Unique viruses that infect Archaea related to eukaryotes

Ian Rambo

The University of Texas at Austin

Valerie De Anda

University of Texas at Austin <https://orcid.org/0000-0001-9775-0737>

Marguerite Langwig

The University of Texas at Austin

Brett Baker (✉ brett_baker@utexas.edu)

University of Texas Austin

Research Article

Keywords: Asgard archaea, viral genomes, Caudovirales-like structural proteins

Posted Date: October 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-960995/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Unique viruses that infect Archaea related to eukaryotes

Ian M. Rambo¹, Valerie De Anda¹, Marguerite V. Langwig¹⁺ and Brett J. Baker^{1,2*}

¹Department of Marine Science, The University of Texas at Austin, Port Aransas, TX

²Department of Integrative Biology, The University of Texas at Austin, Austin, TX

⁺Current address: Department of Bacteriology and Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI

^{*}Corresponding author. Email: brett_baker@utexas.edu

Asgard archaea are globally distributed, newly described microbes related to eukaryotes. Despite their importance, Asgard viruses have not been described. Here we characterize seven double-stranded DNA (dsDNA) viral genomes that infected Lokiarchaeota, Helarchaeota, and Thorarchaeota in deep-sea hydrothermal sediments. These viruses code for *Caudovirales*-like structural proteins, as well as proteins distinct from those described in archaeal viruses. They contain genes common in eukaryotic nucleocytoplasmic large DNA viruses (NCLDV), and appear to be capable of semi-autonomous genome replication, repair, epigenetic modifications, and transcriptional regulation. Moreover, Helarchaeota viruses may hijack host ubiquitin systems similar to eukaryotic viruses. Recovery of these Asgard viral genomes reveals they contain features of both prokaryotic and eukaryotic viruses, and provides insights into their roles in the ecology and evolution of their hosts.

Asgard archaea are globally distributed microbes that are related to eukaryotes ^{1,2}. Their genomic composition indicates they are descendants of the archaeal host that gave rise to the first eukaryotic common ancestor ³. Asgard biodiversity has expanded greatly in recent years due to the recovery of genomes from a range of marine and terrestrial aquatic sediments ⁴. An anaerobic, slow-growing Asgard lineage, Lokiarchaeota, has recently been cultured and appears to have syntrophic dependencies with bacteria ⁵. This supports multiple genome-based metabolic inferences of Asgard-bacteria interactions which are thought to have led to the formation of the first mitochondria-containing eukaryotic cell ^{5,6}. It is also hypothesized that interactions with viruses contributed to the origin of complex cellular life. This is based on the nucleus-like viral factories of some NCLDV ⁷ and bacteriophages ^{8,9}, which allow for replication within the host cytoplasm and the decoupling of transcription and translation by mRNA capping ¹⁰. Putative viral proteins have been identified within Lokiarchaeota genomes ¹¹, suggesting a role of viruses in the exchange of genetic elements and the evolution of Asgards. Type I and Type III CRISPR-Cas immune systems have been described in several Asgard phyla (Odinarchaeota, Thorarchaeota, Lokiarchaeota, and Helarchaeota) ¹², yet no Asgard-linked virus genomes have been characterized to date.

Fig. 1 | Protein clustering network of virus genomes recovered from Guaymas Basin sediments compared to previously described viruses. A monopartite network analysis constructed with vContact2 v0.9.19 and Cytoscape v3.8.0 for taxonomic assignment of Guaymas Basin UViGs (see Methods). Nodes represent individual genomes and edges indicate similarity among genomes within a Viral Cluster ($n = 770$). Asgard-linked viruses are labeled Fenrir, Nidhogg, and Ratatoskr. Vedfolnir and Sköll were classified as outliers. Created with BioRender.com.

Fig. 2 | Virus-host linkages to Asgard archaea. a, Lokiarchaeota, **b**, Helarchaeota, and **c**, Thorarchaeota MAGs are linked by CRISPR-Cas systems to UViGs. **d**, Overview of complete circular Nidhogg virus genome. Coding regions point right for (+) and left for (-) sense, and are drawn to scale. For **a**, **b**, and **c**, CRISPR spacer blastn-short alignments to a UViG (indicated by phage-like icons) are shown with arrows representing 100% identity (solid) and 95%-99.9% identity (dashed). The Lokiarchaeota, Helarchaeota, and Omnitrophica MAGs were recovered in this study. The Chloroflexi and Thorarchaeota genomes were recovered previously¹³. Predicted CRISPR-Cas systems of Asgard MAGs are represented by arrows colored according to *cas* Type signature, which point right for (+) and left for (-) sense. Arrows at the end of CRISPR arrays indicate sense in the same fashion. Each *cas* cassette displayed is located on a different scaffold. See Supplementary Text for abbreviations. Created with BioRender.com.

To explore the role of viruses in Asgard archaea, we searched ~5 TB of assembled metagenomic sequence data from hydrothermal vent-associated sediments in Guaymas Basin (~2,000 m water depth, Gulf of California)^{13–15}. From this, we recovered 6,756 uncultivated virus genomes (UViGs) (Fig. 1) estimated to be of high- and medium-quality (see Methods). The UViGs were linked to CRISPR spacer sequences from Guaymas Basin metagenome-assembled genomes (MAGs) to determine which viruses infected Asgard cells. This revealed seven dsDNA UViGs that infected Lokiarchaeota, Helarchaeota, and Thorarchaeota (Fig. 2, Supplementary File 1). Interestingly, two Asgard-linked viruses also infected *Chloroflexi* and *Omnitrophica* bacteria reconstructed from Guaymas Basin, suggesting a close association of Asgards with these bacteria. One UViG is an integrated provirus in the *Chloroflexi* genome with CRISPR spacer links to Thorarchaeota, while the other six are non-integrated and classified as lytic. We tentatively name these UViGs after Norse mythological creatures: “Vedfolnir” (*Chloroflexi* provirus linked to Thorarchaeota), “Fenrir” (Lokiarchaeota), “Sköll” (Lokiarchaeota), “Nidhogg” (Helarchaeota), and “Ratatoskr” (Helarchaeota and *Omnitrophica*). Their genome sizes (~21.9–117.5 kilobases (kb)) fall within the known range of archaeal dsDNA viruses¹⁶ (Fig. 2, Supplementary File 1). Nidhogg and Fenrir viruses recovered from the same sediment core (Meg22) but at different depths (10–12 and 12–14 cm) have highly similar nucleotide identities (99.98 and 99.90% identity for Nidhogg and Fenrir, respectively). Nidhogg viruses were closed into complete circular genomes (Fig. 2d; Extended Data Fig. 1; Supplementary File 2; see Methods) and the two linear Fenrir viruses are predicted to be >90% complete (Supplementary File 2).

In order to classify and understand the evolutionary histories of Asgard viruses, we identified DNA polymerase B (PolB) proteins in the seven UViGs. This protein is broadly distributed among all domains of life and DNA viruses¹⁷. PolB protein sequences from Fenrir, Nidhogg, and Ratatoskr were aligned with PolB from cellular organisms, phages, archaeal and eukaryotic viruses, and NCLDV¹⁸ (Supplementary File 3). Phylogenetic analyses of these sequences revealed that Fenrir, Nidhogg, and Ratatoskr are distinct from previously characterized viruses and have a complex evolutionary history (Fig. 3a). The close relationship of Fenrir PolB with Gammaproteobacteria suggests that these genes were horizontally transferred from bacteria, while Nidhogg PolB appears to be phylogenetically related to haloarchaeal myoviruses (Fig. 3A). Ratatoskr PolB is phylogenetically related to Heimdallarchaeota, the Asgard phylum thought to be most closely related to eukaryotes² (Fig. 3b). To determine the family-level taxonomy of Asgard viruses, we characterized their viral protein family (VPF) composition¹⁹ (Supplementary File 4). All seven Asgard UViGs contain proteins with family-level relatedness to *Myoviridae*, *Siphoviridae*, and *Podoviridae*, as well as NCLDVs and hyperthermophilic archaeal viruses (Fig. 3c). Despite Asgard UViGs not clustering with existing reference viruses (Fig. 1), their VPF membership ratios may indicate novel families or genera of *Caudovirales* (Fig. 3c). This is plausible given the associations of cosmopolitan and domain-specific *Caudovirales* with archaea^{16,20}. However, the lack of clear affiliation with previously described viruses highlights the novelty of these newly recovered Asgard viruses.

Fig. 3 | Taxonomic placement of Asgard viruses based on phylogeny and protein composition.
a, Maximum-likelihood phylogeny of DNA polymerase B generated using LG+F+R10 model. Circles on tree branches represent ultrafast bootstrap supports ≥ 95 . Viruses described in this paper are highlighted in gold. **b**, Ratatoskr placement within the Asgard Superphylum and Crenarchaeota group surrounded by the dashed box in **a**. Bootstraps are shown by values on branches. **c**, Membership ratios of VPFs shared among Asgard viruses. A, B and E in the legend indicate archaeal, bacterial, and/or eukaryotic hosts.

We identified hallmark proteins of *Caudovirales* in Asgard UViGs, including minor head protein, baseplate J, tail fibers, portal protein, and terminase large subunit (Fig. 4). Putative major capsid proteins (MCPs), which encode a viral protein shell, were identified in Ratatoskr, Nidhogg, and Vedfolnir. An HK-97-fold MCP was identified in Ratatoskr. Putative MCPs detected in Nidhogg and Vedfolnir (PhANNs, 80% and 90% confidence, respectively) have structural similarities to contractile phage tail sheath-like proteins (>99.9% probability, e-value <4.4e-24) and phage procapsid protease (99.5% probability, e-value 1.6e-13) (HHPred, Supplementary File 5). Fenrir and Ratatoskr both encode HNH endonucleases that potentially cleave DNA into genome-length units during packaging, and may operate in concert with their terminase large subunit and portal proteins²¹. However, putative MCPs were not identified in Fenrir and Sköll genomes, which may be due to the prevalence of novel virion architectures in archaea-specific viruses¹⁶. MCPs could be absent due to the tight evolutionary connections of capsid-containing viruses and capsidless selfish elements (CSE). Capsid-containing archaeal viruses have undergone multiple transitions to and from CSEs via the gain and loss of capsid genes as seen in eukaryotic viruses, but still parasitize the genetic information of their hosts^{22,23}.

To understand how Asgard viruses impact their hosts, we searched for genes present in prokaryotic viruses and NCLDV s involved in genome replication, nucleotide metabolism, transcription, and host-virus interactions^{17,18,24} (Fig. 4). Asgard viruses encode core DNA replication genes present in prokaryotic and eukaryotic viruses (PolB, archaeo-eukaryotic primase, clamp loaders, RNase H, and ATP-dependent DNA ligases)¹⁷. Nidhogg viruses have genes predicted to be involved in autonomous genome replication and proofreading including an ATP-dependent DNA ligase (*ligD*) with a transcription factor SII-like zinc finger domain, RNase HI/reverse transcriptase-like protein, PolB, and primase-polymerase (PrimPol). Viral PrimPol-like proteins are predicted to have additional roles in DNA and/or RNA priming, as well as damage-tolerant DNA polymerase activity²⁵. These genes have not been described in archaeal viruses to date, and have only been identified in corynebacterium virus BFK20 and *Acanthamoeba polyphaga* mimivirus²⁶. The Ratatoskr UViG encodes an archaeal-eukaryotic DNA primase for priming activity in DNA replication²⁵, as well as PolB. Fenrir and Nidhogg encode AAA ATPases with similarity to replication factor C small subunit (*rfcS*) and large subunit (*rfcL*) that may function as clamp loaders. Sköll has proliferating cell nuclear antigen (PCNA) that is most closely related to Lokiarchaeota from deep-sea sediments (36.5% protein similarity). PCNA is an essential replication processivity factor in eukaryotes and archaea²⁷ and is present in some of their viruses¹⁷. Sköll also codes for a protein with similarity to a ribonucleotide reductase (ribonucleoside diphosphate reductase, beta subunit) and hedgehog/intein superfamily domains. Ribonucleotide reductase is coded by a conserved core gene for nucleotide metabolism in NCLDV s¹⁸. The presence of a hedgehog/intein (Hint) protein domain in the ribonucleotide reductase may indicate that this gene was transferred to Sköll from a previous host. Fenrir viruses may regulate the dNTP pool during replication via a haloacid dehalogenase (HAD) superfamily protein similar to cytosolic 5'(3') deoxyribonucleotidase. Fenrir also contains a deoxynucleotide monophosphate kinase (DNMP kinase) similar to a large lake algae virus deoxynucleoside monophosphate kinase (bitscore = 115.9, Extended Data Fig. 2) and several *Mimiviridae* DNMP kinases (Supplementary File 5). Fenrir, Nidhogg, and Ratatoskr viruses possess DNA repair genes, which is a primary characteristic of NCLDV s. This is in contrast to small viruses, which lack these genes²⁸. Fenrir, Nidhogg, and Ratatoskr viruses also encode ERCC4 endonuclease domain proteins with similarity to African swine fever endonuclease EP364R. Nidhogg and Fenrir code ERCC4 endonucleases

154 similar to archaeal XPF 3' flap repair endonucleases. These XPF-like endonucleases contain helix-
155 hairpin-helix domains, enabling non-specific DNA binding²⁹.

156 In addition to core DNA replication machinery and repair genes, Fenrir and Nidhogg
157 viruses have a variety of mechanisms for gene expression and post-translational modification.
158 Nidhogg may regulate transcription via a homeodomain protein. Homeodomain proteins are
159 eukaryotic DNA-binding transcription factors also present in *Mimiviridae*^{24,30}. Nidhogg viruses
160 code a Ro60-like protein with TROVE and von Willebrand factor A domains. Ro60 may form a
161 ribonucleoprotein complex with non-coding Y RNA to interact with proteins and other RNAs³¹.
162 Nidhogg viruses may also repair, splice, and edit transcripts via RNA ligase³² (Fig. 4). Fenrir
163 codes a winged helix-turn-helix domain protein, which is a common DNA binding motif found in
164 archaeal virus transcription factors²⁰. Translation machinery was not identified in these UViGs,
165 suggesting they utilize host ribosomes and eukaryotic initiation factors present in the Asgard hosts.
166 However, Nidhogg and Fenrir may process viral protein precursors with encoded peptidases.
167 Nidhogg is predicted to code a peptidase with similarity to Vaccinia virus I7L/C57-family
168 processing endopeptidase, while Fenrir appears to be capable of producing an M42-family
169 aminopeptidase (Fig. 4).

170 Several Asgard phyla code homologs of eukaryotic ubiquitin protein modification systems
171 ². Eukaryotic viruses have been shown to hijack these systems to aid in various stages of viral
172 propagation³³. Nidhogg codes three ubiquitin-activating (ThiF family/E1-like) protein homologs
173 (Extended Data Fig. 1) that could utilize Helarchaeota ubiquitin systems (E1-like activating
174 enzyme, E2-like conjugating enzyme, and ubiquitin-like protein domain) (Supplementary File 6)
175 and repurpose them for processes such as replication, transcription, virion formation, and release
176 ³³. Phylogenetic analyses of Nidhogg proteins suggest they are novel E1 ubiquitin-activating
177 homologs that are ancestral to Helarchaeota E1-like and eukaryotic ubiquitin-like modifier
178 activating 5 proteins (Extended Data Fig. 3; Supplementary File 7). This is interesting given that
179 this type of host-virus interaction has only been described in eukaryotic systems.

Nidhogg, and Ratatoskr also encode hexosyltransferases, a class of glycosyltransferases (GTs) that may confer the ability to modify hexoses on DNA, proteins, or lipids in order to evade host defense systems and inhibit host molecular functions³⁷. GTs have been identified in phages and NCLDV^{37,38}, but are not well-documented in archaeal viruses. Fenrir encodes two overlapping genes with sequence homology to *Sulfolobus islandicus* rod-shaped virus 1 uncharacterized GT (SIRV1-GT) and alpha-1,2-mannosyltransferase (WbdA). This suggests that Fenrir could produce mannose glycan chains that are terminated by an encoded SAM-dependent methyltransferase (WbdD)³⁹ without utilizing host machinery. Nidhogg encodes a GT with hits to SIRV1-GT, as well as a glycogen synthase (GlgA) homolog which may form glucan chains (Fig. 4a).

Given the relatedness of Asgards to eukaryotes, it is interesting that Asgard-infecting viruses have replication and regulation machinery akin to eukaryotic NCLDV³⁷s (PCNA, PolB, ribonucleotide reductase, DNA ligase, DNMP kinase, and replication factor C-like AAA ATPases). However, mRNA capping genes that allow for uncoupling of transcription and translation in NCLDV³⁷s do not appear to be present in these viruses. In addition, many of their characteristics are distinct from previously studied archaeal viruses. For example, the complete Nidhogg genomes code for homeodomain and PrimPol proteins which are lacking in archaeal viruses. Nidhogg viruses also contain genes for ubiquitin protein modification systems which are used by eukaryotic viruses for propagation but have not been previously described in archaeal viruses. These Asgard viruses as a whole encode structural proteins reminiscent of those in *Caudovirales*, yet most of their protein composition is distinct from this order. They appear to have both archaeal and eukaryotic viral characteristics, which is consistent with the evolutionary position of their hosts. This first description of Asgard-linked viruses advances our understanding of the roles of viruses in the ecology and evolution of Asgards.

Methods

Metagenomic assembly and binning. The sampling, assembly, and binning methods for the genomic data analyzed in this study are described in detail in references^{13–15}. Samples were collected from Guaymas Basin (GB) in the Gulf of California, Mexico (27°N 0.388, 111°W 24.560) from a depth of ~2,000 m using polycarbonate cores. Whole community DNA from ≥10 g of sediment was extracted using the DNeasy PowerSoil and MoBio PowerMax kits (QIAGEN, Hilden, Germany). DNA concentrations were quantified using a QUBIT 2.0 fluorometer (ThermoFisher, Singapore) and metagenomic sequencing was performed at the Michigan State University RTSF Genomics Core and the Joint Genome Institute. Sequencing was performed in a 2×150 bp and 2×125 bp paired end format. Sequences described in¹⁴ were trimmed and quality controlled using Sickle v1.33⁴⁰ and assembly was performed using IDBA-UD v1.0.9 (--seed_kmer 55)⁴¹. In¹³, sequences were quality controlled and assembled by the Joint Genome Institute (JGI). In¹⁴ binning of individual GB assemblies (scaffolds ≥2 kb) was performed using CONCOCT v0.4.0⁴² and MetaBat v2.12.1⁴³. CONCOCT was used with default settings and MetaBat was run with the following parameters: --minCVSum 0 --saveCls -d -v --minCV 0.1 -m 2000. Consensus MAGs produced from these two binning tools were determined using DAS Tool v1.0⁴⁴ with default settings. In¹³, binning was performed using ESOM, Anvi'o v2.2.2, and MetaBat v1, and consensus MAGs were obtained using DAS Tool v1.0. CheckM v1.0.11 and v1.0.5⁴⁵ was used to determine MAG completeness and contamination (Table S2). Binned and unbinned contigs ≥2 kb were annotated with IMG/M v5.0⁴⁶. Annotations were consolidated and mapped to MAGs using the

MetaGaia metabolic_profile.py script (<https://github.com/valdeanda/MetaGaia>). Additional annotation of Asgard MAGs was performed with InterProScan v5.31-70.0⁴⁷.

CRISPR-Cas system prediction and subtyping. CRISPR arrays were predicted in GB MAGs with CRISPRDetect v2.4⁴⁸ using the following parameters: -minimum_no_of_repeats 3 -minimum_repeat_length 23 -right_flank_length 500 -left_flank_length 500 -array_quality_score_cutoff 3. Spacer sequences were extracted from the output GFF3 file using a custom Python script, and de-replicated using CD-HIT-EST v4.8.1⁴⁹ (parameters: -c 1.0 -n 11 -b 20). CRISPRDetect and IMG/M CRISPR Recognition Tool v1.8.2⁵⁰ results were compared to verify CRISPR array predictions.

Cas cassette prediction and subtyping were performed for MAGs using CRISPRcasIdentifier v1.0.0⁵¹, with combinations of Classification and Regression Trees (CART), Support Vector Machine (SVM), and Extremely Randomized Trees (ERT) algorithms for classification and regression (parameters: --mode combined --regressors ERT SVM CART --classifiers ERT SVM CART --class-probabilities --hmm-sets HMM1 HMM2 HMM3 HMM4 HMM5). Subtypes were assigned via a consensus-based approach in which all combinations of CART, SVM, and ERT classifiers/regressors provided the same subtype assignment, with that subtype probability exhibiting the highest mean and lowest standard deviation of the predictions. *Cas* cassette annotations that served as the basis for subtype predictions were compared against IMG/M and InterProScan results, and subtype predictions were verified by ensuring the presence of key effector complex genes^{52,53}.

Viral contig recovery and annotation. VIBRANT v1.2.1 was used to recover uncultivated viral genomes (UViGs) as previously described⁵⁴. In brief, contigs ≥ 10 kb containing >4 open reading frames (ORFs) predicted by Prodigal v2.6.3⁵⁵ were annotated using the KOALA-KOFAM v2019-03-20⁵⁶ Pfam v32⁵⁷, and virus orthologous groups (VOG) release 94⁵⁸ databases with hmmsearch via HMMER v3.3.2⁵⁹. For contigs containing a provirus, the proviral region was extracted and retained if it exhibited a clear shift from the host genome, met the aforementioned sequence length and ORF requirements, and contained a putative integrase. Lytic and provirus UViGs designated in VIBRANT as mid- and high-quality UViGs based on Virus Orthologous Groups (VOGs), nucleotide replication proteins, and enrichment of viral hallmark genes (e-value and bitscore cutoffs of $1e-5$ and 30, respectively) were retained. VirSorter v2.2.3⁶⁰ was run against the same assemblies to assure the quality of retained VIBRANT UViGs (parameters: --include-groups dsDNAphage,ssDNA,NCLDV --min-length 10000 --min-score 0.5 --viral-gene-required --hallmark-required-on-short). To reduce the chance of false positives given the sensitive VirSorter2 minimum quality score cutoff of 0.5, UViGs also identified with VIBRANT were retained if they had VirSorter2 quality scores ≥ 0.95 , contained viral hallmark genes, and encoded $>3x$ more viral than cellular genes. UViG completeness and quality were assessed with CheckV v0.8.1⁶¹ using HMM and amino acid identity metrics (parameters: end_to_end; CheckV database v1.0). Complete circular genomes based on direct terminal repeats⁶¹ and complete kmer matches⁵⁴ were verified with BLASTN⁶². Pairwise alignment-based gene annotation was performed with DIAMOND v2.0.4.142⁶³ using BLASTP⁶² against the NCBI NR database (downloaded October 4, 2020). HMM-based annotations were completed using InterProScan v5.31-70.0⁴⁷, IMG/M v5.0⁴⁶, and nucleocytoplasmic virus orthologous group HMMs (NCVOGs)⁶⁴ with HMMER v3.3.2⁵⁹. Candidate UViGs were verified to contain hypothetical gene enrichments. Artificial neural network-based classification of phage structural proteins was performed with the PhANNs v0.3

web server ⁶⁵. Protein homology of putative major capsid proteins assigned by PfANNs was assessed using the HHpred web server ⁶⁶ against the PDB_mmCIF70_17_May, SCOPe70_2.07, Pfam-Av34, and UniProt-SwissProt-viral70_23_August_2020 databases (HHBlits=>UniRef30 multiple sequence alignment (MSA) generation method, 3 MSA generation iterations, 20% min coverage of MSA hits, local alignment, no MAC realignment, secondary structure scoring during alignment, 1e-6 cutoff for MSA generation, 0% min seq identity of MSA hits with query). Visualization of the circular Nidhogg genome was performed with the CGView server ⁶⁷.

Protein clustering and taxonomic comparisons against publicly available viruses. Protein clustering networks for family and genus-level taxonomic assignment of the 6,756 GB UViGs were constructed using vContact2 v0.9.19 ⁶⁸ using all-vs-all DIAMOND alignments, ClusterONE ⁶⁹ and MCL ⁷⁰ (parameters: --pcs-mode MCL --vcs-mode ClusterONE --seed-method unused_nodes --similarity match --merge-method single --min-size 2 --rel-mode Diamond). Reference viruses were obtained from the RefSeq v201 prokaryotic virus database, eukaryotic viruses from RefSeq v202 viral database ⁷¹ (n = 11,082) and huge phage genomes (n = 361) ⁷². Protein clustering networks were visualized in Cytoscape v3.8.0 ⁷³ using an edge-weighted spring embedded model. UViG family and genus-level taxonomic classifications were performed with VPF-Class v1.0 ¹⁹ via reference viral protein families (VPFs) using a minimum e-value cutoff of 1e-4. VPF classification files from May 20, 2020 were used to build the reference database. Waffle charts representing VPF membership ratios were created in RStudio v1.3.959 ⁷⁴ using R v4.0.0 ⁷⁵ and the ‘waffle’ package v1.0.1 ⁷⁶.

CRISPR-based host-virus linking. Host-virus linkages were resolved by aligning MAG CRISPR spacers to viral contigs ⁷⁷ using blastn-short with BLAST+ v2.5.0 ⁶² (percent identity $\geq 95\%$, e-value $\leq 1e-5$, 100% query coverage, word size = 8, maximum target sequences = 10000000, allowing 1 mismatch). An undirected network of significant virus-host CRISPR spacer matches from GB samples was constructed with NetAn v1.0 ⁷⁸ and visualized in RStudio v1.3.959 ⁷⁴ with VisNetwork v2.0.9 (<https://github.com/datastorm-open/visNetwork>).

Phylogenetic analysis. A phylogenetic tree of DNA polymerase family B (PF00136) was constructed to better resolve the evolutionary relationships of Asgard-linked viruses. *Caudovirales* and *Herpesvirales* sequences from PF00136 were de-replicated at 100% similarity using CD-HIT ⁴⁹ (word-size = 5). Bacterial sequences were clustered at 50% similarity with CD-HIT (word-size = 3). Eukaryote, NCLDV, and archaea references were selected as previously described ⁷⁹. A total of 746 sequences were aligned with MAFFT v7.457 (parameters: --maxiterate 1000) ⁸⁰. Poorly-aligned sequences were manually removed with Geneious Prime v2021.0.3 and re-aligned with MAFFT v7.450. Alignment columns comprising >50% gaps were removed. Maximum-likelihood phylogenetic trees were constructed with IQTree v2.0.3 ⁸¹ and model selection was performed with ModelFinderPlus ⁸². Two trees were constructed as follows: 1) LG+F+R10 with 1000 ultrafast bootstraps (parameters: -m MFP -bb 1000 -bnni -alrt 1000 -wsl -wsr -wbt -mset LG, LG+C20, LG+C40, LG+C60, LG4M, LG4X -mrates ,R4,R5,R8,R10,G4,G8 -mfreq ,F) and 2) LG+F+R10 with the C60 complex mixture model over 1000 ultrafast bootstraps (parameters: -m MFP -bb 1000 -bnni -alrt 1000 -wsl -wsr -wbt -mset LG, LG+C20, LG+C40, LG+C60, LG4M, LG4X -mrates ,R4,R5,R8,R10,G4,G8 -mfreq ,F), followed by a posterior mean site frequency model approximation using the ultrafast bootstrap guide trees and LG+C60+F+R10 model.

A maximum likelihood phylogeny of DNMP kinase was constructed to resolve the placement of Lokiarchaeota virus Fenrir sequences. Reference bacterial and viral DNMP kinase sequences were downloaded from Uniprot, along with top BLASTP hits to Fenrir virus DNMP kinase (word size = 6, 1e-10 maximum e-value, bitscore $\geq$ 90). Uniprot sequences were clustered at 97% identity (for bacteria) and de-replicated at 100% identity (for viruses) with CD-HIT (word-size = 5). Sequences were aligned with MUSCLE v3.8.1551⁸³. The MSA was curated in Geneious Prime v2021.0.3 by manually removing poorly-aligned sequences and re-aligning with MAFFT. Alignment columns comprising >50% gaps were removed. A maximum-likelihood phylogenetic tree of 241 sequences was constructed with RAxML v8.2.11 (HPC-PTHREADS-AVX, parameters: -m PROTGAMMAAUTO -N 1000 -p 44 -x 44 -f a) using an LG likelihood model with fixed base frequencies.

The ubiquitin-activating enzyme E1-like tree was constructed from 11 protein sequences belonging to the NEDD8-activating enzyme E1 catalytic subunit family (IPR030468), 218 sequences belonging to the ubiquitin-activating E1 enzyme (IPR035985), 14 viral sequences obtained from NCBI (search parameters: ThiF family protein[Protein Name] OR ubiquitin-activating enzyme [Protein Name] AND viruses[Filter]), 133 sequences derived from Lokiarchaeota and Helarchaeota based on InterProScan v5.31-70.0 results, and 3 Nidhogg virus sequences. Sequences were aligned with MAFFT v7.457⁸⁰. The MSA was curated in Geneious Prime v2021.0.3 by manually removing poorly-aligned sequences and re-aligning with MAFFT v7.457. Alignment columns comprising >50% gaps were removed. The phylogeny was constructed with IQTree v1.6.1⁸¹ and the LG+R8 model was selected with ModelFinderPlus⁸² (parameters: -m MFP -bb 1000 -bnni -seed 42). All phylogenetic trees described in this study were visualized in iTol v6.3⁸⁴.

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

Data availability

The genomic sequences associated with the study have been deposited in NCBI under BioProject PRJNA743331.

References

1. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
2. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
3. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).
4. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nat Microbiol* **5**, 887–900

376 (2020).

377 5. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**,
378 519–525 (2020).

379 6. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the eukaryotic cell
380 based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148
381 (2019).

382 7. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Curr. Opin.*
383 *Microbiol.* **31**, 44–49 (2016).

384 8. Chaikerasitak, V. *et al.* Assembly of a nucleus-like structure during viral replication in
385 bacteria. *Science* **355**, 194–197 (2017).

386 9. Malone, L. M. *et al.* A jumbo phage that forms a nucleus-like structure evades CRISPR-Cas
387 DNA targeting but is vulnerable to type III RNA-based immunity. *Nat Microbiol* **5**, 48–55
388 (2020).

389 10. Bell, P. J. L. Evidence supporting a viral origin of the eukaryotic nucleus. *Virus Res.* **289**,
390 198168 (2020).

391 11. Krupovic, M., Dolja, V. V. & Koonin, E. V. The LUCA and its complex virome. *Nat. Rev.*
392 *Microbiol.* **18**, 661–670 (2020).

393 12. Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2
394 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

395 13. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and
396 biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999
397 (2018).

398 14. Castelle, C. J. *et al.* Protein Family Content Uncovers Lineage Relationships and Bacterial

399 Pathway Maintenance Mechanisms in DPANN Archaea. *Front. Microbiol.* **12**, 660052
400 (2021).

401 15. Langwig, M. V. *et al.* Large-scale protein level comparison of Deltaproteobacteria reveals
402 cohesive metabolic groups. *ISME J.* (2021).

403 16. Prangishvili, D. *et al.* The enigmatic archaeal virosphere. *Nat. Rev. Microbiol.* **15**, 724–739
404 (2017).

405 17. Kazlauskas, D., Krupovic, M. & Venclovas, Č. The logic of DNA replication in double-
406 stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res.*
407 **44**, 4551–4564 (2016).

408 18. Iyer, L. M., Aravind, L. & Koonin, E. V. Common origin of four diverse families of large
409 eukaryotic DNA viruses. *J. Virol.* **75**, 11720–11734 (2001).

410 19. Pons, J. C. *et al.* VPF-Class: Taxonomic assignment and host prediction of uncultivated
411 viruses based on viral protein families. *Bioinformatics* (2021).

412 20. Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V. Viruses
413 of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res.* **244**,
414 181–193 (2018).

415 21. Kala, S. *et al.* HNH proteins are a widespread component of phage DNA packaging
416 machines. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6022–6027 (2014).

417 22. Koonin Eugene V. & Dolja Valerian V. Virus World as an Evolutionary Network of Viruses
418 and Capsidless Selfish Elements. *Microbiol. Mol. Biol. Rev.* **78**, 278–303 (2014).

419 23. Iranzo Jaime, Koonin Eugene V., Prangishvili David, Krupovic Mart & Sandri-Goldin R. M.
420 Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between
421 Viruses and Capsidless Mobile Elements. *J. Virol.* **90**, 11043–11055 (2016).

- 422 24. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic
423 DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution.
424 *Viol. J.* **6**, 223 (2009).
- 425 25. Guillian, T. A., Keen, B. A., Brissett, N. C. & Doherty, A. J. Primase-polymerases are a
426 functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Res.* **43**,
427 6651–6664 (2015).
- 428 26. Gupta, A., Lad, S. B., Ghodke, P. P., Pradeepkumar, P. I. & Kondabagil, K. Mimivirus
429 encodes a multifunctional primase with DNA/RNA polymerase, terminal transferase and
430 translesion synthesis activities. *Nucleic Acids Res.* **47**, 6932–6945 (2019).
- 431 27. MacNeill, S. A. PCNA-binding proteins in the archaea: novel functionality beyond the
432 conserved core. *Curr. Genet.* **62**, 527–532 (2016).
- 433 28. Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G. & Raoult, D. Mimivirus:
434 leading the way in the discovery of giant viruses of amoebae. *Nat. Rev. Microbiol.* **15**, 243–
435 254 (2017).
- 436 29. Doherty, A. J., Serpell, L. C. & Ponting, C. P. The Helix-Hairpin-Helix DNA-Binding
437 Motif: A Structural Basis for Non-Sequence-Specific Recognition of DNA. *Nucleic Acids*
438 *Res.* **24**, 2488–2497 (1996).
- 439 30. Iyer, L. M., Balaji, S., Koonin, E. V. & Aravind, L. Evolutionary genomics of nucleo-
440 cytoplasmic large DNA viruses. *Virus Res.* **117**, 156–184 (2006).
- 441 31. Sim, S., Hughes, K., Chen, X. & Wolin, S. L. The Bacterial Ro60 Protein and Its Noncoding
442 Y RNA Regulators. *Annu. Rev. Microbiol.* **74**, 387–407 (2020).
- 443 32. Ho, C. K., Wang, L. K., Lima, C. D. & Shuman, S. Structure and mechanism of RNA ligase.
444 *Structure* **12**, 327–339 (2004).

- 445 33. Tang, Q., Wu, P., Chen, H. & Li, G. Pleiotropic roles of the ubiquitin-proteasome system
446 during viral propagation. *Life Sci.* **207**, 350–354 (2018).
- 447 34. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage orphan
448 DNA methyltransferases: insights from their bacterial origin, function, and occurrence.
449 *Appl. Environ. Microbiol.* **79**, 7547–7555 (2013).
- 450 35. Jeudy, S. *et al.* Exploration of the propagation of transpovirons within Mimiviridae reveals a
451 unique example of commensalism in the viral world. *ISME J.* **14**, 727–739 (2020).
- 452 36. Agarkova, I. V., Dunigan, D. D. & Van Etten, J. L. Virion-associated restriction
453 endonucleases of chloroviruses. *J. Virol.* **80**, 8114–8123 (2006).
- 454 37. Markine-Goriaynoff, N. *et al.* Glycosyltransferases encoded by viruses. *J. Gen. Virol.* **85**,
455 2741–2754 (2004).
- 456 38. Piacente, F., Gaglianone, M., Laugieri, M. E. & Tonetti, M. G. The Autonomous
457 Glycosylation of Large DNA Viruses. *Int. J. Mol. Sci.* **16**, 29315–29328 (2015).
- 458 39. Hagelueken, G. *et al.* A coiled-coil domain acts as a molecular ruler to regulate O-antigen
459 chain length in lipopolysaccharide. *Nat. Struct. Mol. Biol.* **22**, 50–56 (2014).
- 460 40. Joshi, N. & Sickles, F. J. A sliding-window, adaptive, quality-based trimming tool for FastQ
461 files. (2011).
- 462 41. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for
463 single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**,
464 1420–1428 (2012).
- 465 42. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat.*
466 *Methods* **11**, 1144–1146 (2014).
- 467 43. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient

genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

44. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* vol. 3 836–843 (2018).

45. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

46. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

47. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

48. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).

49. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

50. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

51. Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C. P. L. F. & Backofen, R. CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. *Gigascience* **9**, (2020).

52. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

53. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).

491 54. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and
492 curation of microbial viruses, and evaluation of viral community function from genomic
493 sequences. *Microbiome* **8**, 90 (2020).

494 55. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
495 identification. *BMC Bioinformatics* **11**, 119 (2010).

496 56. Aramaki, T. *et al.* KofamKOALA: KEGG ortholog assignment based on profile HMM and
497 adaptive score threshold. *Bioinformatics*. **36**, 2251-2252 (2020)).

498 57. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**,
499 D427–D432 (2019).

500 58. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups
501 (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids*
502 *Res.* **45**, D491–D498 (2017).

503 59. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

504 60. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA
505 and RNA viruses. *Microbiome* **9**, 37 (2021).

506 61. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled
507 viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).

508 62. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
509 (2009).

510 63. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
511 DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

512 64. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics.
513 *Nature* **578**, 432–436 (2020).

514 65. Cantu, V. A. *et al.* PhANNs, a fast and accurate tool and web server to classify phage
515 structural proteins. *Plos Comp Biol.* **16**, e1007845 (2020).

516 66. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a
517 New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).

518 67. Grant, J. R. & Stothard, P. The CGView Server: a comparative genomics tool for circular
519 genomes. *Nucleic Acids Res.* **36**, W181–W184 (2008).

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

522 69. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-
523 protein interaction networks. *Nat. Methods* **9**, 471–472 (2012).

524 70. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
525 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

526 71. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology
527 Information. *Nucleic Acids Res.* **37**, D5–15 (2009).

528 72. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth’s ecosystems. *Nature* **578**,
529 425–431 (2020).

530 73. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular
531 interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

532 74. RStudio Team. RStudio: Integrated Development Environment for R. (2019).

533 75. R Core Team. R: A Language and Environment for Statistical Computing. (2020).

534 76. Rudis, B. & Gandy, D. waffle: Create waffle chart visualizations in R. (2016).

535 77. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance
536 in natural microbial communities. *Science* **320**, 1047–1050 (2008).

78. De Anda, V. *et al.* Understanding the Mechanisms Behind the Response to Environmental Perturbation in Microbial Mats: A Metagenomic-Network Based Approach. *Front. Microbiol.* **9**, 2606 (2018).
79. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19585–19592 (2019).
80. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
81. 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).
82. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
83. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
84. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

Acknowledgments: We thank Dr. Daniel Tamarit and Dr. Thijs Ettema for advice about this research.

Funding:

Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grant 73592LPI (<https://doi.org/10.46714/735925LPI>) (BJB)

Simons Foundation early career award 687165 (BJB)

National Science Foundation Awards OCE-0647633 (APT)

563 **Author contributions:**
564 Conceptualization: VDA and BJB
565 Data curation: IMR, VDA, MVL
566 Funding acquisition: BJB
567 Investigation: IMR, VDA
568 Methodology: IMR, VDA, BJB
569 Project administration: BJB
570 Resources: BJB
571 Supervision: VDA, BJB
572 Visualization: IMR, VDA
573 Writing – original draft: IMR, VDA, MVL, BJB
574 Writing – review & editing: IMR, VDA, MVL, BJB
575
576 **Competing interests:** The authors declare that they have no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DataS1.xlsx](#)
- [DataS2.xlsx](#)
- [DataS3.xlsx](#)
- [DataS4.xlsx](#)
- [DataS5.xlsx](#)
- [DataS6.xlsx](#)
- [DataS7.xlsx](#)
- [RamboetalNatureMicroSupplementaryInformation.pdf](#)