nature microbiology

Article

https://doi.org/10.1038/s41564-023-01579-5

Double-stranded RNA sequencing reveals distinct riboviruses associated with thermoacidophilic bacteria from hot springs in Japan

Received: 3 July 2023

Accepted: 8 December 2023

Published online: 17 January 2024

Check for updates

Syun-ichi Urayama[1,2] ✉, Akihito Fukudome[3], Miho Hirai[4], Tomoyo Okumura[5], Yosuke Nishimura[6], Yoshihiro Takaki[4], Norio Kurosawa[7], Eugene V. Koonin[8], Mart Krupovic[9] & Takuro Nunoura[6]

Metatranscriptome sequencing expanded the known diversity of the bacterial RNA virome, suggesting that additional riboviruses infecting bacterial hosts remain to be discovered. Here we employed double-stranded RNA sequencing to recover complete genome sequences of two ribovirus groups from acidic hot springs in Japan. One group, denoted hot spring riboviruses (HsRV), consists of viruses with distinct RNA-directed RNA polymerases (RdRPs) that seem to be intermediates between typical ribovirus RdRPs and viral reverse transcriptases. This group forms a distinct phylum, *Artimaviricota*, or even kingdom within the realm *Riboviria*. We identified viruses encoding HsRV-like RdRPs in marine water, river sediments and salt marshes, indicating that this group is widespread beyond extreme ecosystems. The second group, denoted hot spring partiti-like viruses (HsPV), forms a distinct branch within the family *Partitiviridae*. The genome architectures of HsRV and HsPV and their identification in bacteria-dominated habitats suggest that these viruses infect thermoacidophilic bacteria.

Recent metagenomics and metatranscriptomics analyses transformed the study of viromes. These approaches that do not require laborious virus cultivation have become the principal source of virus discovery[1]. Indeed, numerous virus groups across all taxonomic levels have been discovered. In particular, the diversity of RNA viruses that, in the current virus taxonomy, comprise the kingdom *Orthornavirae* within the realm *Riboviria* has expanded more than an order of magnitude through global metatranscriptome surveys[2–9].

Only one hallmark gene encoding the RNA-directed RNA polymerase (RdRP) is conserved across the entire kingdom *Orthornavirae*. Therefore, detection of the RdRP, typically using search methods based on sequence profiles, is the principal approach employed in metatranscriptome mining for riboviruses, and phylogenetic analysis of the RdRP is the basis of ribovirus taxonomy. Before the advent of massive metatranscriptome analysis, the viruses in this kingdom have been classified into 5 large phyla corresponding to

[1]Department of Life and Environmental Sciences, Laboratory of Fungal Interaction and Molecular Biology (donated by IFO), University of Tsukuba, Tsukuba, Japan. [2]Microbiology Research Center for Sustainability (MiCS), University of Tsukuba, Tsukuba, Japan. [3]Howard Hughes Medical Institute, Department of Biology and Department of Molecular and Cellular Biochemistry, Indiana University, Bloomington, IN, USA. [4]Super-cutting-edge Grand and Advanced Research (SUGAR) Program, Japan Agency for Marine Science and Technology (JAMSTEC), Yokosuka, Japan. [5]Marine Core Research Institute, Kochi University, Nankoku, Kochi, Japan. [6]Research Center for Bioscience and Nanoscience (CeBN), JAMSTEC, Yokosuka, Japan. [7]Faculty of Science and Engineering, Soka University, Hachioji, Japan. [8]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [9]Institut Pasteur, Université Paris Cité, Archaeal Virology Unit, Paris, France. ✉e-mail: urayama.shunichi.gn@u.tsukuba.ac.jp

major clades in the RdRP phylogeny[10]. Metatranscriptome studies largely validated the robustness of these phyla and additionally identified several candidate smaller phyla. The diversity of riboviruses across the lower taxonomy ranks demonstrated a nearly uniform increase, for example, roughly fivefold in one study that provided quantitative estimates[8].

Metatranscriptome mining yielded qualitative insights into the global view of the RNA virome. Traditionally, riboviruses have been recognized as the major component of the eukaryote virome, whereas the viromes of bacteria and archaea were dominated by DNA viruses[11,12]. For many years, only two small families of RNA viruses, each infecting a narrow range of bacteria, have been known: *Leviviridae* (single-stranded RNA (ssRNA) bacteriophages) and *Cystoviridae* (double-stranded RNA (dsRNA) bacteriophages). Metatranscriptome analyses revealed a much greater diversity of leviviruses than previously suspected, elevating this family to the rank of the class *Leviviricetes* that includes multiple orders and families[8,13–15]. The family *Cystoviridae* was substantially expanded as well[8]. For uncharacterized groups of viruses without a close relationship to any known groups, host assignment becomes a challenge. Nevertheless, several lines of evidence including (nearly) exclusive co-occurrence with bacteria, prediction of multiple virus genes preceded by prokaryote-type (Shine–Dalgarno (SD)) ribosome-binding sequences (RBS), identification of virus-encoded cell wall degrading enzymes, and most notably, targeting by reverse transcriptase (RT)-containing type III CRISPR systems strongly suggest that several previously uncharacterized groups of riboviruses infect prokaryotes[8]. Thus, the diversity of riboviruses infecting bacteria has been substantially underestimated and additional groups of such viruses most probably remain to be discovered.

Long dsRNA is a molecular marker of RNA virus infection[16]. The recently developed method of Fragmented and primer-Ligated DsRNA Sequencing (FLDS) made it possible to capitalize on the presence of (nearly) identical terminal sequences in genome segments of the same virus. This information enables one to identify multisegmented RNA virus genomes even if they did not show sequence similarity to known viruses[17–19]. Here we used FLDS to identify riboviruses associated with microbial consortia dominated by bacteria and archaea in several acidic hot springs in Japan. This analysis resulted in the identification of two distinct groups of riboviruses with multisegmented RNA genomes with organization typical of bacterial riboviruses.

Composition of small subunit ribosomal RNA and identification of RNA virus

To determine the composition of active microbial consortia in the hot spring water samples, total ssRNA sequencing reads were mapped on the small subunit (SSU) ribosomal RNA (rRNA) sequences from the Silva database (SILVA SSU v.138) using phyloFlash[20] (Fig. 1 and Supplementary Text). All samples were dominated by prokaryotes, with the H4, H5, Y66 and Oi samples, where RNA viruses were identified, containing <1% of eukaryotic SSU rRNA reads (Extended Data Table 1).

In FLDS, potential complete genomes of multipartite RNA viruses were obtained from samples H4, H5, Y66 and Oi (Extended Data Table 2). For the samples from the other stations, sequence libraries were successfully constructed except for the Ob sample, but no contigs representing potential complete genomes of RNA viruses in FLDS read mapping[18] were obtained.

Bipartite RNA virus from the hot spring and other ecosystems

FLDS of the Oi sample (79.3 °C, pH 2.2) yielded three populations of contigs (Fig. 2a) which collectively recruited ~50% of the clean FLDS reads from the Oi library. Among the contigs, we identified similar 5′- and 3′-terminal sequences (Fig. 2b), a characteristic feature of segmented RNA viruses[21]. On the basis of the similarity of the 5′- and 3′-terminal sequences, lengths of the segments and gene content, we concluded that two sets of contigs constituted genomes of a distinct group of bipartite RNA viruses. The segments were denoted RNA1, RNA2 and RNA2* (Supplementary Text and Extended Data Table 3). In total, we obtained complete sequences for 4, 4 and 2 divergent variants of segments RNA1, RNA2 and RNA2*, respectively (Fig. 2a). The similarity between the termini of the segments precluded assignment of all sets of segments to particular virus strains. However, segments RNA1a and RNA2a were most abundant and had longer conserved terminal sequences and were thus assigned to the same virus strain with a bisegmented genome.

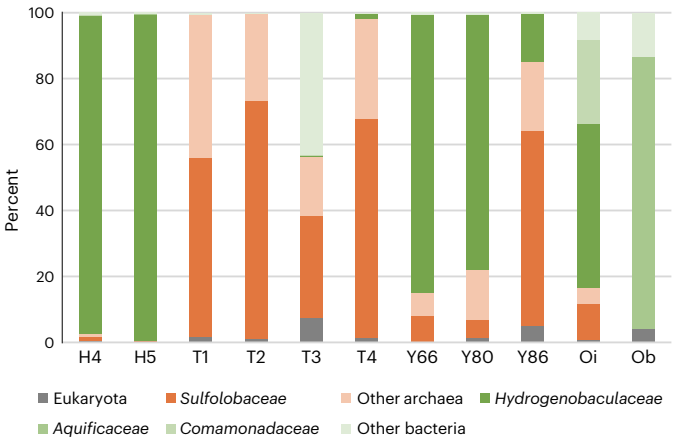


Fig. 1 | Composition of microbiomes associated with the hot spring samples. The composition was analysed on the basis of the mapped sequence reads on the rRNA sequences using Phyloflash. Details are given in Extended Data Table 1.

RNA1, RNA2 and RNA2* harboured 4–5, 5–6 and 5–7 open reading frames (ORFs), respectively (Fig. 2a). None of the predicted proteins encoded by these RNAs showed significant similarity (BLASTP *E*-value = 5×10^{-03}) to any protein sequences in public databases. Even the most sensitive profile–profile searches using HHpred yielded no significant (HHpred probability >90%) hits for any of the predicted proteins. However, HHpred searches queried with the amino acid sequence of ORF4 from the RNA1 segment produced a partial hit to several RdRPs. Although the hits were not significant (HHpred probability <90%) and encompassed only a small region of the RdRP (~15% of the target profile), the aligned region covered the diagnostic RdRP motifs B (SGxxxT, x – any amino acid) and C (GDD) (Extended Data Fig. 1a), so we pursued this clue further. However, despite several attempts, we were unable to convincingly identify RNA1_ORF4 of HsRV as an RdRP (Supplementary Text). Thus, we set out to enrich the sequence diversity of RNA1_ORF4 by reanalyzing the entire FLDS dataset. To this end, unmapped sequence reads were assembled and RNA1_ORF4 protein sequences were used as queries to search against the assembled contigs using BLASTX. This search yielded 10 additional RNA1_ORF4-like sequences encoded by H5_contig_1 from H5 and Oi_contigs_1–9 from Oi samples (*E*-value ≤ 1×10^{-05}) (Extended Data Table 4). The additional homologues detected in this search were combined with the 4 initially identified RNA1_ORF4 sequences and the produced multiple sequence alignment (MSA) was used as a query in an HHpred search against the PDB70 database. This search yielded significant hits (probability >90%) to various ribovirus RdRPs, although the aligned region remained limited (~15% of the target profiles). Collectively, these searches suggested that RNA1_ORF4 homologues are highly divergent RdRPs.

Using the MSA that included the identified RNA1_ORF4 homologues, a high-quality (average per-residue Local Distance Difference Test (pLDDT) = 90.7) AF2 model of the putative RdRP was obtained (Fig. 2c). Examination of this model revealed a topology typical of the

https://doi.org/10.1038/s41564-023-01579-5

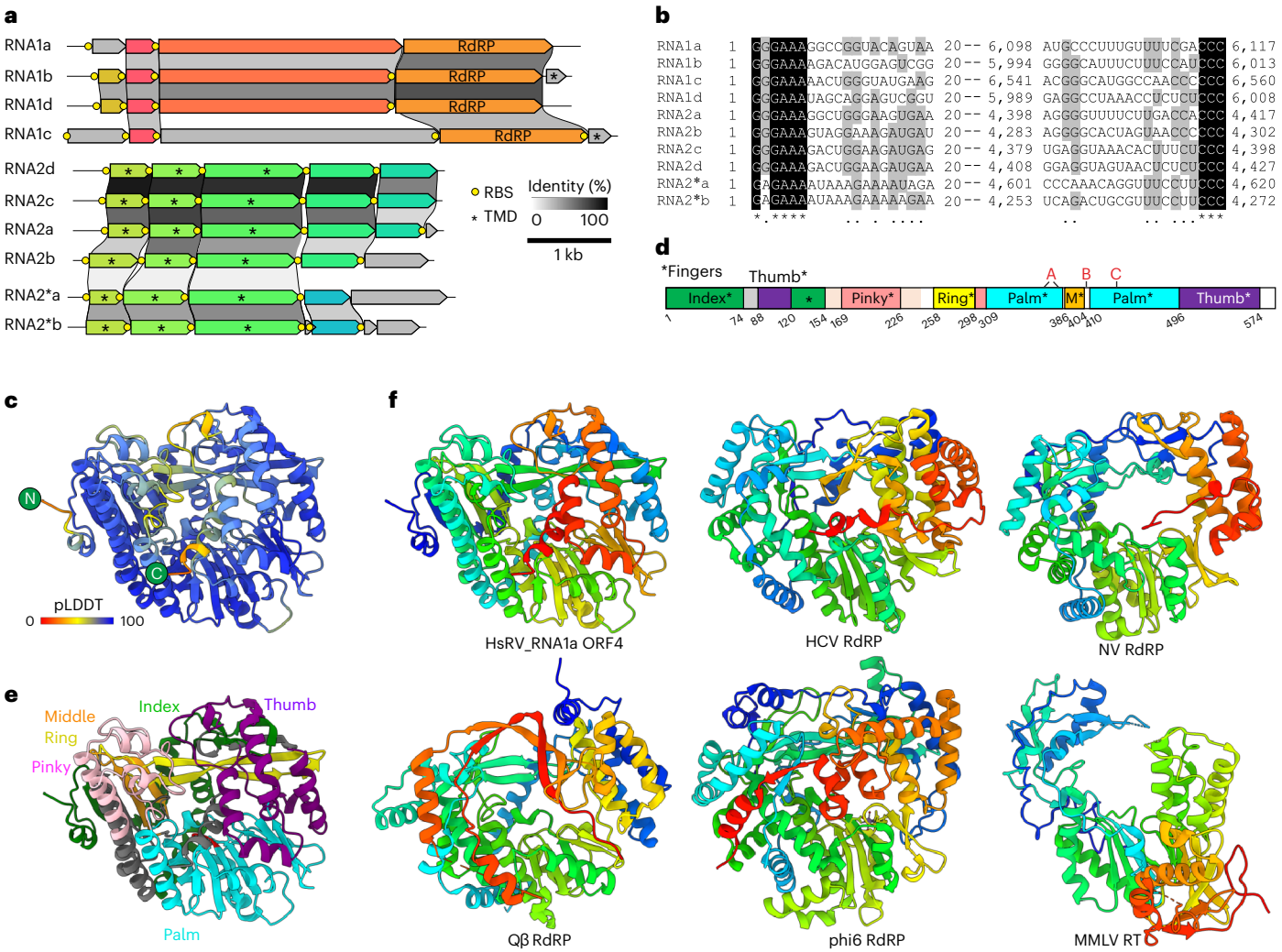


Fig. 2 | Unusual bipartite RNA virus genomes from the Oi hot spring. **a**, Genome organization and conservation of the three genomic segments (RNA1, RNA2 and RNA2*) of HsRV. ORFs encoding homologous proteins are shown as arrows with identical colours. Yellow circles represent predicted SD RBS. Asterisks denote putative genes encoding predicted transmembrane domain (TMD)-containing proteins. **b**, MSA of the 5′- and 3′-terminal regions of the coding strands of reconstructed genome segments. Black shading, 100% nucleotide identity; grey shading, >50% nucleotide identity. **c**, Quality assessment of the AlphaFold2 model of the HsRV RdRP. The structural model is coloured on the basis of the pLDDT scores (average pLDDT = 90.7), with the colour key shown at the bottom left corner. **d**,**e**, Domain organization of the HsRV RdRP. **d**, Schematic representation of the domain organization, with exact coordinates of each subdomain, including the five 'Fingers', indicated. M, middle finger. The positions of the motifs A, B and C are indicated. **e**, The structural model of HsRV RdRP coloured using the same scheme as in **d**. **f**, Comparison of the HsRV RdRP with homologues from other RNA viruses, including hepatitis C virus (HCV; PDB: 6GP9), Norwalk virus (NV; PDB: 1SH0), Qβ (PDB: 3MMP), phi6 (PDB: 1HHS) as well as RT from Moloney murine leukaemia virus (MMLV; PDB: 4MH8). The structures are coloured using the rainbow scheme, from blue N terminus to red C terminus.

palm-domain polymerases, with readily discernible 'Fingers', 'Palm' and 'Thumb' subdomains (Fig. 2d,e) and overall architecture similar to that of viral RdRPs (Fig. 2f), albeit with some unique structural features. In particular, the RNA1_ORF4 model displayed an extended and highly ordered 'Fingers' subdomain, with the 'fingertips' forming a 5-stranded β-sheet that is missing in other RdRPs and interacts with the 'Thumb' subdomain. The conserved motifs B and C identified by HHpred were located within the Palm subdomain, at positions equivalent to those in other RdRPs. Structural superposition of the Palm subdomains from different RdRPs allowed identification of the third core motif, A, in RNA1_ORF4 (see below). Thus, we concluded that RNA1_ORF4 encodes an RdRP and provisionally named the discovered bipartite virus 'hot spring RNA virus (HsRV)', with the strain harbouring segments RNA1a and RNA2a denoted HsRV1. The four RdRPs encoded by the complete RNA1 segments shared 37 to 75% pairwise amino acid sequence identity and thus appear to represent four distinct virus species (or even higher taxa). To characterize the diversity of HsRV-related RdRP in our FLDS data, the minor contigs including the aforementioned 10 sequences were analysed (Extended Data Fig. 2a). This analysis yielded several contigs with a high (>90%) identity to HsRV_RNA1b RdRP. In addition, several contigs with moderate (>60%) identity to HsRV_RNA1a or _RNA1b were detected. Y66 and Y86 also included a few contigs related to HsRV RdRP.

The sequence profile of the HsRV RdRP was used to search the previously described FLDS sequence data from coastal seawater samples[19], leading to the identification of two additional contigs (GenBank accessions: BDQA01000957 and BDQA01004869) encoding incomplete HsRV-like RdRPs. Searches against the IMG/VR database queried with these RdRPs yielded significant hits (E-value ≤ 1×10^{-05}) to three additional putative RdRPs encoded by apparently complete or near-complete 5.3–5.6-kb-long genome segments (Ga0456180_000042, Ga0393213_00017, Ga0169446_00510; Fig. 3a, Extended Data Fig. 1b, Table 5 and Supplementary Text).

https://doi.org/10.1038/s41564-023-01579-5

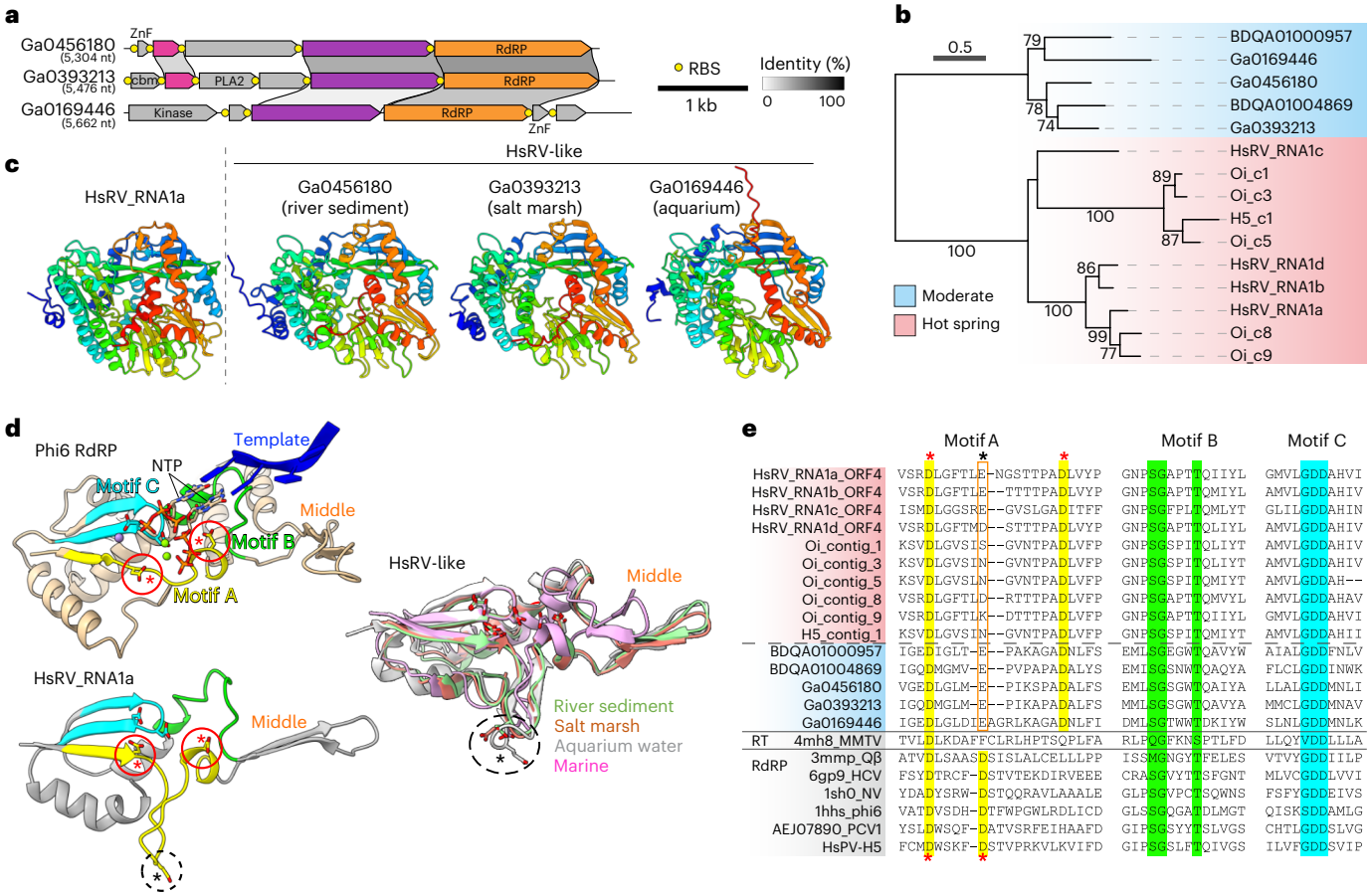


Fig. 3 | HsRV-like viruses from moderate environments. a, RdRP-encoding segments of HsRV-like viruses from non-extreme aquatic ecosystems. ORFs encoding homologous proteins are shown as arrows with identical colours. **b**, Maximum-likelihood phylogeny of the HsRV-like RdRPs encoded by viruses from extreme (pink) and moderate (blue) ecosystems. Node support was assessed using the SH-aLRT, with the corresponding values (%) shown on the branches. The scale bar represents the number of substitutions per site. **c**, Comparison of the HsRV RdRP with the homologues encoded by viruses from moderate aquatic ecosystems. The model was produced using AlphaFold2. The models are coloured using the rainbow scheme, from blue N terminus to red C terminus. **d**, Comparison of the catalytic cores encompassing the conserved RdRP motifs A (yellow), B (green) and C (cyan). Top: the structure of bacteriophage phi6 RdRP with the substrate nucleoside triphosphates (NTP) and template RNA strand (blue ribbon). Bottom: the HsRV RdRP. Middle: structurally superposed HsRV-like RdRPs from moderate ecosystems. The NTP and active site residues of motifs A and C are shown using the stick representation. The conserved aspartate residues of motif A are circled, with structurally equivalent residues indicated with red asterisks, whereas the non-conserved residue located in the loop facing away from the motif C in HsRV and related RdRP is indicated with the black asterisk. **e**, MSA of the conserved motifs of HsRV-like RdRPs from extreme (red shading) and moderate (blue shading) ecosystems with the corresponding regions from RdRPs and RT from other viruses (grey shading), including Moloney murine leukaemia virus (MMLV), hepatitis C virus (HCV), Norwalk virus (NV), PCV1 and hot spring partiti-like virus H5 (HsPV-H5). The sequences are indicated with the PDB or GenBank accession numbers. The conserved residues are shaded yellow, green and cyan, respectively, matching those in **d**. The conserved aspartate residues of motif A are highlighted in yellow, with structurally equivalent residues indicated with red asterisks, whereas the non-conserved residue in HsRV-like RdRPs located at the equivalent position as the second aspartate in other RdRPs is indicated with the black asterisk.

Ga0456180, Ga0393213 and Ga0169446 originate from floodplain (river sediments), salt marsh and aquarium samples, respectively. Phylogenetic analysis of HsRV-like RdRPs showed clear separation between viruses from the hot spring and those from moderate aquatic environments (Fig. 3b). Collectively, these results indicate that HsRV-like viruses are broadly distributed in both hot springs and non-extreme aquatic ecosystems.

Structural similarities between HsRV-like RdRPs and RTs

AF2 models of the three HsRV-like RdRPs from moderate ecosystems showed clear structural similarity with the HsRV RdRP, including the extended 'Fingers' subdomain (Fig. 3c). Another signature feature of these proteins is an unusual, extended RdRP motif A. In the canonical motif A, the two conserved Asp residues involved in catalysis and substrate discrimination[22,23], respectively, are separated by 4–5 residues and bracket the catalytic GDD residues of motif C (Fig. 3d,e). By contrast, in HsRV-like RdRPs, the second Asp residue of motif A is not conserved, and the corresponding residue is located in a loop facing perpendicularly away from motif C, suggesting that it cannot perform the same function. However, all analysed HsRV-like RdRPs contain an Asp (Asp*) which is located 12–14 residues away from the first Asp of motif A (Fig. 3e). Despite the extended spacing in the protein sequence, Asp* occupies a position equivalent to that of the second Asp of the canonical motif A (Fig. 3d,e) and is likely to be its counterpart involved in substrate discrimination.

We next performed structural clustering on the basis of the pairwise DALI *Z*-scores of the HsRV-like RdRPs together with selected RdRPs of other riboviruses, including putative phyla of RNA phages identified in recent metatranscriptome analyses[7,8,24] and RT encoded by eukaryotic viruses of the order *Ortervirales*[25] as well as non-viral RTs from bacteria and eukaryotes (Fig. 4a). The HsRV-like RdRPs from both hot springs and moderate aquatic ecosystems formed a tight cluster, underscoring their relatedness despite high sequence divergence. All previously known viral RdRPs formed a clade in the structure-based

https://doi.org/10.1038/s41564-023-01579-5

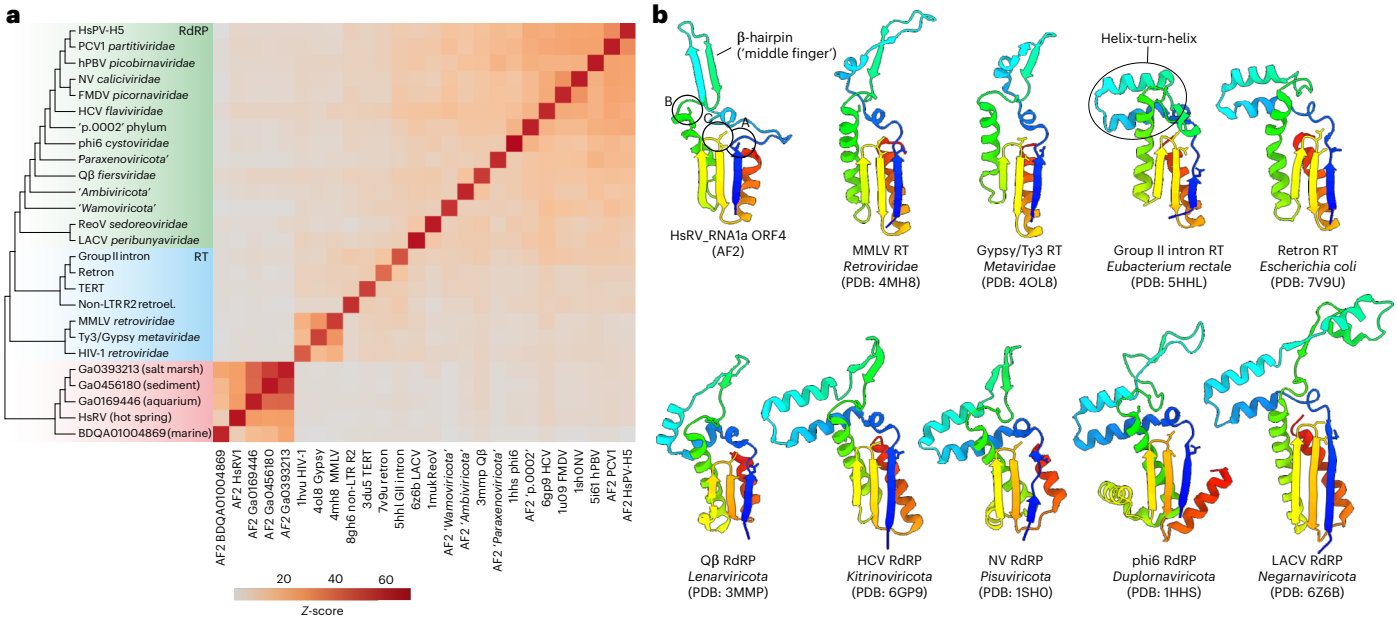


Fig. 4 | Structural relationships between RdRPs and RTs. a, Matrix and cluster dendrogram were constructed on the basis of the pairwise *Z*-score comparisons calculated using DALI. Different protein groups are highlighted with different background colours on the dendrogram: green, RdRPs from previously characterized viruses; blue, viral and non-viral RTs; red, HsRV-like RdRPs. The colour scale indicates the corresponding *Z*-scores. hPBV, human picobirnavirus; FMDV, foot-and-mouth disease virus; ReoV, reovirus; LACV, La Crosse virus; HIV-1, human immunodeficiency virus 1; TERT, telomerase RT; non-LTR R2 retroel., non-long terminal repeat R2 retroelement; AF2, AlphaFold2 model. For experimentally determined structures, the corresponding PDB accession numbers are indicated at the bottom of the matrix. **b**, Structural comparison of the core domain of RdRPs and RT encompassing the conserved motifs A–C. The structures are coloured using the rainbow scheme, from blue N terminus to red C terminus.

dendrogram, but the HsRV-like RdRPs remained separated from those (Fig. 4a). The two viral RdRP clusters were interspersed with the RTs, such that the viral RTs were the closest structural neighbours of the HsRV-like RdRPs. This result confirms the extreme divergence of the HsRV-like RdRPs and might reflect a closer relationship to viral RTs. This unexpected link was strengthened by the comparison of the 'Palm' subdomain of HsRV-like RdRPs with homologues from other riboviruses as well as viral and non-viral RTs. In RdRPs of riboviruses from 5 established phyla[10], the first β-strand (blue in Fig. 4b) containing motif A and the motif B-containing α-helix are separated by a characteristic helix-turn-helix (HTH) region followed by a β-hairpin corresponding to the 'Middle' finger subdomain (Fig. 2d,e). However, the HTH motif is absent in both the HsRV-like RdRPs and viral RTs. Notably, non-viral RTs, such as those from group II introns or retrons, contain the HTH motif but lack the β-hairpin region, which is compatible with the intermediate position of RTs between the two clades of viral RdRPs. Thus, the HsRV-like RdRPs might comprise an evolutionary intermediate between viral RdRPs and RTs. A BLASTN search against the metagenomic DNA sequences obtained from the hot springs did not detect HsRV-like sequences, suggesting that HsRV-like viruses are bona fide riboviruses that lack a DNA intermediate stage (Supplementary Text).

A thermoacidophilic partiti-like virus

Analysis of the FLDS RNA sequencing data from the stations H4 (68.8 °C, pH 3.2), H5 (69.7 °C, pH 3.1) and Y66 (68.7 °C, pH 2.7) revealed a bipartite virus genome unrelated to HsRV (Fig. 5a, Extended Data Table 2 and Fig. 2b). The genomic segments, RNA1 and RNA2, shared conserved 5′ terminal sequences and encoded one and two proteins, respectively (Fig. 5b). ORF1 of RNA1 was unambiguously identified as an RdRP, yielding significant BLASTP hits to RdRPs of members of the *Partitiviridae* family, with the best hit being to the unclassified Driatsky virus (QIS87951; *E*-value = 1×10^{-95}). We denoted this virus as hot spring partiti-like virus (HsPV). The similarity between the termini of the segments precluded assignment of all sets of segments to particular virus strains. However, on the basis of co-occurrence in the same sample and similar abundances, segment pairs RNA1_a and RNA2_b from sample H5 could be assigned to the same virus strain, HsPV1. Phylogenetic analysis of the RdRP sequence from diverse classified and unclassified partiti-like viruses showed that HsPVs and Driatsky virus (see below) were nested within genPartiti.0029 (Fig. 5c), a highly diverse, unclassified group defined in a recent metatranscriptome study[8]. The genPartiti.0029, including HsPV and Driatsky virus and several other subclades, formed a deep clade separate from all other partitiviruses. Thus, genPartiti.0029 can be considered a separate sister family to the bona fide *Partitiviridae*. AF2 modelling yielded an HsPV RdRP model closely similar to that of the RdRP of the deltapartitivirus pepper cryptic virus 1 (PCV1; Fig. 5d and Extended Data Fig. 3a), which was confirmed by DALI *Z*-score-based clustering (Fig. 4a), where the two viruses formed a clade next to picobirnaviruses.

Structural modelling of RNA2 ORF1 of different HsPV strains and Driatsky virus yielded a high-quality model (pLDDT = 78.8), with only the terminal regions being of lower quality (Extended Data Fig. 3b and Supplementary Text). Structure similarity searches against the PDB database using DALI produced significant hits to capsid proteins (CPs) of partitiviruses and picobirnaviruses[26–28], with the best match (*Z*-score = 8.2) to the CP of PCV1 (Fig. 5e; PDB ID: 7ncr; *Deltapartitivirus*). Thus, the RdRP phylogeny and structural similarity of the CPs indicate that HsPV is related to members of the family *Partitiviridae*. The phylogenetic relationship between amino acid sequences of HsPVs is shown in Extended Data Fig. 4.

HsRV and HsPV probably infect prokaryotic hosts

All samples in which HsRV and HsPV were detected nearly exclusively contained rRNA sequences from prokaryotes, with eukaryotic presence being below 1%. This is consistent with eukaryotes being unable to thrive in polyextremophilic conditions combining high temperatures and acidic pH. The microbial communities in all 4 samples (H4, H5,

https://doi.org/10.1038/s41564-023-01579-5

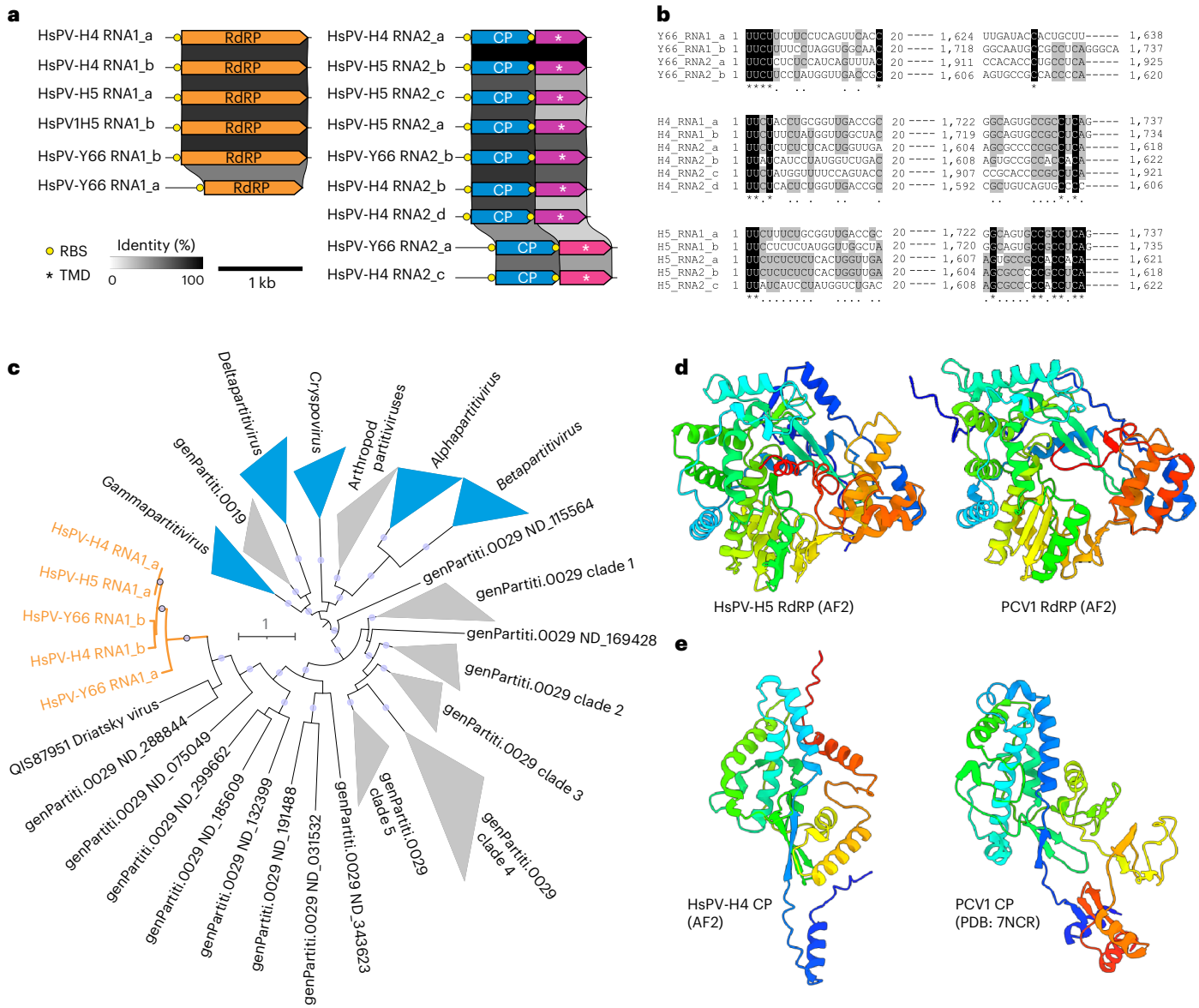


Fig. 5 | A thermoacidophilic partiti-like virus. a, Genome organization and conservation of the two genome segments, RNA1 and RNA2, of HsPV. ORFs encoding homologous proteins are shown as arrows with identical colours. Yellow circles represent predicted SD RBS. Asterisks denote putative genes encoding predicted TMD-containing proteins. **b**, MSA of the 5′- and 3′-terminal regions of the coding strands of reconstructed genome segments. Black shading, 100% nucleotide identity; grey shading, >50% nucleotide identity. **c**, Maximum-likelihood phylogeny of the RdRP proteins from representative members of the family *Partitiviridae* and related sequences (including all HsPV strains, shown in orange). Clades corresponding to the official *Partitiviridae* genera are shown in blue, whereas those corresponding to unclassified groups are in grey. Node supports were assessed using the SH-aLRT; circles indicate nodes with ≥90% supports. The scale bar represents the number of substitutions per site. GenBank accession numbers of used sequences are shown in Supplementary Text. **d**, Comparison of the RdRP from HsPV-H5 with a homologue from deltapartitivirus PCV1. **e**, Comparison of the CP from HsPV-H4 with a homologue from deltapartitivirus PCV1. The structures are coloured using the rainbow scheme, from blue N terminus to red C terminus. The HsPV RdRP and CP structures coloured on the basis of the pLDDT quality scores can be found in Extended Data Fig. 3.

Y66 and Oi) were dominated by bacteria (Supplementary Text). Thus, HsRV and HsPV most probably infect bacteria. To test this inference, we predicted ribosome-binding SD motifs in all HsRV and HsPV strains. SD motifs are essential for translation initiation in many prokaryotes, and their conservation is a diagnostic feature of prokaryotic genes that has been used to assign bacterial hosts to several groups of RNA viruses, namely, picobirnaviruses and partitiviruses, including genPartiti.0019 and genPartiti.0029 (refs. 8,29). Analysis of the HsRV and HsPV genomes showed that nearly every gene in these viruses is preceded by an SD motif (Figs. 2a, 3a and 5a and Extended Data Table 6), further suggesting that both HsRV and HsPV infect prokaryotic hosts. Bacteria of the genus *Hydrogenobaculum* (family *Aquificaceae*) were predominant (>95%) in samples H4 and H5 and highly abundant in Y66 (>85%), suggesting that HsPV detected in all three samples infects *Hydrogenobaculum* sp.

No CRISPR spacers matching the HsRV and HsPV genomes were identified in the public databases or the 919 CRISPR spacer sequences obtained by metagenomic DNA sequencing of the hot spring samples (Supplementary Text). Nevertheless, the lack of eukaryotes in the hot spring samples, contrasted by the dominance of bacteria, together with the presence of typical prokaryotic SD motifs upstream of the predicted virus genes and the polycistronic organization of the viral genomes, strongly suggest that HsRV and HsPV are viruses of thermophilic bacteria.

Table 1 | Characteristics of hot spring water samples

Code	Geographical coordinates	Area	Temp (°C)	pH	DO (mg l^{-1})	H_2S (mM)	Sampling date	Site characteristics
H4	31° 54′ 07.5″ N, 130° 50′ 06.2″ E	Hayashida	68.8	3.2	2.1	1.3	10 Mar 2017	Transparent water pool with sulfur precipitates
H5	31° 54′ 07.5″ N, 130° 50′ 06.2″ E	Hayashida	69.7	3.1	2.0	1.8	10 Mar 2017	Transparent water pool with sulfur precipitates
T1	31° 54′ 37.7″ N, 130° 49′ 00.6″ E	Tearai	92.1	2.9	–	0.0	09 Mar 2017	Yellowish grey water pool with active venting
T2		Tearai	95.9	2.1	–	0.0	09 Mar 2017	Yellowish grey vent pool
T3		Tearai	94.4	2.4	0.0	0.0	09 Mar 2017	Slightly grey water vent pool
T4		Tearai	92.8	2.7	0.0	0.0	09 Mar 2017	Yellowish grey water vent pool
Y66	31° 55′ 03.8″ N, 130° 48′ 40.4″ E	Yunoike	68.7	2.7	2.1	0.0	10 Mar 2017	Yellowish grey vent pool
Y80		Yunoike	75–86[a]	2.5	1.5	0.0	10 Mar 2017	Muddy small vent pool
Y86		Yunoike	86.5	2.5	0.0	0.0	10 Mar 2017	Muddy boiling vent pool
Oi	32° 44′ 25.3″ N, 130° 15′ 48.4″ E	Unzen	79.3	2.2	0.0	0.4	18 Nov 2015	Yellowish grey vent pool
Ob	32° 43′ 33.0″ N, 130° 12′ 24.7″ E	Obama	72.8	7.9	0.0	0.0	17 Nov 2015	Transparent water pool under hot spring water tank

[a]There were temperature gradients in the pool site: surface layer 75.0 °C; bottom layer 81.6 °C, 80.3 °C, 85.9 °C; middle layer 81.0 °C.

Discussion

The discovery of the HsRV-like group of riboviruses recapitulates previous findings of several small groups of riboviruses that are predicted to infect bacteria and might become distinct phyla[7,8]. However, the RdRPs of HsRV and its relatives seem to deviate from the RdRP consensus farther than any of the other recently discovered putative phyla, with none of which they appear to be affiliated, and possess unusual (predicted) structural features that appear to link them to viral RTs. Whether this connection reflects an intermediate position of the HsRV-like viruses between the kingdoms *Orthornavirae* and *Pararnavirae*, or results from convergent evolution, remains uncertain and should be clarified by sequencing and structural analysis of additional members of this group, or possibly, other groups of riboviruses with similar features. Furthermore, although we did not detect any evidence of the formation of DNA copies of the genomes of HsRV-like viruses, it will be of interest to determine whether their RdRPs possess RT activity, as shown for some viral RdRPs[30]. Regardless, HsRV-like viruses are strong candidates for a separate phylum in the kingdom *Orthornavirae*, which we propose to name '*Artimaviricota*' after the potential link to viral RTs (arti) and 'artima' which means 'close' in Lithuanian, or even a third kingdom within the realm *Riboviria*.

This report is a proof of concept for the discovery of multiple, perhaps many groups of riboviruses with unexpected properties by obtaining complete genomes of segmented riboviruses from meta-dsRNA-seq data and mining metatranscriptomes from habitats with distinct conditions. Information on non-RdRP segments is unavailable for most of the RNA virus lineages identified only from metatranscriptomes, whereas riboviruses that are distantly related to known RNA viruses can be missed altogether. Our approach helps to overcome these limitations and contributes to a more complete characterization of RNA viromes.

Methods

Sample collection

A total of 11 samples were collected from five hot springs regions in southern Japan, in proximity to active volcanoes (Table 1 and Supplementary Text), according to the instructions of Unzen City, Unzen Nature Conservation Bureau and private companies that maintain each hot spring region. Temperature, pH and dissolved oxygen (DO) were measured in situ by using a multiple electrode sensor (D-55, Horiba). H_2S concentration was calculated from the spectrophotometric absorbance at 680 nm of methylene blue formed from a reaction with *N*,*N*-dimethyl-*p*-phenylenediamine in $FeCl_2$-HCl solution. Typical measurement errors are 0.1 for pH, 0.1 mg l^{-1} for DO and 5% for H_2S. Dissolved chemicals and water isotope ratios of the geothermal waters were also measured and are summarized in Supplementary Text.

Most of the sampling sites were characterized by high temperatures above 65 °C, acidic pH (2–3, except for Site Ob with a slightly alkaline pH of 7.9) and lower level of DO with accompanying grey mud or light-yellow sulfur deposits. At each sampling station, ~10 l of hot spring water was collected in a sterilized plastic bag and then filtered with 0.2-µm-pore-size cellulose acetate membrane filters in 47 mm diameter (Advantecn) within 0.5–3 h after sampling. The filters were stored at −80 °C until nucleic acid extraction.

RNA extraction

Cells collected on a portion of the 0.2-µm-pore-size filters corresponding to ~2 l of hot spring water were pulverized in a mortar in liquid nitrogen and suspended in dsRNA extraction buffer (20 mM Tris–HCl, pH 6.8, 200 mM NaCl, 2 mM EDTA, 1% SDS and 0.1% (v/v) β-mercaptoethanol) or TRIzol buffer for ds- and ssRNA purification, respectively. For dsRNA purification, total nucleic acids were manually extracted with SDS-phenol. dsRNA was purified using the cellulose resin chromatography method[16,31]. The remaining DNA and ssRNA were removed by DNase I (Invitrogen) and S1 nuclease (Invitrogen) treatment[19]. For ssRNA purification, the ssRNA fraction was collected using the TRIzol Plus RNA purification kit (Invitrogen) according to manufacturer protocol. The ssRNA fraction was treated with DNase I (Invitrogen) and concentrated using the RNA Clean and Concentrator-5 kit (Zymoresearch).

Complementary DNA synthesis

Complementary DNA (cDNA) was synthesized from purified dsRNA and ssRNA as described previously[19]. In brief, purified dsRNA was physically fragmented into ~1.5 kbp and adapter oligonucleotide (U2: 5′-GAC GTA AGA ACG TCG CAC CA-3′) was ligated to the 3′-end of fragmented dsRNAs. After heat denaturation with an oligonucleotide primer (U2-comp: 5′-TGG TGC GAC GTT CTT ACG TC-3′), that has complementary sequence to the adapter oligonucleotide, cDNA was synthesized using SMARTer RACE 5′/3′ kit (Takara Bio). ssRNA was converted into cDNA using SMARTer Universal Low Input RNA kit according to manufacturer protocol (Takara Bio). After PCR amplification, cDNA was fragmented by a Covaris S220 ultrasonicator.

Illumina sequencing library construction and sequencing

Illumina sequencing libraries were then constructed using KAPA Hyper Prep Kit Illumina platforms (Kapa Biosystems) from the physically shared environmental cDNAs. The libraries were sequenced using the

Illumina MiSeq v3 Reagent kit (600 cycles) with 300-bp paired-end reads on the Illumina MiSeq platform.

Data processing

Trimmed reads were obtained using a custom Perl pipeline script (https://github.com/takakiy/FLDS) from dsRNA raw sequence reads[17]. The clean reads were subjected to de novo assembly using CLC GENOMICS WORKBENCH v.11.0 (Qiagen) with the following parameters: a minimum contig length of 500, word value set to auto and bubble size set to auto. The full-length sequences were manually extracted using CLC GENOMICS WORKBENCH v.11.0 (Qiagen), Genetyx v.14 (Genetyx) and Tablet viewer v.1.19.09.03 (ref. 32) as described previously[33]. In brief, contigs for which both termini were determined to be the ends were identified as full-length sequences. In cases of dominant reads (more than 10 reads) that stopped in the same position around the ends of contigs in the mapping analysis, that position was recognized as the segment (genome) end. In this study, major full-length sequences with >1,000 average coverage were analysed, except for the Oi sample where all full-length sequences were recovered. From ssRNA raw sequence reads, trimmed reads were also obtained using a custom Perl pipeline script (https://github.com/takakiy/FLDS). The resultant clean reads were applied to phyloFlash[20] to identify active microbes in our samples.

Sequence analyses

RNA viral genes were identified using the BLASTX programme against the NCBI non-redundant (nr) database with an *E*-value ≤ 1×10^{-05}. The ribosome-binding SD motifs were identified using Prodigal[34]. Remote homology searches were performed using HHpred against the PDB70, Pfam, UniProt-SwissProt-viral70 and NCBI-CD (conserved domains) databases[35]. MSA of HsRV RNA1_ORF4s was built using MEGA6 (ref. 36). The alignment was then used as input in HHblits 3.3.0, which compared the alignments to the PDB70 (pdb70_from_mmcif_220313) database. Transmembrane domains were predicted using TMHMM[37].

Search for HsRV homologues in public databases

To identify viruses related to HsRV in the IMG/VR database[38], BLASTP searches (*E*-value ≤ 1×10^{-05}) queried with the RdRP sequences encoded by HsRV-like contigs previously deposited to GenBank (accessions: BDQA01000957 and BDQA01004869) were performed on the IMG/VR website (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi?section=Viral&page=findViralGenesBlast). The nucleotide sequences of the contigs encoding the related RdRPs were downloaded and annotated as described above for the HsRVs.

Modelling protein structures with AlphaFold2 and structural comparisons

Structural predictions for HsRV and HsRV-like RdRP amino acid sequences were performed using ColabFold 1.5.1 installed locally through LocalColabFold (https://github.com/YoshitakaMo/localcolabfold). A custom MSA with ten HsRV (HsRV_La~d, H5_contig_1, Oi_contig_1, Oi_contig_3, Oi_contig_5, Oi_contig_8, Oi_contig_9) and five HsRV-like (BDQA01000957, BDQA01004869, Ga0456180, Ga0393213, Ga0169446) RdRP amino acid sequences was used as input. The number of recycles used for HsRV_La ORF4 and HsRV-like RdRP predictions were 6 and 10, respectively. For the core (motifs A–C) region of marine HsRV-like RdRP BDQA01004869 (Fig. 3d), 20 recycles were used. For Fig. 4a, Ambiviricota RdRP model (pLDDT 95, predicted template modeling (pTM) score 0.938) was generated with 3 recycles using a custom MSA of 422 Ambivirus RdRP sequences available at https://github.com/ababaian/serratus/wiki/ambivirus_extended_data (ref. 24). Paraxenoviricota (TARA_132_DCM_0.22-3_k119_33585_1_799) RdRP model (pLDDT 88.6, pTM 0.882) was generated with 20 recycles using a custom MSA of 12 amino acid sequences obtained by running BLASTP against ORFs from 44779_RdRP_contigs available at https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/ZayedWainainaDominguez-Huerta_RNAevolution_Dec2021/Contigs (ref. 7). Similarly, Wamoviricota (84SUR2MMQQ14_2_ERR1712161_contig_61452_3_468) RdRP model (pLDDT 84.5, pTM 0.822) was modelled with 20 recycles using a custom MSA of 6 sequences from the 44779_RdRP_contigs[7] and 56 additional sequences obtained from a BLASTP search against the IMG/VR database. p.0002 (ND_055403_2847-982) RdRP model (pLDDT 84.3, pTM 0.864) was generated with 12 recycles using a custom MSA with 107 p.0002 RdRP sequences kindly provided by Dr Yuri I. Wolf[8]. The RdRPs of HsPV-H5 and PCV1 (GenBank ID: YP_009466859) were modelled using AlphaFold 2 through ColbFold (v.1.5.2)[39,40] with 6 recycles each. For the HsPV-H4 CP modelling, an alignment of RNA2 ORF1 homologues from HsPV-like viruses and Driatsky virus was used as a template with 12 recycles. The obtained model had a medium quality (average pLDDT = 57.3), although the central region was modelled with higher quality (average pLDDT > 70). This model was used as a query in DALI search, which identified the CP of PCV1 (PDB ID: 7ncr) as the best hit with a *Z*-score of 6.5. Thus, to improve the quality of the HsPV-H4 CP model, we repeated the modelling using the same sequence alignment and providing the PDB structure of the PCV1 CP as a template, with 24 recycles. The obtained model had an average pLDDT score of 78.1. Model display, structural alignment, colouring and figure preparation were performed using UCSF ChimeraX software[41].

Phylogenetic analysis

Amino acid sequences of RdRP encoded by identified viruses and viruses related to the family *Partitiviridae* were aligned using MAFFT (G-INS-1)[42]. The ambiguous positions in the alignment were removed using TrimAl (gap threshold 0.2)[43]. The maximum-likelihood tree was constructed using IQ-TREE (v.2.0.6)[44]. The best-fitting substitution model was selected using ModelFinder[45] and was LG + F + R8. Node supports were estimated using the SH-like approximate likelihood-ratio test (SH-aLRT) with 1,000 replicates. For phylogenetic analysis of the HsRV-like RdRPs, the proteins were aligned using PROMALS3D[46] and uninformative positions we removed using TrimAl with the gappyout functions[43]. The final alignment contained 520 positions. The maximum-likelihood tree was constructed using IQ-TREE (v.2.0.6)[44]. The best-fitting substitution model was selected using ModelFinder[45] and was LG + I + G4. Node supports were estimated using SH-aLRT (1,000 replicates).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Datasets obtained in this study have been made available in the GenBank database repository (accession nos. HsRV: BTCN01000001–BTCN01000010; HsPV-H4: BTCO01000001–BTCO01000006; HsPV-H5: BTCP01000001–BTCP01000005; HsPV-Y66: BTCQ01000001–BTCQ01000004; H5_contig_1: BTCR01000001; Oi_contig_1-9: BTCS01000001–BTCS01000009) and Short Read Archive database (accession no. DRA016131). Datasets (PDB70 mmcif_2023-10-24, Pfam v.35, UniProt-SwissProt-viral70_Nov_2021 and NCBI-CD v.3.19) are available at http://ftp.tuebingen.mpg.de/pub/protevo/toolkit/databases/hhsuite_dbs/. Searches using the IMG/VR dataset were available only at https://img.jgi.doe.gov/cgi-bin/vr/main.cgi?section=WorkspaceBlast&page=viralform. Datasets (SILVA SSU v.138, Neo-HMM v.1.1 and RVDB-HMM v.23.0) are publicly available.

Code availability

A custom code used in this study has been made available in a git repository publicly available on GitHub at https://github.com/takakiy/FLDS (Cleanup_FLDS.pl).

References

1. Simmonds, P. et al. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).
2. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
3. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. *mBio* **9**, e02329-18 (2018).
4. Wolf, Y. I. et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* **5**, 1262–1270 (2020).
5. Shi, M., Zhang, Y. Z. & Holmes, E. C. Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res.* **243**, 83–90 (2018).
6. Shi, M. et al. The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
7. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
8. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037 (2022).
9. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
10. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **84**, e00061-19 (2020).
11. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* **479-480**, 2–25 (2015).
12. Nasir, A., Forterre, P., Kim, K. M. & Caetano-Anolles, G. The distribution and impact of viral lineages in domains of life. *Front. Microbiol.* **5**, 194 (2014).
13. Callanan, J. et al. Leviviricetes: expanding and restructuring the taxonomy of bacteria-infecting single-stranded RNA viruses. *Microb. Genomics* **7**, 000686 (2021).
14. Callanan, J. et al. Expansion of known ssRNA phage genomes: from tens to over a thousand. *Sci. Adv.* **6**, eaay5981 (2020).
15. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA bacteriophage diversity. *PLoS Biol.* **14**, e1002409 (2016).
16. Morris, T. J. & Dodds, J. A. Isolation and analysis of double-stranded-RNA from virus-infected plant and fungal tissue. *Phytopathology* **69**, 854–858 (1979).
17. Hirai, M. et al. RNA viral metagenome analysis of subnanogram dsRNA using fragmented and primer ligated dsRNA sequencing (FLDS). *Microbes Environ.* **36**, ME20152 (2021).
18. Urayama, S., Takaki, Y. & Nunoura, T. FLDS: a comprehensive dsRNA sequencing method for intracellular RNA virus surveillance. *Microbes Environ.* **31**, 33–40 (2016).
19. Urayama, S. et al. Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* **18**, 1444–1455 (2018).
20. Gruber-Vodicka, H. R., Seah, B. K. & Pruesse, E. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *mSystems* **5**, e00920 (2020).
21. Yang, Y. et al. Characterization of the first double-stranded RNA bacteriophage infecting *Pseudomonas aeruginosa*. *Sci. Rep.* **6**, 38795 (2016).
22. Venkataraman, S., Prasad, B. & Selvarajan, R. RNA dependent RNA polymerases: insights from structure, function and evolution. *Viruses* **10**, 76 (2018).
23. Te Velthuis, A. J. Common and unique features of viral RNA-dependent polymerases. *Cell. Mol. Life Sci.* **71**, 4403–4420 (2014).
24. Forgia, M. et al. Hybrids of RNA viruses and viroid-like elements replicate in fungi. *Nat. Commun.* **14**, 2591 (2023).
25. Krupovic, M. et al. Ortervirales: new virus order unifying five families of reverse-transcribing viruses. *J. Virol.* **92**, e00515–18 (2018).
26. Luque, D., Mata, C. P., Suzuki, N., Ghabrial, S. A. & Castón, J. R. Capsid structure of dsRNA fungal viruses. *Viruses* **10**, 481 (2018).
27. Byrne, M., Kashyap, A., Esquirol, L., Ranson, N. & Sainsbury, F. The structure of a plant-specific partitivirus capsid reveals a unique coat protein domain architecture with an intrinsically disordered protrusion. *Commun. Biol.* **4**, 1155 (2021).
28. Duquerroy, S. et al. The picobirnavirus crystal structure provides functional insights into virion assembly and cell entry. *EMBO J.* **28**, 1655–1665 (2009).
29. Krishnamurthy, S. R. & Wang, D. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology* **516**, 108–114 (2018).
30. Peyambari, M., Guan, S. & Roossinck, M. J. RdRp or RT, that is the question. *Mol. Biol. Evol.* **38**, 5082–5091 (2021).
31. Okada, R., Kiyota, E., Moriyama, H., Fukuhara, T. & Natsuaki, T. A simple and rapid method to purify viral dsRNA from plant and fungal tissue. *J. Gen. Plant Pathol.* **81**, 103–107 (2015).
32. Milne, I. et al. Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
33. Urayama, S., Takaki, Y., Hagiwara, D. & Nunoura, T. dsRNA-seq reveals novel RNA virus and virus-like putative complete genome sequences from *Hymeniacidon* sp. sponge. *Microbes Environ.* **35**, ME19132 (2020).
34. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
35. Gabler, F. et al. Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinformatics* **72**, e108 (2020).
36. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
37. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
38. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
39. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
40. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
41. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
42. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
43. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
44. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
45. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
46. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol. Biol.* **1079**, 263–271 (2014).

Acknowledgements

We thank NITTETSU MINING CO., LTD KAGOSHIMA GEOTHERMAL FACILITY, NIPPON PAPER LUMBER CO., LTD and Kirishima Iwasaki

Hotel for support for field sampling; S. Kawagucci, M. Yoshida, Y. Yoshida-Takashima, M. Tsuda and F. Kondo for discussions, suggestions, sample collections and preliminary experiments related to this study; and Y. I. Wolf for technical help. This study was supported by JSPS KAKENHI (Grant Nos. 15H05468 to T.O. and 20K20377 to T.N.) and by Grants-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Grant Nos. 22H04879 and 20H05579 to S.U.; 19H05684, 16H06429, 16K21723 and 16H06437 to T.N). This research was also supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute which provided supercomputing resources for protein structure modelling, and by a grant from the Institute for Fermentation, Osaka, Japan. E.V.K. was supported by the Intramural Research Program of the US National Institutes of Health (National Library of Medicine).

Author contributions

All authors had a substantial contribution to this work. S.U. and T.N. were responsible for the design of the work and the acquisition, analysis and interpretation of data, and drafted the initial work. S.U., A.F., E.V.K., M.K. and T.N. substantively revised the work. A.F., Y.N., Y.T. and M.K. performed bioinformatic analysis. M.H. and T.O. performed experiments, and analysed and interpreted the data. S.U., A.F., T.O., Y.N., N.K., E.V.K., M.K. and T.N. wrote the paper.

Competing interests

JAMSTEC holds a patent for the 'Double-stranded RNA fragmentation method and use thereof', with S.U. and T.N. listed as inventors. These patents include European Patent (EP) Registration No. 3363898, registered on 30 November 2022; China Registration No. ZL201680060127.X, registered on 8 February 2022; US Registration No. 10894981, registered on 19 January 2021; and Japanese patent No. 6386678, registered on 17 August 2018. The other authors declare no competing interests.

Additional information

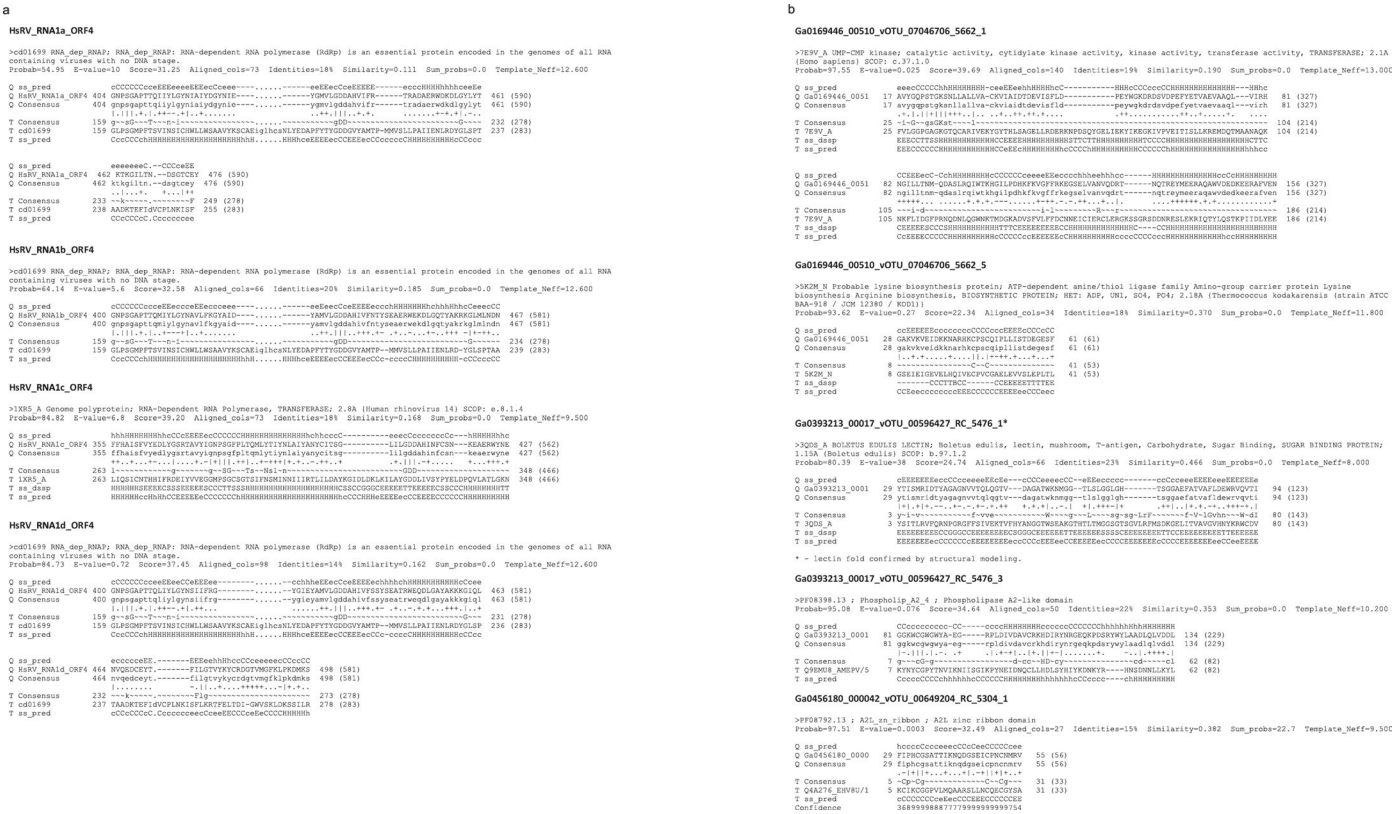
Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01579-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01579-5.

Correspondence and requests for materials should be addressed to Syun-ichi Urayama.

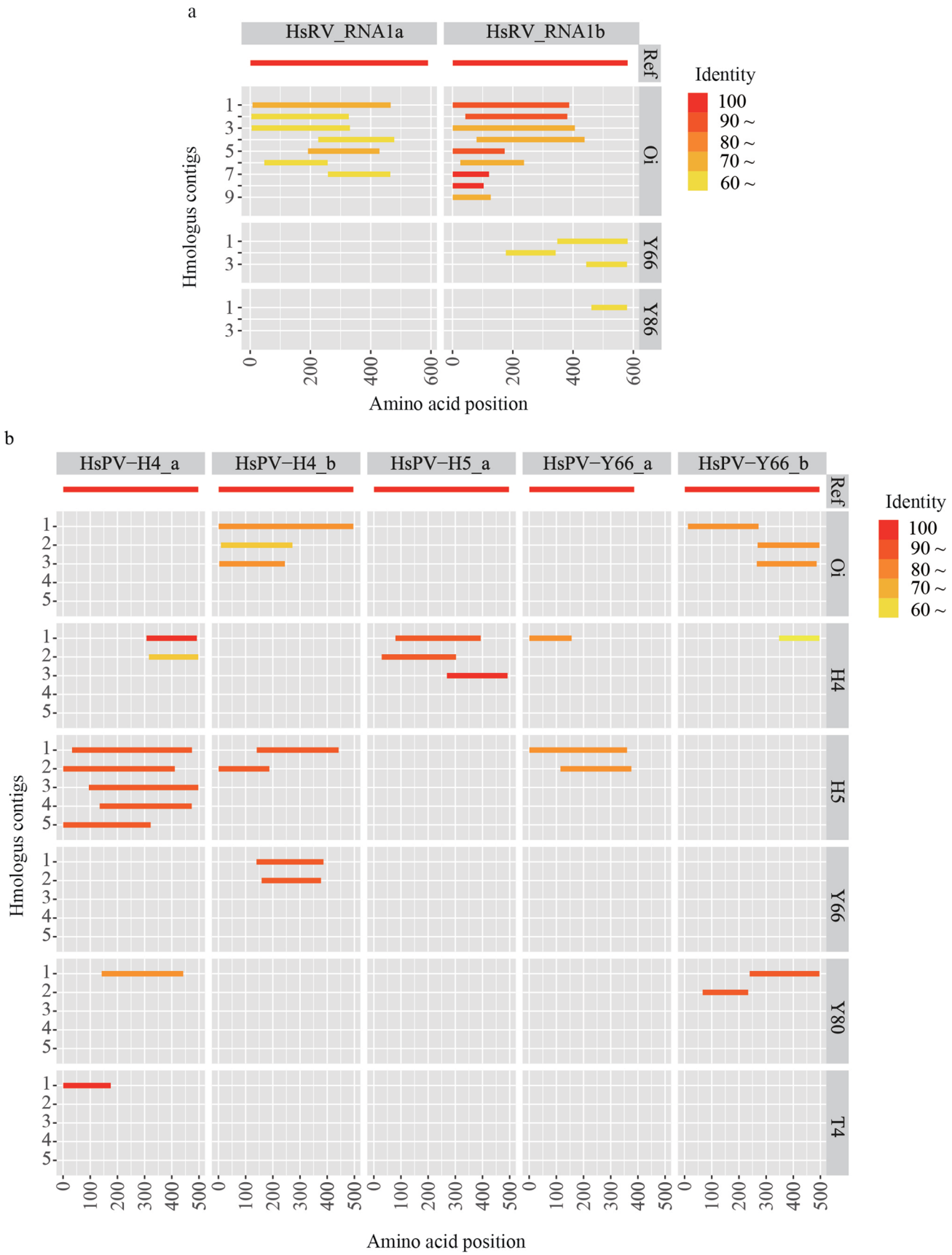
Peer review information *Nature Microbiology* thanks Vanessa Marcelino and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-023-01579-5



Extended Data Fig. 1 | Results of the HHsearch analysis. Results of the HHsearch analysis queried with the **a**, putative ORF4 protein sequences from HsRV virus strains, and **b**, indicated protein sequences encoded by HsRV-like viruses from moderate ecosystems. H(h), α-helix; E(e), β-strand; C(c), coil.

https://doi.org/10.1038/s41564-023-01579-5

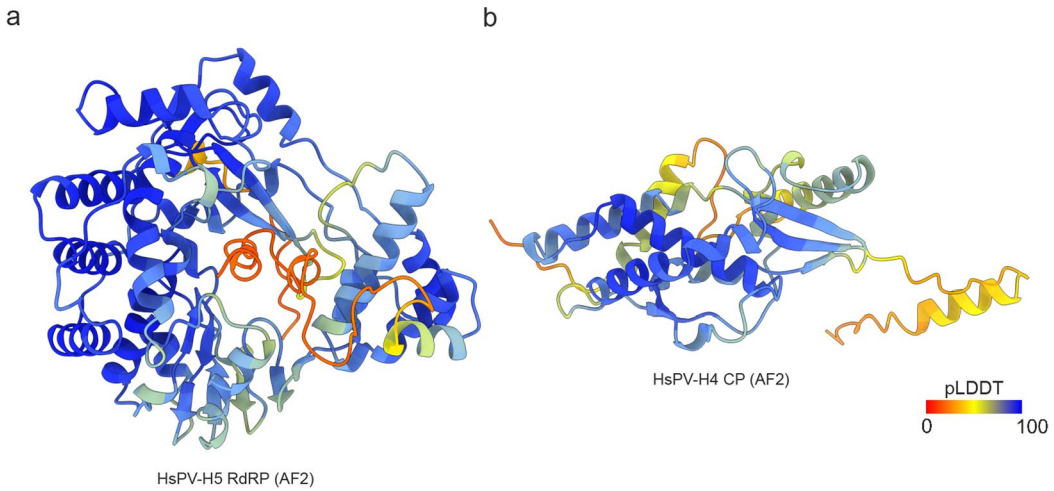


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41564-023-01579-5

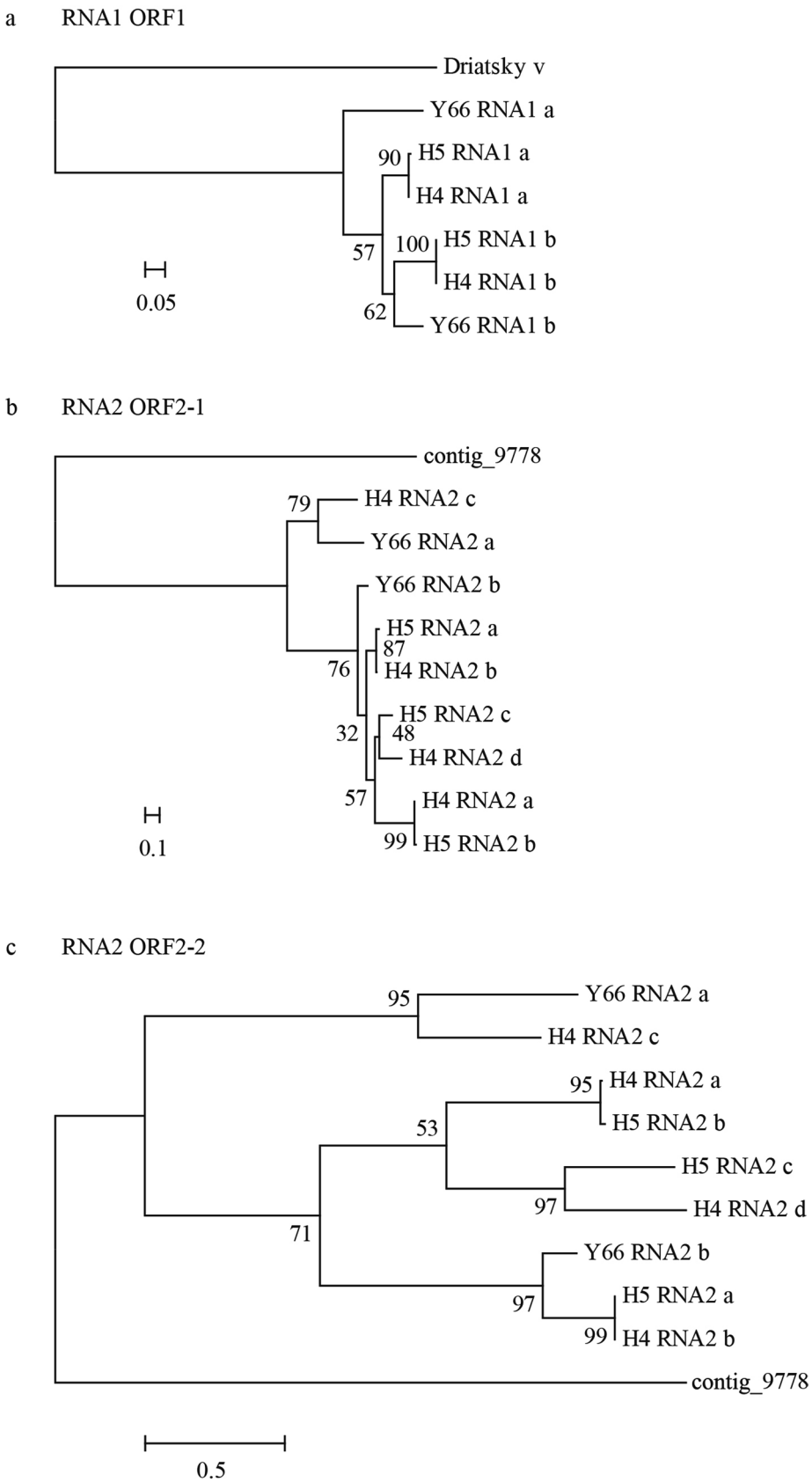
Extended Data Fig. 2 | Distribution of minor contigs related to the RdRPs of HsRV and HsPV. Distribution of minor contigs related to the RdRPs of a, HsRV and b, HsPV. Each bar represents the position of predicted amino acid sequences of contigs aligned to the HsRV or HsPV RdRP shown at the top of the panel, and their identities to the reference RdRP sequences are indicated by the colors in the heatmap. The name of source libraries are shown in the right-side panel. Trimmed reads from each sample were assembled using CLC assembler, followed by the removal of sequences identical to HsRV or HsPV. Using the amino acid sequences of RdRPs from HsRV and HsPV as queries, tBLASTN searches were performed on the remaining contigs. Sequences with > 60% amino acid identity and > 100 aa hit were shown.

https://doi.org/10.1038/s41564-023-01579-5



Extended Data Fig. 3 | pLDDT scores of HsPV RdRP and CP. Quality assessment of the AF2 model of the HsPV a, RdRP and b, CP. The structural model is colored based on the pLDDT scores, with the color key shown at the bottom right corner.

https://doi.org/10.1038/s41564-023-01579-5



Extended Data Fig. 4 | HsPV phylogeny. Maximum-likelihood trees of each ORF encoded by HsPVs and related sequences. Sequences were aligned using MEGA6. The ambiguous positions in the alignment were removed using TrimAl. The maximum likelihood tree was constructed using RAxML. The best-fitting substitution model was selected by ProtTest. Numbers indicate the percentage bootstrap support from 1,000 RAxML bootstrap replicates. We used RAxML with the **a**, LG+G+I+F model for ORF1, **b**, LG+G model for ORF2-1 and **c**, LG+G+I+F model for ORF2-2.

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 1 | Relative abundances of rRNA reads in ssRNA seq of representative microbial lineages

	H4	H5	T1	T2	T3	T4	Y66	Y80	Y86	Oi	Ob
Eukaryota	0.59	0.19	1.64	0.99	7.31	1.51	0.11	1.53	5.16	0.75	4.02
Sulfolobaceae	1.00	0.23	54.50	72.32	31.19	66.31	7.89	5.28	59.04	10.94	0.00
other Archaea	1.01	0.19	43.23	26.29	17.83	30.36	7.06	15.11	20.99	4.83	0.14
Hydrogenobaculaceae	96.57	98.90	0.13	0.12	0.27	1.61	84.24	77.54	14.47	49.84	0.01
Aquificaceae	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.09	82.54
Comamonadaceae	0.05	0.00	0.00	0.00	0.38	0.00	0.01	0.00	0.00	25.21	0.00
other Bacteria	0.78	0.47	0.50	0.29	43.03	0.32	0.69	0.56	0.34	8.32	13.29

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 2 | Classification of NGS reads

Sample	Library type	Raw reads (pair)	% of removed	% of rRNA	% of RNA viral (candidate)	% of other reads
H4	dsRNA	930,952	31.8	9.8	47.7	10.6
H4	ssRNA	68,766	9.9	81.7	0.1	8.3
H5	dsRNA	1,009,144	10.3	3.6	81.8	4.3
H5	ssRNA	57,932	8.4	87.3	0.1	4.2
T1	dsRNA	1,276,072	42.1	34.2	0.0	23.7
T1	ssRNA	77,818	12.1	12.1	0.0	75.8
T2	dsRNA	1,069,831	19.2	80.0	0.0	0.8
T2	ssRNA	26,373	12.0	30.4	0.0	57.6
T3	dsRNA	1,203,878	17.0	81.2	0.0	1.8
T3	ssRNA	56,252	14.5	35.3	0.0	50.1
T4	dsRNA	1,433,410	62.5	19.1	0.0	18.4
T4	ssRNA	71,737	18.6	7.7	0.0	73.7
Y66	dsRNA	1,168,048	28.9	38.0	4.0	29.1
Y66	ssRNA	100,556	10.5	81.2	0.0	8.2
Y80	dsRNA	1,059,410	20.3	61.5	0.0	18.2
Y80	ssRNA	100,547	8.5	60.7	0.0	30.8
Y86	dsRNA	1,043,072	20.6	56.0	0.0	23.4
Y86	ssRNA	44,189	15.2	19.4	0.0	65.4
Ob	dsRNA	286,638	99.2	0.1	0.0	0.8
Ob	ssRNA	142,279	45.4	23.2	0.0	31.4
Oi	dsRNA	236,876	28.2	4.9	32.9	34.1
Oi	ssRNA	89,726	39.9	38.4	0.2	21.5

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 3 | Result of CDS clustering using a standard BLAST-mcl pipeline

Cluster No. *[1]	Sequence No.	Virus	ORFs	Length	E-value	Identity
Cluster 1	1	HsRV	RNA1a_ORF1	134		
Cluster 2	2	HsRV	RNA1b_ORF1	110		
		HsRV	RNA1d_ORF1	104	1.00E-30	37
Cluster 3	1	HsRV	RNA1c_ORF1	246		
Cluster 4	4	HsRV	RNA1a_ORF2	128		
		HsRV	RNA1b_ORF2	130	3.00E-26	34
		HsRV	RNA1c_ORF2	123	1.00E-14	26
		HsRV	RNA1d_ORF2	127	2.00E-31	42
Cluster 5	3	HsRV	RNA1a_ORF3	965		
		HsRV	RNA1b_ORF3	933	0	33
		HsRV	RNA1d_ORF3	931	0	33
Cluster 6	1	HsRV	RNA1c_ORF3	1115		
Cluster 7	1	HsRV	RNA1b_ORF5	77		
Cluster 8	4	HsRV	RNA1a_ORF4	590		
		HsRV	RNA1b_ORF4	581	3.00E-125	37
		HsRV	RNA1c_ORF4	591	3.00E-128	37
		HsRV	RNA1d_ORF4	581	1.00E-134	39
Cluster 9	1	HsRV	RNA1c_ORF5	89		
Cluster 10	6	HsRV	RNA2a_ORF1	162		
		HsRV	RNA2b_ORF1	195	2.00E-20	33
		HsRV	RNA2c_ORF1	168	5.00E-76	70
		HsRV	RNA2d_ORF1	167	2.00E-77	70
		HsRV	RNA2*a_ORF1	136	8.00E-10	36
		HsRV	RNA2*b_ORF1	176	5.00E-06	29
Cluster 11	6	HsRV	RNA2a_ORF2	203		
		HsRV	RNA2b_ORF2	203	2.00E-83	62
		HsRV	RNA2c_ORF2	205	6.00E-113	79
		HsRV	RNA2d_ORF2	205	2.00E-123	80
		HsRV	RNA2*a_ORF2	257	5.00E-07	40
		HsRV	RNA2*b_ORF2	252	4.00E-08	38
Cluster 12	6	HsRV	RNA2a_ORF3	397		
		HsRV	RNA2b_ORF3	391	4.00E-120	48
		HsRV	RNA2c_ORF3	397	0	69
		HsRV	RNA2d_ORF3	397	0	68
		HsRV	RNA2*a_ORF3	442	3.00E-07	23
		HsRV	RNA2*b_ORF3	435	4.00E-09	25
Cluster 13	4	HsRV	RNA2a_ORF4	269		
		HsRV	RNA2b_ORF4	238	2.00E-09	32
		HsRV	RNA2c_ORF4	269	2.00E-112	63
		HsRV	RNA2d_ORF4	268	7.00E-112	62
		HsRV	RNA2*b_ORF5	203		
Cluster 14	3	HsRV	RNA2*a_ORF4	179	7.00E-24	34
		HsRV	RNA2*a_ORF5	380	1.00E-06	26
Cluster 15	1	HsRV	RNA2*b_ORF4	41		
Cluster 16	3	HsRV	RNA2c_ORF5	231		
		HsRV	RNA2d_ORF5	225	1.00E-74	59
		HsRV	RNA2a_ORF5	197	8.00E-12	31
Cluster 17	1	HsRV	RNA2b_ORF5	250		
Cluster 18	1	HsRV	RNA2a_ORF6	43		
Cluster 19	1	HsRV	RNA2*b_ORF6	51		
Cluster 20	1	HsRV	RNA2*b_ORF6	167		
Cluster 21	6	HsPV	H4_RNA1_a_ORF1	497		
		HsPV	H4_RNA1_b_ORF1	496	0	84
		HsPV	H5_RNA1_a_ORF1	497	0	99
		HsPV	H5_RNA1_b_ORF1	496	0	84
		HsPV	Y66_RNA1_a_ORF1	386	0	76
		HsPV	Y66_RNA1_b_ORF1	496	0	86
Cluster 22	9	HsPV	H4_RNA2_a_ORF1	252		
		HsPV	H4_RNA2_b_ORF1	251	2.00E-123	73
		HsPV	H4_RNA2_c_ORF1	254	9.00E-86	52
		HsPV	H4_RNA2_d_ORF1	251	7.00E-135	74
		HsPV	H5_RNA2_a_ORF1	252	6.00E-133	75
		HsPV	H5_RNA2_b_ORF1	252	0	99
		HsPV	H5_RNA2_c_ORF1	251	1.00E-125	75
		HsPV	Y66_RNA2_a_ORF1	249	1.00E-82	52
		HsPV	Y66_RNA2_b_ORF1	251	2.00E-136	73
Cluster 23	9	HsPV	H4_RNA2_a_ORF2	202		
		HsPV	H4_RNA2_b_ORF2	203	1.00E-32	38
		HsPV	H4_RNA2_c_ORF2	204	3.00E-18	24
		HsPV	H4_RNA2_d_ORF2	203	1.00E-50	45
		HsPV	H5_RNA2_a_ORF2	203	1.00E-32	38
		HsPV	H5_RNA2_b_ORF2	202	5.00E-146	98
		HsPV	H5_RNA2_c_ORF2	203	1.00E-53	46
		HsPV	Y66_RNA2_a_ORF2	207	3.00E-15	24
		HsPV	Y66_RNA2_b_ORF2	203	6.00E-33	36

[*1]The CDSs were clustered using a standard BLAST-mcl pipeline [BLASTP (v2.9.0) with default options, hits selected based on E-value < 1e-10, MCL clustering (v.14-137) with an inflation value of 2.8].

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 4 | RNA virus and virus-like genomes identified in this study

Group	Segment / Contig	Accession	Length (nt)	Ave. Cove.	Status	Top Hit (public DB)
HsRV	RNA1a	BTCN01000001	6,117	1,621	full	No hit
	RNA1b	BTCN01000002	6,013	401	full	No hit
	RNA1c	BTCN01000003	6,560	66	full	No hit
	RNA1d	BTCN01000004	6,008	47	full	No hit
	RNA2a	BTCN01000005	4,417	2,518	full	No hit
	RNA2b	BTCN01000006	4,302	1,605	full	No hit
	RNA2c	BTCN01000007	4,398	395	full	No hit
	RNA2d	BTCN01000008	4,427	199	full	No hit
	RNA2*a	BTCN01000009	4,620	49	full	No hit
	RNA2*b	BTCN01000010	4,272	44	full	No hit
HsRV-relates	Oi_contig_1	BTCS01000001	984	11	partial	RdRP [Ribovi ria::Orthornavirae (FAM010882)]
	Oi_contig_2	BTCS01000002	617	113	partial	No hit
	Oi_contig_3	BTCS01000003	3,057	13	partial	RdRP [Riboviria::Orthornavirae (FAM010882)]
	Oi_contig_4	BTCS01000004	622	40	partial	No hit
	Oi_contig_5	BTCS01000005	3,243	9	partial	RdRP [Riboviria (FAM004495)]
	Oi_contig_6	BTCS01000006	531	4	partial	No hit
	Oi_contig_7	BTCS01000007	580	14	partial	No hit
	Oi_contig_8	BTCS01000008	1,766	10	partial	No hit
	Oi_contig_9	BTCS01000009	5,001	86	partial	No hit
	H5_contig_1	BTCR01000001	2,030	17	partial	RdRP [Riboviria (FAM004495)]
HsPV (Y66)	RNA1_a	BTCQ01000001	1,638	2,997	full	RdRP [Driatsky virus]
	RNA1_b	BTCQ01000002	1,737	2,321	full	RdRP [Driatsky virus]
	RNA2_a	BTCQ01000003	1,925	2,674	full	No hit
	RNA2_b	BTCQ01000004	1,620	1,532	full	No hit
HsPV (H4)	RNA1_a	BTCO01000001	1,737	25,387	full	RdRP [Driatsky virus]
	RNA1_b	BTCO01000002	1,734	11,450	full	RdRP [Driatsky virus]
	RNA2_a	BTCO01000003	1,618	36,444	full	No hit
	RNA2_b	BTCO01000004	1,622	32,460	full	No hit
	RNA2_c	BTCO01000005	1,921	3,690	full	No hit
	RNA2_d	BTCO01000006	1,606	1,338	full	No hit
HsPV (H5)	RNA1_a	BTCP01000001	1,737	45,541	full	RdRP [Driatsky virus]
	RNA1_b	BTCP01000002	1,735	34,386	full	RdRP [Driatsky virus]
	RNA2_a	BTCP01000003	1,621	55,398	full	No hit
	RNA2_b	BTCP01000004	1,618	51,336	full	No hit
	RNA2_c	BTCP01000005	1,622	22,136	full	No hit

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 5 | HHsearch hits for the IMG/VR virus proteins

Protein	Annotation	HHsearch profile matched	HHsearch probability
Ga0169446_00510_vOTU_07046706_5662_1	Predicted kinase	7E9V_A UMP-CMP kinase; catalytic activity, cytidylate kinase activity, kinase activity, transferase activity, TRANSFERASE; 2.1A {Homo sapiens} SCOP: c.37.1.0	97.55
Ga0169446_00510_vOTU_07046706_5662_4	RdRP		58.29
Ga0169446_00510_vOTU_07046706_5662_5	Potential zinc finger protein	5K2M_N Probable lysine biosynthesis protein; ATP -dependent amine/thiol ligase family Amino-group carrier protein Lysine biosynthesis Arginine biosynthesis, BIOSYNTHETIC PROTEIN; HET: ADP, UN1, SO4, PO4; 2.18A {Thermococcus kodakarensis (strain ATCC BAA-918 / JCM 12380 / KOD1)}	93.62
Ga0393213_00017_vOTU_00596427_RC_5476_3	Phospholipase A2	PF08398.13 ; Phospholip_A2_4 ; Phospholipase A2-like domain	95.08
Ga0393213_00017_vOTU_00596427_RC_5476_6	RdRP	5I62_A Potential RNA -dependent RNA polymerase; dsRNA, replication, transcription, insertion loop, viral protein; 2.001A {Human picobirnavirus (strain Human/Thailand/Hy005102/ -)}	39.43
Ga0456180_000042_vOTU_00649204_RC_5304_1	Potential zinc finger protein	PF08792.13 ; A2L_zn_ribbon ; A2L zinc ribbon domain	97.51
Ga0456180_000042_vOTU_00649204_RC_5304_5	RdRP		18.95

https://doi.org/10.1038/s41564-023-01579-5

Extended Data Table 6 | Detected RBS motif

Gene	Gene start	Gene end	Start codon	RBS motif	RBS spacer	GC%	Length, aa	# TMD
HsPV-H4_RNA1_a_1	196	1689	AUG	AGGAG	5-10bp	0.489	497	
HsPV-H4_RNA1_b_1	194	1684	AUG	GGAGG	5-10bp	0.518	496	
HsPV-H4_RNA2_a_1	194	952	AUG	AGGAG	5-10bp	0.519	252	
HsPV-H4_RNA2_a_2	949	1557	AUG	GGAG/GAGG	5-10bp	0.53	202	2
HsPV-H4_RNA2_b_1	198	953	AUG	AGGAG	5-10bp	0.541	251	
HsPV-H4_RNA2_b_2	950	1561	AUG	GGA/GAG/AGG	5-10bp	0.521	203	2
HsPV-H4_RNA2_c_1	481	1245	AUG	AGGAG	5-10bp	0.527	254	
HsPV-H4_RNA2_c_2	1242	1856	AUG	GGAG/GAGG	5-10bp	0.532	204	2
HsPV-H4_RNA2_d_1	190	945	AUG	AGGAGG	5-10bp	0.525	251	
HsPV-H4_RNA2_d_2	942	1553	AUG	GGAG/GAGG	5-10bp	0.513	203	2
HsPV-H5_RNA1_a_1	196	1689	AUG	AGGAG	5-10bp	0.489	497	
HsPV-H5_RNA1_b_1	195	1685	AUG	GGAGG	5-10bp	0.52	496	
HsPV-H5_RNA2_a_1	194	952	AUG	AGGAGG	5-10bp	0.539	252	
HsPV-H5_RNA2_a_2	949	1560	AUG	GGA/GAG/AGG	5-10bp	0.521	203	2
HsPV-H5_RNA2_b_1	194	952	AUG	AGGAG	5-10bp	0.519	252	
HsPV-H5_RNA2_b_2	949	1557	AUG	GGAG/GAGG	5-10bp	0.524	202	2
HsPV-H5_RNA2_c_1	198	953	AUG	AGGAG	5-10bp	0.532	251	
HsPV-H5_RNA2_c_2	950	1561	AUG	GGA/GAG/AGG	5-10bp	0.511	203	2
HsPV-Y66_RNA1_a_1	467	1627	AUG	GGAGG	5-10bp	0.526	386	
HsPV-Y66_RNA1_b_1	193	1683	AUG	AGGAGG	3-4bp	0.516	496	
HsPV-Y66_RNA2_a_1	486	1235	AUG	GGAG/GAGG	5-10bp	0.516	249	
HsPV-Y66_RNA2_a_2	1232	1855	AUG	GGA/GAG/AGG	5-10bp	0.516	207	2
HsPV-Y66_RNA2_b_1	193	948	AUG	GGAG/GAGG	5-10bp	0.519	251	
HsPV-Y66_RNA2_b_2	945	1556	GUG	GGA/GAG/AGG	5-10bp	0.493	203	2
HsRV_RNA1a_1	309	713	AUG	AATAA	6bp	0.407	134	
HsRV_RNA1a_2	706	1092	AUG	None	None	0.37	128	
HsRV_RNA1a_3	1097	3994	AUG	AATAA	15bp	0.378	965	
HsRV_RNA1a_4	4012	5784	AUG	None	None	0.386	590	
HsRV_RNA1b_1	379	711	AUG	GGAG/GAGG	5-10bp	0.435	110	
HsRV_RNA1b_2	708	1100	AUG	GGxGG	5-10bp	0.369	130	
HsRV_RNA1b_3	1103	3904	AUG	GGA/GAG/AGG	5-10bp	0.392	933	
HsRV_RNA1b_4	3916	5661	AUG	GGxGG	5-10bp	0.408	581	
HsRV_RNA1b_5	5710	5943	AUG	None	None	0.385	77	2
HsRV_RNA1c_1	15	755	AUG	GGA/GAG/AGG	5-10bp	0.328	246	
HsRV_RNA1c_2	748	1119	AUG	GGAG/GAGG	5-10bp	0.309	123	
HsRV_RNA1c_3	1122	4469	AUG	GGA/GAG/AGG	5-10bp	0.338	1115	
HsRV_RNA1c_4	4444	6219	GUG	GGAG/GAGG	5-10bp	0.336	591	
HsRV_RNA1c_5	6212	6481	AUG	GGAG/GAGG	5-10bp	0.333	89	3
HsRV_RNA1d_1	406	720	AUG	GGxGG	5-10bp	0.387	104	
HsRV_RNA1d_2	720	1103	AUG	GGxGG	5-10bp	0.385	127	
HsRV_RNA1d_3	1105	3900	GUG	GGA/GAG/AGG	5-10bp	0.393	931	
HsRV_RNA1d_4	3915	5660	AUG	GGA/GAG/AGG	5-10bp	0.41	581	
HsRV_RNA2a_1	422	910	AUG	GGA/GAG/AGG	5-10bp	0.429	162	2
HsRV_RNA2a_2	916	1527	AUG	GGAG/GAGG	5-10bp	0.363	203	4
HsRV_RNA2a_3	1533	2726	AUG	GGA/GAG/AGG	5-10bp	0.403	397	4
HsRV_RNA2a_4	2795	3604	AUG	GGAG/GAGG	5-10bp	0.383	269	
HsRV_RNA2a_5	3606	4199	AUG	None	None	0.37	197	
HsRV_RNA2a_6	4207	4338	AUG	GGAG/GAGG	5-10bp	0.364	43	
HsRV_RNA2b_1	197	784	AUG	GGxGG	5-10bp	0.415	195	2
HsRV_RNA2b_2	860	1471	AUG	GGA/GAG/AGG	5-10bp	0.373	203	2
HsRV_RNA2b_3	1478	2653	AUG	GGA/GAG/AGG	5-10bp	0.429	391	6
HsRV_RNA2b_4	2724	3440	AUG	GGxGG	5-10bp	0.411	238	
HsRV_RNA2b_5	3478	4230	AUG	GGAG/GAGG	5-10bp	0.393	250	
HsRV_RNA2c_1	397	903	AUG	GGA/GAG/AGG	5-10bp	0.448	168	2
HsRV_RNA2c_2	903	1520	AUG	AGGAGG	3-4bp	0.401	205	4
HsRV_RNA2c_3	1526	2719	AUG	GGA/GAG/AGG	5-10bp	0.424	397	7
HsRV_RNA2c_4	2775	3584	AUG	GGAG/GAGG	5-10bp	0.41	269	
HsRV_RNA2c_5	3629	4324	AUG	GGAG/GAGG	5-10bp	0.399	231	
HsRV_RNA2d_1	447	950	AUG	GGAG/GAGG	5-10bp	0.433	167	2
HsRV_RNA2d_2	947	1564	AUG	GGAGG	5-10bp	0.39	205	2
HsRV_RNA2d_3	1569	2762	AUG	GGA/GAG/AGG	5-10bp	0.427	397	6
HsRV_RNA2d_4	2822	3628	AUG	GGAGG	5-10bp	0.416	268	
HsRV_RNA2d_5	3661	4338	AUG	GGAG/GAGG	5-10bp	0.409	225	
HsRV_RNA2*a_1	264	674	AUG	GGAG/GAGG	5-10bp	0.45	136	2
HsRV_RNA2*a_2	667	1440	AUG	GGAG/GAGG	5-10bp	0.376	257	2
HsRV_RNA2*a_3	1445	2773	AUG	GGA/GAG/AGG	5-10bp	0.402	442	5
HsRV_RNA2*a_4	2825	3364	AUG	GGA/GAG/AGG	5-10bp	0.361	179	
HsRV_RNA2*a_5	3378	4520	AUG	None	None	0.392	380	
HsRV_RNA2*b_1	225	755	AUG	None	None	0.392	176	2
HsRV_RNA2*b_2	757	1515	AUG	GGA/GAG/AGG	5-10bp	0.406	252	2
HsRV_RNA2*b_3	1520	2827	AUG	GGAG/GAGG	5-10bp	0.388	435	5
HsRV_RNA2*b_4	2830	2955	AUG	GGAG/GAGG	5-10bp	0.389	41	
HsRV_RNA2*b_5	2912	3523	GUG	GGA/GAG/AGG	5-10bp	0.395	203	
HsRV_RNA2*b_6	3535	3690	AUG	GGA/GAG/AGG	5-10bp	0.353	51	
HsRV_RNA2*b_7	3687	4190	AUG	None	None	0.401	167	
Ga0169446_00510_vOTU_07046706_5662_1	59	1042	UUG	None	None	0.467		
Ga0169446_00510_vOTU_07046706_5662_2	1177	1428	AUG	AGGA	5-10bp	0.508		
Ga0169446_00510_vOTU_07046706_5662_3	1428	2864	AUG	GGA/GAG/AGG	5-10bp	0.495		
Ga0169446_00510_vOTU_07046706_5662_4	2903	4564	AUG	None	None	0.486		
Ga0169446_00510_vOTU_07046706_5662_5	4561	4746	AUG	AGGAG	5-10bp	0.462		
Ga0169446_00510_vOTU_07046706_5662_6	4824	5153	AUG	GGAGG	3-4bp	0.448		
Ga0393213_00017_vOTU_00596427_RC_5476_1	83	454	AUG	AGGAG	5-10bp	0.495		
Ga0393213_00017_vOTU_00596427_RC_5476_2	464	844	AUG	AGGA	5-10bp	0.462		
Ga0393213_00017_vOTU_00596427_RC_5476_3	844	1533	AUG	GGA/GAG/AGG	5-10bp	0.457		
Ga0393213_00017_vOTU_00596427_RC_5476_4	1514	2092	AUG	AGGA	5-10bp	0.489		
Ga0393213_00017_vOTU_00596427_RC_5476_5	2089	3573	AUG	GGAG/GAGG	5-10bp	0.492		
Ga0393213_00017_vOTU_00596427_RC_5476_6	3570	5294	AUG	4Base/6BMM	13-15bp	0.482		
Ga0456180_000042_vOTU_00649204_RC_5304_1	160	330	UUG	GGA/GAG/AGG	5-10bp	0.415		
Ga0456180_000042_vOTU_00649204_RC_5304_2	330	665	AUG	GGAG/GAGG	5-10bp	0.506		
Ga0456180_000042_vOTU_00649204_RC_5304_3	687	1988	AUG	GGAG/GAGG	5-10bp	0.52		
Ga0456180_000042_vOTU_00649204_RC_5304_4	1995	3485	AUG	GGAGG	5-10bp	0.53		
Ga0456180_000042_vOTU_00649204_RC_5304_5	3466	5214	AUG	GGA/GAG/AGG	5-10bp	0.525		

nature portfolio

Corresponding author(s): Syun-ichi Urayama

Last updated by author(s): Nov 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	A custom Perl pipeline script used for read clenup is available at GitHub (https://github.com/takakiy/FLDS). The following commercial programs were used. CLC GENOMICS WORKBENCH version 11.0 (Qiagen Japan, Tokyo, Japan); Genetyx version 14 (Genetyx, Tokyo, Japan) The following open source programs were used. Tablet viewer (version 1.19.09.03); phyloFlash (version 3.4); BLASTX (version 2.2.31+); Prodigal (version 2.6.3); HHpred (online server [no versions]); MEGA6.06; TMHMM (version 2.0); ColabFold 1.5.1; AlphaFold 2 through ColbFold v1.5.2; DALI (online, DaliLite.v5); ChimeraX (version 1.5); trimAl (version 1.4.rev15); RAxML (8.2.10); ProtTest (version 3.4.2); PROMALS3D; IQ-TREE (version 2.0.6); MAFFT (version 7); BLASTP (v2.9.0); HHblits (v3.3.0); ModelFinder (a part of IQ-TREE); BLASTn/p/x (2.12.0+)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Datasets obtained in this study have been available in the GenBank database repository (Accession Nos. HsRV: BTCN01000001-BTCN01000010; HsPV-H4: BTCO01000001-BTCO01000006; HsPV-H5: BTCP01000001-BTCP01000005; HsPV-Y66:BTCQ01000001-BTCQ01000004; H5_contig_1: BTCR01000001; Oi_contig_1-9: BTCS01000001-BTCS01000009) and Short Read Archive database (Accession No. DRA016131). Datasets (PDB70 [mmcif_2023-10-24], Pfam [v35], UniProt-SwissProt-viral70_Nov_2021 and NCBI-CD [v3.19]) are available at http://ftp.tuebingen.mpg.de/pub/protevo/toolkit/databases/hhsuite_dbs/. Searches using the IMG/VR dataset were available only at https://img.jgi.doe.gov/cgi-bin/vr/main.cgi?section=WorkspaceBlast&page=viralform. Datasets (SILVA SSU [version 138], Neo-HMM [v1.1], and RVDB-HMM [v23.0]) are publicly available.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This research does not involve human participants, their data, or biological material.
Reporting on race, ethnicity, or other socially relevant groupings	This research does not involve human participants, their data, or biological material.
Population characteristics	This research does not involve human participants, their data, or biological material.
Recruitment	This research does not involve human participants, their data, or biological material.
Ethics oversight	This research does not involve human participants, their data, or biological material.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study collected microbes in hot spring water and performed sequencing analyses for RNA virus discovery.
Research sample	Microbes in hot spring water.
Sampling strategy	No sample-size calculations were performed.
Data collection	The chemical composition of hot spring water was measured by T.O. Sequencing data were obtained using Illumina Miseq platform by M.H.
Timing and spatial scale	Sample were collected at 09- or 10-Mar-2017 and 17- or 18-Nov-2015. Each sample was collected once.
Data exclusions	Data from two sampling points were not included in analyses since we could not obtain data from these two samples.
Reproducibility	For data analyses, all raw data is available in the GenBank database repository. Reproducibility of environmental samples and sequencing was not confirmed.
Randomization	No randomization was performed and no controlling for covariants is relevant to this study design.
Blinding	Blinding does not apply to this study since it is discovery-oriented.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The weather was sunny or cloudy.
Location	Locations of the samplings are follow; H4: 31°54'07.5"N 130°50'06.2"E H5: 31°54'07.5"N 130°50'06.2"E T1-4: 31°54'37.7"N 130°49'00.6"E Y66, Y80, Y86: 31°55'03.8"N 130°48'40.4"E Oi: 32°44'25.3"N 130°15'48.4"E Ob: 32°43'33.0"N 130°12'24.7"E
Access & import/export	All samples were obtained with the permission of the landowner (or official manager) and in compliance with national law. The issuer are as follow; Unzen City, Unzen Nature Conservation Bureau, Kirishima Iwasaki Hotel, NIPPON PAPER LUMBER CO. LTD. and NITTETSU MINING CO. LTD KAGOSHIMA GEOTHERMAL FACILITY.
Disturbance	Sampling was done with a minimal number of people and collected from ample spring water sources.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	This study did not involve laboratory animals.
Wild animals	This study did not involve wild animals.
Reporting on sex	This study did not involve sex information.
Field-collected samples	A total of 11 samples were collected from five hot springs regions at southern Japan, in close proximity to active volcanoes, according to the instructions of Unzen City, Unzen Nature Conservation Bureau and private companies that maintain each hot spring region. At each sampling station, approximately 10 L of hot spring water was collected in a sterilized plastic bag, and then filtered with 0.2-µm-pore-size cellulose acetate membrane filters in 47 mm diameter (Advantec, Tokyo, Japan) within 0.5-3 hours after sampling. The filters were stored at −80°C until nucleic acid extraction.
Ethics oversight	No ethical approval or guidance was required

Note that full information on the approval of the study protocol must also be provided in the manuscript.