6 gene encoded by (a) G1 and (b) G2 Hodors. The foldseek online search results (<https://search.foldseek.com/search>) and the protein structure alignment of the best hit are shown. Hodor01 from G1 and Hodor10 from G2 are used as examples. Both the predicted G1 and G2 VirB6 were with PDB 7o3v as the structure alignment hit. (c) The transmembrane analyses of the putative VirB6. The details of the transmembrane features are shown. The analyses were performed online with default parameters (<https://services.healthtech.dtu.dk/services/TMHMM-2.0/>). Note that the VirB6 sequences from the 38 curated Hodor genomes were assigned to four protein families based on  $\geq 70\%$  identity across  $\geq 80\%$  of their length.



**Supplementary | Analysis of Hodor HD nuclease and helicase proteins.** (a) The sequence logo of the G1 and G2 HD nucleases, the residues in the red boxes are zoomed in. (b) Sequence structure alignment of G1 and G2 HD nucleases. (c) Sequence alignment of G1 and G2 Hodor helicases. (d) Sequence structure alignment of G1 and G2 helicases. (e) Sequence-based functional annotation of G1 and G2 helicase proteins. Conserved helicase-related domains, including the P-loop NTPase core and the UvrD-like helicase C-terminal domain, are shown as colored blocks. Key conserved helicase residues are indicated by symbols, including a Walker A ATP-binding residue, the DEAD-box helicase motif, and a conserved C-terminal ATP-binding residue. (f) The phylogenetic relatedness of Hodor HD nucleases to other HD nucleases (blue) and relaxases of the MOB<sub>H</sub> family (red). (g) The phylogenetic relatedness of Hodor helicases to other helicases (blue) and relaxases of the MOB<sub>H</sub> family (red).

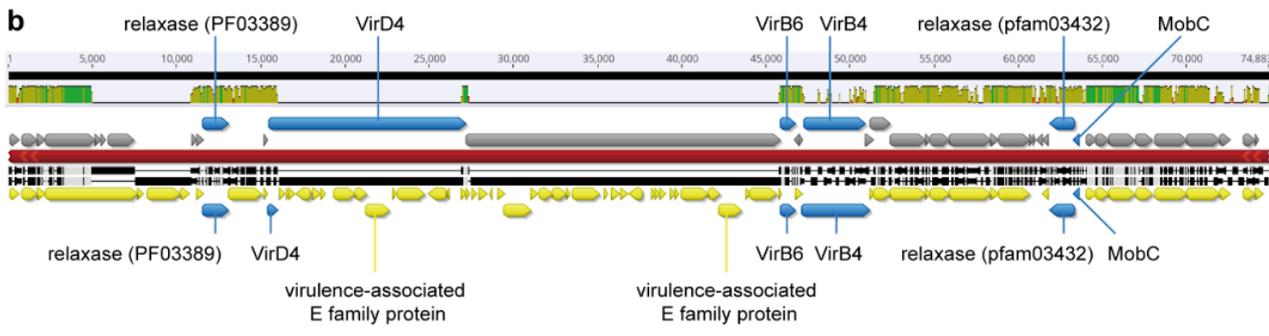
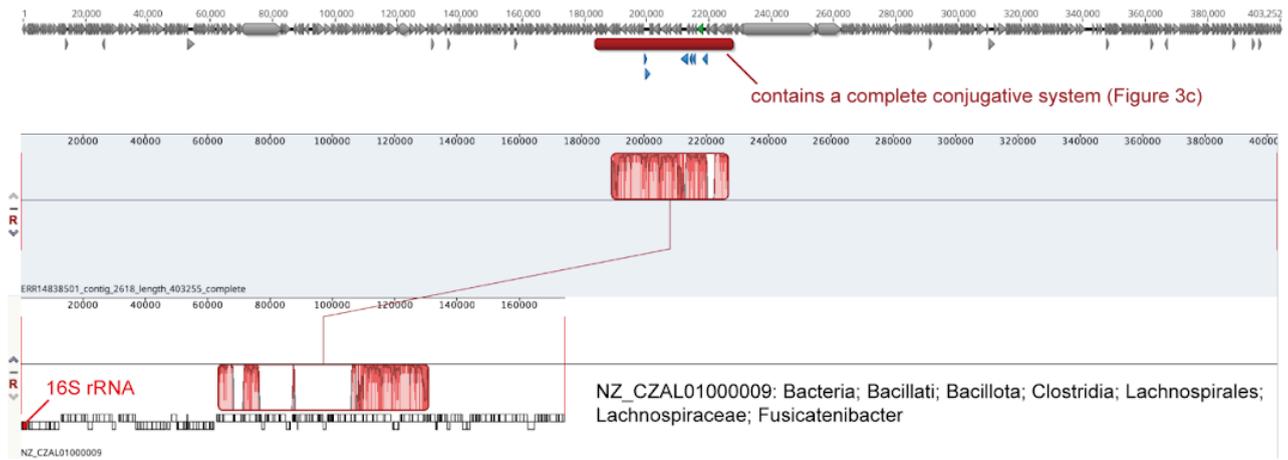


**Supplementary Figure S12 | The detection and comparison of Hodor genomes in metagenomic assemblies of the long-read datasets.** (a) The Hodor genome detected in the sample of individual D02 collected at the first time (T1) is most similar to the curated genome of Hodor06 from G1. (b) The Hodors in D02 collected from both time points (T1 and T2) are almost identical. (c) The Hodor genome detected in the sample of D04 from T1 is most similar to the curated genome of Hodor32 from G2. (d) The Hodors in D04 collected from T1 and T2 have highly similar genomes, with several gene gain and lose events. In detail, the Hodor collected at T2 lacks genes for (1) a prophage antirepressor, (2) a transposase, (3) two transposases and four hypothetical proteins, (4) a DUF5052 protein and two hypothetical proteins, while the Hodor collected at T1 lacks (5) an RNA-guided endonuclease IscB gene. (e) The Hodor in D09 from T1 is most similar to the curated genome of Hodor06 from G1. (f) The Hodor in D09 from T2 is most similar to the curated genome of Hodor33 from G2. The long-read datasets were reported by Wirbel et al. (Wirbel et al., 2026). The genome alignment was performed using Mauve<sup>1</sup> within Geneious Prime Build 2025-03-24<sup>2</sup>. In (a) and (c), the region marked by a pentagon indicates additional complete conjugative apparatus acquired in these genomes.

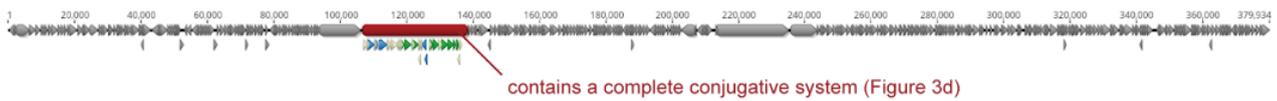


**Supplementary Figure S13 | The detection of complete conjugative systems in two additional curated Hodor genomes is not due to assembly errors.** We retrieved the long reads by BLASTn and mapped them to the region spanning the potential insertion location, and all mapped reads support the assembly. Note that the frequent SNPs are normal, which are due to the indels in long read sequencing. (a) The Hodor genome (Hodor37, which belongs to G1) assembled from the individual D02 at the first sampling point (T1). (b) The Hodor genome (Hodor38, belongs to G2) assembled from the individual D04 at the first sampling point (T1).

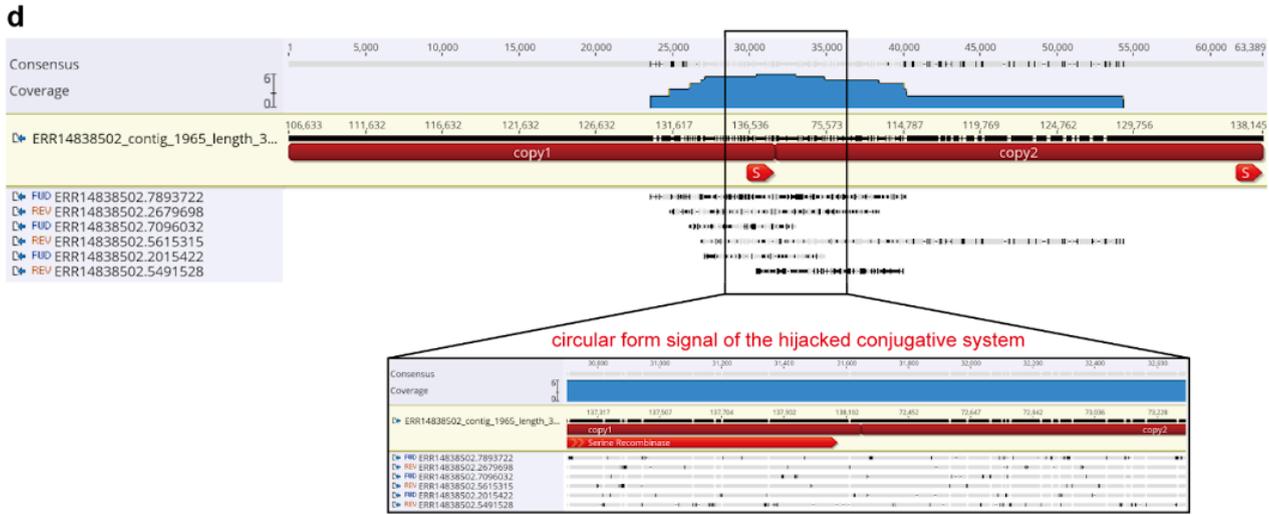
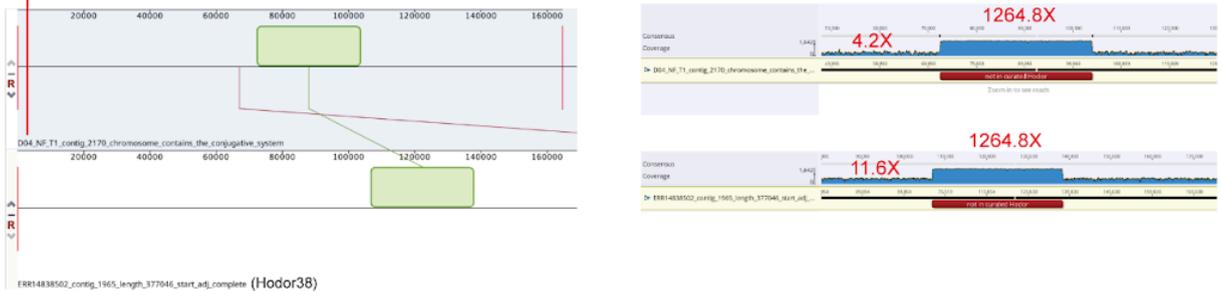
**a** ERR14838501\_contig\_2618\_length\_403255 (Hodor37; D02, T1)



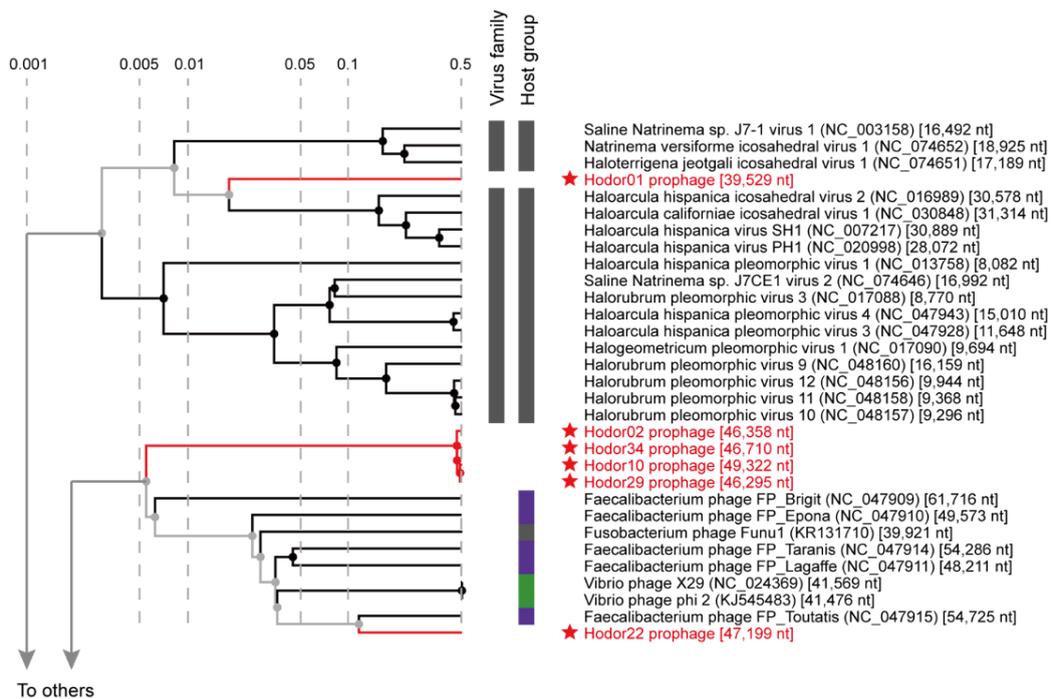
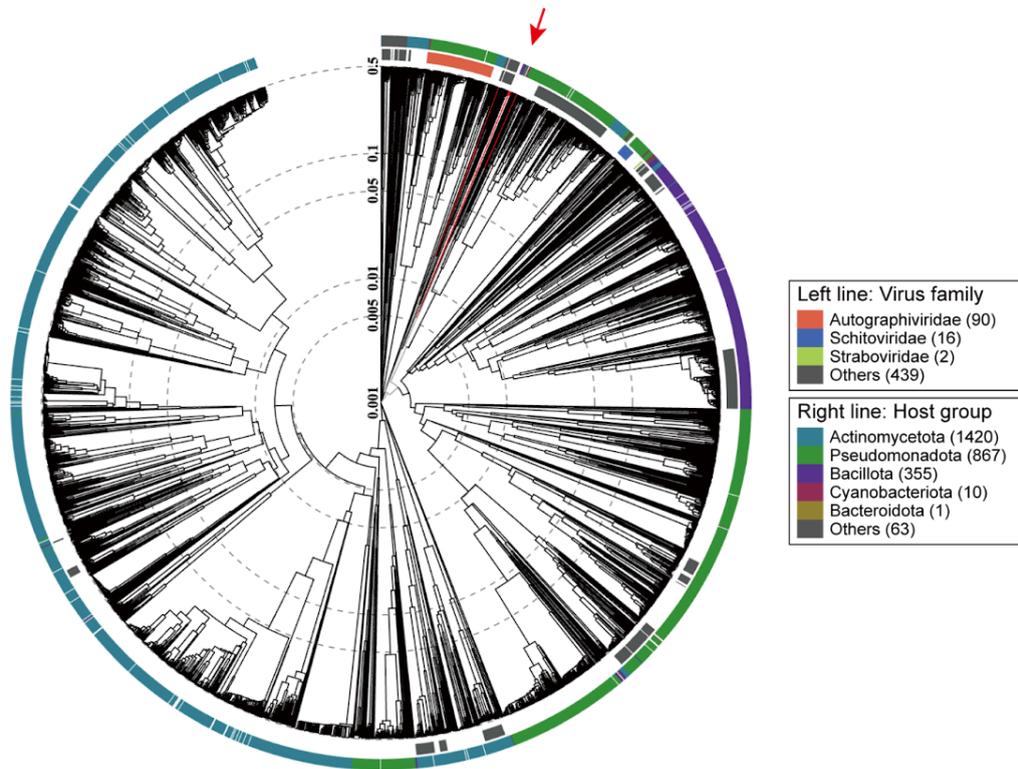
**c** ERR14838502\_contig\_1965\_length\_377046 (Hodor38; D04, T1)



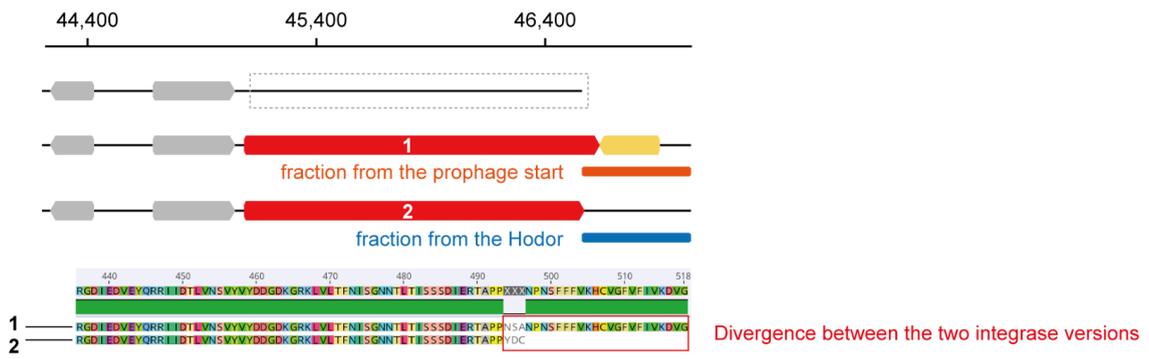
c\_Bacilli;o\_Erysipelotrichales;f\_Erysipelotrichaceae;g\_Holdemanela



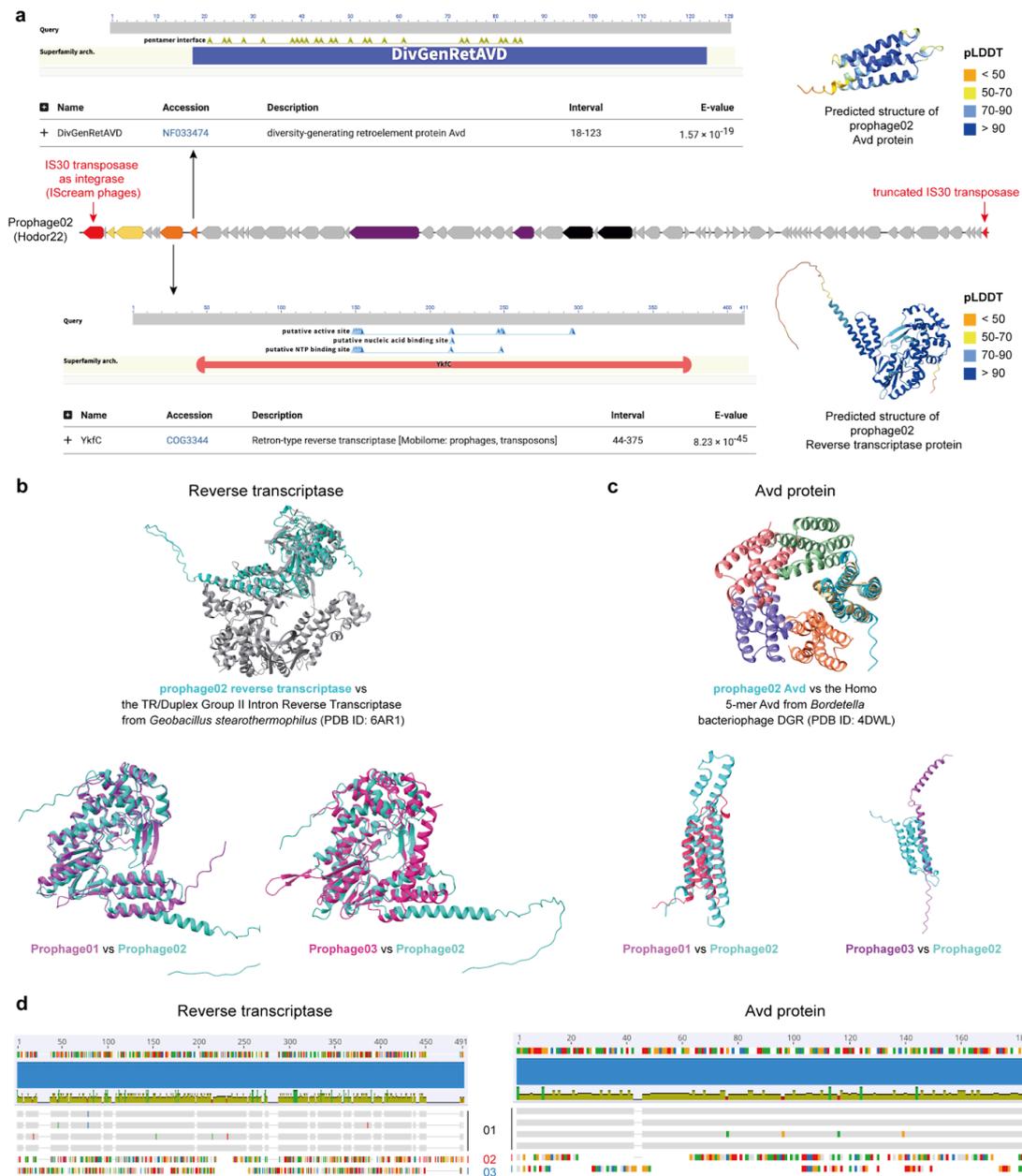
**Supplementary Figure 14 | Chromosomal ICE-like conjugative systems captured by Hodor37 and Hodor38.** (a) The additional conjugative module present in Hodor37 exhibits high sequence similarity to an ICE-like region within the NCBI contig NZ\_CZAL01000009, assigned to the genus *Fusicatenibacter* (Bacillota). The contig harbors a full-length 16S rRNA gene, showing 99.85% identity to *Fusicatenibacter saccharivorans* strain HT03-11 (NR\_114326.1), supporting its taxonomic assignment. (b) Gene synteny comparison between the Hodor37 conjugative region and the ICE-like region in NZ\_CZAL01000009. The conserved plasmid mobility genes are highlighted, including relaxase (PF03432) and MobC homologs, which share  $\geq 95.7\%$  and  $\geq 97.2\%$  amino acid identity, respectively. Two genes encoding virulence-associated E family proteins from NZ\_CZAL01000009 are indicated. (c) The conjugative module in Hodor38 is identical to a region from a 164,710-bp scaffold assembled from long-read sequencing of the same sample, taxonomically assigned to the genus *Holdemanella* (Bacillota). Read mapping indicates markedly elevated sequencing depth across the conjugative region ( $\sim 1264.8\times$ ) relative to both the host scaffold ( $\sim 4.2\times$ ) and the Hodor scaffold ( $\sim 11.6\times$ ), suggesting that this module exists in additional genetic contexts beyond either element. (d) Evidence for a circular intermediate of the conjugative module. Two tandem copies of the region were concatenated to detect reads spanning the junction. Six long sequencing reads traversed the artificial boundary, consistent with the presence of an extrachromosomal circular form of this genetic fraction in the sample.



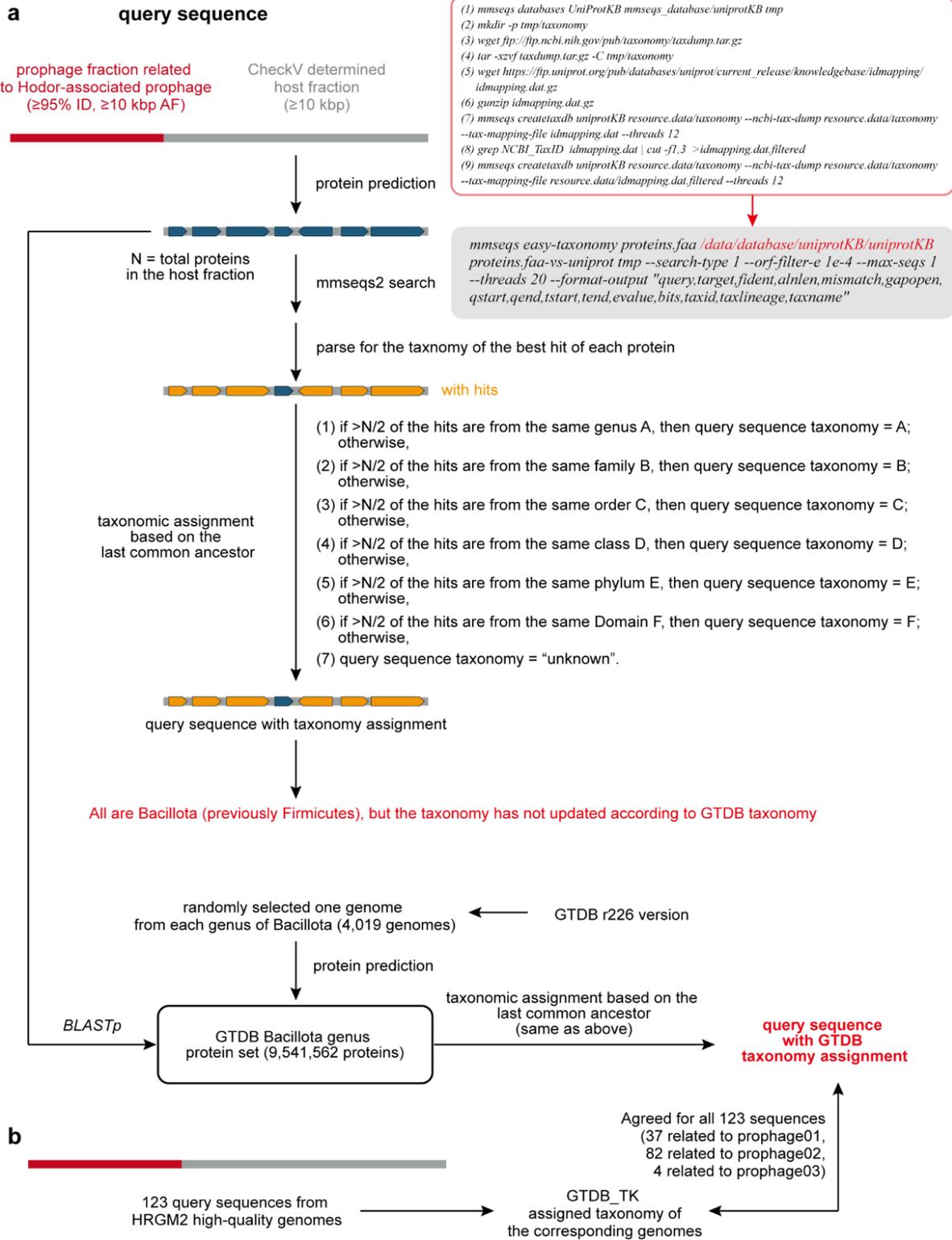
**Supplementary Figure S15 | The taxonomy and potential hosts of Hodor-associated prophages.** The Hodor-associated prophages were uploaded to vitytree (<https://www.genome.jp/vitytree/>)<sup>4</sup> for analyses based on the protein families shared with reference dsDNA phages in the vitytree database. The upper panel shows the position (red arrow) of the Hodor-associated prophages (highlighted in red and stars), with the genome length (nt) shown in brackets. The bottom panel indicates the closely related phages and their taxonomy and bacterial hosts.



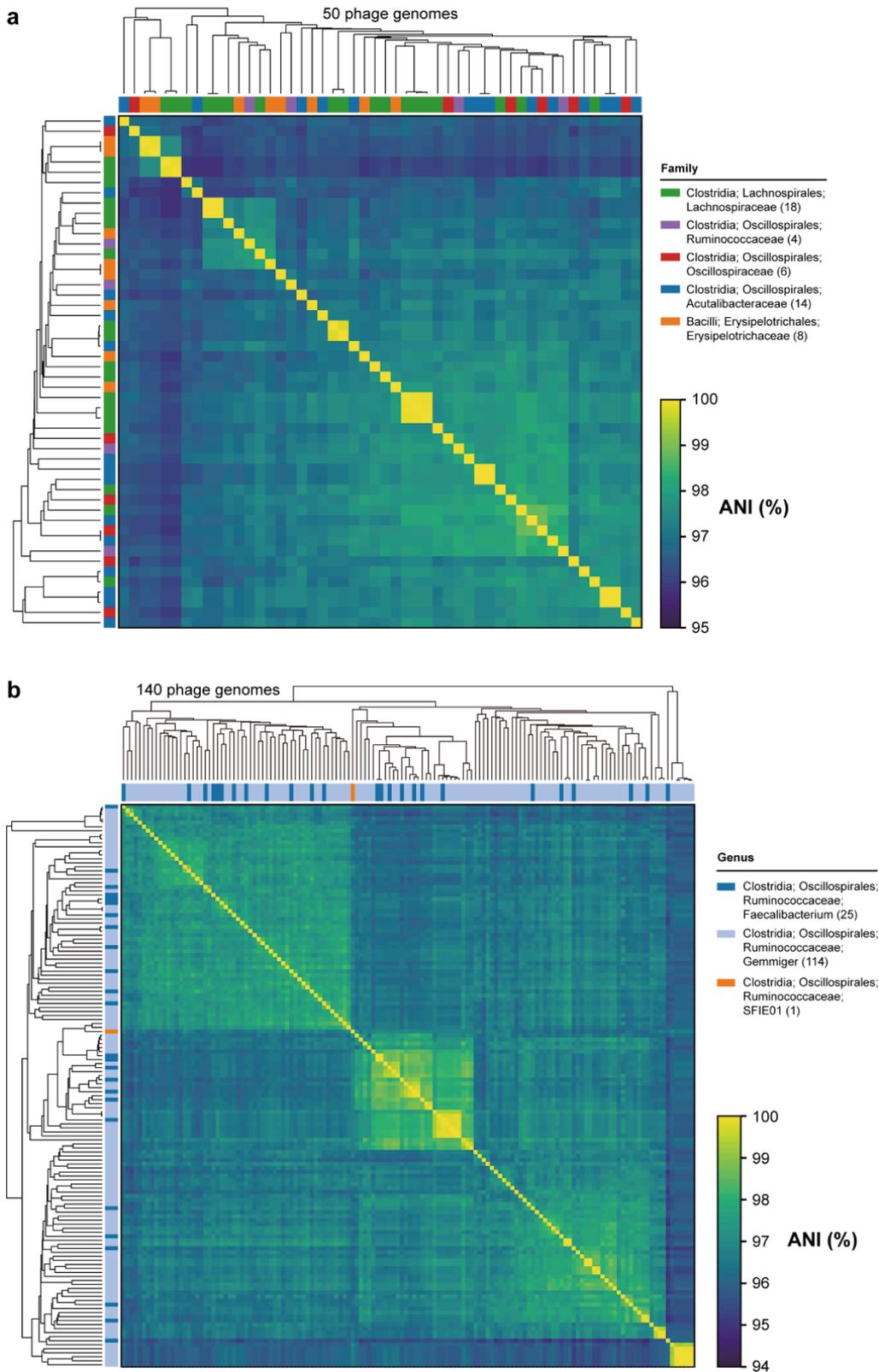
**Supplementary Figure 16 | The integrase genes in prophage01.** The gene could not be predicted in the linear prophage genome (the top illustration), unless a small fraction from the start of the prophage (the middle illustration) or the Hodor genome (the bottom illustration) is added to the end of the linear prophage genome. The red box indicates the divergence between the two versions of the integrase genes.



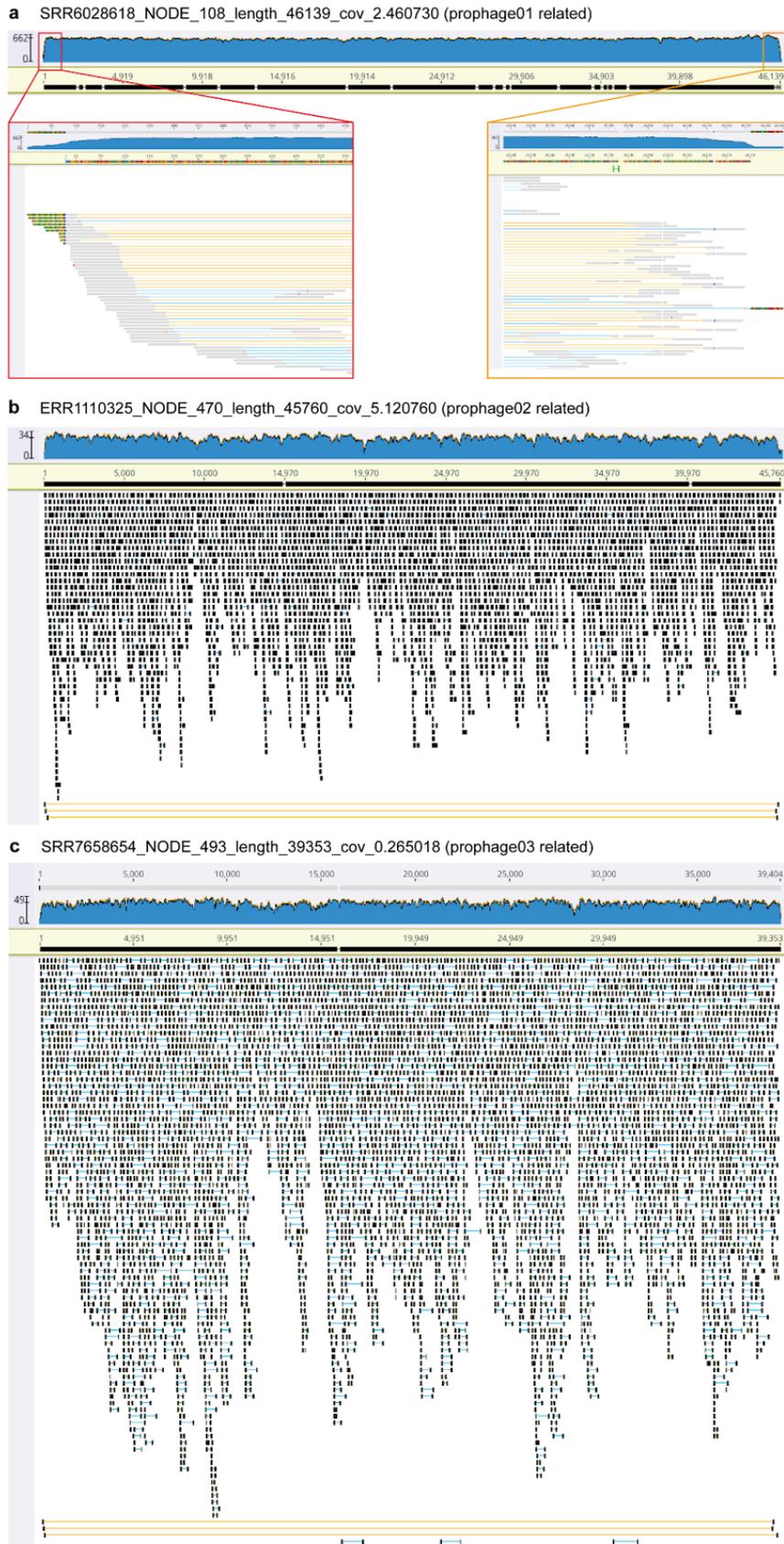
**Supplementary Figure S17 | Reverse transcriptase and Avd proteins identified in Hodor-associated prophages.** (a) The identification of the reverse transcriptase and the Avd protein in proophage02. The protein structures of the predicted genes are shown to the right. (b) The reverse transcriptase annotation in proophage02 was confirmed by structure alignment (vs PDB 6AR1). A similar reverse transcriptase was identified in proophage01 and proophage03. (c) The Avd annotation in proophage02 was confirmed by structure alignment (vs PDB 4DWL). A similar Avd protein was identified in proophage01 and proophage03. (d) Comparison of the reverse transcriptase and the Avd proteins in all three prophage groups.



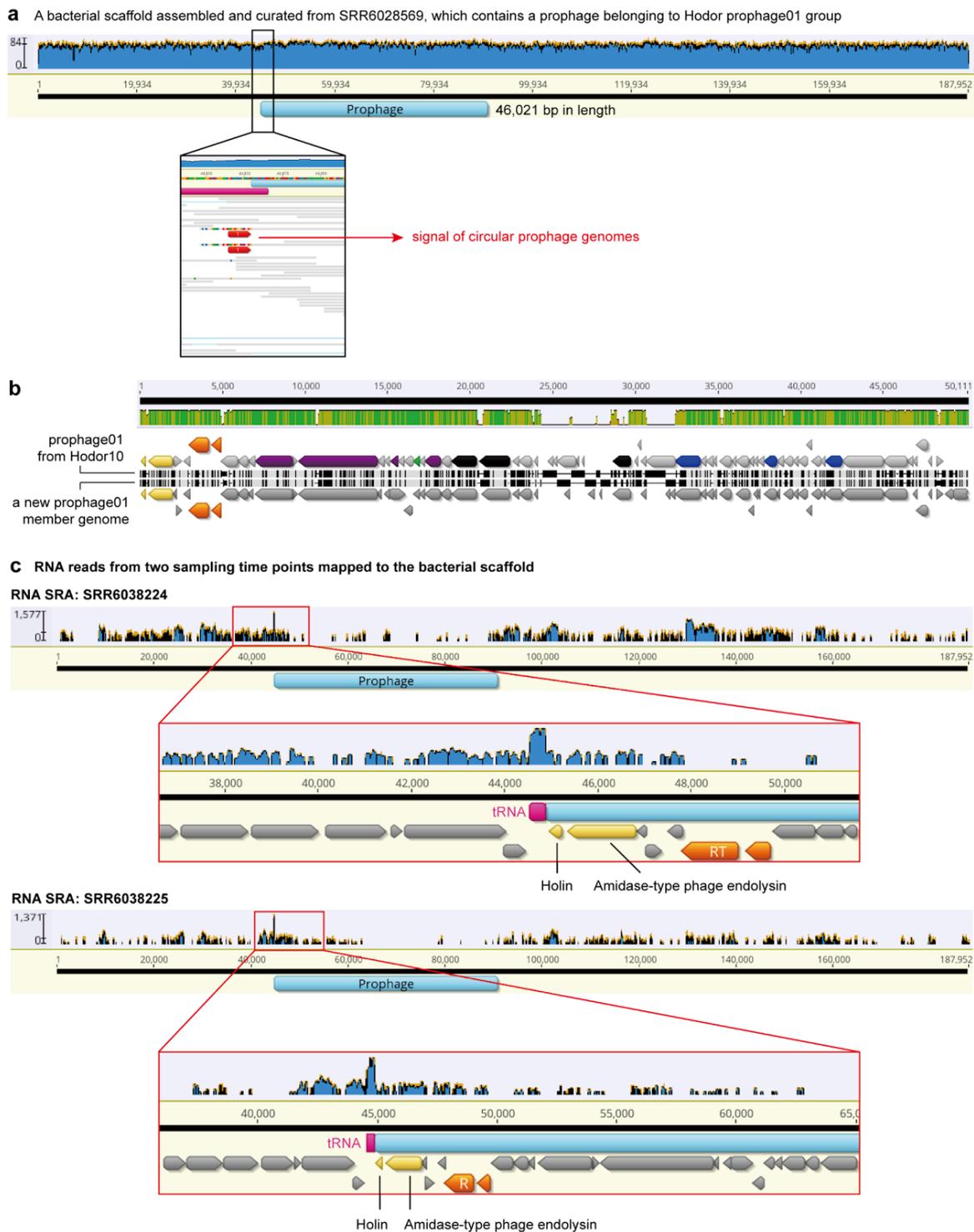
**Supplementary Figure S18 | Taxonomic assignment of host-derived regions flanking Hodor-associated prophages.** (a) Prophage-related regions ( $\geq 95\%$  nucleotide identity across  $\geq 10$  kb alignment fraction) were first delineated from host genomic fragments identified by CheckV (host fraction  $\geq 10$  kb). The protein-coding genes were predicted from the host-associated regions and searched against UniProtKB using mmseqs easy-taxonomy (search-type 1, E-value  $\leq 1e-4$ , best hit retained). Only proteins with significant matches were retained for downstream taxonomic inference. The resulting hits were assigned taxonomy based on UniProt lineage information (based on NCBI taxonomy) and subsequently reconciled with GTDB classification. Across all examined fragments, taxonomic affiliations consistently mapped to members of Bacillota (formerly Firmicutes), although UniProt annotations have not been fully updated to reflect the current GTDB taxonomy. (b) The taxonomic assignment approaches were approved by the Hodor-associated prophage containing scaffolds from the HRGM2 genomes.



**Supplementary Figure S19 | Average nucleotide identity (ANI) of within-cluster phage sequences.** (a) The Hodor-associated prophage01-related cluster (cluster01a in [Figure 5d](#) of the main text). All the phage sequences with host information to the family level and with a minimum length of 40 kbp are included (50 genomes). (b) The Hodor-associated prophage02-related cluster (cluster02 in [Figure 5d](#) of the main text). All the phage sequences with host information to the genus level within the family of Ruminococcaceae and with a minimum length of 40 kbp are included (140 genomes). For each taxon, the number of phage genomes is shown in brackets. The ANI values were extracted from the phage sequence clustering analysis on prophage01 and prophage02, respectively, which was based on an all-vs-all BLASTn search (see [Methods](#) in the main text).



**Supplementary Figure 20 | Reads mapping evidence of Hodor-associated prophages in lytic form.** The paired-end metagenomic reads mapped to the genomes of (a) prophage01, (b) prophage02, and (c) prophage03. The paired-end reads linked by a long yellow line indicate the genome is in circular form.

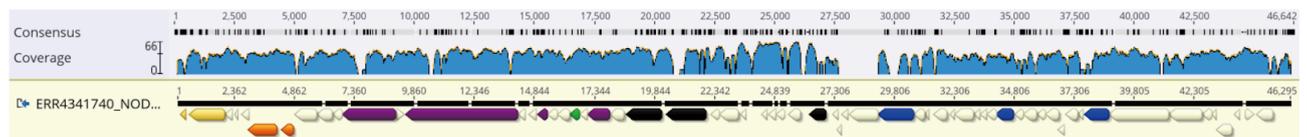


**Supplementary Figure S21 | Metatranscriptomic analysis to show the transcription of lysis-related genes.**

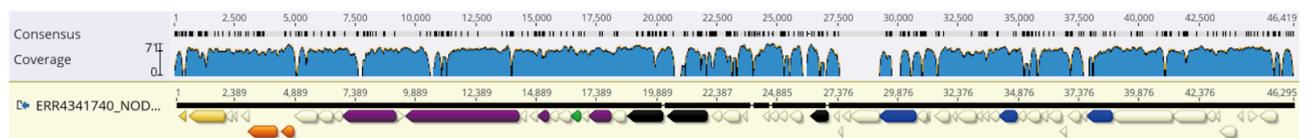
(a) The metagenomic reads mapping profiles of the prophage01 member genome. The zoom-in shows the signal of circular prophage genomes (only a small subset, though) in the sample. (b) The genome alignment of the prophage genome to the prophage genome in Hodor10 (G2). (c) The metatranscriptomic reads mapping profiles of the prophage01 member genome. Two metatranscriptomic SRA samples are shown. The zoom-in shows the transcriptional details, with the lysis-related gene names shown. The genome alignment was performed using Mauve<sup>1</sup> within Geneious Prime Build 2025-03-24<sup>2</sup>.

## a Virion-enriched reads mapped to prophage01 genome in Hodor29 (G2)

(1) Virion-enriched DNA SRA: SRR15465705

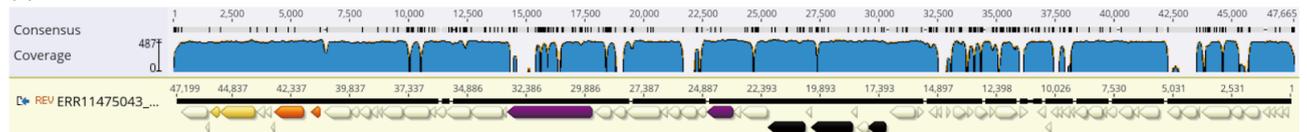


(2) Virion-enriched DNA SRA: SRR15465714

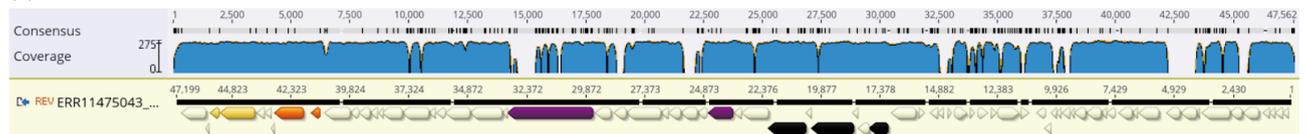


## b Virion-enriched reads mapped to prophage02 genome in Hodor22 (G1)

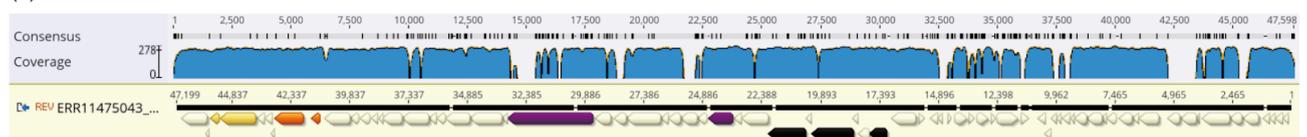
(1) Virion-enriched DNA SRA: SRR15465718



(2) Virion-enriched DNA SRA: SRR15465720



(3) Virion-enriched DNA SRA: SRR15465722



**Supplementary Figure S22 | The mapping profiles of virion-enriched metagenomic reads to prophage genomes in curated Hodor genomes.** (a) Reads from two samples mapped to the prophage01 genome. (b) Reads from three samples mapped to the prophage02 genome. All these reads were from the NCBI project PRJNA755142<sup>5</sup>. As the reference genomes used for mapping were the prophage genomes in the curated Hodor genomes, instead of those reconstructed from the virion-enriched samples, the profiles have some regions with no reads mapped. As prophage03-related sequences were only detected in 19 Logan-assembled samples, and none of them were virion-enriched (see [Supplementary Table 14](#) for details), we thus do not show them here. The mapping was performed using Bowtie2 and visualized in Geneious Prime Build 2025-03-24<sup>2</sup>.

## Supplementary references

1. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394–1403 (2004).
2. Kears, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
3. Yan, A., Butcher, J., Schramm, L., Mack, D. R. & Stintzi, A. Multiomic spatial analysis reveals a distinct mucosa-associated virome. *Gut Microbes* **15**, 2177488 (2023).
4. Nishimura, Y. *et al.* ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
5. Zhang, F. *et al.* Critical Assessment of Whole Genome and Viral Enrichment Shotgun Metagenome on the Characterization of Stool Total Virome in Hepatocellular Carcinoma Patients. *Viruses* **15**, (2022).