



**HAL**  
open science

## A closed *Candidatus Odinarchaeum* genome exposes Asgard archaeal viruses

Daniel Tamarit, Eva F Caceres, Mart Krupovic, Reindert Nijland, Laura Eme,  
Nicholas P Robinson, Thijs J G Ettema

► **To cite this version:**

Daniel Tamarit, Eva F Caceres, Mart Krupovic, Reindert Nijland, Laura Eme, et al.. A closed *Candidatus Odinarchaeum* genome exposes Asgard archaeal viruses. 2021. hal-03368645

**HAL Id: hal-03368645**

**<https://hal.science/hal-03368645>**

Preprint submitted on 6 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

**A closed *Candidatus Odinarchaeum* genome exposes Asgard archaeal viruses**

Daniel Tamarit<sup>1,2</sup>, Eva F. Caceres<sup>3</sup>, Mart Krupovic<sup>4</sup>, Reindert Nijland<sup>5</sup>, Laura Eme<sup>6</sup>, Nicholas P. Robinson<sup>7</sup>, Thijs J. G. Ettema<sup>1</sup>

<sup>1</sup> Laboratory of Microbiology, Wageningen University, 6708WE Wageningen, The Netherlands

<sup>2</sup> Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, SE-75007 Uppsala, Sweden

<sup>3</sup> Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, 75123 Uppsala, Sweden.

<sup>4</sup> Archaeal Virology Unit, Institut Pasteur, Paris, France

<sup>5</sup> Marine Animal Ecology Group, Wageningen University, 6708WE Wageningen, The Netherlands

<sup>6</sup> Laboratoire Écologie, Systématique, Évolution, CNRS, Université Paris-Sud, Université Paris-Saclay, AgroParisTech, 91400 Orsay, France

<sup>7</sup> Division of Biomedical and Life Sciences, Faculty of Health and Medicine, Lancaster University, Lancaster LA1 4YG, United Kingdom

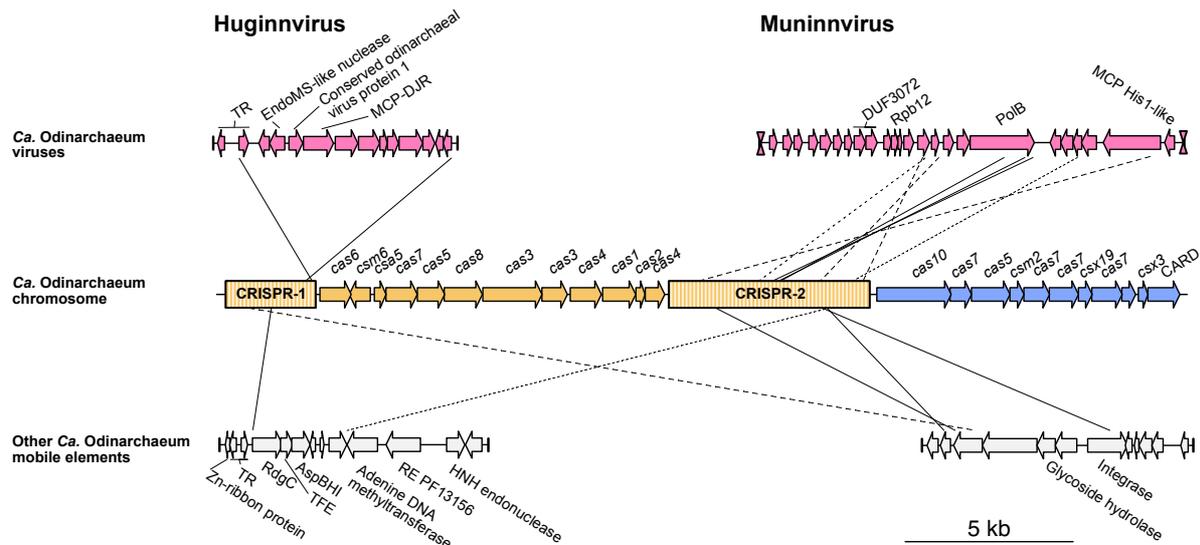
21 **Asgard archaea have recently been identified as the closest archaeal relatives of**  
22 **eukaryotes. Their ecology remains enigmatic, and their virome, completely unknown.**  
23 **Here, we describe the closed genome of *Ca. Odinararchaeum yellowstonii* LCB\_4, and,**  
24 **from this, obtain novel CRISPR arrays with spacer targets to several viral contigs. We**  
25 **find related viruses in sequence data from thermophilic environments and in the**  
26 **genomes of diverse prokaryotes, including other Asgard archaea. These novel viruses**  
27 **open research avenues into the ecology and evolution of Asgard archaea.**

28 Asgard archaea are a diverse group of microorganisms that comprise the closest relatives of  
29 eukaryotes<sup>1-7</sup>. Their genomes were first explored over six years ago<sup>8</sup>, and much of their  
30 physiology and cell biology remains to be studied. While over two hundred draft genomes are  
31 available for this group, the majority is represented by highly fragmented and incomplete  
32 metagenome assembled genomes (MAGs), which has precluded obtaining insights into their  
33 mobile genetic elements (mobilome). Given the central role of Asgard archaea in  
34 eukaryogenesis models, access to their complete genomes and information about their  
35 interactions with viruses are highly relevant.

36 To obtain a complete Asgard archaeal genome, we reassembled the *Odinararchaeota* LCB\_4  
37 genome, currently a 96% complete assembly of 1.46 Mbp distributed in 9 contigs<sup>1</sup>. A promising  
38 reassembly yielded a 1.41 Mbp contig, a 13 kbp contig containing CRISPR-associated (*cas*)  
39 genes, and multiple short contigs harboring mobile element or repeat signatures (see Methods  
40 for details). After contig boundary inspection, we postulated that the first two contigs  
41 represented the entire *Odinararchaeota* LCB\_4 chromosome as these were flanked by similar  
42 CRISPR arrays that extended for several kilobasepairs (Suppl. Fig S1). We successfully  
43 amplified these gaps using long-range PCR, sequenced the resulting amplicons with  
44 Nanopore sequencing, and performed a hybrid assembly, finally generating a single 1.418  
45 Mbp circular contig. Given the high quality of this genome, we suggest recognizing this strain  
46 as *Candidatus Odinararchaeum yellowstonii* LCB\_4 (hereafter 'LCB\_4'), in reference to  
47 Yellowstone National Park, location of the hot spring where it was sampled.

48 The LCB\_4 genome contains a unique disposition of CRISPR-Cas genes (Fig. 1), including  
49 neighboring Type I-A and Type III-D *cas* gene clusters, separated by a 6.1 kb-long Type I-A  
50 CRISPR array, and further followed by another 2.7 kb-long Type I-A CRISPR array, with a  
51 total of 144 CRISPR spacers across both arrays. Nine of these spacers targeted, with 100%  
52 identity and query coverage, four putative mobile element contigs obtained in the same  
53 assembly that were not part of the closed genome (Fig. 1). Two of these contigs contained  
54 genes encoding common mobile element proteins, such as restriction endonucleases and  
55 integrases, but did not contain any viral signature genes (Suppl. Table S1). A third contig  
56 represented a complete, circular viral genome (Suppl. Fig. 1E) encoding transcriptional  
57 regulators, an endonuclease, and a double jelly-roll major capsid protein (DJR-MCP), typical  
58 of tailless icosahedral viruses (Fig. 1; Suppl. Fig. S2A; Suppl. Table S1). This specific protein  
59 was previously found in a study of the DJR-MCP family, and was tentatively classified as  
60 belonging to an "Odin group" since it was found in the same metagenome as *Ca.*  
61 *Odinararchaeum* LCB\_4<sup>9</sup>. The present complete recovery of LCB\_4 CRISPR arrays allowed us  
62 to confirm that this circular contig indeed represents a virus associated with *Ca.*  
63 *Odinararchaeum*, for which we suggest the name Huginnvirus, in reference to Odin's raven,  
64 Huginn ("thought").

65



66

67 **Figure 1. *Ca. Odinarchaeum* LCB\_4 CRISPR-Cas system and mobile elements.** CRISPR-Cas  
68 systems in the *Ca. Odinarchaeum* LCB\_4 chromosome (center) were colored according to their type  
69 classification (orange: I-A; blue: III-D). Full contigs representing mobile elements are shown at the  
70 corners, with vertical lines representing contig boundaries. Viral terminal inverted repeats are  
71 represented by hourglass symbols. Connecting lines represent full-coverage spacer (35-42 nt) BlastN  
72 hits against mobile element targets (0 mismatches: full; 1 mismatch: dashed; 2 mismatches: dotted).  
73 TR=Transcriptional regulator; MCP=Major Capsid Protein; DJR=Double Jelly-Roll; TFE=Transcription  
74 Initiation Factor E; RE=Restriction enzyme; PolB=Family B DNA polymerase.

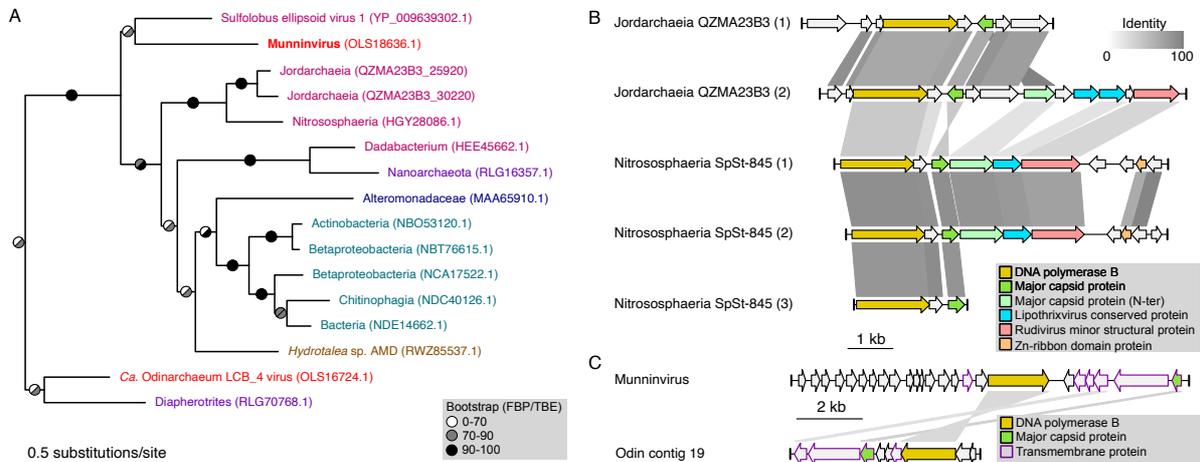
75

76 Furthermore, 3 spacers yielded 100% identity (and a further 5 spacers if 1-2 mismatches were  
77 permitted) and 100% query coverage matches against a 12.7 kb-long contig recovered by the  
78 new *Ca. Odinarchaeum* LCB\_4 assembly (Fig. 1). All 3 hits targeted an ORF encoding a  
79 primed family B DNA Polymerase (PolB), a gene frequently observed in archaeal viruses.  
80 Further inspection of this contig revealed genes encoding a zinc-ribbon protein and a His1-  
81 family MCP (Suppl. Fig S2B; Suppl. Table S1), conserved in spindle-shaped viruses<sup>10</sup>. This  
82 contig was flanked by ca. 80 nucleotides-long terminal inverted repeats, a typical signature of  
83 viruses with linear dsDNA genomes replicated by protein-primed PolBs<sup>11</sup>. Thus, this contig  
84 represents a complete Asgard archaeal viral genome for which we suggest the name  
85 Muninnvirus, in relation to Odin's raven, Muninn ("memory").

86 We further queried the PolB sequence from the Muninnvirus genome through phylogenetic  
87 analysis, finding that it is closely related to a homolog in a *Sulfolobus ellipsoid virus 1* (SEV1)<sup>12</sup>  
88 (Fig. 2A), recently isolated from a Costa Rican hot spring. No other genes were found to be  
89 shared between Muninnvirus and SEV1, indicative of recent horizontal transfer of *polB* in at  
90 least one of these viruses. Interestingly, other close homologs included multiple sequences  
91 that were likewise obtained from hot springs or hydrothermal vents (Fig. 2A). Two of these hits  
92 were part of an Asgard archaeal MAG (QZMA23B3), and a third one belonged to a MAG  
93 (SpSt-845) originally classified as Bathyarchaeota. A phylogenomic analysis that indicated  
94 that the QZMA23B3 belonged to the recently described Asgard archaeal class Jordarchaeia<sup>7</sup>,  
95 and that SpSt-845 in fact belonged to the Nitrososphaeria (Suppl. Fig. S3). Two additional  
96 PolB sequences from the same Nitrososphaerial MAG were found to be highly similar (>80%  
97 identity) to HGY28086.1 (and were thus not included in the phylogenetic analysis). The five  
98 PolB homologs were encoded in contigs containing SIRV2-family MCP genes (Fig. 2B; Suppl.  
99 Fig. S2C; Suppl. Table S1), exclusive to archaeal filamentous viruses with linear dsDNA  
100 genomes and classified into the realm Adnaviria<sup>13</sup>. Both the Jordarchaeia and the  
101 Nitrososphaeria contigs displayed high conservation in synteny and protein sequences,  
102 indicating high contig completeness and recent diversification (Fig. 2B). Notably, none of the  
103 known archaeal viruses with SIRV2-family MCP encodes its own PolB, suggesting that the

104 group identified herein represents a new archaeal virus family. However, while we detected  
 105 CRISPR arrays in the MAGs where these viral contigs were identified, we could not find  
 106 accurate spacer matches (query coverage > 90%, identity > 90%) to these viral sequences  
 107 and therefore the identity of the hosts of these thermophilic viruses remains unclear.

108



109

110 **Figure 2. Discovery of additional Asgard archaeal mobile elements.** (A) Phylogeny of DNA  
 111 polymerase B obtained with IQTree2 under the Q.pfam+C60+R4+F+PMSF model. Color: *Ca.*  
 112 *Odinarchaeum* LCB\_4 MAG (red); sequences obtained from hot springs (pink); hydrothermal vents  
 113 (purple); marine water (dark blue); Chatahoochee river (USA) (light blue); mine drainage (brown).  
 114 Branch support value are Felsenstein Bootstrap Proportions (left) and Transfer Bootstrap Expectation  
 115 (right). The tree presented here is a clade of the full tree shown in Suppl. Fig. S4. (B and C) Comparison  
 116 between the viral contigs (B) of *Jordarchaeia* QZMA23B3 and *Nitrososphaeria* SpSt-845, and (C) of  
 117 *Muninnvirus* and another viral contig in the *Ca. Odinarchaeum* LCB\_4 assembly. Similarity lines  
 118 represent BlastP hits with E-value lower than 1e-5, and percent identity as shown in the upper-right  
 119 legend.

120

121 The PolB phylogeny further suggests that a clade of viral sequences found in MAGs from  
 122 mesophiles evolved from a likely thermophile-infecting ancestor. While none of the mentioned  
 123 mobile elements share other proteins in common with *Muninnvirus*, a more distant relative of  
 124 the *Muninnvirus* PolB sequence was found in a contig from the same assembly. Like  
 125 *Muninnvirus*, this sequence encoded a His1-like MCP, and a gene encoding a transmembrane  
 126 protein of unknown function (Fig. 2C). The latter two genes surrounded another gene encoding  
 127 a relatively long protein (580 and 553 amino acid residues) with multiple transmembrane  
 128 helices and complex predicted structures (Suppl. Fig. S2D), with no detectable similarity, but  
 129 possibly related functions.

130 The CRISPR-Cas system of *Ca. Odinarchaeum yellowstonii* LCB\_4 is likely its primary  
 131 antiviral defense system. We could find no homologs for DISARM<sup>14</sup> or other recently  
 132 discovered antiviral systems<sup>15,16</sup> in its genome. The retention of numerous CRISPR spacers  
 133 against these mobile elements is significant and indicates coevolutionary dynamics with  
 134 viruses from multiple families.

135 Our findings highlight the benefits of improving the quality of Asgard archaeal genomes. The  
 136 discovery of viruses of thermophilic Asgard archaea opens the door to the study of the Asgard  
 137 archaeal mobilome, promising exciting new insights into the ecology, physiology and evolution  
 138 of the closest archaeal relatives of eukaryotes.

139

140

141

## 142 **ACKNOWLEDGEMENTS**

143 We thank Laura Wenzel for discussions on hybrid assemblies, and Raymond Staals and John  
144 van der Oost for helpful comments on CRISPR-Cas systems. This research was funded by  
145 the Swedish Research Council (International Postdoc grant 2018-00669 to DT), the European  
146 Research Council (ERC consolidator grant 817834 to TJGE), the Swedish Foundation for  
147 Strategic Research (SSF-FFL5 to TJGE) and a Wellcome Trust collaborative award  
148 (203276/Z/16/Z to TJGE). NR was supported by start-up funds from BLS, Lancaster University  
149 and a Leverhulme Research Project Grant (RPG-2019-297). MK was supported by the  
150 Agence Nationale de la Recherche (ANR-20-CE20-0009-02) and Ville de Paris  
151 (Emergence(s) project MEMREMA). LE received funding from the European Research  
152 Council (ERC Starting Grant 803151).

153

## 154 **DATA AVAILABILITY**

155 Raw Nanopore amplicon reads and complete *Ca. Odinarchaeum* LCB\_4 assembly are  
156 available at NCBI under Bioproject PRJNAXXXXXXX. Additional data and supporting  
157 alignments and trees can be found in Zenodo, under the link: XXXXXXXXX.

158

## 159 **METHODS**

160

### 161 ***Odinarchaeum* LCB\_4 genome reassembly**

162

163 To reassemble the *Odinarchaeum* LCB\_4 genome (Suppl. Fig. S1A), its corresponding  
164 Illumina reads<sup>17</sup> (BioSample SAMN04386028) were mapped against Asgard MAGs<sup>4</sup> using  
165 Minimap2<sup>18</sup> v2.2.17. Mapped reads were extracted and assembled with Unicycler<sup>19</sup> v0.4.4.  
166 This assembly obtained a 1.406 Mb contig, which was not predicted as circular despite both  
167 of its contig boundaries ending in CRISPR arrays (Suppl. Fig. S1B). Additional short (< 13 kb)  
168 contigs were not considered part of the main chromosome if they represented mobile elements  
169 (with signatures such as differing coverage, circularity, CRISPR spacer hits and/or presence  
170 of typical mobile element genes), rRNA genes from different organisms, or CRISPR arrays.  
171 The latter two were expected due to the conservation of sequences such as rRNA gene  
172 sequences and CRISPR repeats. After removing these contigs, only one additional contig of  
173 10.6 kb containing type I-A *cas* genes remained. Given that the 1.406 Mb contig ended in type  
174 I-A CRISPR arrays, we hypothesized that these two contigs could represent the entire circular  
175 chromosome of *Ca. Odinarchaeum* LCB\_4. In parallel, we assembled the Illumina reads with  
176 MEGAHIT<sup>20</sup> v.1.1.3 with options “--k-min 57 --k-max 147 --k-step 12”. While highly  
177 fractionated, this assembly found an alternative solution for the sequences involved in the  
178 contig borders of the previous assembly (Suppl. Fig. S1B). In this case, the type-I-A *cas* genes  
179 were surrounded by two separate CRISPR arrays. Moreover, four consecutive spacers in the  
180 innermost side of one of the CRISPR arrays in this assembly were identical to the outermost  
181 spacers of the CRISPR array present at the border of the 1.406 Mb contig in the Unicycler  
182 assembly (Suppl. Fig. S1B). These results suggested a specific disposition for the two  
183 aforementioned contigs.

184

### 185 **Long-range PCR and Nanopore sequencing**

186

187 Four regions were selected for long-range Polymerase Chain Reaction (lrPCR): two contig  
188 gaps, corresponding to CRISPR arrays, and two control regions spanning ca. 5 kb of the rRNA

189 operon and ca. 10 kb of a ribosomal protein gene cluster (Suppl. Table S2). Primers were  
190 designed using the Sigma-Aldrich OligoEvaluator™  
191 (<http://www.oligoevaluator.com/OligoCalcServlet>) and synthesized by Integrated DNA  
192 Technologies, Inc. MDA-amplified environmental DNA isolated from the Lower Culex Basin at  
193 Yellowstone National Park (USA)<sup>17</sup> was then amplified with Herculase polymerase (Agilent).  
194 Amplification of control and gap regions was then performed following the parameters shown  
195 in Suppl. Table S2. Products were separated on a 0.8% agarose gel in 1XTBE stained with  
196 SYBR-Gold, and purified using a Qiagen Spin purification kit following the manufacturers  
197 instructions. Purified PCR fragments were pooled and used to construct a library with the  
198 ligation kit SQK-LSK109. Sequencing was performed on an Oxford Nanopore MinION Mk1C  
199 sequencer using an R9.4.1 flow cell. Raw sequence data was basecalled using Guppy v4.2.2  
200 High Accuracy. Reads were separated in two bins at 3-9 kb (subsampling to 30X) and 9-12  
201 kb, and processed to obtain consensus sequences using Decona ([https://github.com/Saskia-  
202 Oosterbroek/decona](https://github.com/Saskia-Oosterbroek/decona)) v0.1.2 (-c 0.85 -w 6 -i -n 25 -M -r). Both control regions, comprising the  
203 rRNA and ribosomal protein operons, were 100% identical to the corresponding nucleotide  
204 sequences of the published assembly.

205

### 206 **Hybrid assembly**

207

208 Reads were filtered using NanoFilt v.2.6.0 with options “-q 10 -l 1000”. We used these filtered  
209 Nanopore reads and the mapped Illumina reads to perform a hybrid assembly with Unicycler  
210 0.4.4, which resolved both the main chromosomal contig and a viral contig (Huginnvirus) as  
211 circular (Suppl. Fig S1DE). Read mapping was performed using bowtie2<sup>21</sup> v.2.3.5.1 for  
212 Illumina reads and minimap2<sup>18</sup> v.2.17.r941 for Nanopore reads. A local cumulative GC skew  
213 minimum (Suppl. Fig S1E), together with low R-Y (purine minus pyrimidine), M-K (amino minus  
214 keto) and cumulative AT skew values, was selected as a potential replication origin and the  
215 circular contig was permuted to set this position as nucleotide +1.

216

### 217 **Annotation**

218

219 CRISPR arrays were detected and classified using CRISPRdetect<sup>22</sup> v2.4 online, and *cas*  
220 genes were detected and classified through CRISPRcasIdentifier<sup>23</sup> v1.1.0. Proteins were  
221 classified into COG families<sup>24</sup> based on 5 best local Blastp<sup>25</sup> hits to the same COG, and domain  
222 annotation was performed through InterProScan<sup>26</sup> v5.48-83.0. Mobile element protein  
223 annotation was performed using HHsearch<sup>27</sup> against Pfam<sup>28</sup>, PDB<sup>29</sup>, SCOPe<sup>30</sup>, CDD<sup>31</sup> and  
224 Uniprot<sup>32</sup> viral protein sequence databases. Synteny plots were performed with genoPlotR<sup>33</sup>.  
225 Structural predictions were performed with RoseTTAFold<sup>34</sup> through the Robetta portal.

226

### 227 **Phylogenetics**

228

229 Reference DNA polymerase B sequences were obtained from Kim et al.<sup>35</sup> and used for  
230 Psiblast<sup>36</sup> v2.10.0+ against the NR database and a group of Asgard archaeal MAGs from Eme  
231 et al.<sup>4</sup>. Sequences with over 70% similarity were removed with CD-Hit<sup>37</sup> v4.7. The remaining  
232 sequences were aligned with Mafft-linsi<sup>38</sup> v7.450 and columns with over 50% gaps were  
233 removed using trimAl<sup>39</sup> v1.4.rev22. Additionally, sequences with over 50% gaps in the trimmed  
234 alignment were removed. Maximum-likelihood trees were reconstructed using IQ-TREE<sup>40</sup>  
235 v2.0-rc1 using ModelFinder<sup>41</sup> with all combinations of the empirical models LG, JTT, WAG and  
236 Q.pfam with site-class mixtures (none, C20, C40, C60), rate heterogeneity (none, G4 and R4)

237 and frequency (none, F) parameters. Using the previous tree as a guide, a PMSF<sup>42</sup>  
238 approximation of the selected model was used to reconstruct a tree with 100 non-parametric  
239 bootstrap pseudo-replicates, which was then interpreted both as the standard Felsenstein  
240 bootstrap proportions and as transfer bootstrap expectation<sup>43</sup>.

241 To assess the taxonomy of selected MAGs, all archaeal GTDB<sup>44</sup> representative sequences  
242 (as of August 2021) were retrieved and supplemented with Asgard archaeal sequences from  
243 the Hermod<sup>45</sup>, Sif<sup>5</sup>, Wukong<sup>6</sup> and Jord<sup>7</sup> groups. Together with query sequences, GToTree<sup>46</sup>  
244 v1.5.45 was then used to reconstruct a tree with parameters “-H Archaea -D -G 0.2”.

245

246

## 247 SUPPLEMENTARY MATERIAL

248

249 **Supplementary Table S1. Annotation of mobile elements.** Yellow cells represent  
250 annotation of key viral proteins.

251 **Supplementary Table S2. Methods: primers and IrPCR cycling parameters**

252 **Supplementary Figure S1. Obtaining a closed *Ca. Odinarchaeum* LCB\_4 genome.** (A)  
253 Summary methodology for the reassembly, refinement and closing of the *Ca. Odinarchaeum*  
254 LCB\_4 genome. (B) Schematic of the assembly status before long-range PCR, indicating the  
255 presence of gaps and the agreement between two separate assemblies, which guided primer  
256 design. (C) Purified PCR products; lane 1: DNA ladder, 2: Positive control ca. 5 kb rRNA gene  
257 cluster; 3: Positive control ca. 10 kb ribosomal protein gene cluster; 4-5: first gap closing, at  
258 distances of ca. 5 and 5.5 kb; 6-8: second gap closing, at distances of ca. 4, 4.5 and 5 kb. (D)  
259 Genome map of *Ca. Odinarchaeum* LCB\_4, including (from inside out): GC skew (line) and  
260 cumulative GC skew (histogram); GC content; Crick strand genes; Watson strand genes;  
261 Nanopore reads coverage capped at 1500X; Illumina read coverage (light: proper pairs,  
262 NM<3) capped at 50X; repeats; chromosome. (E) Comparison between previous assembly  
263 and new assembly for Huginnvirus, indicating circularity. Similarity lines represent two single  
264 BlastN hits with up to 1 mismatches. (F) Genomic patterns of the *Ca. Odinarchaeum* LCB\_4  
265 indicating a potential origin of replication at position 959350.

266 **Supplementary Figure S2. Predicted structure of selected proteins.** Comparisons  
267 between the structures of (A) DJR-MCPs (left: Huginnvirus: OLS18934.1; right: *Sulfolobus*  
268 turreted icosahedral virus 1: 3J31); (B) His1-like MCPs (left: Muninnvirus: OLS18630.1; right:  
269 His1 virus: YP\_529533.1); (E) SIRV2-like MCPs (left: *Jordarchaeia* QZMA23B3:  
270 QZMA23B3\_25900; right: *Sulfolobus islandicus* rod-shaped virus 2 (SIRV-2): 3J9X) and (F)  
271 transmembrane proteins (left: Muninnvirus: OLS18631.1; right: *Ca. Odinarchaeum* LCB\_4  
272 virus: OLS16720). All structures predicted with RoseTTAFold are color-coded according to  
273 their error estimate (Å). (C,D) Given the high error estimates for the predicted structures of  
274 His1-like MCPs (B), we append HHsearch results for OLS18630.1 (Muninnvirus) and  
275 OLS18934.1 (*Ca. Odinarchaeum* LCB\_4 MAG), the latter of which shows a tandem  
276 duplication (Regions 1 and 2) of the His1-like MCP. H(h),  $\alpha$ -helix; E(e),  $\beta$ -strand; C(c), coil.

277 **Supplementary Figure S3. Taxonomic placement of archaeal MAGs.** Phylogenomic tree  
278 obtained with FastTree including two archaeal MAGs (arrows) containing viral contigs within  
279 all current GTDB Archaea representatives (see methods). Branch colors represent  
280 Nitrososphaeria (red), Asgard archaea (light blue), and *Jordarchaeia* (purple). All placements  
281 are supported with branch support values of 1.0. Full tree can be found in data repository.

282 **Supplementary Figure S4. PoIB phylogeny.** Midpoint-rooted full tree corresponding to Fig.  
283 2A. Taxon labels shown in Fig 2A are colored with the same pattern, and their corresponding  
284 branches are colored red. Support values are transfer bootstrap expectation (left) and  
285 Felsenstein bootstrap proportions (right).

286 **REFERENCES**

- 287 1 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic  
288 cellular complexity. *Nature* **541**, 353-358, doi:10.1038/nature21031 (2017).
- 289 2 Williams, T. A., Cox, C. J., Foster, P. G., Szollosi, G. J. & Embley, T. M.  
290 Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol*  
291 **4**, 138-147, doi:10.1038/s41559-019-1040-x (2020).
- 292 3 Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the  
293 origin of eukaryotes. *Nat Rev Microbiol* **15**, 711-723, doi:10.1038/nrmicro.2017.133  
294 (2017).
- 295 4 Eme, L. *et al.* Inference and reconstruction of the Skadiarchaeal ancestry of  
296 eukaryotes. *Manuscript under revision* (2021).
- 297 5 Farag, I. F., Zhao, R. & Biddle, J. F. "Sifarchaeota," a Novel Asgard Phylum from  
298 Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic  
299 Methylophony. *Appl Environ Microbiol* **87**, doi:10.1128/AEM.02584-20 (2021).
- 300 6 Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with  
301 eukaryotes. *Nature* **593**, 553–557 (2021).
- 302 7 Sun, J. E., P.N.; Gagen, E.J.; Woodcroft, B.J.; Hedlund, B.P.; Woyke, T.; Hugenholtz,  
303 P.; Rinke, C. . Recoding of stop codons expands the metabolic potential of two novel  
304 Asgardarchaeota lineages. *ISME Communications* **1**, 30 (2021).
- 305 8 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and  
306 eukaryotes. *Nature* **521**, 173-179, doi:10.1038/nature14447 (2015).
- 307 9 Yutin, N., Backstrom, D., Ettema, T. J. G., Krupovic, M. & Koonin, E. V. Vast  
308 diversity of prokaryotic virus genomes encoding double jelly-roll major capsid  
309 proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**, 67,  
310 doi:10.1186/s12985-018-0974-y (2018).
- 311 10 Krupovic, M., Quemin, E. R., Bamford, D. H., Forterre, P. & Prangishvili, D.  
312 Unification of the globally distributed spindle-shaped viruses of the Archaea. *J Virol*  
313 **88**, 2354-2358, doi:10.1128/JVI.02941-13 (2014).
- 314 11 Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V.  
315 Viruses of archaea: Structural, functional, environmental and evolutionary genomics.  
316 *Virus Res* **244**, 181-193, doi:10.1016/j.virusres.2017.11.025 (2018).
- 317 12 Wang, H. *et al.* Novel Sulfolobus Virus with an Exceptional Capsid Architecture. *J*  
318 *Virol* **92**, doi:10.1128/JVI.01727-17 (2018).
- 319 13 Krupovic, M. *et al.* Adnaviria: a New Realm for Archaeal Filamentous Viruses with  
320 Linear A-Form Double-Stranded DNA Genomes. *J Virol* **95**, e0067321,  
321 doi:10.1128/JVI.00673-21 (2021).
- 322 14 Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-  
323 phage activities. *Nat Microbiol* **3**, 90-98, doi:10.1038/s41564-017-0051-0 (2018).
- 324 15 Bernheim, A. *et al.* Prokaryotic viperins produce diverse antiviral molecules. *Nature*  
325 **589**, 120-124, doi:10.1038/s41586-020-2762-2 (2021).
- 326 16 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial  
327 pangenome. *Science* **359**, doi:10.1126/science.aar4120 (2018).
- 328 17 Baker, B. J. *et al.* Genomic inference of the metabolism of cosmopolitan subsurface  
329 Archaea, Hadesarchaea. *Nat Microbiol* **1**, 16002, doi:10.1038/nrmicrobiol.2016.2  
330 (2016).
- 331 18 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,  
332 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
- 333 19 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial  
334 genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**,  
335 e1005595, doi:10.1371/journal.pcbi.1005595 (2017).

- 336 20 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by  
337 advanced methodologies and community practices. *Methods* **102**, 3-11,  
338 doi:10.1016/j.ymeth.2016.02.020 (2016).
- 339 21 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*  
340 *Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 341 22 Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C. & Brown, C. M.  
342 CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**,  
343 356, doi:10.1186/s12864-016-2627-0 (2016).
- 344 23 Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. & Backofen, R.  
345 CRISPRcasIdentifier: Machine learning for accurate identification and classification  
346 of CRISPR-Cas systems. *Gigascience* **9**, doi:10.1093/gigascience/giaa062 (2020).
- 347 24 Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial  
348 genome coverage and improved protein family annotation in the COG database.  
349 *Nucleic Acids Res* **43**, D261-269, doi:10.1093/nar/gku1223 (2015).
- 350 25 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**,  
351 421, doi:10.1186/1471-2105-10-421 (2009).
- 352 26 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.  
353 *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).
- 354 27 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein  
355 annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).
- 356 28 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**,  
357 D412-D419, doi:10.1093/nar/gkaa913 (2021).
- 358 29 Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D  
359 structures of biological macromolecules for basic and applied research and education  
360 in fundamental biology, biomedicine, biotechnology, bioengineering and energy  
361 sciences. *Nucleic Acids Res* **49**, D437-D451, doi:10.1093/nar/gkaa1038 (2021).
- 362 30 Chandonia, J. M., Fox, N. K. & Brenner, S. E. SCOPe: classification of large  
363 macromolecular structures in the structural classification of proteins-extended  
364 database. *Nucleic Acids Res* **47**, D475-D481, doi:10.1093/nar/gky1134 (2019).
- 365 31 Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids*  
366 *Res* **48**, D265-D268, doi:10.1093/nar/gkz991 (2020).
- 367 32 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*  
368 **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).
- 369 33 Guy, L., Kultima, J. R. & Andersson, S. G. genoPlotR: comparative gene and genome  
370 visualization in R. *Bioinformatics* **26**, 2334-2335, doi:10.1093/bioinformatics/btq413  
371 (2010).
- 372 34 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a  
373 three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754  
374 (2021).
- 375 35 Kim, J. G. *et al.* Spindle-shaped viruses infect marine ammonia-oxidizing  
376 thaumarchaea. *Proc Natl Acad Sci U S A* **116**, 15645-15650,  
377 doi:10.1073/pnas.1905682116 (2019).
- 378 36 Schaffer, A. A. *et al.* Improving the accuracy of PSI-BLAST protein database  
379 searches with composition-based statistics and other refinements. *Nucleic Acids Res*  
380 **29**, 2994-3005, doi:10.1093/nar/29.14.2994 (2001).
- 381 37 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-  
382 generation sequencing data. *Bioinformatics* **28**, 3150-3152,  
383 doi:10.1093/bioinformatics/bts565 (2012).

- 384 38 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version  
385 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780,  
386 doi:10.1093/molbev/mst010 (2013).
- 387 39 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for  
388 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*  
389 **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).
- 390 40 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic  
391 Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534,  
392 doi:10.1093/molbev/msaa015 (2020).
- 393 41 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermin, L. S.  
394 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*  
395 **14**, 587-589, doi:10.1038/nmeth.4285 (2017).
- 396 42 Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity  
397 with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic  
398 Estimation. *Syst Biol* **67**, 216-235, doi:10.1093/sysbio/syx068 (2018).
- 399 43 Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big  
400 data. *Nature* **556**, 452-456, doi:10.1038/s41586-018-0043-0 (2018).
- 401 44 Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea.  
402 *Nat Biotechnol* **38**, 1079-1086, doi:10.1038/s41587-020-0501-8 (2020).
- 403 45 Zhang, J. W. *et al.* Newly discovered Asgard archaea Hermodarchaeota potentially  
404 degrade alkanes and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA  
405 pathway. *ISME J* **15**, 1826-1843, doi:10.1038/s41396-020-00890-x (2021).
- 406 46 Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**,  
407 4162-4164, doi:10.1093/bioinformatics/btz188 (2019).  
408