

# Characterization of Mollivirus kamchatka , the first modern representative of the proposed Molliviridae family of giant viruses.

Eugene Christo-Foroux, Jean-Marie Alempic, Audrey Lartigue, Sébastien Santini, Karine Labadie, Matthieu Legendre, Chantal Abergel, Jean-Michel Claverie

# ▶ To cite this version:

Eugene Christo-Foroux, Jean-Marie Alempic, Audrey Lartigue, Sébastien Santini, Karine Labadie, et al.. Characterization of Mollivirus kamchatka , the first modern representative of the proposed Molliviridae family of giant viruses.. Journal of Virology, 2020, 10.1101/844274 . hal-02464533

# HAL Id: hal-02464533 https://amu.hal.science/hal-02464533

Submitted on 3 Feb 2020

**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	Characterization of Mollivirus kamchatka, the first modern representative
3	of the proposed Molliviridae family of giant viruses.
4	
5	Running title: characterization of the first modern mollivirus
6	
7	
8	Eugene Christo-Foroux <sup>#</sup> , Jean-Marie Alempic <sup>a</sup> , Audrey Lartigue <sup>a</sup> , Sebastien Santini <sup>a</sup> ,
9	Karine Labadie <sup>b</sup> , Matthieu Legendre <sup>a</sup> , Chantal Abergel <sup>a</sup> , Jean-Michel Claverie <sup>#a</sup>
10	
11	
12	a Aix Marseille Univ., CNRS, IGS, Information Génomique & Structurale (UMR7256),
13	Institut de Microbiologie de la Méditerranée (FR 3489), Marseille, France
14	
15	b Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Évry, France
16	
17	# Correspondance to: Eugene.christo-foroux@igs.cnrs-mrs.fr, jean-michel.claverie@univ-amu.fr
18	
19	
20	Keywords: Paleovirology; Kamchatka; Comparative Genomics; Nucleo-cytoplasmic Large DNA
21	viruses; Genome Structure.
22	
23	

### 24 Abstract

25 Microbes trapped in permanently frozen paleosoils (permafrost) are the focus of increasing 26 researches in the context of global warming. Our previous investigations led to the discovery and 27 reactivation of two Acanthamoeba-infecting giant viruses, Mollivirus sibericum and Pithovirus 28 sibericum from a 30,000-year old permafrost layer. While several modern pithovirus strains have 29 since been isolated, no contemporary mollivirus relative was found. We now describe Mollivirus 30 kamchatka, a close relative to M. sibericum, isolated from surface soil sampled on the bank of the 31 Kronotsky river in Kamchatka. This discovery confirms that molliviruses have not gone extinct and 32 are at least present in a distant subarctic continental location. This modern isolate exhibits a 33 nucleo-cytoplasmic replication cycle identical to that of *M. sibericum*. Its spherical particle (0.6-µm 34 in diameter) encloses a 648-kb GC-rich double stranded DNA genome coding for 480 proteins of 35 which 61 % are unique to these two molliviruses. The 461 homologous proteins are highly 36 conserved (92 % identical residues in average) despite the presumed stasis of M. sibericum for the 37 last 30,000 years. Selection pressure analyses show that most of these proteins contribute to the 38 virus fitness. The comparison of these first two molliviruses clarify their evolutionary relationship 39 with the pandoraviruses, supporting their provisional classification in a distinct family, the 40 Molliviridae, pending the eventual discovery of intermediary missing links better demonstrating 41 their common ancestry.

43

# 44 Importance

45	Virology has long been viewed through the prism of human, cattle or plant diseases leading to a
46	largely incomplete picture of the viral world. The serendipitous discovery of the first giant virus
47	visible under light microscopy (i.e., >0.3 $\mu$ m in diameter), mimivirus, opened a new era of
48	environmental virology, now incorporating protozoan-infecting viruses. Planet-wide isolation
49	studies and metagenomes analyses have shown the presence of giant viruses in most terrestrial
50	and aquatic environments including upper Pleistocene frozen soils. Those systematic surveys have
51	led authors to propose several new distinct families, including the Mimiviridae, Marseilleviridae,
52	Faustoviridae, Pandoraviridae, and Pithoviridae. We now propose to introduce one additional
53	family, the <i>Molliviridae</i> , following the description of <i>M. kamchatka</i> , the first modern relative of <i>M</i> .
54	sibericum, previously isolated from 30,000-year old arctic permafrost.

56

## 57 Introduction

58 Studies started about 30 years ago, have provided multiple evidence that soils frozen since the late 59 Pleistocene and predominantly located in arctic and subarctic Siberia, do contain a wide diversity 60 of microbes that can be revived upon thawing (1-3) after tens of thousands of years. These studies 61 culminated by the regeneration of a plant from 30,000 year-old fruit tissue (4). Inspired by those 62 studies, we then isolated from a similar sample two different Acanthamoeba-infecting large DNA 63 viruses, named Pithovirus sibericum and Mollivirus sibericum, demonstrating the ability for these 64 viruses, and maybe many others, to remain infectious after similarly long periods of stasis in 65 permafrost (5-6). Several modern relatives to the prototype Pithovirus have since been isolated 66 and characterized, leading to the emergence of the new proposed Pithoviridae family (7-10). In 67 contrast, and despite the increasing sampling efforts deployed by several laboratories, no other 68 relative of *M. sibericum* was found. Without additional isolates, its classification as a prototype of a 69 new family, or as a distant relative of the pandoraviruses (with which it shared several 70 morphological features and 16% of its gene content) (6) remained an open question. Here we 71 report the discovery and detailed characterization of the first modern *M. sibericum* relative, 72 named *Mollivirus kamchatka*, after the location of the Kronotsky river bank where it was retrieved. 73 The comparative analysis of these first two molliviruses highlights their evolutionary processes and 74 suggests their provisional classification into their own family, the Molliviridae, distinct from the 75 Pandoraviridae, pending the eventual discovery of intermediary missing links clearly establishing 76 their common ancestry.

77

#### 79 Results

#### 80 Virus isolation

81	The original sample consisted of about 50 ml of vegetation-free superficial soil scooped (in sterile
82	tubes) from the bank of the Kronotsky river (coordinates: N 54°32′59′′ E 160°34′55′′) on july 6 <sup>th</sup> ,
83	2017. Before being stored at 4°C, the sample was transported in a backpack for a week at ambient
84	temperature (5° C up to 24° C). This area corresponds to a continental subarctic climate: very cold
85	winters, and short, cool to mild summers, low humidity and little precipitation. Back in the
86	laboratory, few grams of the sample were used in the Acanthamoeba co-cultivation procedure
87	previously described (6). After a succession of passages and enrichment on Acanthamoeba
88	castellanii cultures, viral particles were produced in sufficient quantity to be recovered and
89	purified.
90	Virion morphology and ultrastructure
90 91	<b>Virion morphology and ultrastructure</b> As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles.
90 91 92	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M</i> .
<ol> <li>90</li> <li>91</li> <li>92</li> <li>93</li> </ol>	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M.</i> <i>sibericum</i> - appear approximately spherical, 600 nm in diameter, lined by an internal lipid
<ol> <li>90</li> <li>91</li> <li>92</li> <li>93</li> <li>94</li> </ol>	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M</i> . <i>sibericum</i> - appear approximately spherical, 600 nm in diameter, lined by an internal lipid membrane and enclosed in a 20 nm-thick electron-dense thick tegument covered with a mesh of
<ol> <li>90</li> <li>91</li> <li>92</li> <li>93</li> <li>94</li> <li>95</li> </ol>	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M</i> . <i>sibericum</i> - appear approximately spherical, 600 nm in diameter, lined by an internal lipid membrane and enclosed in a 20 nm-thick electron-dense thick tegument covered with a mesh of fibers (Fig. 1).
<ol> <li>90</li> <li>91</li> <li>92</li> <li>93</li> <li>94</li> <li>95</li> <li>96</li> </ol>	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M</i> . <i>sibericum</i> - appear approximately spherical, 600 nm in diameter, lined by an internal lipid membrane and enclosed in a 20 nm-thick electron-dense thick tegument covered with a mesh of fibers (Fig. 1). Analysis of the replication cycle
<ol> <li>90</li> <li>91</li> <li>92</li> <li>93</li> <li>94</li> <li>95</li> <li>96</li> <li>97</li> </ol>	Virion morphology and ultrastructure As for <i>M. sibericum</i> , light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of <i>M</i> . <i>sibericum</i> - appear approximately spherical, 600 nm in diameter, lined by an internal lipid membrane and enclosed in a 20 nm-thick electron-dense thick tegument covered with a mesh of fibers (Fig. 1). Analysis of the replication cycle The replication cycle in <i>A. castellanii</i> cells was monitored using TEM and light microscopy of DAPI-

99 host cells infected by *M. sibericum*, was similarly observed upon *M. kamchatka* replication (6, 11).

100 After entering the amoeba cell through phagocytosis, *M. Kamchatka* virions are found gathered in

101 large vacuoles individually or in groups of 2-6 particles. Multiple nuclear events occur during the

102 infection starting with the drift of the host cell nucleolus to the periphery of the nucleus 4-5h post

103 infection (p.i.). 7h p.i., the nucleus appears filled with numerous fibrils that may correspond to viral

genomes tightly packed in DNA-protein complexes (Fig. 2A). About 30% of the nuclei observed at
that time exhibit a ruptured nuclear membrane (Fig. 2B). Besides those internal nuclear events, we
observed a loss of vacuolization within the host cell from 4h p.i. to the end of the cycle, in average
9h p.i.. Large viral factories are formed in the cytoplasm at the periphery of the disorganized
nucleus (Fig. 2). These viral factories displays the same characteristics as those formed during *M*. *sibericum* infections, involving an active recycling of membrane fragments (11).

110 Comparative genomics

111 DNA prepared from purified *M. kamchatka* particles was sequenced using both Illumina and 112 Oxford Nanopore platforms. *M. kamchatka* genome, a linear double stranded DNA molecule 113 (dsDNA), was readily assembled as a unique sequence of 648,864 bp. The read coverage was 114 uniform throughout the entire genome except for a 10 kb terminal segment presumably repeated 115 at both ends and exhibiting twice the average value. The *M. kamchatka* genome is thus 116 topologically identical to that of a *M. sibericum*, slightly larger in size (when including both terminal 117 repeats), and with the same global nucleotide composition (G+C= 60%). This similarity was 118 confirmed by the detailed comparison of their genome sequences exhibiting a global collinearity 119 solely interrupted by a few insertions and deletions.

120 Prior to the comparison of their gene contents, *M. sibericum* and *M. kamchatka* were both 121 annotated using the same stringent procedure that we previously developed to correct for gene 122 overpredictions suspected to occur in G+C rich sequences such as those of pandoraviruses (12-14). 123 A total of 495 and 480 genes were predicted for *M. sibericum* and *M. kamchatka*, with the 124 encoded proteins ranging from 51 to 2171 residues and from 57 to 2176 residues, respectively. 125 *M. kamchatka* predicted protein sequences were used in similarity search against the non-126 redundant protein sequence database (15) and the re-annotated *M. sibericum* predicted 127 proteome. Out of the 480 proteins predicted to be encoded by *M. kamchatka* 463 had their closest 128 homologs in *M. sibericum*, with 92% identical residues in average. After clustering the paralogs,

129 these proteins corresponded to 434 distinct genes clusters delineating a first estimate of the 130 mollivirus core gene set. Four hundred and eleven of these clusters contained a single copy (singletons) gene for each strain. Remarkably, 290 of the 480 (60.4 %) M. kamchatka-encoded 131 132 proteins did not exhibit a detectable homolog among cellular organisms or previously sequenced 133 viruses (excluding *M. sibericum*). Those will be referred to as "ORFans". Among the 190 proteins 134 exhibiting significant ( $E<10^{-5}$ ) matches in addition to their *M. sibericum* counterparts, 78 (16% of 135 the total gene content) were most similar to Pandoravirus predicted proteins, 18 (3.7%) to proteins 136 of other virus families, 51 (10.6%) to A. castellanii proteins, 24 (5%) to proteins of other 137 eukaryotes, 17 (3.5%) to bacterial proteins and 2 (0.4%) to proteins of Archaea (Fig. 3). 138 The interpretation of these statistics are ambiguous as, on one hand, the large proportion of 139 "ORFans" (>60%) is characteristic of what is usually found for the prototypes of novel giant virus 140 families (7). On the other hand, the closest viral homologs are not scattered in diverse previously 141 defined virus families, but mostly belongs to the Pandoraviridae (78/96=81%) (Fig. 3). The two 142 molliviruses thus constitute a new group of viruses with their own specificity but with a 143 phylogenetic affinity with the pandoraviruses, as previously noticed (7). The proportion of M. 144 kamchatka proteins with best matching counterparts in Acanthamoeba confirms the high gene 145 exchange propensity with the host, already noticed for *M. sibericum* (6). 146 Recent evolutionary events since the *M. sibericum/M. kamchatka* divergence 147 We investigated the evolutionary events specific of each of the molliviruses by focusing on proteins

148 lacking reciprocal best matches between the two strains. We found 63 such cases of which 10

149 corresponded to unilateral strain-specific duplications of genes, and 53 were unique to a given

150 strain. These unique genes (Table 1 & 2) result from gains or losses in either of mollivirus strains

151 (20 in *M. Kamchatka*, 33 in *M. sibericum*). The likely origins of these strain-specific genes

- 152 (horizontal acquisition, *de novo* creation (13, 14), or differential loss) are listed in Table 1 and Table
- 153 **2**.

154 Six M. kamchatka proteins, absent from M. sibericum, have homologs in pandoraviruses suggesting common gene ancestors (and loss in *M. sibericum*) or horizontal acquisitions. According 155 156 to its embedded position within the pandoravirus phylogenetic tree, only one anonymous protein 157 (mk 165) could be interpreted as a probable horizontal transfer from pandoraviruses (Fig. 4A). 158 Another candidate, mk 92, shares 75% of identical residues with a pandoravirus DNA 159 methyltransferase (pger cds 559). However the very long branch associated to the P. dulcis 160 homolog (eventually due to a non-orthologous replacement) raises some doubt as for the origin of 161 the *M. kamchatka* gene (Fig. 4B). 162 Two *M. kamchatka*-specific proteins, encoded by adjacent genes (mk 466, mk 467), have 163 homologs in Acanthamoeba, suggesting potential host-virus exchanges. Phylogenetic 164 reconstruction did not suggest a direction for the transfer of mk 466. However, since the unique 165 homolog of mk 467 is found in Acanthamoeba (and not in other eukaryotes), the corresponding 166 gene probably originated from a close relative of *M. kamchatka* and was recently transferred to its 167 host. 168 Three proteins unique to *M. sibericum* have homologs in pandoraviruses (Table 2), suggesting 169 common gene ancestors (and loss in *M. kamchatka*) or horizontal acquisitions. One protein 170 (ms 14) has a unique homolog in Acanthamoeba, suggesting a virus to host exchange. The 171 homolog of ms\_312 in Cavenderia fasciculata (16) might be the testimony of past interactions 172 between molliviruses and deeply rooted ancestor of the Amoebozoa clade. 173 The above analyses of the genes unique to each mollivirus indicate that if horizontal transfer may contribute to their presence, it is not the predominant mechanisms for their acquisition. We 174 175 then further investigated the 12 genes unique to *M. kamchatka* and 26 genes unique to *M.* 176 sibericum (i.e. "strain ORFans" w/o homolog in the databases) by computing three independent 177 sequence properties (Fig. 5): the codon adaptation usage index (CAI), the G+C composition, and 178 the ORF length.

179	With an average CAI value of 0.26, the strain-specific ORFans appear significantly different
180	from the rest of the mollivirus genes (mean = 0.40, Wilcoxon test p< 2 $10^{-7}$ ). These genes also
181	exhibit a significantly lower G+C content (56% for <i>M. kamchatka</i> and 57% for <i>M. sibericum</i> ) than
182	the rest of the genes (60.5% for both viruses), also closer to the value computed for intergenic
183	regions (54% in average for both viruses). Moreover, the strain-specific ORFans are smaller in
184	average compared to the rest of the genes (115bp/378bp for <i>M. kamchatka</i> and 122bp/369bp for
185	<i>M. sibericum</i> ). Altogether, those results suggest that <i>de novo</i> gene creation might occur in the
186	intergenic regions of molliviruses as already postulated for pandoraviruses (13, 14).
187	New predicted protein functions in <i>M. kamchatka</i>
188	Sixty four of the <i>M. kamchatka</i> predicted proteins exhibit sequence motifs associated to known
189	functions. Fifty nine of them are orthologous to previously annotated genes in <i>M. sibericum</i> (6).
190	This common subset confirms the limited complement of DNA processing and repair enzymes
191	found in molliviruses: mainly a DNA polymerase: mk_287, a primase: mk_236, and 3 helicases:
192	mk_291, mk_293, mk_351. <i>M. kamchatka</i> confirms the absence of key deoxynucleotide synthesis
193	pathways (such as thymidylate synthase, thymidine kinase and thymidylate kinase), and of a
194	ribonucleoside-diphosphate reductase (present in pandoraviruses). The five <i>M. kamchatka</i> specific
195	proteins (Table 1) associated to functional motifs or domain signature correspond to:
196	- two proteins (mk_93 and mk_104) containing a type of zinc finger (Ring domain) mediating
197	protein interactions,
198	- one protein (mk_469) with similarity to the (BI)-1 like family of small transmembrane proteins,
199	- one predicted LexA-related signal peptidase (mk_166),
200	- one DNA methyltransferase (mk_92).
201	Evaluation of the selection pressure exerted on mollivirus genes

202 The availability of two distinct strains of mollivirus allows the first estimation of the selection

203 pressure exerted on their shared genes during their evolution. This was done by computing the

ratio  $\omega$ =dN/dS of the rate of non-synonymous mutations (dN) over the rate of synonymous mutations (dS) for pairs of orthologous genes.  $\omega$  values much lesser than one are associated to genes the mutation of which have the strongest negative impact on the virus fitness. The high sequence similarity of proteins shared by *M. sibericum* and *M. kamchatka* allowed the generation of flawless pairwise alignments and the computation of highly reliable  $\omega$  values for most (*i.e.* 397/411) of their orthologous singletons.

210 Fourteen singleton pairs were not taken into account in the selection pressure analysis 211 because of their either identical or quasi-identical sequences (11 of them), or unreliable pairwise 212 alignments (3 of them). For the 397 gene pairs retained in the analysis, the mean  $\omega$  value was 0.24 213  $\pm$  0.14 (Fig. 6). This result corresponds to a strong negative selection pressure indicating that most 214 of the encoded proteins greatly contribute to the molliviruses' fitness. Together with the high level 215 of pairwise similarity (92%) of their proteins, this also indicates that *M. kamchatka* evolved very 216 little during the last 30,000 years and that the *M. sibericum* genome was not prominently damaged 217 during its cryostasis in permafrost.

218 The analysis restricted to the 244 pairs of ORFan-coding genes resulted in a very similar  $\omega$ 219 value of 0.29± 0.15 (Fig. 6). This indicates that although homologs of these proteins are only found 220 in molliviruses, they have the same impact on the virus fitness than more ubiguitous proteins. This 221 confirms that they do encode actual proteins, albeit with unknown functions. In contrast, four 222 orthologous pairs (ms 160/mk 141; ms 280/mk 262; ms 171/mk 151; ms 430/mk 411; 223 ms 60/mk 48) exhibit  $\omega$  value larger than one. Those ORFans either are under positive selection 224 for maintaining their functions, or newly created gene products undergoing refinement or 225 pseudogenization.

We further examined the selection pressure of proteins-coding genes with homologs in pandoraviruses. We used their 10 sequenced genomes to generate the corresponding gene clusters (Fig. 7). The 90 clusters shared by both virus groups included 64 singletons (single copy

gene present in all viruses), among which 55 were suitable for dN/dS computations. The mean  $\omega$ value (0.17 ± 0.1) was very low, indicating that these genes, forming a "super core" gene set common to the molliviruses and pandoraviruses, are under an even stronger negative selection pressure than those constituting the provisional (most likely overestimated) mollivirus core gene set .

#### 234 Genomic inhomogeneity

235 The original genome analysis of Lausannevirus (a member of the Marseilleviridae family) (17)

revealed an unexpected non-uniform distribution of genes according to their annotation.

237 "Hypothetical" genes (i.e. mostly ORFans) were segregated from "annotated" (i.e. mostly non-

238 ORFans) in two different halves of the genome. In a more recent work, we noticed a similar bias in

the distribution of Pandoravirus core genes (13). The availability of a second mollivirus isolate gave

240 us the opportunity to investigate this puzzling feature for yet another group of Acantamoeba-

241 infecting virus. In Fig. 8, we plotted the distribution of three types of genes: 1- those with

homologs in A. castellanii (n= 55 for M. sibericum and n= 51 for M. kamchatka), 2- those belonging

to the super core set shared by both molliviruses and pandoraviruses (n=64), 3- those unique to

either mollivirus strains (n=26 for *M. sibericum*, n=12 for *M. kamchatka*). These plots reveal a

strong bias in the distribution of the super core vs. ORFan genes (Fig. 8). The first half of the *M*.

*sibericum* genome exhibits 90% of its ORFans while the second half contains most of the members

of the super core gene set. In contrast, genes eventually exchanged with the host display a more

248 uniform distribution. The lack of an apparent segregation in the distribution of ORFans in the *M*.

249 kamchatka genome might be due to their underprediction as no transcriptome information is

available for this strain. Fig. 9 shows that there is also a strong bias in the distribution of single-

251 copy genes vs. those with paralogs in either M. sibericum and/or M. kamchatka. Altogether, these

analyses suggest that the two genome halves follow different evolutionary scenario, the first half

concentrating the genomic plasticity (*de novo* gene creation, gene duplication), the other half
 concentrating the most conserved, eventually essential, gene content.

- 255
- 256 **Discussion**

257 Following the discovery of their first representatives, each families of giant (e.g. Mimiviridae, 258 Pithoviridae, Pandoraviridae) and large (e.g. Marseilleviridae) viruses infecting acanthamoeba have 259 expanded steadily, suggesting they were relatively abundant and present in a large variety of 260 environments. One noticeable exception has been the molliviruses, the prototype of which 261 remained unique after its isolation from 30,000-year old permafrost. The absence of M. sibericum 262 relatives from the large number of samples processed by others and us since 2014, raised the 263 possibility that they might have gone extinct, or might be restricted to the Siberian arctic. Our isolation of a second representative of the proposed Molliviridae family, M. kamtchatka, at a 264 265 location more than 1,500 km from the first isolate and enjoying a milder climate, is now refuting 266 these hypotheses. Yet, the planet-wide ubiquity of these viruses remains to be established, in 267 contrast to other acanthamoeba-infecting giant viruses (7). Even when present, mollivirus-like 268 viruses appear to be in very low abundance, as judged from the very small fraction of 269 metagenomics reads they represent in total sample DNA for *M. kamchatka* (about 0.02 part per 270 million) as well as for *M. sibericum* (about one part per million)(6). Another possibility would be 271 that the actual environmental host is not an acanthamoeba, the model host used in our laboratory. 272 However, evidences of specific gene exchanges with acanthamoeba (including a highly conserved 273 homolog major capsid protein) (6, 18, 19) make this explanation unlikely. We conclude that 274 members of the proposed Molliviridae family are simply less abundant than other acanthamoeba-275 infecting viruses, a conclusion further supported by the paucity of Mollivirus-related sequences in 276 the publicly available metagenomics data (data not shown).

277 As always the case, the characterization of a second representative of a new virus representative opened new opportunities of analysis. Unfortunately, the closeness of *M*. 278 279 kamchatka with M. sibericum limited the amount of information that could be drawn from their 280 comparison. For instance, the number of genes shared by the two isolates is probably a large 281 overestimate of the "core" gene set characterizing the whole family. On the other hand, the 282 closeness of the two isolates allowed an accurate determination of the selection pressure 283  $(\omega = dN/dS)$  exerted on many genes, showing that most of them, including mollivirus ORFans, 284 encode actual proteins the sequence of which are under strong negative selection and thus 285 contribute to the virus fitness. Given the partial phylogenetic affinity (i.e. 90 shared gene clusters) 286 of the mollivirus with the pandoraviruses, we also assessed the selection pressure exerted on 55 of 287 these "super core" genes, and found them under even stronger negative selection (Fig. 7). This 288 suggests that this super core gene set might have been present in a common ancestor to both 289 proposed families.

290 If we postulate that *M. sibericum* underwent into a complete stasis when it became frozen in 291 permafrost while *M. kamchatka* remained in contact with living acanthamoeba, we could consider 292 the two viral genomes to be separated by at least 30,000 years of evolution (eventually more if 293 they are not in a direct ancestry relationship)(20). The high percentage of identical residues (92%) 294 in their proteins corresponds to a low substitution rate of 1.7 10<sup>-6</sup> amino acid change/position 295 /year. This is an overestimate since the two viruses probably started to diverge from each other 296 longer than 30,000-year ago. This value is nevertheless comparable with estimates computed for 297 poxviruses (21) given the uncertainty on the number of replicative cycles occurring per year. The 298 high level of sequence similarity of *M. kamchatka* with *M. sibericum* also indicates that the later 299 did not suffer much DNA damage during its frozen stasis, even in absence of detectable virus-300 encoded DNA repair functions.

301 Horizontal gene transfers with the host were suggested by the fact that 51 proteins shared by 302 the two mollivirus strains exhibited a second best match in acanthamoeba. Because no homolog is 303 detected in other eukaryotes for most of them, these transfers probably occurred in the mollivirus-304 to-host direction. The clearest case is that of a major capsid protein homolog (mk 314, ml 347) 305 sharing 64% identical residues with a predicted acanthamoeba protein (locus: XP 004333827). 306 Two other genes encoding proteins that have also homologs in molliviruses flank the 307 corresponding host gene. However, the corresponding viral genes are not collinear in *M. sibericum* 308 or *M. kamchatka* and were probably transferred from a different, yet unknown mollivirus strain. 309 The presence of a 100% conserved major capsid protein homolog in the genome of *M. kamchatka* 310 and *M. sibericum* is itself puzzling. Such protein (with a double-jelly roll fold) is central to the 311 structure of icosahedral particles (22). Consistent with its detection in *M. sibericum* virions (6), its 312 conservation in *M. kamchatka*, suggests that it still plays a role in the formation of the spherical 313 mollivirus particles, while it has no homolog in the pandoraviruses. Inspired by previous 314 observations made on the unrelated Lausannevirus genome (17), we unveiled a marked 315 asymmetry in the distribution of different types of protein-coding genes in the Mollivirus genomes. 316 As shown in Fig. 8 the left half of the genome concentrates most of the genes coding for strain-317 specific ORFans while the right half concentrates most of super core genes shared with 318 pandoraviruses. This asymmetry is even stronger for the multiple copy genes while single-copy 319 genes are uniformly distributed along the genome (Fig. 9). The molliviruses thus appear to confine 320 their genomic "creativity" (de novo creation and gene duplication) in one-half of their genome, 321 leaving the other half more stable. An asymmetry in the distribution of the core genes was 322 previously noticed in the pandoravirus genomes (13). Such features might be linked to the 323 mechanism of replication that is probably similar for the two virus families. Further studies are 324 needed to investigate this process. The asymmetrical genomic distribution of pandoravirus core 325 genes and mollivirus super core genes might be a testimony of their past common ancestry.

326	Despite their differences in morphology, as well as in virion and genome sizes, the comparative		
327	analysis of the prototype <i>M. sibericum</i> and of the new isolate <i>M. kamchatka</i> confirms their		
328	phylogenetic affinity with the Pandoraviruses (Fig. 3, Fig. 10). However, it remains unclear whether		
329	this is due to a truly ancestral relationship between them, or if it is only the consequence of		
330	numerous past gene exchanges favored by the use of the same cellular host. From the perspective		
331	of the sole DNA polymerase sequence, the two known molliviruses do cluster with the		
332	pandoraviruses, albeit at a larger evolutionary distance than usually observed between members		
333	of the same virus family (Fig. 10). In absence of an objective threshold, and pending the		
334	characterization of eventual "missing links", we thus propose to classify <i>M. sibericum</i> and <i>M.</i>		
335	kamchatka as members of the proposed Molliviridae family, distinct from the Pandoraviridae.		
336			
337	Materials and Methods		
338	Virus isolation		
339	We isolated <i>M. kamchatka</i> from muddy grit collected near Kronotski Lake, Kamchatka (Russian		
340	Ederation N :54 32 59 E :160 34 55) The sample was stored for twenty days in pure rice medium		
	rederation will 34 32 39, E .100 34 33). The sample was stored for twenty days in pure fice medium		
341	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a		
341 342	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of		
<ul><li>341</li><li>342</li><li>343</li></ul>	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin		
<ul><li>341</li><li>342</li><li>343</li><li>344</li></ul>	<ul> <li>(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of</li> <li>Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation</li> </ul>		
<ul> <li>341</li> <li>342</li> <li>343</li> <li>344</li> <li>345</li> </ul>	<ul> <li>(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of</li> <li>Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious</li> </ul>		
<ul> <li>341</li> <li>342</li> <li>343</li> <li>344</li> <li>345</li> <li>346</li> </ul>	<ul> <li>(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of</li> <li>Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious phenotypes was centrifuged for 1 hour at 16,000xg at room temperature. Two T75 flasks were</li> </ul>		
<ul> <li>341</li> <li>342</li> <li>343</li> <li>344</li> <li>345</li> <li>346</li> <li>347</li> </ul>	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious phenotypes was centrifuged for 1 hour at 16,000xg at room temperature. Two T75 flasks were seeded with 60,000 cells/cm <sup>2</sup> and infected with the resuspended viral pellet. Infected cells were		
<ul> <li>341</li> <li>342</li> <li>343</li> <li>344</li> <li>345</li> <li>346</li> <li>347</li> <li>348</li> </ul>	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious phenotypes was centrifuged for 1 hour at 16,000xg at room temperature. Two T75 flasks were seeded with 60,000 cells/cm <sup>2</sup> and infected with the resuspended viral pellet. Infected cells were cultured in the same conditions as described below. We confirmed the presence of viral particles		
<ul> <li>341</li> <li>342</li> <li>343</li> <li>344</li> <li>345</li> <li>346</li> <li>347</li> <li>348</li> <li>349</li> </ul>	(23) at room temperature. An aliquot of the pelleted sample triggered an infected phenotype on a culture of <i>Acanthamoeba castellanii Neff</i> (ATCC30010TM) cells adapted to 2,5µg/mL of Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone-yeast-extract-glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious phenotypes was centrifuged for 1 hour at 16,000xg at room temperature. Two T75 flasks were seeded with 60,000 cells/cm <sup>2</sup> and infected with the resuspended viral pellet. Infected cells were cultured in the same conditions as described below. We confirmed the presence of viral particles by light microscopy.		

350 Validation of the presence of *M. kamchatka* in the original sample

351	To confirm the origin of the <i>M. kamchatka</i> isolate from the soil of the Kronotsky river bank,
352	DNA was extracted from the sample and sequenced on an Illumina platform, leading to
353	340,320,265 pair-ended reads (mean length 150bp). These metagenomics reads were then
354	mapped onto the genome sequence of <i>M. kamchatka</i> . Seven matching (100 % Identity) pair-ended
355	reads (hence 14 distinct reads) were detected, indicating the presence of virus particles in the
356	original sample, although at very low concentration. However, the very low probability of such
357	matches by chance (p<10 <sup>-63</sup> ) together with the scattered distribution of these matches along the
358	viral genome, further demonstrate the presence of <i>M. kamchatka</i> in the original sample.

359

#### 360 Virus Cloning

361 Fresh A. castellanii cells were seeded on a 12-well culture plate at a final concentration of 362 70,000cells/cm<sup>2</sup>. Cell adherence was controlled under light microscopy after 45 minute and viral 363 particles were added to at a multiplicity of infection (MOI) around 50. After 1 h, the well was 364 washed 15 times with 3 mL of PPYG to remove any viral particle in suspension. Cells were then 365 recovered by gently scrapping the well, and a serial dilution was performed in the next three wells 366 by mixing 200µL of the previous well with 500µL of fresh medium. Drops of 0.5µL of the last 367 dilution were recovered and observed by light microscopy to confirm the presence of a unique A. 368 *castellanii* cell. The 0.5µL droplets were then distributed in each well of three 24-well culture plate. 369 Thousand uninfected A. castellanii cells in 500µL of PPYG were added to the wells seeded with a 370 single cell and incubated at 32°C until witnessing the evidence of a viral production from the 371 unique clone. The corresponding viral clones were recovered and amplified prior purification, DNA 372 extraction and cell cycle characterization by electron microscopy.

#### 373 Virus mass production and purification

A total number of 40 T75 flasks were seeded with fresh *A. castellanii* cells at a final

375 concentration of 60,000 cells/cm<sup>2</sup>. We controlled cell adherence using light microscopy after 45

minute and flasks were infected with a single clone of *M. kamchatka* at MOI=1. After 48h hours of
incubation at 32°C, we recovered cells exhibiting infectious phenotypes by gently scrapping the
flasks. We centrifuged for 10 min at 500×g to remove any cellular debris and viruses were pelleted
by a 1-hour centrifugation at 6,800xg. The viral pellet was then layered on a discontinuous cesium
chloride gradient (1,2g/cm²/ 1,3g/cm²/ 1,4g/cm²/ 1,5g/cm²) and centrifuged for 20h at 103,000xg.
The viral fraction produced a white disk, which was recovered and washed twice in PBS and stored
at 4 °C or -80 °C with 7,5% DMSO.

383

#### 384 Infectious cycle observations using TEM

385 Twelve T25 flasks were seeded with a final concentration of 80,000 cells/cm<sup>2</sup> in PPYG medium 386 containing antibiotics. In order to get a synchronous infectious cycle eleven flasks were infected by 387 freshly produced *M. kamchatka* at substantial MOI=40. The *A. castellanii* infected flasks were fixed 388 by adding an equal volume of PBS buffer with 5 % glutaraldehyde at different time points after the 389 infection : 1h pi, 2h pi, 3h pi, 4h pi, 5h pi, 6h pi, 7h pi, 8h pi, 9h pi, 10h pi and 25h pi. After 45min of 390 fixation at room temperature, cells were scrapped and pelleted for 5min at 500xg. Then cells were 391 resuspended in 1ml of PBS buffer with 2,5 % glutaraldehyde and stored at 4°C. Each sample was 392 coated in 1mm<sup>3</sup> of 2 % low melting agarose and embedded in Epon-812 resin. Optimized osmium-393 thiocarbohydrazide-osmium (OTO) protocol was used for staining the samples: 1h fixation in PBS 394 with 2 % osmium tetroxide and 1.5 % potassium ferrocyanide, 20 min in water with 1 % 395 thiocarbohydrazide, 30min in water with 2 % osmium tetroxide, overnight incubation in water with 396 1 % uranyl acetate and finally 30min in lead aspartate. Dehydration was made using an increasing 397 concentration of ethanol (50 %, 75 %, 85 %, 95 %, 100 %) and cold dry acetone. Samples were 398 progressively impregnated with an increasing mix of acetone and Epon-812 resin mixed with DDSA 399 0,34v/v and NMA 0,68v/v (33 %, 50 %, 75 % and 100%). Final molding was made using a hard 400 Epon-812 mix with DDSA 0,34v/v NMA 0,68v/v and 0,031v/v of DMP30 accelerator and hardened

in the oven at 60°C for 5 days. Ultrathin sections (90nm thick) were observed using a FEI Tecnai G2
operating at 200kV.

403 **DNA Extraction** 

*M. kamchatka* genomic DNA was extracted from approximately 5 x 10<sup>9</sup> purified virus particles using Purelink Genomic extraction mini kit according to the manufacturer's recommendation. Lysis was performed with in a buffer provided with the kit and extra DTT at a final concentration of

407 1mM.

408 Genome sequencing and assembly

409 *M. kamchatka* genome was assembled using Spades (24) with a stringent K-mer parameter

410 using both various iteration steps (k = 21,41,61,81,99,127), the "-careful" option to minimize

411 number of mismatches in the final contigs, and the "-nanopore" option to use long reads.

#### 412 Annotation of *Mollivirus sibericum* and *Mollivirus kamchatka*

413 A stringent gene annotation of *M. sibericum* was performed as previously described (13) using 414 RNA-seq transcriptomic data (6). Stranded RNA-seq reads were used to accurately annotate 415 protein-coding genes. Stringent gene annotation of *M. kamchatka* was performed w/o RNA seq 416 data but taking into account protein similarity with *M. sibericum*. Gene predictions were manually 417 curated using the web-based genomic annotation editing platform Web Apollo (25). Functional 418 annotations of protein-coding genes of both genomes were performed using a two-sided approach 419 as already previously described (13). Briefly, protein domains were searched with the CD-search 420 tool (26) and protein sequence searching based on the pairwise alignment of hidden Markov 421 models (HMM) was performed against the Uniclust30 database using HHblits tool (27). Gene 422 clustering was done using Orthofinder's default parameters (28) adding the "-M msa -oa" option. 423 Strict orthology between pairs of proteins was confirmed using best reciprocal blastp matches. 424 **Selection Pressure Analysis** 

- 425 Ratios of non-synonymous (dN) over synonymous (dS) mutation rates for pairs of orthologous
- 426 genes were computed from MAFFT global alignment (29) using the PAML package and codeml with
- 427 the « model = 2 » (30). A strict filter was applyied to the dN/dS ratio: dN > 0, dS > 0,  $dS \le 2$  and
- 428 dN/dS  $\leq$  10. The computation of the Codon Adaptation Index (CAI) of both Mollivirus was
- 429 performed using the cai tool from the Emboss package (31).
- 430 Metagenome sequencing, assembly and annotation
- 431 All metagenomic sample were sequenced with DNA-seq paired-end protocol on Illumina HiSeq
- 432 platform at Genoscope producing 16 datasets of 2x150bp read length. Raw reads quality was
- 433 evaluated with FASTQC (32). Identified contaminants were removed and remaining reads were
- 434 trimmed on the right end using 30 as quality threshold with BBTools (33). Assemblies of filtered
- 435 data sets were performed using MEGAHIT (34) with the following options: "--k-list 33,55,77,99,127
- 436 --min-contig-len 1000". All filtered reads were then mapped to the generated contigs using
- 437 Bowtie2 (35) with the "--very-sensitive" option.
- 438 Availability of data
- The *M. kamtchatka* annotated genome sequence is freely available from the public through the
  Genbank repository (<u>URL://www.ncbi.nlm.nih.gov/genbank/</u>) under accession number XXXXX.

441

## 442 Acknowledgements

We are deeply indebted to our volunteer collaborator Alexander Morawitz for collecting the Kamchatka soil sample. We thank N. Brouilly, F. Richard and A. Aouane (imagery platform, Institut de Biologie du Développement de Marseille Luminy) for their expert assistance. E Christo-Foroux is the recipient of a DGA-MRIS scholarship (201760003). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 832601) and from the FRM prize "Lucien Tartois" to C.

- 449 Abergel. The funding bodies had no role in the design of the study, analysis, and interpretation of
- 450 data and in writing the manuscript.

### 451 **Competing interests**

452 The authors declare that they have no competing interests

# 454 Figures

455



456

457 **Fig 1.** Ultrathin section TEM image of a neo-synthetized *M. kamchatka* particle in the cell

458 cytoplasm 7h post infection. The structure of the mature particles appear identical to that of *M*.

459 sibericum.





- 463 **Fig 2.** Ultrathin section TEM image of *A. castellanii* cell 7 to 10 hours post infection by *M.*
- 464 kamchatka. (A) Viral factory exhibiting fibrils (F), a nascent viral particle (V), and surrounding
- 465 mitochondria (M). Fragments of the ruptured nuclear membrane are visible as dark bead strings.
- 466 (B) Details of a nuclear membrane rupture through which fibrils synthetized in the nucleus (N) are
- 467 shed into the cytoplasm (C).





- 472 Fig 3. Distribution of the best-matching NR homologs of *M. kamchatka* predicted proteins. Best-
- 473 matching homologous proteins were identified using BLASTP (E value <10<sup>-5</sup>) against the non-
- 474 redundant (NR) database (15) (after excluding *M. sibericum*). Green shades are used for
- 475 eukaryotes, red shades for viruses.

bioRxiv preprint first posted online Nov. 15, 2019; doi: http://dx.doi.org/10.1101/844274. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

477



478

480	Fig 4. Eventual gene transfers from a pandoravirus to <i>M. kamchatka</i> . Both phylogenetic trees were
481	computed from the global alignments of orthologous protein sequences using MAFFT (29). IQtree
482	(36) was used to determine the optimal substitution model (options: « -m TEST » and « –bb
483	1000 »). (A) Protein mk_165 (no predicted function). (B) Predicted methyltransferase mk_92. Both
484	M. kamchatka protein sequences appear embedded within the pandoravirus trees. In (A), the long
485	branch leading to the <i>M. kamchatka</i> homolog suggests its accelerated divergence after an ancient
486	acquisition from a pandoravirus. In (B), the long branch leading to the P. dulcis homolog might
487	alternatively be interpreted as a non-orthologous replacement of the ancestral pandoravirus
488	version of the gene.





- 492 Fig 5. Genomic features of strain-specific ORFans. (A) Codon adaption index (CAI). (B) G+C content.
- 493 (C) Protein length. Box plots show the median, the 25th and 75th percentiles. P-value are
- 494 calculated using the Wilcoxon test.





- 497 **Fig 6.** Selection pressure among different classes of genes. Values of  $\omega$  (i.e. dN/dS) were computed
- 498 from the alignments of homologous coding regions in *M. kamchatka* and *M. sibericum*. (A)
- 499 Distribution of calculated  $\omega$  values (n=397). (B) Box plots of the  $\omega$  ratio among ORFan genes
- 500 (n=243) and non ORFan genes (n=154). Box plots show the median, the 25th, and 75th percentiles.
- 501 All p-value are calculated using the Wilcoxon test.
- 502

503



504

**Fig 7.** Comparison of the mollivirus and pandoravirus core gene contents. (A) The distribution of the protein clusters shared by all pandoraviruses (black), the two Molliviruses (pink), and by both virus groups (super core genes) (blue). (B) Box plot of  $\omega$  values calculated from the alignment of molliviruses core genes (pink), and super core genes (blue). Box plots show the median, the 25th, and 75th percentiles.











Fig 9. Distribution of single-copy vs. multiple copy genes along mollivirus genomes. Single copy
genes (in blue) in both strains are uniformly distributed in contrast to genes with paralogs (pink) in
at least one strain that cluster in the left half of the genomes. (A) *M. sibericum* (n=48). (B) *M. kamchatka* (n= 46).

525

Tree scale: 0.1 🛏



526

Fig 10. Phylogeny of DNA polymerase B of large and giant dsDNA viruses. This neighbor-joining tree
was computed (JTT substitution model, 100 resampling) on 397 amino acid positions from an
alignment of 42 sequences computed by MAFFT (29). Branches with boostrap values <60% were</li>
collapsed.

#### 531 Tables

532

# 533 **Table 1.** Status of the protein-coding genes unique to *M. kamchatka*

ORF ID	Predicted function	Putative evolutionary scenario
mk_25	None (ORFan)	de novo creation or loss in M. sibericum
mk_92	DNA methyltransferase	loss in <i>M. sibericum</i> (present in some pandoraviruses)
mk_93	Ring domain	loss in M. sibericum (present in some pandoraviruses)
mk_104	Ring domain	loss in M. sibericum (present in some pandoraviruses)
mk_127	None (ORFan)	de novo creation or loss in M. sibericum
mk_159	None	loss in M. sibericum (present in some pandoraviruses)
mk_165	None	HGT from Pandoravirus
mk_166	Peptidase	loss in M. sibericum (present in some pandoraviruses)
mk_172	None (ORFan)	de novo creation or loss in M. sibericum
mk_182	None (ORFan)	de novo creation or loss in M. sibericum
mk_231	None (ORFan)	de novo creation or loss in M. sibericum
mk_313	None (ORFan)	de novo creation or loss in M. sibericum
mk_369	None (ORFan)	de novo creation or loss in M. sibericum
mk_415	None (ORFan)	de novo creation or loss in M. sibericum
mk_441	None (ORFan)	de novo creation or loss in M. sibericum
mk_466	None (ORFan)	de novo creation or loss in M. sibericum
mk_467	None	loss in M. sibericum (HGT to Acanthamoeba)
mk_469	B1-1 like	loss in M. sibericum (present in Acanthamoeba)
mk_476	None (ORFan)	de novo creation or loss in M. sibericum
mk_478	None (ORFan)	de novo creation or loss in M. sibericum

	535	Table 2. Status of the	protein-coding genes	unique to M. sibericum
--	-----	------------------------	----------------------	------------------------

ORF ID	Predicted function	Putative evolutionary scenario
ms_1	None (ORFan)	De novo creation or loss in M. kamchatka
ms_3	None (ORFan)	De novo creation or loss in M. kamchatka
ms_5	None (ORFan)	De novo creation or loss in M. kamchatka
ms_7	None (ORFan)	De novo creation or loss in M. kamchatka
ms_8	None (ORFan)	De novo creation or loss in M. kamchatka
ms_13	None (ORFan)	De novo creation or loss in M. kamchatka
ms_14	None	HGT to Acanthamoeba
ms_38	None (ORFan)	De novo creation or loss in M. kamchatka
ms_42	None (ORFan)	De novo creation or loss in M. kamchatka
ms_53	None (ORFan)	De novo creation or loss in M. kamchatka
ms_64	None	loss in M. Kamchatka (present in Noumeavirus)
ms_109	None (ORFan)	De novo creation or loss in M. kamchatka
ms_120	None (ORFan)	De novo creation or loss in M. kamchatka
ms_136	None (ORFan)	De novo creation or loss in M. kamchatka
ms_138	None (ORFan)	De novo creation or loss in M. kamchatka
ms_139	None (ORFan)	De novo creation or loss in M. kamchatka
ms_144	None (ORFan)	De novo creation or loss in M. kamchatka
ms_157	None (ORFan)	De novo creation or loss in M. kamchatka
ms_159	None (ORFan)	De novo creation or loss in M. kamchatka
ms_166	None (ORFan)	De novo creation or loss in M. kamchatka
ms_172	None (ORFan)	De novo creation or loss in M. kamchatka
ms_190	None	loss in <i>M. kamchatka</i> (present in some pandoraviruses)
ms_193	None (ORFan)	De novo creation or loss in M. kamchatka
ms_246	None (ORFan)	De novo creation or loss in M. kamchatka
ms_258	None (ORFan)	De novo creation or loss in M. kamchatka
ms_311	Zinc-finger domain	loss in M. kamchatka (present in Gossypium hirsutum)
ms_312	Zinc-finger domain	loss in M. kamchatka (present in Cavenderia fasciculata)
ms_313	None	loss in <i>M. kamchatka</i> (present in some pandoraviruses)
ms_464	None (ORFan)	De novo creation or loss in M. kamchatka
ms_465	None (ORFan)	De novo creation or loss in M. kamchatka
ms_479	DNA methyltransferase	loss in <i>M. kamchatka</i> (present in some pandoraviruses)
ms_494	None (ORFan)	De novo creation or loss in M. kamchatka
ms_495	None (ORFan)	De novo creation or loss in M. kamchatka

536

#### 538 References

- 539 1. Shi T, Reeves RH, Gilichinsky DA, Friedmann EI, 1997. Characterization of Viable Bacteria from
- 540 Siberian Permafrost by 16S rDNA Sequencing. Microb Ecol 33:169-179.
- 541 https://doi.org/10.1007/s002489900019.
- 542 2. Vishnivetskaya T, Kathariou S, McGrath J, Gilichinsky D, Tiedje JM 2000. Low-temperature
- 543 recovery strategies for the isolation of bacteria from ancient permafrost sediments. Extremophiles
- 544 4:165-173. https://doi.org/10.1007/s007920070031.
- 545 3. Graham DE, Wallenstein MD, Vishnivetskaya TA, Waldrop MP, Phelps TJ, Pfiffner SM, Onstott TC,
- 546 Whyte LG, Rivkina EM, Gilichinsky DA, Elias DA, Mackelprang R, VerBerkmoes NC, Hettich RL,
- 547 Wagner D, Wullschleger SD, Jansson JK. 2012. Microbes in thawing permafrost: the unknown
- 548 variable in the climate change equation. ISME J 6:709-712.
- 549 https://doi.org/10.1038/ismej.2011.163.
- 4. Yashina S, Gubin S, Maksimovich S, Yashina A, Gakhova E, Gilichinsky D. 2012. Regeneration of
- 551 whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. Proc Natl Acad Sci
- 552 U S A 109:4008-4013. https://doi.org/10.1073/pnas.1118386109.
- 553 5. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L,
- 554 Bruley C, Couté Y, Rivkina E, Abergel C, Claverie JM. 2014. Thirty-thousand-year-old distant relative
- 555 of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A
- 556 111:4274-4279. https://doi.org/ 10.1073/pnas.1320670111.
- 557 6. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot, M, Alempic JM, Ramus C, Bruley C,
- 558 Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie JM. 2015. In-depth study of
- 559 Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proc Natl Acad Sci U
- 560 S A 112:E5327-E5335. https://doi.org/10.1073/pnas.1510795112.

- 561 7. Abergel C, Legendre M, Claverie JM. The rapidly expanding universe of giant viruses: Mimivirus,
- 562 Pandoravirus, Pithovirus and Mollivirus. 2015. FEMS Microbiol Rev 39:779-796.
- 563 https://doi.org/10.1093/femsre/fuv037.
- 564 8. Levasseur A, Andreani J, Delerce J, Bou Khalil J, Robert C, La Scola B, Raoult D. 2016. Comparison
- 565 of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution. Genome Biol
- 566 Evol 8:2333-2339. https://doi.org/10.1093/gbe/evw153.
- 567 9. Andreani J, Aherfi S, Bou Khalil JY, Di Pinto F, Bitam I, Raoult D, Colson P, La Scola B. 2016.
- 568 Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. Viruses
- 569 8:300. https://doi.org/10.3390/v8110300.
- 570 10. Bertelli C, Mueller L, Thomas V, Pillonel T, Jacquier N, Greub G. 2017. Cedratvirus lausannensis -
- 571 digging into Pithoviridae diversity. Environ Microbiol 19:4022-4034.
- 572 https://doi.org/10.1111/1462-2920.13813.
- 573 11. Quemin ER, Corroyer-Dulmont S, Baskaran A, Penard E, Gazi AD, Christo-Foroux E, Walther P,
- 574 Abergel C, Krijnse-Locker J. 2019. Complex membrane remodeling during virion assembly of the
- 575 30,000-year-old Mollivirus sibericum. J Virol 93(13). pii: e00388-19.
- 576 https://doi.org/10.1128/JVI.00388-19.
- 577 12. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L,
- 578 Bruley C, Garin J, Claverie JM, Abergel C. 2013. Pandoraviruses: amoeba viruses with genomes up
- to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281-286.
- 580 https://doi.org/10.1126/science.1239181.
- 13. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic JM, Beucher L, Philippe N, Bertaux
- 582 L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie JM. 2018. Diversity and evolution of the
- 583 emerging Pandoraviridae family. Nat Commun 9:2285.
- 584 https://doi.org/10.1038/s41467-018-04698-4.

- 585 14. Legendre M, Alempic JM, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S, Couté Y,
- 586 Abergel C, Claverie JM. 2019. Pandoravirus celtis illustrates the microevolution processes at work
- 587 in the giant Pandoraviridae genomes. Front Microbiol 10:430. https://doi.org/
- 588 10.3389/fmicb.2019.00430.
- 589 15. NCBI Resource Coordinators. 2018. Database resources of the National Center for
- 590 Biotechnology Information. Nucleic Acids Res 46:D8-D13. https://doi.org/10.1093/nar/gkx1095.
- 591 16. Heidel AJ, Lawal HM, Felder M, Schilde C, Helps NR, Tunggal B, Rivero F, John U, Schleicher M,
- 592 Eichinger L, Platzer M, Noegel AA, Schaap P, Glöckner G. 2011. Phylogeny-wide analysis of social
- 593 amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res
- 594 21:1882-1891. https://doi.org/10.1101/gr.121137.111.
- 595 17. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub G. 2011.
- 596 Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol 13:1454-1466.
- 597 https://doi.org/10.1111/j.1462-2920.2011.02446.x.
- 598 18. Maumus F, Blanc G. 2016. Study of gene trafficking between acanthamoeba and giant viruses
- 599 suggests an undiscovered fof amoeba-infecting viruses. Genome Biol
- 600 Evol 8:3351-3363. https://doi.org/10.1093/gbe/evw260.
- 601 19. Chelkha N, Levasseur A, Pontarotti P, Raoult D, Scola B, Colson P. 2018. A phylogenomic study of
- 602 Acanthamoeba polyphaga draft genome sequences suggests genetic exchanges with giant viruses.
- 603 Front Microbiol 9:2098. https://doi.org/10.3389/fmicb.2018.02098.
- 604 20. Duchêne S, Holmes EC. 2018. Estimating evolutionary rates in giant viruses using
- ancient genomes. Virus Evol 4:vey006. https://doi.org/10.1093/ve/vey006.
- 606 21. Hughes AL, Irausquin S, Friedman R. 2010. The evolutionary biology of poxviruses. Infect Genet
- 607 Evol 10:50-59. https://doi.org/10.1016/j.meegid.2009.10.001.
- 608 22. San Martín C, van Raaij MJ. 2018. The so far farthest reaches of the double jelly roll capsid
- 609 protein fold. Virol J 15:181. https://doi.org/10.1186/s12985-018-1097-1.

- 610 23. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. 2011. Distant Mimivirus relative with a
- 611 larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A
- 612 108:17486-17491. https://doi.org/10.1073/pnas.1110889108.
- 613 24. Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov AS, Lesin V, Nikolenko S, Pham
- 614 S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev MA, Pevzner P. 2012. SPAdes: A
- 615 New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol
- 616 19:455-477. https://doi.org/10.1089/cmb.2012.0021.
- 617 25. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes I, Elsik C,
- 618 Lewis S. 2019. Apollo: Democratizing genome annotation. PLoS Comput Biol 15: e1006790.
- 619 https://doi.org/10.1371/journal.pcbi.1006790.
- 620 26. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz
- 621 M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D,
- 622 Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222-
- 623 226. https://doi.org/10.1093/nar/gku1221.
- 624 27. Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative
- 625 protein sequence searching by HMM-HMM alignment. Nat Methods. 9:173-175.
- 626 https://doi.org/10.1038/nmeth.1818.
- 627 28. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
- 628 comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157.
- 629 https://doi.org/10.1186/s13059-015-0721-2.
- 630 29. Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple
- 631 sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
- 632 https://doi.org/10.1093/nar/gkf436.
- 633 30. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-
- 634 1591. https://doi.org/10.1093/molbev/msm088.

- 635 31. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software
- 636 suite. Trends Genet 16:276-277. https://doi.org/10.1016/s0168-9525(00)02024-2.
- 637 32. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available
- 638 online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc .
- 639 33. Bushnell B, Rood J, Singer E. 2017. BBMerge Accurate paired shotgun read merging via
- 640 overlap. PLoS ONE 12: e0185056. https://doi.org/10.1371/journal.pone.0185056.
- 641 34. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: An ultra-fast single-node solution
- 642 for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics
- 643 31:1674-1676. https://doi.org/10.1093/bioinformatics/btv033.
- 644 35. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods
- 645 9:357-359. https://doi.org/10.1038/nmeth.1923.
- 646 36. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic
- 647 algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol 32:268-274.
- 648 https://doi.org/10.1093/molbev/msu300.
- 649 37. Wickham H. ggplot2: Elegant Graphics for Data Analysis, p 33-74. 2016. Springer-Verlag New

650 York.