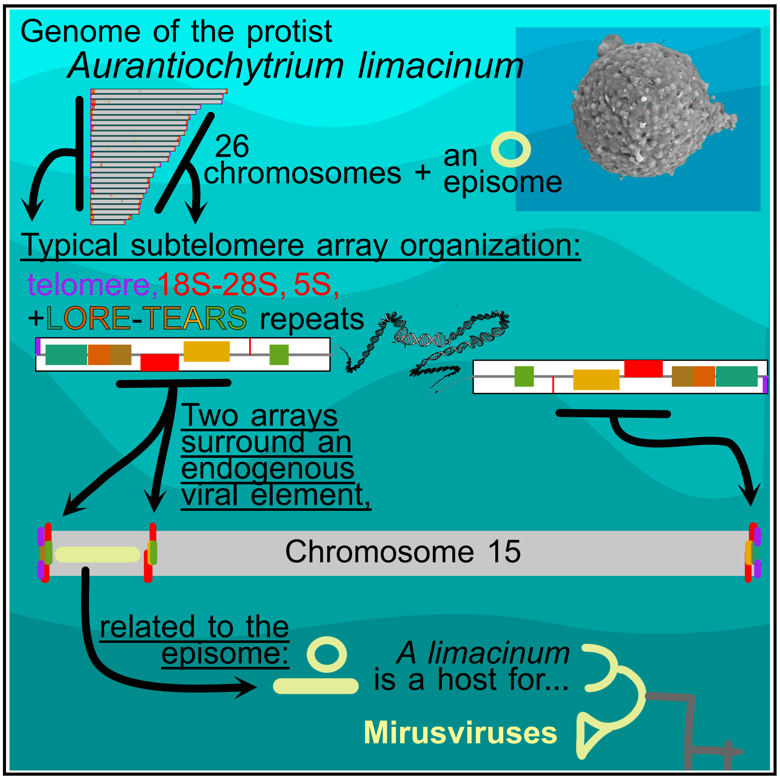
Report

Current Biology

The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses

Graphical abstract


Genome of the protist
Aurantiochytrium limacinum
26 chromosomes + an episome
Typical subtelomere array organization:
telomere,18S-28S, 5S,
+LORE-TEARS repeats
Two arrays surround an endogenous viral element,
Chromosome 15
related to the episome:
A limacinum is a host for...
Mirusviruses

Highlights

- The *Aurantiochytrium limacinum* genome has long repeats and rDNAs in all subtelomeres
- Two mirusvirus genomes are in *A. limacinum*, including a high-copy, circular episome
- The second mirusvirus region is integrated between two subtelomeric rRNA gene arrays
- The first known host for mirusviruses suggests dynamic host-virus genome interactions

Authors

Jackie L. Collier, Joshua S. Rest, Lucie Gallot-Lavallée, ..., Gina V. Filloramo, Anna M.G. Novák Vanclová, John M. Archibald

Correspondence

jackie.collier@stonybrook.edu (J.L.C.),
joshua.rest@stonybrook.edu (J.S.R.)

In brief

Collier, Rest, et al. find that the marine thraustochytrid *Aurantiochytrium limacinum* has arrays of rRNA genes and long repeated sequences in the chromosomal subtelomeres. They also identify it as the first known mirusvirus host, with both a circular mirusvirus episome and an integrated mirusvirus genome between two rDNA and long repeat arrays.

Collier et al., 2023, Current Biology *33*, 1–9
December 4, 2023 © 2023 Elsevier Inc.
https://doi.org/10.1016/j.cub.2023.10.009

CellPress

Current Biology

CellPress

Report

The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses

Jackie L. Collier,[1,8,9,11,12,*] Joshua S. Rest,[2,8,10,11,*] Lucie Gallot-Lavallée,[3] Erik Lavington,[2] Alan Kuo,[4] Jerry Jenkins,[4,5] Chris Plott,[4,5] Jasmyn Pangilinan,[4] Chris Daum,[4] Igor V. Grigoriev,[4,6] Gina V. Filloramo,[3] Anna M.G. Novák Vanclová,[7] and John M. Archibald[3]

[1]School of Marine and Atmospheric Sciences, Stony Brook University, Nicolls Road, Stony Brook, NY 11794, USA
[2]Department of Ecology and Evolution, Stony Brook University, Nicolls Road, Stony Brook, NY 11794, USA
[3]Department of Biochemistry & Molecular Biology, Dalhousie University, College Street, Halifax, NS B3H 4R2, Canada
[4]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, CA 94720, USA
[5]HudsonAlpha Institute for Biotechnology, Genome Way Northwest, Huntsville, AL 35806, USA
[6]Department of Plant and Microbial Biology, University of California Berkeley, University Avenue, Berkeley, CA 94720, USA
[7]Faculty of Science, Charles University, Opletalova, 110 00 Staré Město, Czechia
[8]These authors contributed equally
[9]X(formerly Twitter): @Collier_Lab_SBU
[10]X (formerly Twitter): @JoshuaRest
[11]X (formerly Twitter): @labyrinthulea
[12]Lead contact
*Correspondence: jackie.collier@stonybrook.edu (J.L.C.), joshua.rest@stonybrook.edu (J.S.R.)

https://doi.org/10.1016/j.cub.2023.10.009

SUMMARY

Viruses are the most abundant biological entities in the world's oceans, where they play important ecological and biogeochemical roles. Metagenomics is revealing new groups of eukaryotic viruses, although disconnected from known hosts. Among these are the recently described mirusviruses, which share some similarities with herpesviruses.[1] 50 years ago, "herpes-type" viral particles[2] were found in a thraustochytrid member of the labyrinthulomycetes, a diverse group of abundant and ecologically important marine eukaryotes,[3,4] but could not be further characterized by methods then available. Long-read sequencing has allowed us to connect the biology of mirusviruses and thraustochytrids. We sequenced the genome of the genetically tractable model thraustochytrid *Aurantiochytrium limacinum* ATCC MYA-1381 and found that its 26 linear chromosomes have an extraordinary configuration. Subtelomeric ribosomal DNAs (rDNAs) found at all chromosome ends are interspersed with long repeated sequence elements denoted as long repeated-telomere and rDNA spacers (LORE-TEARS). We identified two genomic elements that are related to mirusvirus genomes. The first is a ~300-kbp episome (circular element 1 [CE1]) present at a high copy number. Strikingly, the second, distinct, mirusvirus-like element is integrated between two sets of rDNAs and LORE-TEARS at the left end of chromosome 15 (LE-Chr15). Similar to metagenomically derived mirusviruses, these putative *A. limacinum* mirusviruses have a virion module related to that of herpesviruses along with an informational module related to nucleocytoplasmic large DNA viruses (NCLDVs). CE1 and LE-Chr15 bear striking similarities to episomal and endogenous latent forms of herpesviruses, respectively, and open new avenues of research into marine virus-host interactions.

RESULTS AND DISCUSSION

We performed short-read 454 and long-read nanopore sequencing of the *Aurantiochytrium limacinum* ATCC MYA-1381 nuclear genome, which independently yielded assemblies of 60.93 and 63.71 Mbp, respectively (Figure S1A; Tables S1 and S2). Neither assembly was rich in repetitive sequence (~4%, mostly simple repeats) (Data S1A and S1B). Multiple metrics indicate that both assemblies are highly complete. 99.66% of RNA sequencing (RNA-seq) reads mapped to the 454 assembly and 96% to the nanopore assembly, and we detected 91.4% and 87.9% of Eukaryota BUSCO genes in the 454 and nanopore assemblies, respectively (8.6% and 12.1% missing BUSCOs, respectively; Data S1C).

The 26 largest nanopore contigs likely represent complete or nearly complete *A. limacinum* physical chromosomes. These contigs range in size from ~1.02 to ~4 Mbp (Figure 1) and total 61.41 Mbp (96.4% of the complete nanopore assembly); they align with 37 of the longest 454 scaffolds (the primary 454 assembly) containing 59.93 Mbp (98.4%) of the total 454 assembly (Figures S1A–S1C; Tables S1 and S2). The nanopore contig sizes are consistent with our examination of the genome by pulsed-field gel electrophoresis (Figure S1D). This genome structure is similar to other stramenopiles for which both chromosome number and genome size are known: three

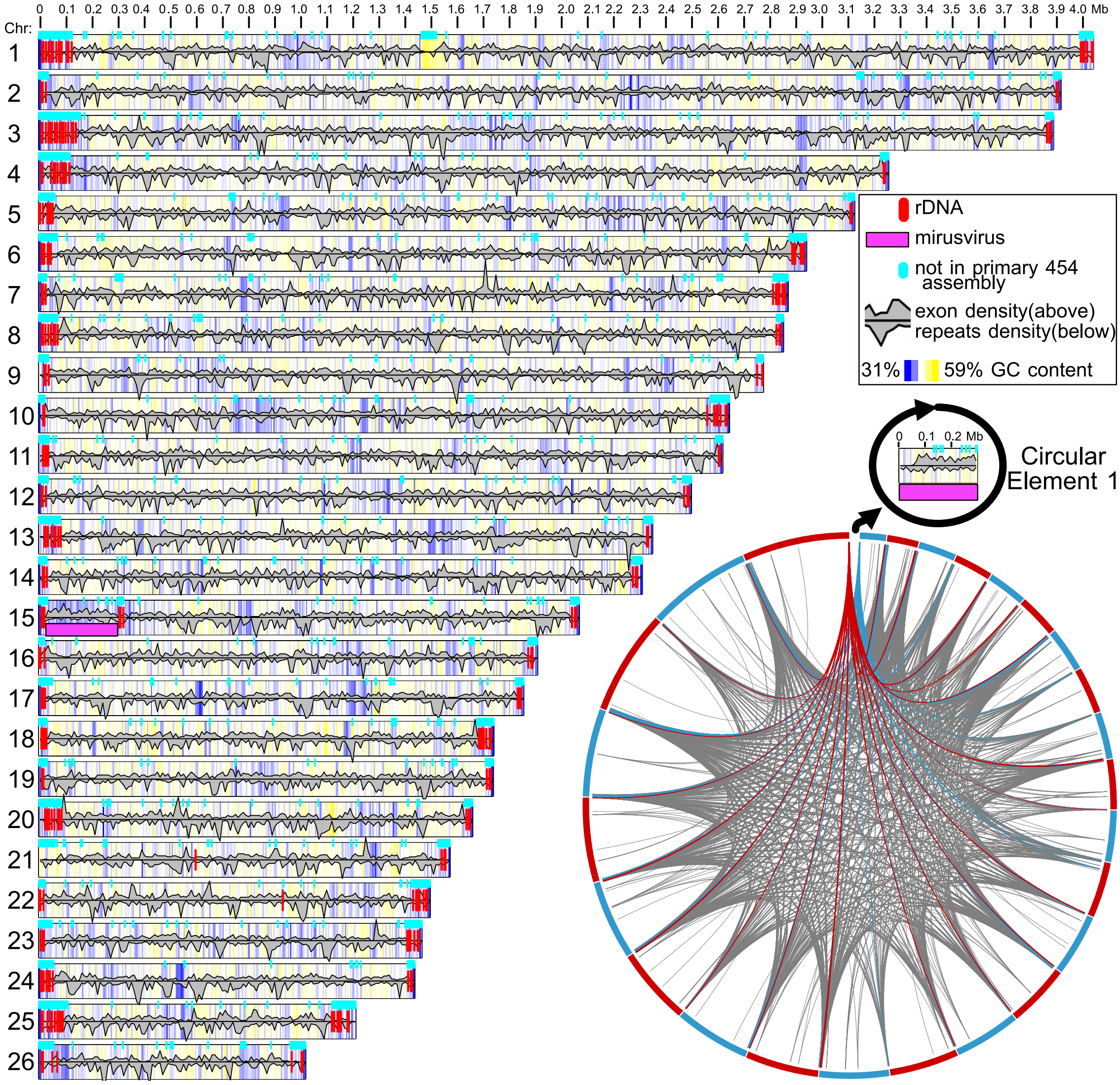


Figure 1. Size and select features of the 26 linear chromosomes and circular element 1 in *Aurantiochytrium limacinum* ATCC MYA-1381

CE1 is predicted to be circular but is displayed as linear. A scale in megabases is provided along the top of the plot.[54] Vertical red lines represent locations of predicted[55] rRNA genes. Cyan boxes represent regions that did not align[56] with the primary 454 assembly (see Figures S1B and S1C). GC content (5 kbp windows) is indicated by background color, with darker shades of blue indicating regions of lower GC content and darker shades of yellow indicating regions with higher GC content; low GC content at the linear chromosome ends reflect telomeric repeats. The gray density plot above the midline of each chromosome indicates the relative density of exons, based on mapping of predicted exons from the 454 assembly to each nanopore chromosome. The gray density plot below the midline (reflected so that higher values form valleys) indicates the relative density of repetitive sequences.[57] Inset: chord plot showing matching sequence regions of at least 1 kbp between contigs.[58] Two sets of chords are colored (arbitrarily red and blue) to highlight the directional nature of the repeats at the scaffold ends. A blank space in the chord plot represents the shortest scaffold, which has no matching sequence regions on other scaffolds.
See also Figures S1 and S2, Tables S1 and S2, and Data S1 and S2.

diatoms and a eustigmatophyte have smaller genomes (31–36.5 Mbp) with similar numbers of chromosomes (24–33), whereas the oomycete *Phytophthora sojae* has an 82.6 Mbp genome with ∼12–14 chromosomes.[5–7]

Aurantiochytrium chromosomes have subtelomeric rDNAs interspersed with long tandem repeats

Most *A. limacinum* chromosomes were assembled telomere to telomere. Among the 52 predicted chromosome ends, 39 terminate with telomeric repeats of sequence TTAGG ∼500 bp in length (mean 480 bp, median 499 bp) (Figure 2). The telomeric repeats identified in the nanopore assembly are slightly shorter than the TTAGGG repeats of vertebrates and other eukaryotes, including protists such as the photosynthetic stramenopile *Pelagomonas calceolata*,[7] but identical to TTAGG repeats reported from diverse insects and a few other eukaryotes.[8] Telomeric repeats are missing from 13 of the chromosome ends; this likely reflects assembly issues (Figure S1B; Table S1).

All assembled subtelomeric regions of the *A. limacinum* chromosomes unexpectedly contain rRNA gene clusters

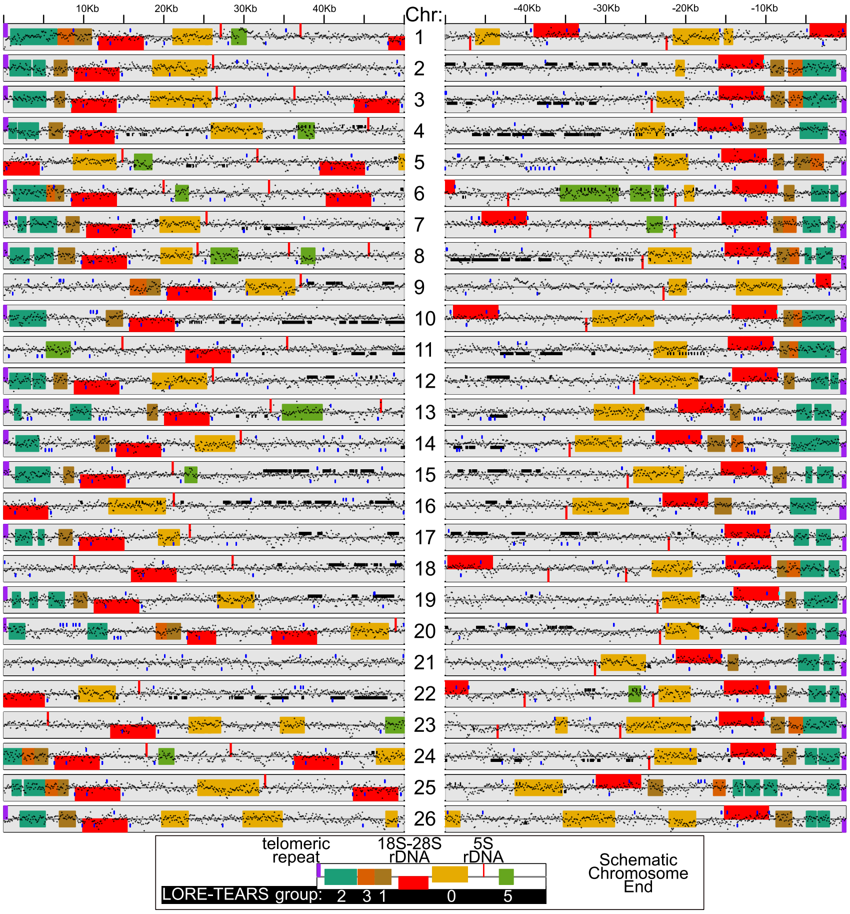


Figure 2. Focused view of the terminal 50 kbp of both ends of the *Aurantiochytrium limacinum* chromosomes

Locations of predicted rRNA genes as in Figure 1 (red; see inset). Purple boxes on ends of chromosomes indicate telomeric repeats. Five classes of LORE-TEARS (0, 1, 2, 3, and 5) are shown (see inset for key). G-quadruplexes are plotted as blue lines; note the regular positioning of G-quadruplexes within and around the rDNAs. Black lines are exons from the 454 assembly mapped using BLAST. GC content is plotted along the centerline (range: 31%–59%). Bottom inset: schematic view of a typical arrangement of elements at the ends of a chromosome, with a large 18S-5.8S-28S rDNA cluster transcribed toward the telomere and a small 5S rDNA transcribed away from the telomere. LORE-TEARS elements are colored and labeled by their sequence similarity group.

See also Figure S1, Tables S1 and S3, and Data S1.

interspersed with long repeated sequences (Figures 1 and 2). These elements are evident as extensive sequence matches between the ends of contigs (Figure 1, inset; in contrast to the 454 assembly; Figure S1E). Specifically, in 37 of the 39 nanopore contig ends with telomeres, an 18S-28S rRNA gene cluster (small subunit or 18S rRNA, ITS1, 5.8S rRNA, ITS2, and large subunit or 28S rRNA; average length = 5,551 bp) transcribed toward the telomere is found ~9.4 kilobase pair (kbp) (median) from the telomeric repeat (Figure 2). In 30 of these 37 contig ends, a 5S rRNA gene transcribed away from the telomeric repeat is found ~10 kbp (median) further from the telomere. In 17 of these 30 cases, no other rRNA genes were identified, and the 454 assembly scaffold mapped to the approximate location of the 5S gene. The remaining contig ends vary from this pattern. In some cases, both ends of a contig have the same organization (Chr17, Chr15, and Chr6), but, more commonly, the two ends are different. Only one rRNA gene (a 5S on Chr3) was found in the opposite orientation, and only three 5S genes (on Chr21, Chr22, and Chr15) and one 18S-28S rRNA gene cluster (on Chr15 associated with an endogenous mirusvirus, see below) were identified in the more central regions of nanopore assembly contigs (Figure 1).

Between each of these telomeric and subtelomeric elements are characteristic long tandem repeats (Figure 2; Table S3). We call these long repeated telomere and rDNA spacers (LORE-TEARS), which are built from repeated 366–529-bp units and lacking similarity to sequences in GenBank. Several distinct LORE-TEARS families occur in regular positions with respect to the chromosome ends. Just downstream of the 28S rRNA genes, there is usually one "group 1" element containing ~4 repeated units, each ~406 bp long. Closer to the telomere, there is usually at least one "group 2" element with ~6 repeated units, each ~366 bp long. Usually upstream of the 18S gene nearest to the telomere and between it and the nearest 5S gene is a "group 0" element, containing ~9 repeated units of ~385 bp. Where two consecutive 5S rRNA genes are detected, there is often a "group 5" element between them (~5 repeats of a ~421 bp unit). Among the 15 chromosomes with telomeric repeats assembled at both ends, seven have a "group 3" element (~3 repeats of a ~529 bp unit) between the group 1 and group 2 elements at only one end, while one has a group 3 element between the group 1 and group 2 elements at both ends, and seven have no group 3 elements. We detected G-quadruplexes associated with rRNA genes and some LORE-TEARS, which is consistent with a regulatory function for these elements at the chromosome ends[9–12] (Figure 2; Data S1D).

The organization of rRNA genes in the subtelomeres of the *A. limacinum* chromosomes suggests a specific relationship with telomeric processes. This arrangement is highly unusual, both in the nature of the repeats and their location. Eukaryotic rRNA genes are most commonly organized in a few large tandem arrays (e.g., one in yeast and five in humans), and 5S genes and 18S-28S gene clusters are typically not associated with one another.[13,14] Subtelomeric rDNA tandem repeats have been found in plants,[15,16] in some metazoans (in aphids at one telomere of the X chromosome[17]), and in the protist parasite *Giardia*.[18,19] The multicellular stramenopile *Saccharina japonica* (the kelp kombu) has a typical tandem 45S array in the middle of one chromosome and a tandem 5S array at the subtelomere of another.[20] In all these cases, rRNA gene arrays reside on the ends of only one or a few chromosomes. Unlinked 18S-28S rRNA gene clusters (i.e., not in tandem arrays) are found in the red alga *Cyanidioschyzon merolae*[21,22] and in several apicomplexan parasites,[23] including *Plasmodium falciparum*.[24] The 5S and 18S-28S coding regions in the nanopore assembly of *A. limacinum* are more closely spaced than is usual for organisms where this linkage occurs but not as tightly linked as in the brown alga (stramenopile) *Scytosiphon lomentaria* and some other protists, where the 5S is just downstream of the 18S-28S.[25,26]

The most similar subtelomeric architecture to that of *A. limacinum* is found in the microsporidian parasites *Encephalitozoon cuniculi* and *E. intestinalis*, which have one subtelomeric, divergently transcribed 18S-28S rRNA gene cluster near the end of each of their 11 chromosomes separated from the telomeric repeats by two types of telomere-associated repeat elements (TAREs) with ∼30–70 bp repeat units.[27] However, the 5S rRNA genes are not subtelomeric in *Encephalitozoon* spp. Subtelomeric 18S-28S rRNA gene clusters are also a feature of the endosymbiotically derived "nucleomorph" genomes of cryptomonads[28,29] and chlorarachniophytes.[30] Brugère et al.[31] suggested that the subtelomeric location of rDNA might be related to selective pressure associated with genome reduction, but the ∼62 Mbp genome of *A. limacinum* is not notably small among free-living stramenopiles, suggesting that genomic streamlining is not a factor here. The ends of chromosomes tend to be different from internal portions in exhibiting a higher frequency of recombination,[32,33] lower level of gene expression,[34] and higher rate of sequence evolution.[35] The selective forces and molecular mechanisms acting to maintain the consistent structure and homogeneous rDNA and LORE-TEARS sequences at the chromosome ends of *A. limacinum* offer new avenues for future research, particularly if similar arrangements are found broadly in labyrinthulomycetes or other unexplored corners of protist diversity. The rRNA genes, LORE-TEARS, and/or associated subtelomeric sequences in *A. limacinum* may be involved in chromosome end maintenance and replication, including the maintenance of rDNA stability and/or nucleolar structure,[14] comparable to the repetitive subtelomeric sequences that are functionally important in other species.[36,37]

Episomal and endogenous mirusvirus-like genetic elements

Viral detection software[38,39] suggested large regions with viral content at two locations in the genome (Data S2A and S2B). The first, a ∼300-kbp element dubbed circular element 1 (CE1), is present in both genome assemblies (Figure S2B; Tables S1 and S2) and is consistent with a ∼350- kbp band in the pulsed-field gel electrophoresis (Figure S1D). CE1 is predicted to be circular (Figure S2A) and has read coverage in both the 454 and nanopore assemblies ∼9× higher than the other scaffolds and contigs, suggesting that it is present at higher copy number than the 26 chromosomes (Table S1) and is maintained as an episome. CE1 lacks the rRNA genes, LORE-TEARS, and telomeric repeats found on the chromosomes (Figure 1). GC content and mapped transcript abundance are similar to the chromosomes (Table S1), but a smaller proportion of predicted genes have functional annotations, orthologous gene assignments, and predicted introns, and CE1 encodes no obvious BUSCO proteins (Figure S2C). Of the 177 predicted genes on CE1, 128 are ORFans (i.e., do not hit any known proteins), 21 have best BLAST hits to bacteria, 22 to eukaryotes, four to archaea, and one to viruses (when excluding the thraustochytrid *Hondaea fermentalgiana*; see below) (Data S2C).

The second area of viral gene content is a 296-kbp region (i.e., all of scaffold_35) at the left end of Chr15 (LE-Chr15) and provides a striking nexus with the subtelomere structure. LE-Chr15 comprises the telomeric repeats, followed by a set of rDNAs and LORE-TEARS, followed by the viral region, and then a second set of rDNAs and LORE-TEARS (Figure 1); this is the only place in the assembly with internal (non-telomeric) arrays of rRNA genes and LORE-TEARS. The GC content of the putative viral integrant is 41.6%, slightly lower than the rest of the chromosome (45.0%), and consistent with a foreign origin. The viral elements detected on CE1 and LE-Chr15 are related but distinct from one another (Figure S2D). Comparing CE1's 177 predicted proteins with the 152 proteins encoded by the virus-like region of LE-Chr15, only 48 are each other's reciprocal best BLAST hits (Data S2C), but their shared homologs generally branch together in phylogenetic trees (Figures 3 and S3).

Sequence similarity searches using proteins characteristic of various groups of viruses as queries (via blastp and HMMsearch) against the *A. limacinum* genome revealed many genes on CE1 and LE-Chr15 related to key genes of the *Mirusviricota*[1] (Table 1; Data S2C). For example, we detected *Mirusviricota*-like major capsid protein (MCP) and terminase coding regions on CE1 and LE-Chr15 (the terminase protein packs the freshly synthesized genome into newly formed capsids). Mirusvirus-like homologs were detected on both CE1 and LE-Chr15 for the remaining virion proteins as well (i.e., capsid maturation protease, portal protein, triplex 1, and triplex 2), as were other core mirusvirus genes, including a heliorhodopsin (CE1), a histone H3 (CE1), TATA-binding protein (CE1 and LE-Chr15), subunits alpha and beta of ribonucleotide reductase (LE-Chr15), and additional proteins of unknown function (Table 1; Data S2C). We also detected core nucleocytoplasmic large DNA virus (NCLDV) informational genes, such as family B DNA polymerase (DNAPol or PolB; identified previously by Gallot-Lavallée and Blanc[40]), DNA-dependent RNA polymerase large subunit (RNAPL), and superfamily II helicase proteins (Table 1; Data S2C). We found no sequence similarity between CE1 or LE-Chr15 and the lytic large DNA virus previously reported to infect the thraustochytrid *Sicyoidochytrium minutum* (SmDNAV).[41,42]

For *A. limacinum* genes that have detectable homologs in mirusviruses, NCLDVs and/or herpesviruses, we conducted phylogenetic analyses (Figures 3 and S3). Some virus-like genes on CE1 and LE-Chr15 are found only in mirusviruses (MCP; Figure 3C; terminases, Figure S3A). For genes with broader distribution, phylogenetic analyses support relationships of several CE1 and LE-Chr15 viral genes to mirusviruses, as well as to nucleocytoviruses, which share several informational genes with mirusviruses. The resolvase, helicase, and DNAPol trees show *A. limacinum* viral sequences branching specifically with mirusviruses (Figures 3A, S3B, and S3C), whereas the topoisomerase and nuclease trees show relatedness of *A. limacinum* viral sequences to nucleocytoviruses (Figures 3B and S3D). In contrast, the *A. limacinum* viral TATA-binding proteins group with archaeal sequences, rather than mirusviruses (Figure S3E). Homologs of thraustochytrid

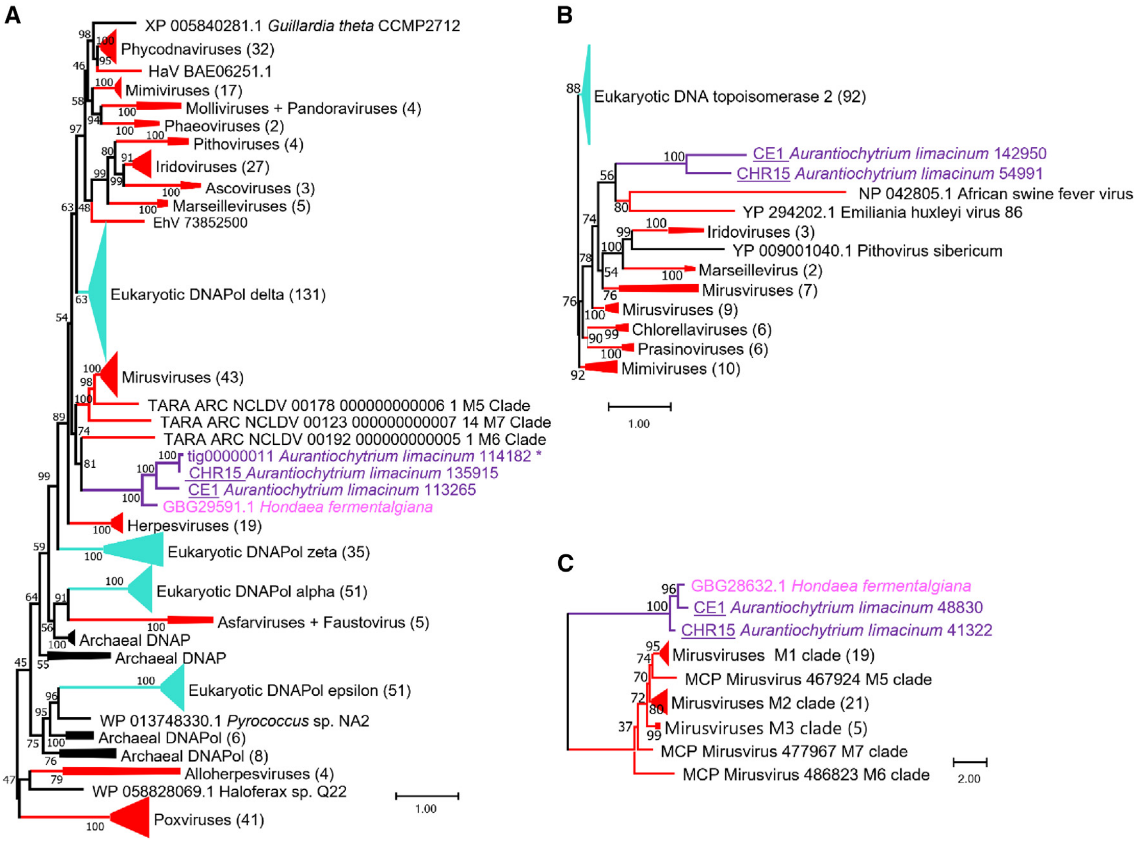

A
XP 005840281.1 Guillardia theta CCMP2712
Phycodnaviruses (32)
HaV BAE06251.1
Mimiviruses (17)
Molliviruses + Pandoraviruses (4)
Phaeoviruses (2)
Pithoviruses (4)
Iridoviruses (27)
Ascoviruses (3)
Marseilleviruses (5)
EhV 73852500
Eukaryotic DNAPol delta (131)
Mirusviruses (43)
TARA ARC NCLDV 00178 000000000006 1 M5 Clade
TARA ARC NCLDV 00123 000000000007 14 M7 Clade
TARA ARC NCLDV 00192 000000000005 1 M6 Clade
tig00000011 Aurantiochytrium limacinum 114182 *
CHR15 Aurantiochytrium limacinum 135915
CE1 Aurantiochytrium limacinum 113265
GBG29591.1 Hondaea fermentalgiana
Herpesviruses (19)
Eukaryotic DNAPol zeta (35)
Eukaryotic DNAPol alpha (51)
Asfarviruses + Faustovirus (5)
Archaeal DNAP
Archaeal DNAP
Eukaryotic DNAPol epsilon (51)
WP 013748330.1 Pyrococcus sp. NA2
Archaeal DNAPol (6)
Archaeal DNAPol (8)
Alloherpesviruses (4)
WP 058828069.1 Haloferax sp. Q22
Poxviruses (41)
1.00
B
Eukaryotic DNA topoisomerase 2 (92)
CE1 Aurantiochytrium limacinum 142950
CHR15 Aurantiochytrium limacinum 54991
NP 042805.1 African swine fever virus
YP 294202.1 Emiliania huxleyi virus 86
Iridoviruses (3)
YP 009001040.1 Pithovirus sibericum
Marseillevirus (2)
Mirusviruses (7)
Mirusviruses (9)
Chlorellaviruses (6)
Prasinoviruses (6)
Mimiviruses (10)
1.00
C
GBG28632.1 Hondaea fermentalgiana
CE1 Aurantiochytrium limacinum 48830
CHR15 Aurantiochytrium limacinum 41322
Mirusviruses M1 clade (19)
MCP Mirusvirus 467924 M5 clade
Mirusviruses M2 clade (21)
Mirusviruses M3 clade (5)
MCP Mirusvirus 477967 M7 clade
MCP Mirusvirus 486823 M6 clade
2.00

Figure 3. Phylogenetic relationships of *Aurantiochytrium limacinum* viral proteins

(A) Phylogenetic tree of virus-like family B DNA polymerase (DNAPol) proteins encoded on CE1 and LE-Chr15 of *A. limacinum* (purple) and homologs in *Hondaea fermentalgiana* (pink). Note that in addition to the homologs found in CE1 and LE-Chr15, a viral-like DNAPol is also found in *A. limacinum* tig00000011; it shows signs of pseudogenization.

(B) Phylogeny of viral DNA topoisomerases rooted with eukaryotic homologs.

(C) Phylogenetic tree of mirusvirus major capsid proteins (MCPs) and their homologs in *A. limacinum* and *H. fermentalgiana*.

(A–C) Viral sequences are in red, eukaryotic homologs are in cyan, and bacterial/archaeal sequences are in black; scale bars indicate inferred number of amino acid substitutions per site and ultrafast bootstrap support values are indicated at each node.

See also Figure S3.

viral and cellular arylsulfatase genes are detected only in various bacteria (Figure S3F). The RNAPL genes of CE1 and LE-Chr15 are particularly unusual. As seen in some other viruses (e.g., *Pithovirus sibericum* [YP_009001268.1, YP_009001052.1] and other pithoviruses) and many archaea,[43] the RNAPL coding region is split: the N- and C-terminal domains are encoded by separate ORFs located far apart from one another on both CE1 and LE-Chr15. The same split RNAPL is also observed in *Hondaea fermentalgiana*. The CE1, LE-Chr15, and *H. fermentalgiana* homologs branch robustly together in independent phylogenies of both RNAPL domains (Figures S3G and S3H), but the precise evolutionary origin(s) of RNAPL in *A. limacinum* is unclear from the data at hand. Unlike most of the *Mirusviricota* contigs previously described,[1] subunit B of DNA-dependent RNA polymerase is not found encoded in CE1 and LE-Chr15. On balance, these data suggest that CE1 and the viral element of LE-Chr15 are of mirusvirus ancestry (although a distinct clade) and confirm in a living host the main features of mirusvirus genomes previously described by metagenomics.[1]

Many types of viruses have mechanisms to maintain their genome in the host without production of viral particles, either endogenized or as episomes. Endogenous NCLDVs have been identified in stramenopiles, including oomycetes and brown algae, as well as in chlorophytes and other lineages.[44–48] It is particularly intriguing that some herpesviruses maintain latent infections as episomes,[49] whereas others maintain latent infections by integrating at host chromosome telomeres.[50] Whether these parallels are coincidental or hint at shared replication biology between *A. limacinum* mirusviruses and herpesvirus (in addition to related virion morphogenesis genes) needs to be assessed in future investigations.

Mirusvirus virion particles have yet to be isolated. Our data show that *A. limacinum* is a probable natural host and that a single host has been infected by multiple distinct mirusviruses. Both CE1 and LE-Chr15 could be latent viral genomes capable of yielding viral particles: both have an apparently complete virion module and full-length DNAPol along with other informational module genes, and viral particles consistent with an endogenous “herpes-type” virus have previously been identified in

Please cite this article in press as: Collier et al., The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.009

Table 1. Viral gene content in *Aurantiochytrium limacinum*

Viral gene ancestry	Viral gene	CE1	LE-Chr15
Mirusvirus virion module proteins	major capsid protein (HK97 fold)	+	+
	HK97 capsid maturation protease	+	+
	portal protein	+	+
	terminase	+	+
	triplex protein 1	+	+
	triplex protein 2	+	+
Other typical mirusvirus proteins (most shared with nucleocytoviricota)	heliorhodopsin	+	–
	histone H3	+	–
	ribonucleoside-diphosphate reductase α	–	+
	ribonucleoside-diphosphate reductase β	–	+
	Holliday junction resolvase	+	+
	TATA-binding protein	+	+
	family B DNA polymerase (DNAPol)	+	+
	RNA polymerase large subunit (RNAPL)[a]	+	+
	superfamily II helicase proteins	+	+
Nucleocytoviricota module proteins	major capsid protein (jelly roll fold)	–	–
	minor capsid protein	–	–
	adenovirus-type cysteine protease	–	–
	packaging ATPase	–	–

Legend: presence (+) and absence (–) of select viral proteins on CE1 and LE-Chr15 relative to key *Mirusviricota* and *Nucleocytoviricota* (NCLDV) proteins. See also Data S2.

[a] The RNAPL coding region is split into two discrete ORFs, as in pithoviruses and many archaea.

thraustochytrids.[2,51] Neither CE1 nor LE-Chr15 show clear signs of genomic erosion (such as high repeat or intron content or internal duplications) previously reported in green algal endogenous viral elements originating from NCLDVs.[47] It is also noteworthy that CE1 encodes proteins with ParA (Aurli_135839) and Fic (Aurli_13050) domains, which have been associated with plasmid segregation.[52] This may speak to how CE1 is maintained as an episomal element. In contrast, the mirusvirus genome in LE-Chr15 appears to have integrated via subtelomeric recombination. To our knowledge, the *A. limacinum* genome provides the first example of co-occurring, related episomal and endogenous viral elements (the high-copy CE1 and LE-Chr15, respectively). The presence of close homologs of the *A. limacinum* mirusvirus genes in the genome assembly of another thraustochytrid, *H. fermentalgiania*[53] (Figures 3 and S3) suggests that mirusviruses have been associated with this protist lineage for some time. Broader application of long-read sequencing will reveal whether *A. limacinum*'s subtelomeric structure and mirusviruses are unique, a general feature of labyrinthulomycetes, or distributed more widely across eukaryotic diversity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Strain and cultivation
- METHOD DETAILS
  - ○ 454 genome sequencing, assembly, and annotation
  - ○ Nanopore genome sequencing and assembly
  - ○ Comparison of short- and long-read assemblies
  - ○ rDNA, G-quadruplex, and tandem repeats
  - ○ Viral gene detection and phylogenetics
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.10.009.

ACKNOWLEDGMENTS

The work (proposal: 10.46936/10.25585/60007256) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, was supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. The construction and analysis of the *crtIBY* mutants in the Collier and Rest labs was supported by a grant from the Gordon and Betty Moore Foundation (GBMF4982). Research in the Archibald Lab, including Oxford Nanopore Technologies nanopore long-read sequencing, was supported by a grant from the Gordon and Betty Moore Foundation (GBMF5782) and a Discovery Grant (RGPIN-2019-05058) from the Natural Sciences and Engineering Research Council of Canada. Thanks to Bruce A. Curtis for technical support with nanopore sequencing and feedback on the final manuscript. Special thank you to the three reviewers who helped us improve this manuscript.

AUTHOR CONTRIBUTIONS

J.L.C. and J.S.R. conceived the study, performed the analyses, and drafted the manuscript. E.L. built the combined Canu assembly and performed initial comparative analyses. C.D., J.J., J.P., C.P., A.K., and I.V.G. performed the short-read sequencing, assembly, and annotation. G.V.F. and A.M.G.N.V. carried out nanopore long-read sequencing and performed the pulsed-field gel electrophoresis. J.L.C., J.S.R., and L.G.-L. performed detailed comparative genomic analyses of virus-like sequences. L.G.-L. and J.M.A. performed and interpreted phylogenetic reconstructions. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 7, 2023
Revised: September 8, 2023
Accepted: October 6, 2023
Published: October 31, 2023

REFERENCES

1. Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., Krupovic, M., et al. (2023). Mirusviruses link herpesviruses to giant viruses. Nature *616*, 783–789. https://doi.org/10.1038/s41586-023-05962-4.
2. Kazama, F.Y., and Schornstein, K.L. (1972). Herpes-type virus particles associated with a fungus. Science *177*, 696–697.
3. Fossier Marchan, L., Lee Chang, K.J., Nichols, P.D., Mitchell, W.J., Polglase, J.L., and Gutierrez, T. (2018). Taxonomy, ecology and biotechnological applications of thraustochytrids: a review. Biotechnol. Adv. *36*, 26–46.
4. Collier, J.L., and Rest, J.S. (2019). Swimming, gliding, and rolling toward the mainstream: cell biology of marine protists. Mol. Biol. Cell *30*, 1245–1248.
5. Filloramo, G.V., Curtis, B.A., Blanche, E., and Archibald, J.M. (2021). Re-examination of two diatom reference genomes using long-read sequencing. BMC Genomics *22*, 379.
6. Diner, R.E., Noddings, C.M., Lian, N.C., Kang, A.K., McQuaid, J.B., Jablanovic, J., Espinoza, J.L., Nguyen, N.A., Anzelmatti, M.A., Jr., Jansson, J., et al. (2017). Diatom centromeres suggest a mechanism for nuclear DNA acquisition. Proc. Natl. Acad. Sci. USA *114*, E6015–E6024.
7. Guérin, N., Ciccarella, M., Flamant, E., Frémont, P., Mangenot, S., Istace, B., Noel, B., Romac, S., Bachy, C., Gachenot, M., et al. (2021). Genomic adaptation of the picoeukaryote *Pelagomonas calceolata* to iron-poor oceans revealed by a chromosome-scale genome sequence. https://doi.org/10.1101/2021.10.25.465678.
8. Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X., and Chen, J.J.-L. (2008). The telomerase database. Nucleic Acids Res. *36*, D339–D343.
9. Juranek, S.A., and Paeschke, K. (2012). Cell cycle regulation of G-quadruplex DNA structures at telomeres. Curr. Pharm. Des. *18*, 1867–1872.
10. Paeschke, K., Juranek, S., Simonsson, T., Hempel, A., Rhodes, D., and Lipps, H.J. (2008). Telomerase recruitment by the telomere end binding protein-β facilitates G-quadruplex DNA unfolding in ciliates. Nat. Struct. Mol. Biol. *15*, 598–604.
11. Biffi, G., Tannahill, D., and Balasubramanian, S. (2012). An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. J. Am. Chem. Soc. *134*, 11974–11976.
12. Wang, F., Tang, M.-L., Zeng, Z.-X., Wu, R.-Y., Xue, Y., Hao, Y.-H., Pang, D.-W., Zhao, Y., and Tan, Z. (2012). Telomere- and telomerase-interacting protein that unfolds telomere G-quadruplex and promotes telomere extension in mammalian cells. Proc. Natl. Acad. Sci. USA *109*, 20413–20418.
13. Kobayashi, T. (2011). Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. Cell. Mol. Life Sci. *68*, 1395–1403.
14. Torres-Machorro, A.L., Hernández, R., Alderete, J.F., and López-Villaseñor, I. (2009). Comparative analyses among the *Trichomonas vaginalis*, *Trichomonas tenax*, and *Tritrichomonas foetus* 5S ribosomal RNA genes. Curr. Genet. *55*, 199–210.
15. Dvořáčková, M., Fojtová, M., and Fajkus, J. (2015). Chromatin dynamics of plant telomeres and ribosomal genes. Plant J. *83*, 18–37.
16. Roa, F., and Guerra, M. (2012). Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evol. Biol. *12*, 225.
17. Criniti, A., Simonazzi, G., Cassanelli, S., Ferrari, M., Bizzaro, D., and Manicardi, G.C. (2009). Distribution of heterochromatin and rDNA on the holocentric chromosomes of the aphids *Dysaphis plantaginea* and *Melanaphis pyraria* (Hemiptera: Aphididae). Eur. J. Entomol. *106*, 153–157.
18. Tůmová, P., Uzlíková, M., Wanner, G., and Nohýnková, E. (2015). Structural organization of very small chromosomes: study on a single-celled evolutionary distant eukaryote *Giardia intestinalis*. Chromosoma *124*, 81–94.
19. Xu, F., Jex, A., and Svärd, S.G. (2020). A chromosome-scale reference genome for *Giardia intestinalis* WB. Sci. Data *7*, 38.
20. Liu, L., Yang, Q.-F., Dong, W.-S., Bi, Y.-H., and Zhou, Z.-G. (2017). Characterization and physical mapping of nuclear ribosomal RNA (rRNA) genes in the haploid gametophytes of *Saccharina japonica* (Phaeophyta). J. Appl. Phycol. *29*, 2695–2706.
21. Maruyama, S., Misumi, O., Ishii, Y., Asakawa, S., Shimizu, A., Sasaki, T., Matsuzaki, M., Shin-i, T., Nozaki, H., Kohara, Y., et al. (2004). The minimal eukaryotic ribosomal DNA units in the primitive red alga *Cyanidioschyzon merolae*. DNA Res. *11*, 83–91.
22. Matsuzaki, M., Misumi, O., Shin, I., T., Maruyama, S., Takahara, M., Miyagishima, S.-Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., et al. (2004). Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature *428*, 653–657.
23. Torres-Machorro, A.L., Hernández, R., Cevallos, A.M., and López-Villaseñor, I. (2010). Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny? FEMS Microbiol. Rev. *34*, 59–86.
24. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature *419*, 498–511.
25. Kawai, H., Nakayama, T., Inouye, I., and Kato, A. (1997). Linkage of 5S ribosomal DNA to other rDNAs in the chromophytic algae and related taxa. J. Phycol. *33*, 505–511.
26. Kawai, H., Muto, H., Fujii, T., and Kato, A. (1995). A LINKED 5S rRNA GENE IN *Scytosiphon Lomentaria* (Scytosiphonales, Phaeophyceae). J. Phycol. *31*, 306–311.
27. Mascarenhas Dos Santos, A.C., Julian, A.T., Liang, P., Juárez, O., and Pombert, J.-F. (2023). Telomere-to-Telomere genome assemblies of human-infecting *Encephalitozoon* species. BMC Genomics *24*, 237.
28. Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.G. (2001). The highly reduced genome of an enslaved algal nucleus. Nature *410*, 1091–1096.
29. Kim, J.I., Tanifuji, G., Jeong, M., Shin, W., and Archibald, J.M. (2022). Gene loss, pseudogenization, and independent genome reduction in non-photosynthetic species of *Cryptomonas* (Cryptophyceae) revealed by comparative nucleomorph genomics. BMC Biol. *20*, 227.
30. Suzuki, S., Shirato, S., Hirakawa, Y., and Ishida, K.-I. (2015). Nucleomorph genome sequences of two chlorarachniophytes, *Amorphochlora amoebiformis* and *Lotharella vacuolata*. Genome Biol. Evol. *7*, 1533–1545.
31. Brugère, J.F., Cornillot, E., Méténier, G., Bensimon, A., and Vivarès, C.P. (2000). *Encephalitozoon cuniculi* (microspora) genome: physical map

and evidence for telomere-associated rDNA units on all chromosomes. Nucleic Acids Res. *28*, 2026–2033.

32. Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D., and Jacob, H.J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. *14*, 528–538.
33. McKim, K.S., Howell, A.M., and Rose, A.M. (1988). The effects of translocations on recombination frequency in *Caenorhabditis elegans*. Genetics *120*, 987–1001.
34. Liu, T., Rechtsteiner, A., Egelhofer, T.A., Vielle, A., Latorre, I., Cheung, M.-S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P., et al. (2011). Broad chromosomal domains of histone modification patterns in *C. elegans*. Genome Res. *21*, 227–236. https://doi.org/10.1101/gr.115519.110.
35. Perry, J., and Ashworth, A. (1999). Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. *9*, 987–989.
36. Tashiro, S., Nishihara, Y., Kugou, K., Ohta, K., and Kanoh, J. (2017). Subtelomeres constitute a safeguard for gene expression and chromosome homeostasis. Nucleic Acids Res. *45*, 10333–10349.
37. Scherf, A., Figueiredo, L.M., and Freitas-Junior, L.H. (2001). *Plasmodium* telomeres: a pathogen's perspective. Curr. Opin. Microbiol. *4*, 409–414.
38. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37.
39. Aylward, F.O., and Moniruzzaman, M. (2021). ViralRecall-A flexible command-line tool for the detection of giant virus signatures in 'omic data. Viruses *13*, 150.
40. Gallot-Lavallée, L., and Blanc, G. (2017). A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. Viruses *9*, https://doi.org/10.3390/v9010017.
41. Takao, Y., Nagasaki, K., and Honda, D. (2007). Squashed ball-like dsDNA virus infecting a marine fungoid protist *Sicyoidochytrium minutum* (Thraustochytriaceae, Labyrinthulomycetes). Aquat. Microb. Ecol. *49*, 101–108.
42. Murakoshi, Y., Shimeki, T., Honda, D., and Takao, Y. (2021). Draft genome sequence of *Sicyoidochytrium minutum* DNA virus Strain 001. Microbiol. Resour. Announc. *10*, e0041821.
43. Langer, D., Hain, J., Thuriaux, P., and Zillig, W. (1995). Transcription in archaea: similarity to that in Eucarya. Proc. Natl. Acad. Sci. USA *92*, 5768–5772.
44. Leonard, G., Labarre, A., Milner, D.S., Monier, A., Soanes, D., Wideman, J.G., Maguire, F., Stevens, S., Sain, D., Grau-Bové, X., et al. (2018). Comparative genomic analysis of the "pseudofungus" *Hyphochytrium catenoides*. Open Biol. *8*, 170184, https://doi.org/10.1098/rsob.170184.
45. Hannat, S., Pontarotti, P., Colson, P., Kuhn, M.-L., Galiana, E., La Scola, B., Aherfi, S., and Panabières, F. (2021). Diverse trajectories drive the expression of a giant virus in the oomycete plant pathogen *Phytophthora parasitica*. Front. Microbiol. *12*, 662762.
46. Delaroque, N., and Boland, W. (2008). The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. BMC Evol. Biol. *8*, 110.
47. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145.
48. Filée, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? Virology *466–467*, 53–59.
49. Cohen, J.I. (2020). Herpesvirus latency. J. Clin. Invest. *130*, 3361–3369.
50. Osterrieder, N., Wallaschek, N., and Kaufer, B.B. (2014). Herpesvirus genome integration into telomeric repeats of host cell chromosomes. Annu. Rev. Virol. *1*, 215–235.
51. Kazama, F.Y., and Schornstein, K.L. (1973). Ultrastructure of a fungus herpes-type virus. Virology *52*, 478–487.
52. Łobocka, M., and Gągała, U. (2020). Prophage P1: an example of a prophage that is maintained as a plasmid. In Bacteriophages: Biology, Technology, Therapy, D.R. Harper, S.T. Abedon, B.H. Burrowes, and M.L. McConville, eds. (Springer International Publishing), pp. 1–13.
53. Dellero, Y., Cagnac, O., Rose, S., Seddiki, K., Cussac, M., Morabito, C., Lupette, J., Aiese Cigliano, R., Sanseverino, W., Kuntz, M., et al. (2018). Proposal of a new thraustochytrid genus *Hondaea* gen. nov. and comparison of its lipid dynamics with the closely related pseudo-cryptic genus *Aurantiochytrium*. Algal Res. *35*, 125–141.
54. Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090.
55. Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108.
56. Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. *14*, 1394–1403.
57. Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*. Unit 4.10.
58. Delehelle, F., Cussat-Blanc, S., Alliot, J.-M., Luga, H., and Balaresque, P. (2018). ASGART: fast and parallel genome scale segmental duplications mapping. Bioinformatics *34*, 2708–2714.
59. Honda, D., Yokochi, T., Nakahara, T., Erata, M., and Higashihara, T. (1998). *Schizochytrium limacinum* sp. nov., a new thraustochytrid from a mangrove area in the west Pacific Ocean. Mycol. Res. *102*, 439–448.
60. Yokoyama, R., and Honda, D. (2007). Taxonomic rearrangement of the genus *Schizochytrium* sensu lato based on morphology, chemotaxonomic characteristics, and 18S rRNA gene phylogeny (Thraustochytriaceae, Labyrinthulomycetes): emendation for Schizochytrium and erection of *Aurantiochytrium* and *Oblongichytrium* gen. nov. Mycoscience *48*, 199–211.
61. Collier, J.L. (2018). Labyrinthulomycete DNA extraction protocol protocols.io. https://doi.org/10.17504/protocols.io.n83dhyn.
62. Silva, G.G., Dutilh, B.E., Matthews, T.D., Elkins, K., Schmieder, R., Dinsdale, E.A., and Edwards, R.A. (2013). Combining de novo and reference-guided assembly with scaffold_builder. Source Code Biol. Med. *8*, 23.
63. Ferris, P., Olson, B.J.S.C., De Hoff, P.L., Douglass, S., Casero, D., Prochnik, S., Geng, S., Rai, R., Grimwood, J., Schmutz, J., et al. (2010). Evolution of an expanded sex-determining locus in *Volvox*. Science *328*, 351–354.
64. Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. *8*, 186–194.
65. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. *8*, 175–185.
66. Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res. *8*, 195–202.
67. Collier, J. (2023). Labyrinthulomycete total RNA extraction protocol - hot phenol. https://doi.org/10.17504/protocols.io.q26g7pyo8gwz/v1.
68. Kuo, A., Bushnell, B., and Grigoriev, I.V. (2014). Chapter One. Fungal genomics: sequencing and annotation. In Advances in Botanical Research, F.M. Martin, ed. (Academic Press), pp. 1–52.
69. Grigoriev, I.V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S.J., Salamov, A., et al. (2021). PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. *49*, D1004–D1011.
70. Rius, M., Rest, J.S., Filloramo, G.V., Novák Vanclová, A.M.G., Archibald, J.M., and Collier, J.L. (2023). Horizontal gene transfer and fusion spread carotenogenesis among diverse heterotrophic protists. Genome Biol. Evol. *15*, evad029, https://doi.org/10.1093/gbe/evad029.

71. Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). Erratum to: the Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 256.

72. Wick, R.R., Judd, L.M., and Holt, K.E. (2018). Deepbinner: demultiplexing barcoded Oxford nanopore reads with deep convolutional neural networks. PLoS Comput. Biol. *14*, e1006583.

73. Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. Microb. Genom. *3*, e000132.

74. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

75. Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods *12*, 733–735.

76. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. *38*, 4647–4654.

77. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

78. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

79. Belmonte-Reche, E., and Morales, J.C. (2020). G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. NAR Genom. Bioinform. *2*, lqz005.

80. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421.

81. Eddy, S.R. (2011). Accelerated profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

82. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

83. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. *10*, 210.

84. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

85. R Core Team (2019). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.R-project.org.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
MinION (nanopore) raw reads	Rius et al.[70]	BioProject PRJNA680238 accessions: SRR13108467, SRR13108466, SRR13108465
Nanopore-based Canu assembly, annotations, HMM files, alignments, and tree files	This paper	https://doi.org/10.5061/dryad.2fqz612t6
454 raw reads, Newbler genome assembly, and annotations	This paper	https://phycocosm.jgi.doe.gov/Aurli1/Aurli1.home.html
CE1 (scaffold_34) sequence and annotation	This paper	https://phycocosm.jgi.doe.gov/pages/getDna.jsf?db=Aurli1&position=scaffold_34:1-297369&strand=+
LE-Chr15 (scaffold_35) sequence and annotation	This paper	https://phycocosm.jgi.doe.gov/pages/getDna.jsf?db=Aurli1&position=scaffold_35:1-295699&strand=+
Experimental models: Organisms/strains		
Aurantiochytrium limacinum ATCC MYA-1381	Yokoyama and Honda[60]	https://www.atcc.org/products/mya-1381
Software and algorithms		
Newbler v 2.6, build:20110517_1502	Silva et al.[62]	N/A
MinKNOW v2.1.12	Oxford Nanopore Technologies	https://nanoporetech.com/
Deepbinner v0.2.0	Wick et al.[72]	https://github.com/rrwick/Deepbinner
Albacore v2.3.1	Oxford Nanopore Technologies	https://nanoporetech.com/
Porechop v0.2.3	Wick et al.[73]	https://github.com/rrwick/Porechop
Canu v1.7.1	Koren et al.[74]	https://github.com/marbl/canu
Nanopolish v0.10.1	Loman et al.[75]	https://github.com/jts/nanopolish
Mauve v2.4.0	Darling et al.[56]	https://darlinglab.org/mauve/download.html
ASGART v2.4.3	Delehelle et al.[58]	https://github.com/delehef/asgart
RNAmmer v1.2	Lagesen et al.[55]	https://services.healthtech.dtu.dk/services/RNAmmer-1.2/
Tandem Repeat Finder v4.09	Benson[77]	https://github.com/Benson-Genomics-Lab/TRF
CD-HIT v4.8.1	Li and Godzik[78]	https://github.com/weizhongli/cdhit/wiki
RepeatMasker 4.1.2	Chen[57]	http://www.repeatmasker.org/
G4-iM Grinder v1.6.1	Belmonte-Reche and Morales[79]	https://github.com/EfresBR/G4iMGrinder
karyoploteR v1.20.3	Gel and Serra[54]	https://github.com/bernatgel/karyoploteR
BLAST+ v2.13.0	Camacho et al.[80]	https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html
HMMER3 v3.3.2	Eddy[81]	http://hmmer.org/
ViralRecall v2.1	Aylward and Moniruzzaman[39]	https://github.com/faylward/viralrecall
VirSorter2 v2.2.3	Guo et al.[38]	https://github.com/jiarong/VirSorter2
MAFFT v7.471	Katoh and Standley[82]	https://mafft.cbrc.jp/alignment/software/
BMGE	Criscuolo and Gribaldo[83]	ftp://ftp.pasteur.fr/pub/GenSoft/projects/BMGE/
IQTree v1.6.3	Nguyen et al. [84]	http://www.iqtree.org/
BUSCO v5.3	Manni et al.[76]	https://busco.ezlab.org/
R v4	R Core Team[85]	https://www.r-project.org/
Other		
DNA extraction protocol	This paper	dx.doi.org/10.17504/protocols.io.n83dhyn
RNA extraction protocol	This paper	dx.doi.org/10.17504/protocols.io.q26g7pyo8gwz/v1

Please cite this article in press as: Collier et al., The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.009

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jackie Collier (jackie.collier@stonybrook.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw 454 sequence reads, Newbler genome assembly, and annotation are available through the JGI Genome Portal (https://phycocosm.jgi.doe.gov/Aurli1/Aurli1.home.html). Raw fast5 MinION sequence reads have been deposited in the Sequence Read Archive database under BioProject: PRJNA680238 (SRA: SRR13108467, SRR13108466, and SRR13108465). The nanopore-based Canu assembly, annotations, HMM files, alignments, and tree files are available through Data Dryad (https://doi.org/10.5061/dryad.2fqz612t6). All other data reported in this paper will be shared by the lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Strain and cultivation

Aurantiochytrium (formerly *Schizochytrium*) *limacinum* Honda et Yokochi ATCC: MYA-1381 (also designated NIBH SR21 or IFO 32693, GenBank: AB022107[59,60]) was isolated from seawater in a mangrove area of the Yap Islands, Micronesia, and routinely maintained in ATCC 790 By+ medium (5 g glucose, 1 g yeast extract, 1 g peptone, 30 g Instant Ocean per liter) at room temperature (25 °C).

METHOD DETAILS

454 genome sequencing, assembly, and annotation

For short-read sequencing, *Aurantiochytrium* ATCC MYA-1381 cultures were grown in 2 liters of ATCC 790 By+ medium distributed in four 2-liter Fernbach flasks (500 ml each) at room temperature (25 °C) without shaking. Cultures were harvested after 7 days, producing 2.4 g wet weight. Genomic DNA was extracted from 0.507 g wet biomass following the protocol dx.doi.org/10.17504/protocols.io.n83dhyn.[61] After the genomic DNA passed Joint Genome Institute (JGI) QA/QC protocols, 454 (454 Life Sciences) sequence reads were generated from both standard unpaired and paired-end libraries. For the unpaired 454 Titanium Rapid library, genomic DNA samples were fragmented via sonication to 400–800 base pairs (bp). These fragments were end-polished and ligated to a set of 454 Y-shape adaptors. The 454 library fragments were then clonally amplified in bulk by capturing them through hybridization on microparticle beads and subjecting them to emulsion-based PCR, resulting in beads that were covered with millions of copies of a single DNA fragment (range 400-800 bp) and where each bead contained a different clonally amplified library fragment. After amplification, the beads were recovered from the emulsions and loaded into the wells of a PicoTiterPlate device (PTP) such that wells contained single DNA beads. The PTP was then inserted into the 454 Genome Sequencer FLX-Titanium instrument (454 Life Sciences) for sequencing, where sequencing reagents were sequentially flowed over the plate and the sequence of the DNA fragments was determined.

For the 454 Titanium Paired-end (PE) library, 15 μg of genomic DNA was sheared by a Hydroshear to ~8 kilobase pair (Kbp) size fragments. The sheared sample was then gel selected for intense 8 Kbp bands, purified, and ligated to 42 bp loxP linkers on either end. These loxP linkers were labeled with biotin, then circularized by the Cre recombinase. As a result, the ends of 8 Kbp fragments were brought together and bridged by a single loxP linker. These circular DNAs were further sheared to 500 bp fragments and the fragments carrying the loxP linkers were recovered by Streptavidin-coated magnetic beads. 454 Titanium adaptors A and B were then ligated to the enriched loxP linker-containing fragments in the same way the unpaired libraries were created. The 454 library fragments were then clonally amplified and sequenced as with the unpaired library.

The short-read assembly was generated using Newbler[62] version 2.6 (build:20110517_1502). The accuracy of the assembly was assessed using 13 Sanger-sequenced fosmid clones. Fosmid DNA was isolated from a single bacterial colony and purified on a Qiagen MaxiPrep column. DNA was sheared to 3-4 Kbp using Adaptive Focused Acoustics technology (Covaris, Woburn, MA, USA) and cloned into the plasmid vector pIK96 as previously described.[63] Universal primers and BigDye Terminator Chemistry (Applied Biosystems) were used for Sanger sequencing randomly selected plasmid subclones to a depth of 10x. The Phred/Phrap/Consed suite of programs were then used for assembling and editing the sequence.[64–66] After manual inspection of the assembled

Please cite this article in press as: Collier et al., The protist *Aurantiochytrium* has universal subtelomeric rDNAs and is a host for mirusviruses, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.009

sequences, finishing was performed both by resequencing plasmid subclones and by walking on plasmid subclones or the fosmid clone using custom primers. All finishing reactions were performed using dGTP BigDye Terminator Chemistry (Applied Biosystems). Finished clones contained no gaps and were estimated to contain less than one error per 10,000 bp. All 13 fosmid clones were aligned to the assembly. In ten of the clones, the assemblies were of high quality, and the overall bp error rate in this group of clones is 0.040% (152 bp discrepant out of 392,816 bp). Three clones (14866, 14868, and 14872) exhibited noteworthy discrepancies: clone 14866 stems from a repetitive region that spans a gap in scaffold00020 and has been collapsed in the assembly; the alignment of clone 14868 indicates that Newbler has inserted scaffolded gaps in the clone region; and clone 14872 falls into a repetitive region that appears to be collapsed in the assembly.

The transcriptome was also sequenced by 454 and used to assess the completeness of the genome assembly as well as to seed and assess the genome annotation. Total RNA was isolated from 1.057 g wet biomass following the protocol dx.doi.org/10.17504/protocols.io.q26g7pyo8gwz/v1.[67] A cDNA library was generated using the cDNA 454 Rapid Library Preparation Kit (Roche). mRNA was purified from total RNA using the Absolutely mRNA purification kit (Stratagene) and chemically fragmented using high heat. The fragmented RNA was reverse transcribed using random hexamers and AMV RT followed by second strand synthesis. The cDNA fragments were treated with end repair and ligated with 454 adapters. The 454 library fragments were then clonally amplified and sequenced as for genome sequencing. The resulting reads were assembled into RNA contigs using Newbler. To assess completeness, RNA sequences from one library (CHBC) were mapped to the genome assembly.

The 454 assembly contained 1662 contigs (N50/L50 233/82.5 Mbp) in 181 scaffolds (N50/L50 10/2.5 Mbp) with 937 assembly gaps (1.5% of the total 60.93 Mbp scaffold length). Scaffold lengths are shown in Table S2. The genome was annotated using the JGI Annotation Pipeline, which detects and masks repeats and transposable elements, predicts genes, characterizes each conceptually translated protein with sub-elements such as domains and signal peptides, chooses a best gene model at each locus to provide a filtered working set, clusters the filtered sets into draft gene families, ascribes functional descriptions (such as GO terms and EC numbers), and creates a JGI genome portal in PhycoCosm (https://phycocosm.jgi.doe.gov/) with tools for public access and community-driven curation of the annotation.[68,69] Draft genome sequence and annotation for *Aurantiochytrium limacinum* ATCC MYA-1381 is available via the PhycoCosm Genome Portal, https://phycocosm.jgi.doe.gov/Aurli1/Aurli1.info.html.

Nanopore genome sequencing and assembly

For long-read sequencing, *Aurantiochytrium* ATCC MYA-1381 and two putative *crtIBY* knockout mutants[70] (designated KO32 and KO33) were cultured for three days in 50 ml ATCC 790 By+ medium. Genomic DNA was extracted as described above. The precipitated DNA was left to dissolve in water by spontaneous diffusion for 48 h at room temperature to avoid shearing and subsequently purified using QIAGEN Genomic-tip 20/G. Agarose gel electrophoresis (1%) was used to visually assess and confirm the integrity of high molecular weight (20+ Kbp) DNA. DNA quality was evaluated using a NanoPhotometer P360 (Implen) to measure A260/280 (~1.8) and A260/230 (2.0-2.2) ratios. The quantity of DNA was calculated using a Qubit 2.0 Fluorometer (ThermoFisher Scientific) with the dsDNA broad range assay kit.

A multiplexed nanopore (MinION, Oxford Nanopore Technologies[71]) sequencing library was prepared using the Oxford Nanopore Technology (ONT) ligation sequencing kit (SQK-LSK109) and the PCR-free native barcoding expansion kit 1-12 (EXP-NBD103) according to the Oxford Nanopore Technologies protocol "1D Native barcoding genomic DNA with EXP-NBD103 and SQK-LSK109" (version NBE_9065_v109_revB_23May2018). Approximately 2 μg of purified genomic DNA per sample were used as input. Unfragmented genomic DNA for the wild-type and putative knockouts was repaired using the NEBNext FFPE DNA repair module (NEB cat. no. M6630) and prepared for adapter ligation using the NEBNext End repair/dA-tailing module (NEB cat. no. E7546) with incubations at 20°C and 65°C for 10 min each. The DNA repaired/end-prepped samples were purified with a 1:1 volume of AMPure XP beads (Beckman), and subjected to an incubation at room temperature for 10 mins; the pelleted beads were subsequently washed twice with 80% ethanol. The DNA was eluted off the beads in 25 μl nuclease free water for 10 min at 37 °C to encourage the elution of long molecules from the beads. The native barcodes NB07, NB08, and NB09 were ligated to the WT, KO32, and KO33 repaired/end-prepped DNA samples, respectively, in a 1.36x scaled ligation reaction. Each native barcoded sample was pooled in approximately equimolar amounts (~1.3 μg each). The 1D barcode sequencing adapters (BAM 1D) were then ligated to the pooled and barcoded DNA using a 1-h incubation at 25 °C. The adapter ligated DNA was purified by a 0.4x AMPure XP bead clean-up including a 10-min incubation at room temperature and two washes using the Long Fragment Buffer mix to enrich for DNA fragments >3 Kbp. The final adapter ligated library was incubated in 15 μl Elution Buffer for 10 min at 37 °C. A total of 1.2 μg of prepared library was loaded on a single MinION R9.4.1 chemistry SpotON flow cell (FLO-MIN106) and sequenced via Oxford Nanopore Technology's MinKNOW software (v2.1.12) without live basecalling. Binning of the raw reads was performed in real time using Deepbinner v0.2.0.[70,72] The raw fast5 MinION data have been deposited in the NCBI SRA database BioProject: PRJNA680238 (WT SRA: SRR13108467; KO32 SRA: SRR13108466; KO33 SRA: SRR13108465).

As described previously,[70] the demultiplexed fast5 files were base called using Albacore v2.3.1, adapters were removed by Porechop v0.2.3,[73] and the resulting data were used for preliminary genome assembly by Canu v1.7.1[74] with parameters adjusted to the expected genome size of 60 Mbp. The resulting consensus sequence was improved by Nanopolish v0.10.1.[75] For wild type, the genome assembly totaled 61.9 Mbp in 55 contigs, while the genomes of KO mutants 32 and 33 both assembled as 62.5 Mbp into 50 and 47 contigs, respectively. Analysis by Mauve v2.4.0[56] revealed locally collinear blocks spanning 35, 29, and 27 contigs among the three assemblies, respectively. The wild-type assembly was the least contiguous, perhaps reflecting its lower mean read length (4913 bp vs 8508 and 7951). In an effort to resolve the differences among the three assemblies and gain greater coverage, reads for all three strains were concatenated into one file as input for Canu v1.7.1, and read trimming and assembly were performed with default settings and an estimated genome size of 60 Mbp. The combined read file had 789085 reads >Q7 and 293668 reads >Q10, mean read length 7124 bp and the longest read was 180475 bp (quality score 9.8). This combined nanopore-based Canu assembly contained 62 contigs and a total of 63.71 Mbp. The 27 contigs described in the main text total 61.77 Mbp. The remaining 35 contigs included five putative mitochondrial contigs and another 30 contigs ranging from 20 Kbp to 230 Kbp, of which 14 represented single reads (Table S1). The same RNAs used to assess the completeness of the 454 assembly were also mapped to this nanopore assembly. This combined assembly was the focal point for the genome-scale investigations described herein because it generally had the longest version of each putative chromosome and had the most contigs with telomeres at both ends.

Comparison of short- and long-read assemblies

We used Mauve v 2.4.0[56] to align the 454 and nanopore assemblies and ASGART v2.4.3[58] to visualize segmental duplications in the two assemblies. BUSCO v5.3[76] was used to evaluate completeness based on the eukaryota_odb10 dataset (Data S1).

rDNA, G-quadruplex, and tandem repeats

rRNA gene locations were predicted using RNAmmer v1.2.[55] Tandem Repeat Finder v4.0977 was used to identify tandem repeats (TR) with maximum repeat unit set as large as possible (2000 bp). Telomeric repeats in the nanopore assembly were identified in this output as repeats with unit 5, 10, or 15 matching the motif TTAGG (or CCTAA). The vast majority of the 29,640 identified TRs were a few bp in length, more than 98% of TR elements were shorter than 200 bp, and the number of TRs declined with increasing TR length until ~350 bp, where a peak appeared. The sequences of the 673 TR elements longer than 299 bp were dereplicated by removing overlapping TRs, keeping the shorter repeat unit and clustering with CD-HIT v4.8.1 with default parameters except sequence identity cutoff 0.8 and -r yes.[78] Manual examination and alignment of the resulting 25 clusters (which excluded 34 singletons) revealed 5 types of TRs with consistent locations relative to rRNA genes and telomeric repeats. Additional repetitive content was identified with RepeatMasker 4.1.2 (Data S1).[57] G-quadruplexes were predicted with G4-iM Grinder v1.6.1[79] using Methods 2 and 3 and filtering for scores greater than 20 (Data S1). Genomic features were visualized using karyoploteR v1.20.3.[54]

Viral gene detection and phylogenetics

Proteins characteristic of various groups of dsDNA viruses, including mirusviruses and nucleocytoviruses, were used to screen the *A. limacinum* assembly using blastp (BLAST+ v2.13.0[80]) and HMM searches (HMMER3 v3.3.2[81]). HMMs were generated from alignments in Gaïa et al.[1] Following the detection of mirusvirus-like structural proteins in CE1 and LE-Chr15, all mirusvirus ORFs (>=90 amino acids; predicted from the mirusvirus contigs from Gaïa et al.[1]) were used as blastp queries to detect additional mirusvirus homologs (Data S2). In parallel, ViralRecall v2.1[39] and VirSorter2 v2.2.3[38] were used to evaluate viral gene content in both the 454 and nanopore *A. limacinum* assemblies (Data S2). The results of similarity searches against nr were then used to characterize additional CE1 and LE-Chr15 proteins.

Unless specified otherwise, *A. limacinum* virus-like proteins were aligned with homologs in diverse viruses, prokaryotes and eukaryotes using MAFFT v7.471[82] with default parameters, and BMGE[83] with default parameters was used to perform site selection. For virus-like family B DNA polymerase (DNAPol) proteins, the sequences were aligned with MAFFT, and sites with less than 20% gaps were retained. For viral DNA topoisomerases, sequences were aligned with MAFFT-linsi. For mirusvirus major capsid proteins (MCP), sequences were aligned with MAFFT-linsi and sites with less than 30% gaps were retained. Maximum likelihood phylogenetic trees were constructed with IQTree[84] v1.6.3 model C60+G4 with 1000 ultrafast bootstrap replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

Descriptive statistics were calculated in R[85] v4, and R packages used to generate figures are noted in the methods sections. Details regarding other statistics can be found in the methods sections or figure/table legends.